Robust Causal Bandits for Linear Models

Zirui Yan, Arpan Mukherjee, Burak Varıcı, and Ali Tajer, *Senior Member, IEEE*

Abstract—The sequential design of experiments for optimizing a reward function in causal systems can be effectively modeled by the sequential design of interventions in causal bandits (CBs). In the existing literature on CBs, a critical assumption is that the causal models remain constant over time. However, this assumption does not necessarily hold in complex systems, which constantly undergo temporal model fluctuations. This paper addresses the *robustness* of CBs to such model fluctuations. The focus is on causal systems with linear structural equation models (SEMs). The SEMs and the *time-varying* pre- and post-interventional statistical models are all *unknown*. Cumulative regret is adopted as the design criteria, based on which the objective is to design a sequence of interventions that incur the smallest cumulative regret with respect to an oracle aware of the entire causal model and its fluctuations. First, it is established that the existing approaches fail to maintain regret sub-linearity with even a few instances of model deviation. Specifically, when the number of instances with model deviation is as few as $T^{\frac{1}{2L}}$, where T is the time horizon and L is the length of the longest causal path in the graph, the existing algorithms will have linear regret in T. For instance, when $T = 10^5$ and $L = 3$, model deviations in 6 out of 10^5 instances result in a linear regret. Next, a robust CB algorithm is designed, and its regret is analyzed, where upper and information-theoretic lower bounds on the regret are established. Specifically, in a graph with N nodes and maximum degree d, under a general measure of model deviation C, the cumulative regret is upper bounded by $\tilde{\mathcal{O}}(d^{L-\frac{1}{2}}(\sqrt{NT} + NC))$ and lower bounded by $\Omega(d^{\frac{L}{2}-2} \max\{\sqrt{T}, d^2C\})$. Comparing these bounds establishes that the proposed algorithm achieves nearly optimal $\tilde{\mathcal{O}}(\sqrt{T})$ regret when C is $o(\sqrt{T})$ and maintains sub-linear regret for a broader range of C.

Index Terms—Causal bandit, robust statistics.

I. Motivation and Overview

CAUSAL bandits provide a rich framework to formalize and analyze the sequential experimental design in causal networks. Such design problems appear in applications that involve a network of interacting components that can causally influence one another. Examples include design of experiments in robotics [1], gene expression networks [2], drug discovery [3], and recommendation systems [4]. In causal systems, *interventions* are experimental mechanisms that facilitate uncovering the cause-effect relationships in causal networks and distinguishing them from the conventional association measures [5]. Sequential design of interventions has the key advantage of bringing data-adaptivity in designing the interventions, resulting in an overall reduced experiment cost and a faster process for forming inferential decisions. For instance, reprogramming the cell via gene perturbation experiments needs a careful design of sequential interventions [6] such that the outcome of one experiment guides the design of the subsequent ones. Causal bandits (CBs) provide a theoretically principled way of sequentially designing interventions to identify the one that maximizes a utility for the causal network. Specifically, the canonical model for this utility is a function of the observations obtained from the DAG. The utility is chosen as the average value of a leaf node, which we denote by the *reward node*. This is a model that is widely used in the causal bandit literature. Accordingly, each intervention mechanism is modeled by an arm, the value of the reward node under an intervention is called the reward, and the sequential selection of the interventions is abstracted as arm selection decisions.

The extent of information available about the causal model critically influences the design of CB algorithms. Broadly, there are two central pieces of information: the causal structure (topology) and the data's pre- and post-intervention statistical models. Depending on the availability of each of these two pieces, there are four possible CB settings. Designing CB algorithms and their improvement over standard bandit algorithms was first demonstrated in [7], [8] for the settings where both the structure and interventional distributions are fully specified. Subsequently, the studies transitioned to more realistic settings and explored how to accommodate unknown structure and distribution information [7], [9], [10], (approximately) known distributions but unknown structure [11], [12], and known structure and unknown distributions [13], [14], [15].

Motivation: We investigate CBs from a new perspective. The existing studies all focus on having a *fixed causal model over time*, which applies to both directions in which the models are known or unknown. In reality, however, large complex causal systems undergo model fluctuations caused by a wide range of reasons such as non-stationarity in the system or heterogeneous data [16], [17], measurement errors [18], selection bias [19], and missing data [20]. Temporal model fluctuations can change the causal structure or statistical models. For instance, in drug discovery, there are multiple observable variables or representation nodes [21], and the model fluctuations due to measurement errors can occur in both the observable variables or their weights to the representation nodes. However, the algorithm for time-invariant settings can be highly susceptible to model fluctuations. For instance, the CB algorithm in [15] enjoys a nearly optimal growth in the horizon T, i.e., $\mathcal{O}(\sqrt{T})$. Nevertheless, it will lose the

optimal rate with even minuscule instances of model deviation. More specifically, as we will show in Section V-C, the regret becomes linear in T if the system experiences model deviations in $T^{\frac{1}{2L}}$ instances, where L is the longest causal path in the graph. For even small values of L, $T^{\frac{1}{2L}}$ will be an extremely small fraction of the instances. For instance, when $T = 10^5$ and $L = 3$, model deviations in 6 out of 10^5 instances result in a linear regret.

Objectives: We pursue four objectives. (1) Under relevant measures of model fluctuations, we design a robust CB algorithm to model deviations over time. (2) We characterize almost matching upper and lower bounds on the regret as a function of model deviation level, time horizon, graph parameters, and the size of the cardinality of the intervention space. (3) We analytically assess the robustness of the relevant existing algorithms, establishing their lack of robustness to model fluctuations. (4) We consider a general intervention setting in which each subset of nodes in the graph is intervenable, and each intervention induces a distinct reward distribution, resulting in an intervention space that grows exponentially with the size of the graph. As the final objective, we show that our algorithm circumvents the exponential growth of the achievable regret with the cardinality of the intervention space and breaks it down to linear growth. Our focus is on the causal graphs that are specified by a linear structural equation model (SEM). We assume that the structure (topology) is fixed but the statistical models undergo temporal fluctuations.

Contribution & key observations: We design a novel weighting methodology for linear regression that takes advantage of the *weighted* exploration bounce. This approach enables us to accurately accommodate the impact of model deviations in our regret analysis. Based on that, we characterize novel time-uniform confidence ellipsoid models for robust linear regression, which may be of broader interest in robust linear bandits. Furthermore, we propose a robust CB algorithm and analyze the compounding effect under the novel confidence ellipsoids, which offers insights into the behavior of our regret bound. When considering a *known* budget of C that captures the level of model deviations over time, the achievable regret is $\tilde{\mathcal{O}}(d^{L-\frac{1}{2}}(\sqrt{NT} + NC))$, where N and d are the number of nodes and maximum degree in the graph, respectively, and L is the length of the largest causal path. Compared to the established lower bound $\Omega(d^{\frac{L}{2}-2} \max\{\sqrt{T}, d^2 C\})$, we observe that both bounds scale polynomially in d and exponentially in L. Furthermore, both bounds scale with $\sqrt{T}$ when the model faces $C = o(\sqrt{T})$ and linear in C when $C = \Omega(\sqrt{T})$. This indicates that our algorithm enjoys *nearly optimal* regret when $C = o(\sqrt{T})$, and it maintains a *sub-linear* regret when the aggregate model deviation is sub-linear, which is the best possible regret order that any algorithm can achieve. The cost incurred to maintain such robustness is that the regret grows linearly with the deviation amount.

Related literature: The earlier studies on CB algorithms assume that both the graph structure and the interventional distributions are known (fully or partially) [22], [23]. More recent studies dispense with one or both of the assumptions. Despite their discrepancies in model and objectives, the common theme in all these studies is that they assumed a fixed causal model. Among the related work that does not make either assumption, [9] incorporates causal learning algorithms to CBs but does not improve upon regret of non-causal bandit algorithms; [7] focuses on atomic interventions; [11] achieves a regret bound that scales with the cardinality of the intervention space; and [14] focuses on binary random variables.

In a different direction, there exist studies that assume that the graph structure is known while the distributions are unknown. The relevant literature includes [13], [14], which focuses on binary random variables. More recently, [15] focuses on linear systems and generalizes the results to the soft intervention settings, continuous random variables, and arbitrarily large intervention spaces. In parallel, [24] uses soft interventions and generalizes to non-linear models but limits to the Gaussian process SEMs in reproducing kernel Hilbert space (RKHS) and intervention space on controllable action variables. Finally, we note that even though we focus on linear SEMs, we observe that our reward is a non-linear function of the unknown parameters. Hence, our CB model fundamentally differs from linear bandits. This is the case even in the CB settings with a fixed model [15]. Nevertheless, we briefly comment on the literature on linear bandits with model misspecification or contamination. These studies assume fixed (permanent) deviation, whereas, in our setting, the deviations can vary over time [25], [26], [27], [28]. Furthermore, in linear bandits with contamination, the contamination is imposed on the observed rewards [29], [30], [31], [32], [33], [34], whereas we are focusing on model deviation.

Notations: For $N \in \mathbb{Z}_+$, we define the set $[N] \triangleq \{1, \dots, N\}$. The Euclidean norm of a vector $\mathbf{X} \in \mathbb{R}^N$ is denoted by $\|\mathbf{X}\|$. For a subset $\mathcal{S} \subseteq [N]$, we define $\mathbf{X}_{\mathcal{S}} \triangleq \mathbf{X} \odot \mathbf{1}(\mathcal{S})$, where $\odot$ denotes the Hadamard product and the vector $\mathbf{1}(\mathcal{S}) \in \{0, 1\}^N$ has 1s at the indices corresponding to $\mathcal{S}$. We denote the i-th column of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ by $[\mathbf{A}]_i$, and the entry at i-th row and j-th column by $[\mathbf{A}]_{i,j}$. The spectral norm of a matrix is denoted as $\|\mathbf{A}\|$. We further define the $\mathbf{A}$-norm for positive semidefinite matrix $\mathbf{A}$ as $\|\mathbf{X}\|_{\mathbf{A}} = \sqrt{\mathbf{X}^\top \mathbf{A} \mathbf{X}}$.

II. CAUSAL BANDIT MODEL

Causal graphical model: Consider a directed acyclic graph (DAG) denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [N]$ denotes the set of nodes, and $\mathcal{E}$ denotes the set of edges, where the ordered tuple $(i, j) \in \mathcal{E}$, indicates that there is a directed edge from i to j. Each node $i \in [N]$ is associated with a random variable X_i. Accordingly, we define the vector $\mathbf{X} \triangleq [X_1, \dots, X_N]^\top$. We consider a *linear* SEM, according to which

$$\mathbf{X} = \mathbf{B}^\top \mathbf{X} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{N \times N}$ is a strictly upper triangular edge weight matrix, and $\boldsymbol{\epsilon} \triangleq (\epsilon_1, \dots, \epsilon_N)^\top$ denotes the exogenous noise variables, with a known mean $\boldsymbol{\nu} \triangleq \mathbb{E}[\boldsymbol{\epsilon}]$. The noise vector $\boldsymbol{\epsilon}$ is 1-sub-Gaussian, and its Euclidean norm is upper bounded by $\|\boldsymbol{\epsilon}\| \leq m_{\epsilon}$. The graph's structure is assumed to be *known*, while the weight matrix $\mathbf{B}$ associated with the graph is *unknown*. For any node $i \in [N]$, we denote the set of parents of i by $\text{pa}(i)$. We denote the maximum in-degree of the graph by $d \triangleq$

$\max_i\{|\mathrm{pa}(i)|\}$ and the length of the longest directed path in the graph by L.

Intervention model: We consider *soft interventions* on the graph nodes. A soft intervention on node $i \in [N]$ alters the conditional distribution of X_i given its parents $\mathbf{X}_{\mathrm{pa}}(i)$, i.e., $\mathbb{P}(X_i|\mathbf{X}_{\mathrm{pa}}(i))$. An intervention can be applied to a subset of nodes simultaneously. If node $i \in \mathcal{V}$ is intervened, the impact of the intervention is a change in the weights of the edges incident on node i. These weights are embedded in $[\mathbf{B}]_i$, i.e., the i-th column of $\mathbf{B}$. We denote the post-intervention weight values by $[\mathbf{B}^*]_i \neq [\mathbf{B}]_i$. Accordingly, corresponding to the interventional weights, we define the interventional weight matrix $\mathbf{B}^*$, which is composed of the columns $\{[\mathbf{B}^*]_i : i \in [N]\}$. Note that soft interventions subsume commonly used stochastic *hard* interventions in which a hard intervention on node i sets $[\mathbf{B}^*]_i = \mathbf{0}$.

Since we allow any arbitrary combination of nodes to be selected for concurrent intervention, there exist 2^N interventional actions to choose from. We define $\mathcal{A} \triangleq 2^{\mathcal{V}}$ as the set of all possible interventions, i.e., all possible subsets of $[N]$. For any intervention $a \in \mathcal{A}$, we define the post-intervention weight matrix $\mathbf{B}_a$ such that columns corresponding to the non-intervened nodes retain their observational values from $\mathbf{B}$, and the columns corresponding to the intervened nodes change to their new interventional values from $\mathbf{B}^*$. The columns of $\mathbf{B}_a$ are specified by

$$[\mathbf{B}_a]_i \triangleq [\mathbf{B}]_i \cdot \mathbb{1}\{i \notin a\} + [\mathbf{B}^*]_i \cdot \mathbb{1}\{i \in a\}, \quad (2)$$

where $\mathbb{1}$ denotes the indicator function. The interventions change the probability models of $\mathbf{X}$. We define $\mathbb{P}_a$ as the probability measure of $\mathbf{X}$ under intervention $a \in \mathcal{A}$. For any given $\mathbf{B}$ and $\mathbf{B}^*$ we assume that $\|[\mathbf{B}_a]_i\| \leq m_B$. Without loss of generality, we assume $m_B = 1$. Due to the boundedness of noise ϵ and column of $\mathbf{B}_a$ matrices, there exists $m \in \mathbb{R}^+$ such that $\|\mathbf{X}\| \leq m$.

Causal bandit model: Our objective is the sequential design of interventions. The set of possible interventions can be modeled as a multi-armed bandit setting with 2^N arms, one arm corresponding to each possible intervention. Following the canonical CB model [8], [22], we designate node N (i.e., the node without a descendant) as the *reward* node. Accordingly, X_N specifies the reward value. We denote the expected reward collected under intervention $a \in \mathcal{A}$ by

$$\mu_a \triangleq \mathbb{E}_a[X_N], \quad (3)$$

where $\mathbb{E}_a$ denotes expectation under $\mathbb{P}_a$. We denote the intervention that yields the highest average reward by $a^* \triangleq \arg\max_{a \in \mathcal{A}} \mu_a$; denote the sequence of interventions by $\{a(t) \in \mathcal{A} : t \in \mathbb{N}\}$; and denote the data generated at time t and under intervention $a(t)$ by $\mathbf{X}(t) = [X_1(t), \dots, X_N(t)]^\top$. The learner's goal is to minimize the average cumulative regret over the time horizon T with respect to the reward accumulated by an oracle aware of the systems model, interventional distributions, and model fluctuations. We define the expected accumulated regret as

$$\mathbb{E}[R(T)] \triangleq T\mu_{a^*} - \sum_{t=1}^{T} \mathbb{E}[X_N(t)]. \quad (4)$$

III. TEMPORAL MODEL FLUCTUATIONS

Due to the size and complexity of the graphical models that represent complex systems, assuming that the observational and interventional models $\mathbf{B}$ and $\mathbf{B}^*$ remain unchanged over time is a strong assumption. These models can undergo temporal variations due to various reasons, such as model misspecifications, stochastic behavior of the system, and adversarial influences. To account for such variations, we refer to $\mathbf{B}_{a(t)}$ as the nominal model of the graph at time t and denote the actual time-varying *unknown* model by $\mathbf{D}_{a(t)}$. Accordingly, we define the deviation of the actual model from the nominal model by

$$\Delta_{a(t)} \triangleq \mathbf{D}_{a(t)} - \mathbf{B}_{a(t)}. \quad (5)$$

To quantify the impact of model deviations on the regret $R(T)$, we specify two measures that capture the extent of deviations. The first measure captures the maximum number of times each node deviates from the nominal model. The second measure provides a budget for the maximum deviation in the linear model that model deviations can inflict over time. Clearly, if the model of node i undergoes deviation at time t under intervention a, we have $\|[\Delta_{a(t)}]_i\| \neq 0$.

Measure 1 [Deviation Frequency (DF)]: This measure accounts for how frequently each node's model can deviate from its nominal model, and over a horizon T it is defined as

$$C_{\mathrm{DF}} \triangleq \max_{i \in [N]} \sum_{t=1}^{T} \max_{a(t) \in \mathcal{A}} \mathbb{1}\{\|[\Delta_{a(t)}]_i\| \neq 0\}. \quad (6)$$

This model is adopted from misspecified bandit literature [26]. To avoid unbounded deviations, we assume that the deviation inflicted on each node at any given time is bounded by a constant $m_c \in \mathbb{R}_+$, i.e.,

$$\max_{i \in [N]} \max_{t \in [T]} \max_{a(t) \in \mathcal{A}} \|[\Delta_{a(t)}]_i\| \leq m_c. \quad (7)$$

Measure 2 [Aggregate Deviation (AD)]: This measure quantifies the aggregate deviation over time. Specifically, we define the maximum aggregate deviation as

$$C_{\mathrm{AD}} \triangleq \max_{i \in [N]} \sum_{t=1}^{T} \max_{a(t) \in \mathcal{A}} \|[\Delta(t)]_i\|. \quad (8)$$

This measure of deviation is also standard in stochastic bandits [34], where the deviation budget is defined as the maximum deviation in the reward that the adversary can inflict over time. We will observe that C_{DF} and C_{AD} impact the regret results similarly. Hence, to unify the results and present them in a way that applies to both measures, we use C to represent the level of model deviation. For measure 1, we define C as the product of a constant factor m_c and C_{DF}, while for measure 2, we define C as C_{AD}. We assume that the model deviation budgets specified by C are known to the learner, allowing the CB algorithm to adapt to the varying levels of model deviation.

IV. ROBUST-LCB ALGORITHM

In this section, we present the details of our algorithm and provide the performance guarantee (regret analysis) in Section V. We also provide theoretical comparisons to the

Algorithm 1 Robust-LCB

1: **Input:** Horizon T, causal graph $\mathcal{G}$, action set $\mathcal{A}$, mean noise vector $\boldsymbol{v}$, deviation budget C
2: **Initialization**: Initialize $\mathbf{B}(0) = \mathbf{B}^*(0) = \mathbf{0}_{N \times N}$ and $\mathbf{V}_i(0) = \mathbf{V}_i^*(0) = \mathbf{I}_N$, $\forall i \in [N]$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4:     Compute $\mathrm{UCB}_a(t)$ according to (21) for $a \in \mathcal{A}$.
5:     Pull $a(t) = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_a(t)$ and observe $\mathbf{X}(t) = (X_1(t), \ldots, X_N(t))^\top$.
6:     **for** $i \in \{1, \ldots, N\}$ **do**
7:         Set $w_i(t)$ as (19), update $[\mathbf{B}(t)]_i$ according to (9) and update $[\mathbf{B}^*(t)]_i$ according to (10).
8:     **end for**
9: **end for**

existing algorithms designed for CBs with fixed models, establishing their lack of robustness against model variations.

Algorithm overview: Identifying the best intervention hinges on determining which of the distributions $\{\mathbb{P}_a : a \in \mathcal{A}\}$ maximizes the expected reward. Nevertheless, these 2^N distributions are unknown. Therefore, a direct approach entails estimating these probability distributions, the complexity of which grows exponentially with N. To circumvent this, we leverage the fact that specifying these distributions has redundancies since all depend on the observational and interventional matrices $\mathbf{B}$ and $\mathbf{B}^*$. These matrices can be fully specified by $2Nd$ scalars, where d is the maximum degree of the causal graph. Hence, at its core, our proposed approach aims to estimate these two matrices.

We design an algorithm that has two intertwined key objectives. One pertains to the *robust estimation* of matrices $\mathbf{B}$ and $\mathbf{B}^*$ when the observations are generated by the non-nominal models. For this purpose, we design a weighted ordinary least squares (W-OLS) estimator. The structure of the estimator and the associated confidence ellipsoids for the estimates are designed to circumvent model deviations effectively. The second objective is designing a decision rule for the sequential selection of the interventions over time. This sequential selection, naturally, is modeled as a multi-armed bandit problem. Therefore, we design an upper confidence bound (UCB)–based algorithm for the sequential selection of the interventions over time. Next, we present the details of the **Robust L**inear **C**ausal **B**andit (Robust-LCB). The steps involved in this algorithm are summarized in Algorithm 1.

Countering model deviations: Our approach to circumventing model deviations is to identify and filter out the samples generated by the non-nominal models. We refer to these samples as *outlier samples*. This facilitates forming estimates for $\mathbf{B}$ and $\mathbf{B}^*$ based on the samples generated by the nominal models. Since the model deviations may happen on multiple nodes simultaneously, the Robust-LCB is designed to identify the nodes undergoing deviation over time and discard the outlier samples generated by these nodes. Such filtration is implemented via assigning time-varying and data-adaptive weights to different nodes such that the weight assigned to node $i \in [N]$ at time $t \in \mathbb{N}$ balance two factors: the probability of node $i \in [N]$ undergoing deviation at t and the contribution of that sample to the estimator. These weights, subsequently, control how the samples from different nodes contribute to estimating $\mathbf{B}$ and $\mathbf{B}^*$.

Robust estimation: We design the *weighted* ordinary least squares (OLS) estimators for $\mathbf{B}$ and $\mathbf{B}^*$, which at time $t \in \mathbb{N}$ are denoted by $\mathbf{B}(t)$ and $\mathbf{B}^*(t)$, respectively. To estimate the observational weights $[\mathbf{B}]_i$, we use the samples from instances at which node i is not intervened. Conversely, to estimate the interventional weights $[\mathbf{B}^*]_i$, we use the samples from the instances at which node i is intervened. By defining $\{w_i(t) \in \mathbb{R}_+ : i \in [N]\}$ as the set of weights assigned to the nodes at time $t \in \mathbb{N}$, i-th columns of these estimates are specified as follows.

$$[\mathbf{B}(t)]_i \triangleq [\mathbf{V}_i(t)]^{-1} \sum_{s \in [t]: i \notin a(s)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s)(X_i(s) - v_i), \quad (9)$$

$$\left[\mathbf{B}^*(t)\right]_i \triangleq \left[\mathbf{V}_i^*(t)\right]^{-1} \sum_{s \in [t]: i \in a(s)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s)(X_i(s) - v_i), \quad (10)$$

where we have defined the *weighted Gram matrices* as

$$\mathbf{V}_i(t) \triangleq \sum_{s \in [t]: i \notin a(s)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s) \mathbf{X}_{\mathsf{pa}(i)}^\top(s) + \mathbf{I}_N, \quad (11)$$

$$\mathbf{V}_i^*(t) \triangleq \sum_{s \in [t]: i \in a(s)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s) \mathbf{X}_{\mathsf{pa}(i)}^\top(s) + \mathbf{I}_N. \quad (12)$$

Furthermore, we define the matrices associated with the squared weights as

$$\widetilde{\mathbf{V}}_i(t) \triangleq \sum_{s \in [t]: i \notin a(s)} w_i^2(s) \mathbf{X}_{\mathsf{pa}(i)}(s) \mathbf{X}_{\mathsf{pa}(i)}^\top(s) + \mathbf{I}_N, \quad (13)$$

$$\widetilde{\mathbf{V}}_i^*(t) \triangleq \sum_{s \in [t]: i \in a(s)} w_i^2(s) \mathbf{X}_{\mathsf{pa}(i)}(s) \mathbf{X}_{\mathsf{pa}(i)}^\top(s) + \mathbf{I}_N. \quad (14)$$

Similarly to (2), we denote the relevant and Gram matrices for node i under intervention $a \in \mathcal{A}$ by

$$\widetilde{\mathbf{V}}_{i,a}(t) \triangleq \mathbb{1}\{i \in a\} \widetilde{\mathbf{V}}_i^*(t) + \mathbb{1}\{i \notin a\} \widetilde{\mathbf{V}}_i(t). \quad (15)$$

Confidence ellipsoids: After performing estimation in each round, we construct the confidence ellipsoids for the OLS estimators $\{\mathcal{C}_i(t) : i \in [N]\}$ for the observational weights and $\{\mathcal{C}_i^*(t) : i \in [N]\}$ for the interventional weights

$$\mathcal{C}_i(t) \triangleq \Big\{\theta \in \mathcal{B}_1:$$
$$\left\|\theta - [\mathbf{B}(t-1)]_i\right\|_{\mathbf{V}_i(t-1)[\widetilde{\mathbf{V}}_i(t-1)]^{-1}\mathbf{V}_i(t-1)} \le \beta_t\Big\}, \quad (16)$$

$$\mathcal{C}_i^*(t) \triangleq \Big\{\theta \in \mathcal{B}_1:$$
$$\left\|\theta - [\mathbf{B}^*(t-1)]_i\right\|_{\mathbf{V}_i^*(t-1)[\widetilde{\mathbf{V}}_i^*(t-1)]^{-1}\mathbf{V}_i^*(t-1)} \le \beta_t\Big\}, \quad (17)$$

where $\mathcal{B}_1$ is the unit ball in $\mathbb{R}^N$ and $\{\beta_t \in \mathbb{R}_+, t \in \mathbb{N}\}$ is a sequence of non-decreasing confidence radii that control the size of the confidence ellipsoids, which we will specify. Accordingly, we define the relevant confidence ellipsoid for node i under intervention $a \in \mathcal{A}$ as

$$\mathcal{C}_{i,a}(t) \triangleq \mathbb{1}\{i \in a\} \mathcal{C}_i^*(t) + \mathbb{1}\{i \notin a\} \mathcal{C}_i(t). \quad (18)$$

Weight designs: Designing the weights $\{w_i(t) : i \in [N]\}$ at time t is instrumental in effectively winnowing out the outlier samples. We select the weights that bring the confidence radius β_t down to nearly constant

$$w_i(t) \triangleq \min\left\{\frac{1}{C}, \frac{1}{C\left\|\mathbf{X}_{\mathsf{pa}(i)}(t)\right\|_{[\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1}}}\right\}, \quad (19)$$

where the weights are inversely proportional to the norm $\|\mathbf{X}_{\mathsf{pa}(i)}(t)\|_{[\widetilde{\mathbf{V}}_{i,a(t)}(t)]}^{-1}$ and deviation budget C, and they are truncated at $1/C$, which ensures that the weights are not arbitrarily large. We refer the term $\|\mathbf{X}_{\mathsf{pa}(i)}(t)\|_{[\widetilde{\mathbf{V}}_{i,a(t)}(t)]}^{-1}$ as *weighted exploration bonus*. A higher exploration bonus means lower confidence in the sample. Setting the weights as the inverse of the exploration bonus avoids potentially significant regret caused by both the stochastic noise and model deviations. We scale the weights proportional to $1/C$ to use smaller weights when the model deviation level is higher.

Intervention selection: We adopt a UCB-based rule for sequentially selecting the interventions. Specifically, at each time t, our algorithm selects the intervention that maximizes a UCB, defined as the maximum value of expected reward when the edge weights are in the confidence ellipsoids $\{\mathcal{C}_{i,a}(t), i \in [N]\}$, under that intervention. Recall the expected reward μ_a for $a \in \mathcal{A}$ defined in (3) is a function of the edge weights $\mathbf{B}_a$, which can be decomposed according to the following lemma.

Lemma 1 [15, Lemma 1]: Consider a linear SEM $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with intervention $a \in 2^{\mathcal{V}}$, whose weight matrix is denoted by $\mathbf{B}_a \in \mathbb{R}^{N \times N}$. Furthermore, consider the function $f(\mathbf{A}) \triangleq \sum_{\ell=0}^{L} [\mathbf{A}^{\ell}]_N$ for $\mathbf{A} \in \mathbb{R}^{N \times N}$, where we denote $\mathbf{A}^{\ell}$ as the ℓ-th power of matrix $\mathbf{A}$. We have

$$\mu_a = \langle f(\mathbf{B}_a), \, \boldsymbol{v} \rangle, \qquad (20)$$

where $\boldsymbol{v} = (v_1, \ldots, v_N)$, and $v_i = \mathbb{E}[\epsilon_i]$ denotes the mean of the noise vector corresponding to node $i \in [N]$.

Thus, for any intervention $a \in \mathcal{A}$, the UCB is naturally defined as

$$\mathrm{UCB}_a(t) \triangleq \max_{\{\forall i \in [N] : [\Theta]_i \in \mathcal{C}_{i,a}(t)\}} \langle f(\Theta), \, \boldsymbol{v} \rangle. \qquad (21)$$

Based on the UCB in (21), at time t, our algorithm selects the intervention that maximizes the UCB,

$$a(t) = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_a(t). \qquad (22)$$

V. REGRET ANALYSIS

In this section, we present the performance guarantees for the proposed Robust-LCB algorithm. We first provide the upper bound on the average cumulative regret in Section V-A. We also establish a minimax lower bound in Section V-B that shows the tightness of our upper bound. By comparing our regret with that of LinSEM-UCB in Section V-C, we evaluate the robustness of our algorithm.

A. Regret Upper Bound

In order to derive the upper bound, we begin by providing a concentration bound for the W-OLS estimator. Notably, we investigate a vector norm that differs from existing work in robust bandits. This norm was first investigated in [35] under the non-stationary setting, and our investigation builds on this to provide novel insights into the robust behavior of the W-OLS.

Lemma 2 (Estimator Concentration): Under a deviation budget C, with a probability at least $1 - 2\delta$, for any node $i \in [N]$ and $t \geq 0$, we have

$$\|[\mathbf{B}(t)]_i - \mathbf{B}_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \leq \beta_t(\delta), \qquad (23)$$

$$and \quad \|[\mathbf{B}^*(t)]_i - \mathbf{B}_i^*\|_{\mathbf{V}_i^*(t)[\widetilde{\mathbf{V}}_i^*(t)]^{-1}\mathbf{V}_i^*(t)} \leq \beta_t(\delta), \qquad (24)$$

where we have defined

$$\beta_t(\delta) \triangleq \sqrt{2\log(1/\delta) + d\log(1 + m^2 t/dC^2)} + 1 + m. \qquad (25)$$

Proof: We will provide the proof corresponding to the observational weights $[\mathbf{B}(t)]_i$, while the proof for the interventional weights $[\mathbf{B}^*(t)]_i$ follows similarly. For any node $i \in [N]$ and $t \geq 0$, we decompose the error in estimation $\|[\mathbf{B}(t)]_i - \mathbf{B}_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)}$ for $t \geq 0$ as follows.

$$\|[\mathbf{B}(t)]_i - [\mathbf{B}]_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \qquad (26)$$

$$= \left\| [\mathbf{V}_i(t)]^{-1} \sum_{s \in [t], i \notin a(t)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s) \Big[\mathbf{X}_{\mathsf{pa}(i)}^{\top}(s)[\mathbf{D}(t)]_i \right.$$

$$\left. + \epsilon_i(s) - v_i \Big] - [\mathbf{B}]_i \right\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \qquad (27)$$

$$\leq \underbrace{\left\| [\widehat{\mathbf{B}}(t)]_i - [\mathbf{B}]_i \right\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)}}_{I_1:\text{ Stochastic and regularization error}} \qquad (28)$$

$$+ \underbrace{\left\| \sum_{s \in [t], i \notin a(t)} \mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^{\top}(s)[\Delta(s)]_i \right\|_{[\widetilde{\mathbf{V}}_i(t)]^{-1}}}_{I_2:\text{ Fluctuation error}}. \qquad (29)$$

where $\widehat{\mathbf{B}}(t)$ refers to the auxiliary estimators which correspond to the ridge regression estimator but with the removal of deviation's impact on the output, i.e.,

$$[\widehat{\mathbf{B}}(t)]_i = [\mathbf{V}_i(t)]^{-1} \sum_{s \in [t], i \notin a(t)} w_i(s) \mathbf{X}_{\mathsf{pa}(i)}(s)$$

$$\times \Big[\mathbf{X}_{\mathsf{pa}(i)}^{\top}(s)[\mathbf{B}]_i + \epsilon_i(s) - v_i \Big]. \qquad (30)$$

The stochastic and regularization errors can be bounded by the following lemma.

Lemma 3: For all node $i \in [N]$, with probability at least $1 - \delta$, $\forall t \geq 0$, we have

$$I_1 = \|[\widehat{\mathbf{B}}(t)]_i - [\mathbf{B}]_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \qquad (31)$$

$$\leq 1 + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{m^2 W(s)}{d}\right)}, \qquad (32)$$

where we define $W(s) = \sum_{s \in [t-1], i \notin a(t)} w_i^2(s)$.

Proof: Note that the weights $w_i(t)$ are predictable, i.e., $\mathcal{F}_i(t-1)$ measurable, if the σ-algebra is defined as

$$\mathcal{F}_i(t) = \sigma(\mathbf{X}_{\mathsf{pa}(i)}(1), \epsilon_i(1), \mathbf{X}_{\mathsf{pa}(i)}(2), \epsilon_i(2), \ldots,$$

$$\mathbf{X}_{\mathsf{pa}(i)}(t), \epsilon_i(t), \mathbf{X}_{\mathsf{pa}(i)}(t+1)), \qquad (33)$$

which is similar to the one used in [36]. This modification of the filtration allows weights to depend on the current value of $\mathbf{X}_{\mathsf{pa}(i)}(t)$. Then the lemma results from [35, Th. 1] with $\mu_t = \lambda_t = 1$. $\blacksquare$

Since $w_i(s) \leq \frac{1}{C}$ for all $s \in [t]$ and $i \in [N]$, we can further upper bound the stochastic and regularization error as follows.

$$I_1 \leq 1 + \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{m^2 t}{dC^2}\right)}. \quad (34)$$

Now we need to bound the fluctuation error I_2, which is bounded as follows.

$$I_2 = \left\| [\widetilde{\mathbf{V}}_i(t)]^{-1/2} \sum_{s\in[t], i\notin a(t)} w_i(s)\mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s)[\Delta_{\mathbf{B}}(s)]_i \right\| \quad (35)$$

$$\leq \sum_{s\in[t], i\notin a(t)} w_i(s) \left\| [\widetilde{\mathbf{V}}_i(t)]^{-1/2}\mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s)[\Delta_{\mathbf{B}}(s)]_i \right\| \quad (36)$$

$$= \sum_{s\in[t], i\notin a(t)} w_i(s) \left\| [\widetilde{\mathbf{V}}_i(t)]^{-1/2}\mathbf{X}_{\mathsf{pa}(i)}(s) \right\| \left| \mathbf{X}_{\mathsf{pa}(i)}^\top(s)[\Delta_{\mathbf{B}}(s)]_i \right| \quad (37)$$

$$\leq m \sum_{s\in[t], i\notin a(t)} w_i(s) \left\| \Delta_{\mathbf{B}}(s)]_i \right\| \left\| \mathbf{X}_{\mathsf{pa}(i)}(s) \right\|_{[\widetilde{\mathbf{V}}_i(t)]^{-1}} \quad (38)$$

$$\leq m \sum_{s\in[t], i\notin a(t)} w_i(s) \left\| \Delta_{\mathbf{B}}(s)]_i \right\| \left\| \mathbf{X}_{\mathsf{pa}(i)}(s) \right\|_{[\widetilde{\mathbf{V}}_i(s)]^{-1}} \quad (39)$$

$$\leq m, \quad (40)$$

where (36) follows from the triangle inequality, (38) follows from the fact that $\|\mathbf{X}_{\mathsf{pa}(i)}(s)\| \leq m$, (39) holds since we have $\|x\|_{[\widetilde{\mathbf{V}}_i(t)]^{-1}} \leq \|x\|_{[\widetilde{\mathbf{V}}_i(s)]^{-1}}$ for any $s \in [t]$ and $x \in \mathbb{R}^N$, and (40) is obtained using the definition of the sequence of weights $\{w_i(s) : s \in [t]\}$. Finally, substituting the results of Lemma 3 and (40), with probability at least $1 - \delta$, for all $t \geq 0$, we have

$$\|[\mathbf{B}(t)]_i - \mathbf{B}_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \leq 1 + m$$
$$+ \sqrt{2\log(1/\delta) + d\log(1 + m^2 t/dC^2)}. \quad (41)$$

Similarly, for the estimators for interventional weights, with probability at least $1 - \delta$, for all $t \geq 0$, we have

$$\left\| [\mathbf{B}^*(t)]_i - \mathbf{B}_i^* \right\|_{\mathbf{V}_i^*(t)[\widetilde{\mathbf{V}}_i^*(t)]^{-1}\mathbf{V}_i^*(t)} \leq 1 + m$$
$$+ \sqrt{2\log(1/\delta) + d\log(1 + m^2 t/dC^2)}. \quad (42)$$

Combining the results in (41) and (42) we complete the proof. ∎

The previous lemma offers high probability error bounds for estimators. Due to the causal structure, these errors accumulate and propagate along the causal path, leading to the reward node N. Consequently, we analyze the compounding impacts of estimation errors and model deviations. This analysis involves examining the eigenvalues of the weighted Gram matrices $\mathbf{V}_{i,a(t)}(t)$ and $\widetilde{\mathbf{V}}_{i,a(t)}(t)$. We introduce the subsequent lemma to show a bound on the accumulated estimation errors on the reward node with proof provided in Section V-B.

Lemma 4: For any given intervention $a \in \mathcal{A}$ matrices $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{M}_i \in \mathbb{R}^{N\times N}$ for all $i \in [N]$, define

$$\Delta_{\mathbf{A}} \triangleq \mathbf{A} - \mathbf{B}_a, \quad \text{and} \quad \Delta_{\mathbf{A}}^{(\ell)} \triangleq \mathbf{A}^\ell - \mathbf{B}_a^\ell. \quad (43)$$

If $\mathbf{A}$ shares the same support with $\mathbf{B}_a$, $\mathbf{M}_i \succeq \mathbf{I}$ and $[\mathbf{M}_i]_j = [\mathbf{M}_i^\top]_j = \mathbf{e}_i$ if $[\mathbf{B}]_{j,i} = 0$, and if the following bound holds

$$\|[\Delta_{\mathbf{A}}]_i\|_{\mathbf{M}_i} \leq \beta, \quad (44)$$

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

$$\left\| [\Delta_{\mathbf{A}}^{(\ell)}]_N \right\| < d^{\frac{\ell-1}{2}}(\beta + 1)^\ell \max_{i\in[N]} \lambda_{\min}^{-\frac{1}{2}}(\mathbf{M}_i). \quad (45)$$

Next, building on the estimation error bounds established in Lemma 2 and the compounding error bounds established in Lemma 4, we derive a unified regret bound that applies to both measures of model deviation. It is noteworthy that the analysis is distinctly different from that in the time-invariant setting since we are facing model fluctuations, for which we have designed novel weights for the W-OLS estimator.

Theorem 1 (Regret Upper Bound): Under a deviation budget C, by setting $\delta = \frac{1}{2NT}$ and $\beta_t(\delta)$ according to Lemma 2, the average cumulative regret of Robust-LCB is upper bounded by

$$\mathbb{E}[R(T)] \leq 2m + \tilde{\mathcal{O}}\left(d^{L-\frac{1}{2}}\left(\sqrt{NT} + NC\right) \right). \quad (46)$$

Proof Sketch: Characterizing the regret bound involves decomposing the regret into two parts, depending on whether the concentration inequality in Lemma 2 holds. If the concentration inequality does not hold, the regret is upper bound by a constant term. Otherwise, the estimation errors are upper bounded in Lemma 2 with their compounding effects on the reward node N upper bounded by Lemma 4. Next, the behavior of the eigenvalues of weighted Gram matrices $\mathbf{V}_{i,a(t)}(t)$ and $\widetilde{\mathbf{V}}_{i,a(t)}$ is investigated to reach the final result. See Appendix C for the detailed proof.

The regret bound derived in Theorem 1 can be decomposed into two parts. The first term recovers the order of the optimal rate achieved in the time-invariant setting. The second term captures the impact of model deviation on the regret bound, that is, the cost of handling unknown model fluctuations. Next, we present a lower bound that confirms the tightness and optimality of our upper bound.

Remark 1: Robust-LCB works when replacing the deviation budget C with an upper bound $\bar{C}$. All the analyses and performance guarantees remain valid when C is substituted with its upper bound $\bar{C}$.

B. Regret Lower Bound

For our analysis of the lower bound, we first show the tightness of the second term of the upper bound that captures the model deviation level C. Building on this insight, we then combine this with the existing lower bound in the time-invariant setting to show the tightness of our regret bound.

Theorem 2 (Regret Lower Bound): For any degree d and graph length L, for any algorithm with knowledge of C, there exists a bandit instance such that the expected regret is at least

$$\mathbb{E}[R(T)] \geq \Omega\left(d^{\frac{L}{2}}C \right). \quad (47)$$

Proof: We construct two instances of causal bandits and demonstrate that under specific deviations, no algorithm can distinguish between them and the initial stage. We consider two linear SEM causal bandit instances sharing the same hierarchical graph $\mathcal{G}$ as shown in Figure 1.

Let us examine the parameterization of the two bandit instances, referred to as $\mathcal{I}_1 = \{\mathbf{B}, \mathbf{B}^*, \epsilon\}$ and $\mathcal{I}_2 = \{\bar{\mathbf{B}}, \bar{\mathbf{B}}^*, \epsilon\}$.
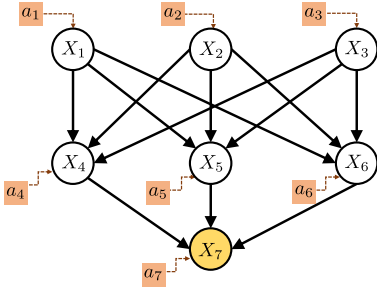
Fig. 1. Example of the hierarchical graph with graph degree $d = 3$ and length $L = 2$.

For the existing edges in graph $(i, j) \in \mathcal{E}$ for $i < j$ and $i, j \in [N]$, we define

$$[\mathbf{B}]_{i,j} = [\bar{\mathbf{B}}]_{i,j} = [\mathbf{B}^*]_{i,j} = [\bar{\mathbf{B}}^*]_{i,j} = \sqrt{\frac{1}{d}}. \quad (48)$$

for $j < N$. For the case $j = N$, we define

$$[\mathbf{B}^*]_{i,j} = -[\bar{\mathbf{B}}^*]_{i,j} = \sqrt{\frac{1}{d}} \quad \text{and} \quad [\mathbf{B}]_{i,j} = [\bar{\mathbf{B}}]_{i,j} = 0. \quad (49)$$

For the noises, we define

$$\epsilon_i \sim \begin{cases} \mathcal{N}(1, 1) & \text{for} \quad i \in \{1, \ldots, d\} \\ \mathcal{N}(0, 1) & \text{for} \quad i \notin \{1, \ldots, d\} \end{cases}. \quad (50)$$

Thus, the only difference between the two bandit instances lies in the intervention weights assigned to the reward node. In the first bandit instance, the optimal action is when the reward node is *intervened*. In contrast, in the second bandit instance, the best action is associated with the reward node being *not intervened*. The regret incurred from choosing the sub-optimal action on the reward node is $d^{L/2}$. Next, consider the scenario where deviations only occur on the reward node during the initial C rounds. Furthermore, consider the scenario in which, at each time $t \in \{1, \ldots, C\}$, the weights are set to 0, and the learner only observes a random noise as the reward. In this case, the learner has no information about the $[\mathbf{B}]_N$ and can only make random guesses. After C rounds, either $\mathbb{E}N_i(C)$ or $\mathbb{E}N_i^*(C)$ is no less then $\frac{C}{2}$. Consequently, there must exist a bandit instance at which the algorithm plays the sub-optimal arm at least $\frac{C}{2}$ times. We conclude that it must incur $\frac{1}{2}d^{\frac{L}{2}}C$ regret with probability at least $1/2$. Ignoring the constant, by using Markov's inequality, we have

$$\mathbb{E}[R(T)] \geq \frac{1}{2} d^{\frac{L}{2}} C \mathbb{P}\left(R(T) \geq \frac{1}{2} d^{\frac{L}{2}} C\right) = \Omega\left(d^{\frac{L}{2}} C\right). \quad (51)$$

∎

Theorem 3: [15, Th. 6] For any degree d and graph length L and any algorithm, there exists a causal bandit instance such that the expected regret is at least

$$\mathbb{E}[R(T)] \geq \Omega\left(d^{\frac{L}{2}-2}\sqrt{T}\right). \quad (52)$$

Combining the results in Theorem 2 and Theorem 3, we can conclude a minimax lower bound in the following corollary.

Corollary 1: For any degree d and graph length L and any algorithm, there exists a causal bandit instance such that the expected regret is at least

$$\mathbb{E}[R(T)] \geq \Omega\left(d^{\frac{L}{2}-2}\max\{\sqrt{T}, d^2 C\}\right). \quad (53)$$

This corollary shows both the lower bounds depend on $\sqrt{T}$, and that the regret bound of our proposed algorithm is tight in terms of T and C. Besides these, however, the achievable and lower bounds have a gap due to mismatching dependence on the number of nodes N and the exact order of the exponential scaling of dimension d with graph length L. The dependence on N arises in the techniques used in Theorem 1 and we conjecture that the dependency of the upper bound on N can be diminishing as T grows. We provide some insights into tightening dependence on N. Let L_i denote the length of the longest causal path that ends at node $i \in [N]$. If we can first bound the cumulative estimation error for the expected value of node i with $L_i = 1$, then we can use induction to bound that for increasing L_i and derive a regret upper bound independent of N. The mismatch in the exact order d with L exists in all the relevant literature, even in simpler settings. For instance, consider linear bandits with time-invariant models with dimension d, which can be considered a special case of our linear causal bandit model by setting $L = 1$ and no model variations. For the widely used optimism in the face of uncertainty linear bandit algorithm (OFUL) in [36], the lower and upper bounds behave according to $\tilde{\mathcal{O}}(\sqrt{dT})$ and $\tilde{\mathcal{O}}(d\sqrt{T})$, respectively. This gap in terms of $\sqrt{d}$ and d matches exactly our gap.

C. Comparison With the Time-Invariant Setting

To highlight the robustness, we compare the analytical results with those of the time-invariant setting [15]. While our results in terms of the deviation budgets are general, for illustration purposes, we consider deviation budgets that scale sub-linearly with respect to T by setting $C = T^\alpha$ for $\alpha \in (0, 1)$.

Confidence ellipsoids: In the time-invariant setting, the confidence ellipsoid radius is set to

$$\beta_T' \triangleq 1 + \sqrt{2\log(2NT) + d\log\left(1 + mT^2/d\right)}. \quad (54)$$

Compared with this, our choice of confidence radius is similar to the time-invariant setting. The main contrast is the inclusion of an additional term m, which accounts for the cost of robustness. The difference in logarithmic terms arises from using different weights and norms in our algorithm.

Regret bounds: In the time-invariant setting, the regret bound scales as [15]

$$\mathbb{E}[R(T)] \leq 2m + \tilde{\mathcal{O}}\left(\beta_T'^L d^{\frac{L}{2}}\sqrt{NT}\right). \quad (55)$$

If we directly apply the algorithm designed for the time-invariant setting to the model fluctuation setting with proper adjustments, we observed it would exhibit model deviation robustness but only for substantially small deviation levels. Specifically, to make the time-invariant algorithm robust, we need to adjust the term β_T' so that it scales linearly with C (see [39, Appendix D] for details). This, in turn, induces a term C^L in the regret bound. Consequently, to preserve a sub-linear regret growth T, α need to fall in the interval $(0, \frac{1}{2L})$, which is a highly restrictive regime of model deviations, especially for graphs with long directed paths. For instance, under the DF measure, for $T = 10^5$ and a graph with $L = 3$, a node cannot be compromised more than 6 samples to maintain sub-linear regret. **This deviation level can be well below the noise level**
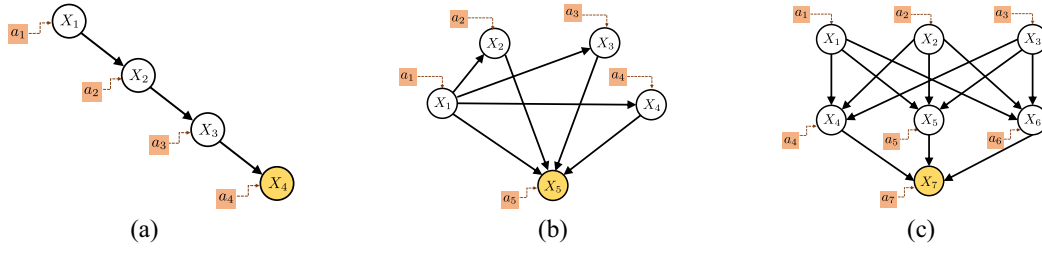
Fig. 2.   (a) Chain graph with $N = 4$. (b) Confounded parallel graph $N = 5$. (c) Hierarchical graph with $d = 3$ and $L = 2$.

of the model, establishing that the algorithms designed for the time-invariant setting lack robustness.

On the other hand, our Robust-LCB algorithm achieves a regret scaling of $T^{\max\{\frac{1}{2}, \alpha\}}$, which preserves the optimal rate $\tilde{\mathcal{O}}(\sqrt{T})$ under the regime $\alpha \leq \frac{1}{2}$, which is independent of other parameters. Furthermore, a non-linear growth of regret in T is achieved for deviation with $\alpha \in (\frac{1}{2}, 1)$, which is a significant improvement over the regime $(0, \frac{1}{2L})$. In the context of the earlier example with $T = 10^5$, this indicates that the algorithm remains $\tilde{\mathcal{O}}(\sqrt{T})$ regret when the model faces $C \leq 316$ outlier samples and achieves sub-linear regret later on. The cost incurred for integrating robustness is reflected in the additional low-order terms we have in our regret bounds. We also remark that by setting $C = 1$, our regret bound recovers the order of the time-invariant setting.

VI. EMPIRICAL EVALUATIONS

In this section, we assess the robustness of the Robust-LCB algorithm and its scaling behavior with the graph parameters. To the best of our knowledge, there is no baseline CB algorithm that can be used as a natural baseline for performance comparisons. Furthermore, soft interventions on continuous variables of a CB model are implemented by only LinSEM-UCB of [15]. Therefore, to assess the robustness, we compare our Robust-LCB algorithm with LinSEM-UCB and the standard non-causal UCB algorithm.

Parameter setting: We consider three types of graph: chain graph (Figure 2(a)), confounded parallel graph (Figure 2(b)) and hierarchical graph (Figure 2(c)). The noise terms are uniformly sampled from $[0, 2]$. We consider the general setting that all nodes that have parents can suffer from model deviations. The norm of both observational weights $\|[\mathbf{B}]_i\|$ and the interventional weights $\|[\mathbf{B}^*]_i\|$ are set to 0.5 and 1 respectively.

- *Chain graph:* The chain graph in Figure 2(a) is a fundamental element of causal graphs. The observational weights $[\mathbf{B}]_{i-1,i}$ and interventional weights $[\mathbf{B}^*]_{i-1,i}$ for node $i \in [N]$ are set to 0.5 and 1, respectively.
- *Hierarchical graph:* For the experiments reported in Figure 3, we use the graph structure with $L = 2$ layers, and each layer has the same number of nodes $d \in \{1, 2, 3, 4, 5\}$. For the experiments illustrated in Figures 4(c) and 5(c), we set the number of nodes in Layer 2 to 3 and Layer 1 to 9. We set the observational and interventional weights for node $i \in [N]$ to $0.5/\sqrt{|\mathsf{pa}(i)|}$ and $1/\sqrt{|\mathsf{pa}(i)|}$, respectively.
- *Confounded parallel graph:* The confounded parallel graph, as shown in Figure 2(b), is a mixture of the parallel
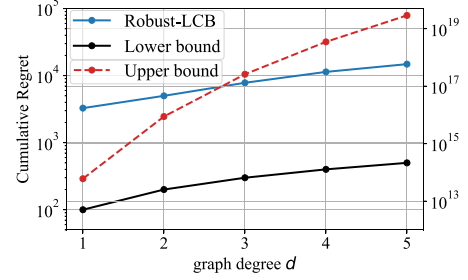

Fig. 3.   Cumulative regret of Robust-LCB with different graph degree d and $L = 2$, $N = Ld + 1$.

graph and confounded graph in [8], where node 1 is the parent of all other nodes, and the reward node N is the child of all other nodes. We set the observational and the interventional weights for nodes $i \in \{2, \ldots, N-1\}$ to 0.5 and 1, respectively. For the reward node N, its parents' observational and interventional weights are set to $0.5/\sqrt{N-1}$ and $1/\sqrt{N-1}$, respectively.

We let the deviations on the model occur at earlier rounds to simulate the worst-case scenario for a given deviation level C. When a deviation occurs on node $i \in [N]$, the weights are deliberately altered to change the optimal action, thereby challenging the algorithm's performance. The simulations are repeated 100 times, and the average cumulative regret is reported.

Scaling behavior with degree d: Figure 3 illustrates the variations of the cumulative regret $\mathbb{E}[R(T)]$ versus the graph degree d when $T = 40000$. We compare the regret of the Robust-LCB algorithm (blue curve with its scale on the left axis), the lower bound characterized in Section V-B (black curve with its scale on the left axis), and the upper bound specified in (46) (red curve with its scale on the right axis). All three curves suggest a polynomial behavior in d, which conforms to our theoretical results.

Comparison of the bounds: Figure 4 compares the cumulative regret of Robust-LCB with that of LinSEM-UCB and UCB under a model deviation level of $C = T^{0.5}$. It demonstrates that only Robust-LCB achieves sub-linear regret, whereas the other two algorithms incur linear regret with respect to the horizon T. Furthermore, it is noteworthy that the LinSEM-UCB always exhibits nearly the worst possible regret as the design of the deviation also showcases the worst case for LinSEM-UCB. In contrast, the regret of UCB tends to be the worst possible outcome as the graph's complexity increases even when the deviation is not designed for it. Nevertheless, these findings imply that the estimators of these algorithms
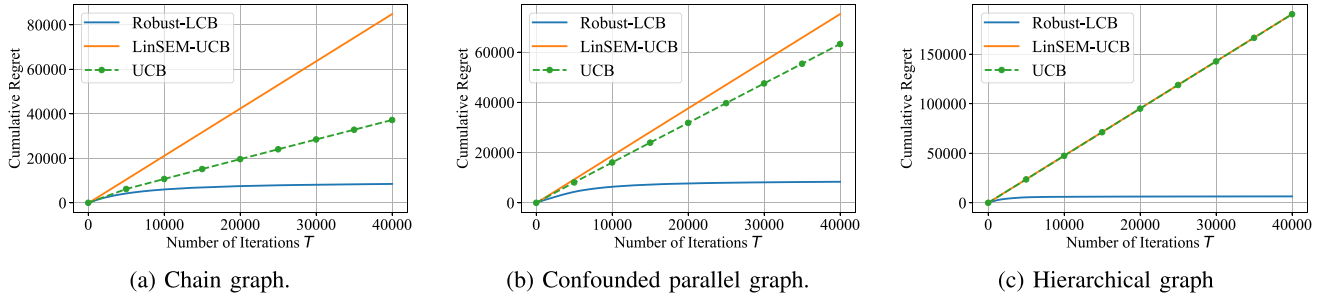
(a) Chain graph.   (b) Confounded parallel graph.   (c) Hierarchical graph

Fig. 4.   Cumulative regret when $C = \sqrt{T}$ for different algorithms.

(a) Chain graph.   (b) Confounded parallel graph.   (c) Hierarchical graph.
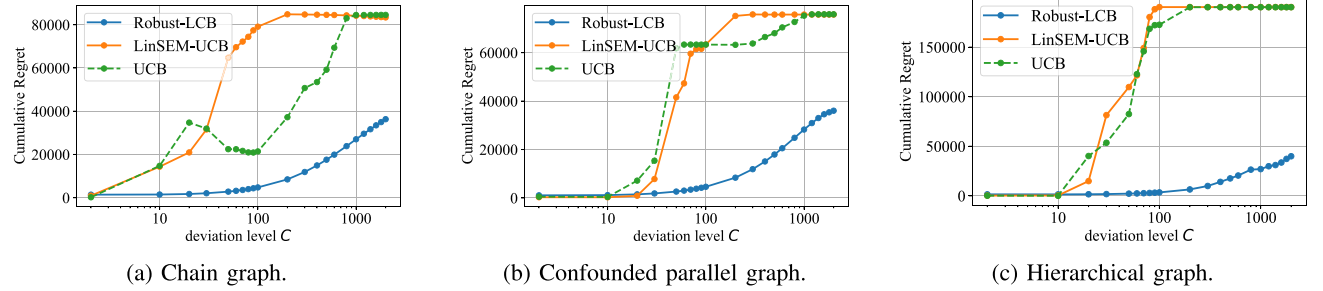
Fig. 5.   Cumulative regret at $T = 40000$ for different deviation levels C.

become ineffective when faced with such deviations, resulting in the selection of a sub-optimal (possibly the worst) arm.

Robustness against C: Figure 5 plots the cumulative regret at $T = 40000$ when the model deviation budget C changes from 2 to 2000. We observe that Robust-LCB and UCB perform slightly better than our algorithm when there is almost no deviation. This can be viewed as the compromise needed to guarantee the robustness of our algorithm. However, both Robust-LCB and UCB begin to fail with even a minor model deviation, as small as $C = 15$ for all three scenarios. Furthermore, they tend to reach nearly the worst possible regret when the deviation level rises high enough (e.g., 200 for the hierarchical graph). In the chain graph scenario, the UCB algorithm exhibits moderate regret when $C \leq 300$. However, as demonstrated in Figure 4(a), it mistakenly treats a non-optimal arm as the optimal one, resulting in linear regret, which happens when $C \geq 10$. In comparison, Robust-LCB outperforms when the deviation is more than negligible, as its regret scales sub-linear with the deviation level C.

VII. CONCLUDING REMARKS

In this paper, we have studied the sequential design of interventions over graphical causal models where both the observational and interventional models are unknown and undergo temporal variations. We have considered the general soft intervention model and have designed an algorithm for identifying the intervention mechanism that optimizes a utility function over the graph while exhibiting robustness against model variations. This objective has been naturally cast and analyzed as a causal bandit problem. We have focused on causal structures described by linear structural equation models. We have analyzed the proposed algorithm from a cumulative regret perspective, where we have characterized the dependence of the regret on the graph parameters. The main

observations are the following. (i) The proposed algorithm maintains sub-linear or nearly optimal regret under a wide range of relevant model deviation measures. This contrasts sharply with the existing algorithms designed for fixed models, which lose their sub-linear regret rate with a minimum level of model variations. (ii) While the cardinality of the intervention space grows exponentially with the graph size, our regret maintains a linear growth in the graph size. (iii) The regret bound depends on the graph structure only through its parameters (maximum degree and the length of the longest causal path). Finally, we have established an information-theoretic lower bound to demonstrate the tightness of our upper bound.

We conclude by providing some potential future directions. The first direction is tightening the gap between the graph-dependent parameters in the upper and lower bounds. Secondly, an alternative model to consider is the setting is the case of non-stationary bandits, in which the temporal variations can be potentially permanent. In such settings, the relevant metric to analyze is the dynamic regret specifically designed for non-stationary bandits. This metric becomes relevant when assuming the absence of a nominal model and the permanent effects of deviations. Lastly, we expect that our insights in this paper, including the effectiveness of weighted exploration bonus, can be extended into more general causal bandit settings, as well as other structured bandit settings involving model deviation. Examples include extending it to causal bandits with general link functions and combinatorial bandits.

APPENDIX A
ADDITIONAL NOTATIONS

First, we provide notations that are useful in our analyses. Since we are dealing with matrices, we denote the singular

values of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, where $M \geq N$, by

$$\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_N(\mathbf{A}). \tag{56}$$

In the proof, we often work with zero-padded vectors and corresponding matrices. 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 *effective* largest and smallest eigenvalues that correspond to effective dimensions of a positive semidefinite matrix $\mathbf{A}$ with rank k by

$$\sigma_{\max}(\mathbf{A}) \triangleq \sigma_1(\mathbf{A}), \quad \text{and} \quad \sigma_{\min}(\mathbf{A}) \triangleq \sigma_k(\mathbf{A}). \tag{57}$$

For a square matrix $\mathbf{U} = \mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{N \times N}$, we denote the *effective* largest and smallest eigenvalues by[1]

$$\lambda_{\max}(\mathbf{U}) \triangleq \lambda_{\max}\left(\mathbf{A}\mathbf{A}^\top\right) = \sigma_{\max}^2(\mathbf{A}), \tag{58}$$

$$\text{and} \quad \lambda_{\min}(\mathbf{U}) \triangleq \lambda_{\min}\left(\mathbf{A}\mathbf{A}^\top\right) = \sigma_{\min}^2(\mathbf{A}). \tag{59}$$

Then we construct data matrices that are highly related to Gram matrices. At time $t \in \mathbb{N}$ and for any node $i \in [N]$, the data matrices $\mathbf{U}_i(t) \in \mathbb{R}^{t \times N}$ and $\mathbf{U}_i^*(t) \in \mathbb{R}^{t \times N}$ consist of the weighted observational and interventional data, respectively. Specifically, for any $s \in [t]$ and $i \in [N]$, we define

$$\left[\mathbf{U}_i^\top(t)\right]_s \triangleq \mathbb{1}\{i \notin a(s)\}\sqrt{w_i(s)}\mathbf{X}_{\mathsf{pa}(i)}^\top(s), \tag{60}$$

$$\text{and} \quad \left[\mathbf{U}_i^{*\top}(t)\right]_s \triangleq \mathbb{1}\{i \in a(s)\}\sqrt{w_i(s)}\mathbf{X}_{\mathsf{pa}(i)}^\top(s). \tag{61}$$

Similarly to (2), we denote the relevant data matrices for node $i \in [N]$ under intervention $a \in \mathcal{A}$ by

$$\mathbf{U}_{i,a}(t) \triangleq \mathbb{1}\{i \in a\}\mathbf{U}_i^*(t) + \mathbb{1}\{i \notin a\}\mathbf{U}_i(t) \tag{62}$$

$$\mathbf{V}_{i,a}(t) \triangleq \mathbb{1}\{i \in a\}\mathbf{V}_i^*(t) + \mathbb{1}\{i \notin a\}\mathbf{V}_i(t). \tag{63}$$

Combining (60) and (63), we have

$$\mathbf{V}_{i,a}(t) = \mathbf{U}_{i,a}^\top(t)\mathbf{U}_{i,a}(t) + \mathbf{I}_N. \tag{64}$$

Similarly we define the data matrices that are related to $\widetilde{\mathbf{V}}_{i,a}(t)$ as

$$\left[\widetilde{\mathbf{U}}_i^\top(t)\right]_s \triangleq \mathbb{1}\{i \notin a(s)\}w_i(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s), \tag{65}$$

$$\text{and} \quad \left[\widetilde{\mathbf{U}}_i^{*\top}(t)\right]_s \triangleq \mathbb{1}\{i \in a(s)\}w_i(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s). \tag{66}$$

The relevant data matrices for node $i \in [N]$ under intervention $a \in \mathcal{A}$ are

$$\widetilde{\mathbf{U}}_{i,a}(t) \triangleq \mathbb{1}\{i \in A\}\mathbf{U}_i^*(t) + \mathbb{1}\{i \notin A\}\mathbf{U}_i(t), \tag{67}$$

$$\text{and} \quad \widetilde{\mathbf{V}}_{i,a}(t) = \widetilde{\mathbf{U}}_{i,a}^\top(t)\widetilde{\mathbf{U}}_{i,a}(t) + \mathbf{I}_N. \tag{68}$$

Define $N_i^*(t)$ as the number of times that node $i \in [N]$ is intervened, and $N_i(t)$ as its complement, i.e.,

$$N_i^*(t) \triangleq \sum_{s=1}^t \mathbb{1}\{i \in a(s)\}, \quad \text{and} \quad N_i(t) \triangleq t - N_i^*(t). \tag{69}$$

Accordingly, for any $i \in [N]$ and $t \in \mathbb{N}$, define

$$N_{i,a}(t) \triangleq \mathbb{1}\{i \in a\}N_i^*(t) + \mathbb{1}\{i \notin a\}N_i(t), \tag{70}$$

[1]For matrix $\mathbf{V} = \mathbf{U} + \mathbf{I}$, we denote the *effective* smallest eigenvalues by $\lambda_{\min}(\mathbf{V}) \triangleq \sigma_{\min}^2(\mathbf{A}) + 1$.

Furthermore, we define the error and its *l*-th power of the estimator of $\mathbf{B}$ matrices as

$$\Delta_a(t) \triangleq \mathbf{B}_a(t) - \mathbf{B}_a, \quad \forall i \in [N], \tag{71}$$

$$\text{and} \quad \Delta_a^{(\ell)}(t) \triangleq \mathbf{B}_a^\ell(t) - \mathbf{B}_a^\ell. \tag{72}$$

APPENDIX B
PROOF OF LEMMA 4

Proof: When it is clear from context, we use the shorthand terms $\mathbf{B}$ for $\mathbf{B}_a$ in this proof. According to the given definitions $\Delta_{\mathbf{A}}^{(\ell)} = [\Delta_{\mathbf{A}} + \mathbf{B}]^\ell - \mathbf{B}^\ell$, each term in the binomial expansion of $\Delta_{\mathbf{A}}^{(\ell)}$ can be represented as a product involving factors of $\Delta_{\mathbf{A}}$ and $\mathbf{B}$. For any $\ell \in [L]$ and $k \in [\ell] \cup \{0\}$, there exist $\binom{\ell}{k}$ terms that consist of the $\Delta_{\mathbf{A}}$ factor appearing k times and the $\mathbf{B}$ factor appearing $(\ell - k)$ times. We denote the set of these product terms by

$$\mathcal{H}_{\ell,k} \triangleq \{\mathbf{H} : \mathbf{H} \text{ has } \Delta_{\mathbf{A}} \ k \text{ times and } \mathbf{B} \ \ell - k \text{ times}\}. \tag{73}$$

Therefore, sets $\mathcal{H}_{\ell,1}, \ldots, \mathcal{H}_{\ell,\ell}$ contain all valid products composed of $\mathbf{B}$ and $\Delta_{\mathbf{A}}$. Hence, we can write the expansion of $\Delta_{\mathbf{A}}^{(\ell)}$ as

$$\Delta_{\mathbf{A}}^{(\ell)} = \sum_{k=1}^\ell \sum_{\mathbf{H} \in \mathcal{H}_{\ell,k}} \mathbf{H}. \tag{74}$$

To bound the norm of $\Delta_{\mathbf{A}}^{(\ell)}$, we first bound the norm of each element in the summation. We show by induction that for any $\ell \in [L]$, $k \in [\ell] \cup \{0\}$, and $\mathbf{H} \in \mathcal{H}_{\ell,k}$,

$$\|[\mathbf{H}]_i\| \leq d^{\frac{\ell-1}{2}} \beta^k \lambda^k \ \forall i \in [N], \tag{75}$$

where $\lambda \triangleq \max_{i \in [N]} \lambda_{\min}^{-1/2}(\mathbf{M}_i)$. Consider $\ell = 1$. For $k = 0$, we have $\mathcal{H}_{1,0} = \{\mathbf{B}\}$ and $\|[\mathbf{B}]_i\| \leq 1$. For $k = 1$, $\mathcal{H}_{1,1} = \{\Delta_{\mathbf{A}}\}$, and

$$\|[\Delta_{\mathbf{A}}]_i\| \leq \|[\Delta_{\mathbf{A}}]_i \odot \mathbf{1}(\mathsf{pa}(i))\|_{\mathbf{M}_i} \lambda_{\min}^{-1/2}(\mathbf{M}_i) \tag{76}$$

$$= \|[\Delta_{\mathbf{A}}]_i\|_{\mathbf{M}_i} \lambda_{\min}^{-1/2}(\mathbf{M}_i) \leq \beta\lambda. \tag{77}$$

Therefore, (75) holds for $\ell = 1$. Now assume that (75) holds for every $1, \ldots, \ell - 1$, for some $\ell \geq 2$. Consider a product term $\mathbf{H} \in \mathcal{H}_{\ell,k}$, for some $k \in [\ell] \cup \{0\}$. The first factor of $\mathbf{H}$ is either $\mathbf{B}_a$ or $\Delta_{\mathbf{A}}$, and we analyze the induction step for each of these possibilities separately.

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

$$\|[\mathbf{H}]_i\|^2 = \left\|\left(\mathbf{B}\bar{\mathbf{H}}\right)_i\right\|^2 = \sum_{u,v \in \mathsf{pa}(i)} [\mathbf{B}]_{u,i}[\mathbf{B}]_{v,i}\bar{\mathbf{H}}_v^\top \bar{\mathbf{H}}_u \tag{78}$$

$$\leq d \sum_{u \in \mathsf{pa}(i)} \left([\mathbf{B}]_{u,i}\right)^2 \left\|\bar{\mathbf{H}}_u\right\|^2 \tag{79}$$

$$\leq d^{\ell-1}\beta^{2k}\lambda^2 \sum_{u \in \mathsf{pa}(i)} \left([\mathbf{B}]_{u,i}\right)^2 \tag{80}$$

$$= d^{\ell-1}\beta^{2k}\lambda^{2k}\underbrace{\|[\mathbf{B}]_i\|^2}_{\leq 1} \leq d^{\ell-1}\beta^{2k}\lambda^{2k}, \tag{81}$$

where in (79) we use Cauchy–Schwarz inequality and the inductive hypothesis (75).

Case 2: If $\mathbf{H}$ starts with $\Delta_{\mathbf{A}}$ represent it by $\mathbf{H} = \Delta_{\mathbf{A}}\bar{\mathbf{H}}$, where $\bar{\mathbf{H}} \in \mathcal{H}_{\ell-1,k-1}$ and $k \in [\ell]$. Similarly to the first case, we have

$$\|[\mathbf{H}]_i\|^2 = \|[\Delta_{\mathbf{A}}\bar{\mathbf{H}}]_i\|^2 \tag{82}$$

$$= \sum_{u,v \in \mathsf{pa}(i)} [\Delta_{\mathbf{A}}]_{u,i}[\Delta_{\mathbf{A}}]_{v,i}[\bar{\mathbf{H}}]_v^\top[\bar{\mathbf{H}}]_u \tag{83}$$

$$\leq d \sum_{u \in \mathsf{pa}(i)} ([\Delta_{\mathbf{A}}]_{u,i})^2 \|[\bar{\mathbf{H}}]_u\|^2 \tag{84}$$

$$\leq d^{\ell-1}\beta^{2k-2}\lambda^{2k-2} \sum_{u \in \mathsf{pa}(i)} ([\Delta_{\mathbf{A}}]_{u,i})^2 \tag{85}$$

$$= d^{\ell-1}\beta^{2k-2}\lambda^{2k-2}\|[\Delta_{\mathbf{A}}]_i\|^2 \tag{86}$$

$$\leq d^{\ell-1}\beta^{2k-2}\lambda^{2k-2}\beta^2\underbrace{\lambda_{\min}^{-1}(\mathbf{M}_i)}_{\leq \lambda^2} \tag{87}$$

$$\leq d^{\ell-1}\beta^{2k}\lambda^{2k}. \tag{88}$$

where (84) we use Cauchy-Schwarz inequality, and in (87) we use the fact that $[\Delta_{\mathbf{A}}]_i = [\Delta_{\mathbf{A}}]_i \odot \mathbf{1}(\mathsf{pa}(i))$ and $\|[\Delta_{\mathbf{A}}]_i\| \leq \|[\Delta_{\mathbf{A}}]_i\|_{\mathbf{M}_i}\lambda_{\min}^{-1/2}(\mathbf{M}_i)$. Taking the square-roots of both sides in (81) and (88) yields

$$\|[\mathbf{H}]_i\| \leq d^{\frac{\ell-1}{2}}\beta^k\lambda^k, \tag{89}$$

which is the desired inequality for all $k \in [\ell] \cup \{0\}$. This completes the proof of (75) by induction. The final result follows by applying (89) to each term in the sum (74).

$$\left\|[\Delta_{\mathbf{A}}^{(\ell)}]_i\right\| = \left\|\sum_{k=1}^{\ell}\sum_{\mathbf{H} \in \mathcal{H}_{\ell,k}}[\mathbf{H}]_i\right\| \tag{90}$$

$$\leq \sum_{k=1}^{\ell}\sum_{\mathbf{H} \in \mathcal{H}_{\ell,k}}\|[\mathbf{H}]_i\| \tag{91}$$

$$\leq d^{\frac{\ell-1}{2}}\sum_{k=1}^{\ell}|\mathcal{H}_{\ell,k}|\beta^k\lambda^k \tag{92}$$

$$= d^{\frac{\ell-1}{2}}\sum_{k=1}^{\ell}\binom{\ell}{k}\beta^k\lambda^k \tag{93}$$

$$< d^{\frac{\ell-1}{2}}(\beta+1)^{\ell}\lambda \tag{94}$$

$$= d^{\frac{\ell-1}{2}}(\beta+1)^{\ell}\max_{i \in [N]}\lambda_{\min}^{-1/2}(\mathbf{M}_i), \tag{95}$$

where (94) is due to the binomial expansion of $(\beta+1)^{\ell}$ and $\mathbf{M}_i \succeq \mathbf{I}$, $\forall i \in [N]$ such that $\lambda \leq 1$. ∎

APPENDIX C
PROOF OF THEOREM 1

The proof mechanism follows the same line of arguments as [15, Th. 5] but with major distinctions. Firstly, we provide a new lemma on the bound on the power of estimation error $\Delta_a^{(\ell)}(t) \triangleq \mathbf{B}_a^{\ell}(t-1) - \mathbf{B}_a^{\ell}$, Furthermore, the effect of the weighted ridge regression, distinct confidence ellipsoids, and the definition of error events, namely $\widetilde{\mathcal{E}}_{i,n}(t)$ and $\widetilde{\mathcal{E}}_{i,n}^*(t)$ are investigated. Finally, we bound a new function of eigenvalues of weighted

Gram matrices $\mathbf{V}_{i,a(t)}(t)$ and $\widetilde{\mathbf{V}}_{i,a(t)}$. To begin with, we first state the lemma which upper bounds the estimation error.

Based on the above estimation error, we start the proof by decomposing the regret in (4) as

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{a^*} - \mu_{a(t)})\right] \tag{96}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \boldsymbol{\epsilon}(t)\rangle\right] \tag{97}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \boldsymbol{\nu}\rangle\right], \tag{98}$$

where the second equation is a result of [15, Lemma 1], which is stated as Lemma 1. The last equation is due to the inner product being a linear function and $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\nu}$.

Now define the error events $\mathcal{E}_i$ and $\mathcal{E}_i^*$ for $i \in [N]$ for each estimator

$$\mathcal{E}_i \triangleq \Big\{\forall t \in [T]:$$
$$\|[\mathbf{B}(t-1)]_i - [\mathbf{B}]_i\|_{\mathbf{V}_i(t)[\widetilde{\mathbf{V}}_i(t)]^{-1}\mathbf{V}_i(t)} \leq \beta_t\Big\}, \tag{99}$$

$$\mathcal{E}_i^* \triangleq \Big\{\forall t \in [T]:$$
$$\|[\mathbf{B}^*(t-1)]_i - [\mathbf{B}^*]_i\|_{\mathbf{V}_i^*(t)[\widetilde{\mathbf{V}}_i^*(t)]^{-1}\mathbf{V}_i^*(t)} \leq \beta_t\Big\}, \tag{100}$$

where the β_t is chosen as

$$\beta_t \triangleq \sqrt{2\log(2NT) + d\log(1 + m^2t/dC^2)} + 1 + m, \tag{101}$$

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.,

$$\mathcal{E}_\cap \triangleq \Big(\bigcap_{i=1}^{N}\mathcal{E}_i\Big)\bigcap\Big(\bigcap_{i=1}^{N}\mathcal{E}_i^*\Big). \tag{102}$$

Then by Lemma 2, we have

$$\mathbb{P}(\mathcal{E}_\cap^c) \leq \sum_{i=1}^{N}\mathbb{P}(\mathcal{E}_i^c) + \sum_{i=1}^{N}\mathbb{P}(\mathcal{E}_i^{*c}) \tag{103}$$

$$\leq \sum_{i=1}^{N}\Big(\frac{1}{2NT} + \frac{1}{2NT}\Big) = \frac{1}{T}. \tag{104}$$

Then we can bound the regret as follows.

$$R(T) \leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap^c\}\sum_{t=1}^{T}\langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \boldsymbol{\nu}\rangle\right] \tag{105}$$

$$+ \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\}\sum_{t=1}^{T}r(t)\right] \tag{106}$$

$$\leq 2mT\mathbb{P}(\mathcal{E}_\cap^c) + \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\}\sum_{t=1}^{T}r(t)\right] \tag{107}$$

$$\leq 2m + \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\}\sum_{t=1}^{T}\langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \boldsymbol{\nu}\rangle\right]. \tag{108}$$

According to the arm selection rules, we have $\text{UCB}_{a^*}(t) \leq \text{UCB}_{a(t)}(t)$. Let $\widetilde{\mathbf{B}}_a$ be the matrix that attains $\text{UCB}_a(t)$, i.e.,

$$\widetilde{\mathbf{B}}_a = \underset{[\Theta]_i \in \mathcal{C}_{i,a}(t)}{\arg\max} \langle f(\Theta), \, \boldsymbol{v} \rangle. \tag{109}$$

Then under the event $\mathcal{E}_\cap$, we have

$$\langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \tag{110}$$

$$\leq \text{UCB}_{a^*}(t) - \langle f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \tag{111}$$

$$\leq \text{UCB}_{a(t)}(t) - \langle f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \tag{112}$$

$$= \langle f(\widetilde{\mathbf{B}}_{a(t)}) - f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle. \tag{113}$$

Subsequently, we use (113) to derive an upper bound for the expectation term in (108), thereby eliminating the dependence on a^*.

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \right] \tag{114}$$

$$\leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \langle f(\widetilde{\mathbf{B}}_{a(t)}) - f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \right] \tag{115}$$

$$\leq \|\boldsymbol{v}\| \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \| f(\widetilde{\mathbf{B}}_{a(t)}) - f(\mathbf{B}_{a(t)}) \| \right] \tag{116}$$

$$\leq \|\boldsymbol{v}\| \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \| f(\widetilde{\mathbf{B}}_{a(t)}) - f(\mathbf{B}_{a(t)}(t)) \| \right]$$

$$+ \|\boldsymbol{v}\| \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \| f(\widetilde{\mathbf{B}}_{a(t)}(t)) - f(\mathbf{B}_{a(t)}) \| \right]. \tag{117}$$

Note that (117) follows from the Cauchy-Schwarz inequality while (117) is due to the triangle inequality. Next, we examine the norm term within the expectation expression in (117). By applying the definition of f, we obtain

$$\left\| f(\widetilde{\mathbf{B}}_{a(t)}) - f(\mathbf{B}_{a(t)}(t)) \right\|$$

$$\leq \sum_{\ell=1}^L \left\| [\widetilde{\mathbf{B}}_{a(t)}^\ell]_N - [\mathbf{B}_{a(t)}^\ell(t)]_N \right\|, \tag{118}$$

$$\left\| f(\widetilde{\mathbf{B}}_{a(t)}(t)) - f(\mathbf{B}_{a(t)}) \right\|$$

$$\leq \sum_{\ell=1}^L \left\| [\mathbf{B}_{a(t)}^\ell(t)]_N - [\mathbf{B}_{a(t)}^\ell]_N \right\|. \tag{119}$$

By the definition of Gram matrices, we have

$$\widetilde{\mathbf{V}}_{i,a(t)}(t) \preceq \mathbf{V}_{i,a(t)}(t), \tag{120}$$

$$\mathbf{V}_{i,a(t)}(t) [\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1} \mathbf{V}_{i,a(t)}(t) \succeq \mathbf{V}_{i,a(t)}(t) \succeq \mathbf{I}_N. \tag{121}$$

We consider the following tuple of matrices for $i \in [N]$

$$\left(\widetilde{\mathbf{B}}_{a(t)}, \mathbf{B}_{a(t)}(t), \mathbf{V}_{i,a(t)}(t) [\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1} \mathbf{V}_{i,a(t)}(t) \right),$$

$$\left(\mathbf{B}_{a(t)}(t), \mathbf{B}_{a(t)}, \mathbf{V}_{i,a(t)}(t) [\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1} \mathbf{V}_{i,a(t)}(t) \right). \tag{122}$$

These tuple of matrices satisfy the condition of Lemma 4. We further define the maximum confidence radius $\beta_T \triangleq \sqrt{2\log(2NT) + d\log(1 + m^2 T/dC^2)} + 1 + m$ and

$$\lambda(t) \triangleq \max_{i \in [N]} \frac{\sqrt{\lambda_{\max}(\widetilde{\mathbf{V}}_{i,a(t)}(t))}}{\lambda_{\min}(\mathbf{V}_{i,a(t)}(t))}, \tag{123}$$

$$\lambda_T \triangleq \mathbb{E}\left[\sum_{t=1}^T \lambda(t) \right]. \tag{124}$$

By using Lemma 4 to upper bound the tow terms in (119), we have

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \langle f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a(t)}), \, \boldsymbol{v} \rangle \right] \tag{125}$$

$$< 2 \sum_{\ell=1}^L d^{\frac{\ell-1}{2}} (\beta_T + 1)^\ell \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cap\} \sum_{t=1}^T \lambda(t) \right] \tag{126}$$

$$\leq 2\lambda_T \sum_{\ell=1}^L d^{\frac{\ell-1}{2}} (\beta_T + 1)^\ell \tag{127}$$

$$\leq 4\lambda_T (\beta_T + 1)^L d^{\frac{L-1}{2}}, \tag{128}$$

where in (128), we use the fact that $\sum_{i=1}^L q^L \leq 2q^L$ for $q \geq 2$ and $\sqrt{d}(\beta_T + 1) > 2$.

a) Bounding $\mathbb{E}[\sum_{i=1}^T \lambda(t)]$: What remains is to bound the term $\mathbb{E}[\sum_{i=1}^T \lambda(t)]$, where $\lambda(t)$ is a function involving the eigenvalues of both Gram matrices $\mathbf{V}_{i,a(t)}(t)$ and $\widetilde{\mathbf{V}}_{i,a(t)}(t)$. To proceed, we define the second-moment matrices and its *effective* largest and smallest eigenvalues as

$$\Sigma_{i,a}(t) \triangleq \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_a}\left[\mathbf{X}_{\text{pa}(i)}(t) \mathbf{X}_{\text{pa}(i)}^\top(t) \right], \tag{129}$$

$$\kappa_{\min} \triangleq \min_{i \in [N], a \in \mathcal{A}, t \in [T]} \sigma_{\min}(\Sigma_{i,a}(t)), \tag{130}$$

$$\kappa_{\max} \triangleq \max_{i \in [N], a \in \mathcal{A}, t \in [T]} \sigma_{\max}(\Sigma_{i,a}(t)), \tag{131}$$

where $\kappa_{\min} > 0$ is guaranteed since there is no deterministic relation between nodes and their patients. These variables are inherent to the system and remain unknown to the learner. Given our focus on the weighted OLS estimator, we also introduce singular values related to auxiliary variables $\mathbf{X}'_{\text{pa}(i)}(t) \triangleq \sqrt{w_i(t)} \mathbf{X}_{\text{pa}(i)}(t)$ and $\widetilde{\mathbf{X}}_{\text{pa}(i)}(t) \triangleq w_i(t) \mathbf{X}_{\text{pa}(i)}(t)$. Accordingly, we define the second weighted moment matrices as follows.

$$\Sigma'_{i,a,w_i}(t) \triangleq \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_a}\left[w_i(t) \mathbf{X}_{\text{pa}(i)}(t) \mathbf{X}_{\text{pa}(i)}^\top(t) \right], \tag{132}$$

$$\widetilde{\Sigma}_{i,a,w_i}(t) \triangleq \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_a}\left[w_i^2(t) \mathbf{X}_{\text{pa}(i)}(t) \mathbf{X}_{\text{pa}(i)}^\top(t) \right]. \tag{133}$$

To bound the singular value of the weighted second moment, we first need uniform bounds for the weights. We find a bound for the norm of $\|\mathbf{X}_{\text{pa}(i)}(t)\|_{[\widetilde{\mathbf{V}}_{i,a}(t)]^{-1}}$ across all $a \in \mathcal{A}$. This yields the following result.

$$\|\mathbf{X}_{\text{pa}(i)}(t)\|_{[\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1}}$$

$$\leq \frac{1}{\lambda_{\min}^{1/2}(\widetilde{\mathbf{V}}_{i,a(t)}(t))} \|\mathbf{X}_{\text{pa}(i)}(t)\| \leq m. \tag{134}$$

Then, the weights can be bounded by

$$\frac{1}{Cm} \leq w_i(t) = \min\left\{ \frac{1}{C}, \frac{1}{C\|\mathbf{X}_{\text{pa}(i)}(t)\|_{[\widetilde{\mathbf{V}}_{i,a(t)}(t)]^{-1}}} \right\} \leq \frac{1}{C}. \tag{135}$$

Subsequently, we can bound the minimum and maximum singular values of matrices $\Sigma'_{i,a,w_i}(t)$ and $\widetilde{\Sigma}_{i,a,w_i}(t)$.

$$\kappa'_{\min} = \frac{1}{Cm} \kappa_{\min} \leq \sigma_{\min}(\Sigma'_{i,a,w_i}(t)) \tag{136}$$

$$\leq \sigma_{\max}\big(\Sigma'_{i,a,w_i}(t)\big) \leq \frac{1}{C}\kappa_{\max} = \kappa'_{\max}, \quad (137)$$

$$\tilde{\kappa}_{\min} = \frac{1}{C^2 m^2}\kappa_{\min} \leq \sigma_{\min}\big(\widetilde{\Sigma}_{i,a,w_i}(t)\big) \quad (138)$$

$$\leq \sigma_{\max}\big(\widetilde{\Sigma}_{i,a,w_i}(t)\big) \leq \frac{1}{C^2}\kappa_{\max} = \tilde{\kappa}_{\max}. \quad (139)$$

Moreover, we have $\|\mathbf{X}'_{\mathsf{pa}(i)}(t)\| \leq m' = \frac{1}{\sqrt{C}}m$ and $\|\widetilde{\mathbf{X}}_{\mathsf{pa}(i)}(t)\| \leq \tilde{m} = \frac{1}{C}m$. In order to proceed, we need upper and lower bounds for the maximum and minimum singular values of $\mathbf{U}_{i,a(t)}(t)$. However, these bounds depend on the number of non-zero rows of $\mathbf{U}_{i,a(t)}(t)$ matrices, which equals to values of the random variable $N_{i,a(t)}(t)$. Firstly, we define the weighted constants

$$\gamma_n \triangleq \max\Big\{\alpha m^2 \sqrt{n}, \alpha^2 m^2\Big\}, \quad (140)$$

$$\gamma'_n \triangleq \max\Big\{\alpha m'^2 \sqrt{n}, \alpha^2 m'^2\Big\}, \quad (141)$$

$$\tilde{\gamma}_n \triangleq \max\Big\{\alpha \tilde{m}^2 \sqrt{n}, \alpha^2 \tilde{m}^2\Big\}, \quad \forall n \in [T]. \quad (142)$$

Then for every $i \in [N]$, $t \in [T]$, and $n \in [t]$, we define the error events corresponding to the maximum and minimum singular values of $\mathbf{U}_i(t)$ and $\widetilde{\mathbf{U}}_i(t)$ as

$$\mathcal{E}_{i,n}(t) \triangleq \Big\{N_i(t) = n$$
$$\text{and } \Big\{\sigma_{\min}(\mathbf{U}_i(t)) \leq \sqrt{\max\{0, n\kappa'_{\min} - \gamma'_n\}}$$
$$\text{or } \sigma_{\max}(\mathbf{U}_i(t)) \geq \sqrt{n\kappa'_{\max} + \gamma'_n}\Big\}\Big\}, \quad (143)$$

$$\mathcal{E}^*_{i,n}(t) \triangleq \Big\{N^*_i(t) = n$$
$$\text{and } \Big\{\sigma_{\min}(\mathbf{U}^*_i(t)) \leq \sqrt{\max\{0, n\kappa'_{\min} - \gamma'_n\}}$$
$$\text{or } \sigma_{\max}(\mathbf{U}^*_i(t)) \geq \sqrt{n\kappa'_{\max} + \gamma'_n}\Big\}\Big\}, \quad (144)$$

$$\widetilde{\mathcal{E}}_{i,n}(t) \triangleq \Big\{N_i(t) = n$$
$$\text{and } \Big\{\sigma_{\min}(\widetilde{\mathbf{U}}_i(t)) \leq \sqrt{\max\{0, n\tilde{\kappa}_{\min} - \tilde{\gamma}_n\}}$$
$$\text{or } \sigma_{\max}(\widetilde{\mathbf{U}}_i(t)) \geq \sqrt{n\tilde{\kappa}_{\max} + \tilde{\gamma}_n}\Big\}\Big\}, \quad (145)$$

$$\widetilde{\mathcal{E}}^*_{i,n}(t) \triangleq \Big\{N^*_i(t) = n$$
$$\text{and } \Big\{\sigma_{\min}(\widetilde{\mathbf{U}}^*_i(t)) \leq \sqrt{\max\{0, n\tilde{\kappa}_{\min} - \tilde{\gamma}_n\}}$$
$$\text{or } \sigma_{\max}(\widetilde{\mathbf{U}}^*_i(t)) \geq \sqrt{n\tilde{\kappa}_{\max} + \tilde{\gamma}_n}\Big\}\Big\}. \quad (146)$$

In other words, the event $\mathcal{E}_{i,n}(t)$ indicates the situation in which either $\sigma_{\min}(\mathbf{U}_i(t))$ or $\sigma_{\max}(\mathbf{U}_i(t))$ violates the established lower and upper bounds. Likewise, $\mathcal{E}^*_{i,n}(t)$, $\widetilde{\mathcal{E}}_{i,n}(t)$ and $\widetilde{\mathcal{E}}^*_{i,n}(t)$ are associated with the singular values of $\mathbf{U}^*_i(t)$, $\widetilde{\mathbf{U}}_i(t)$ and $\widetilde{\mathbf{U}}^*_i(t)$. The next result shows that these events occur with a low probability.

Lemma 5: The probability of the error events $\mathcal{E}_{i,n}(t)$, $\mathcal{E}^*_{i,n}(t)$, $\widetilde{\mathcal{E}}_{i,n}(t)$ and $\widetilde{\mathcal{E}}^*_{i,n}(t)$ defined in (143) to (146) are upper bounded as

$$\mathbb{P}\big(\mathcal{E}_{i,n}(t)\big) \leq d \exp\Big(-\frac{3\alpha^2}{16}\Big), \ \mathbb{P}\big(\mathcal{E}^*_{i,n}(t)\big) \leq d \exp\Big(-\frac{3\alpha^2}{16}\Big), \quad (147)$$

$$\mathbb{P}\big(\widetilde{\mathcal{E}}_{i,n}(t)\big) \leq d \exp\Big(-\frac{3\alpha^2}{16}\Big), \ \mathbb{P}\big(\widetilde{\mathcal{E}}^*_{i,n}(t)\big) \leq d \exp\Big(-\frac{3\alpha^2}{16}\Big). \quad (148)$$

Proof: This is a direct result from [15, Lemma 3], where we use $\kappa'_{\min}, \kappa'_{\max}$ and γ' for events $\mathcal{E}_{i,n}(t)$ and $\mathcal{E}^*_{i,n}(t)$ and using $\tilde{\kappa}_{\min}, \tilde{\kappa}_{\max}$ and $\tilde{\gamma}$ for events $\widetilde{\mathcal{E}}_{i,n}(t)$ and $\widetilde{\mathcal{E}}^*_{i,n}(t)$. The only difference is that one need to use the following lemma, which is an immediate result of [15, Lemma 6].

Lemma 6: Consider matrices $\mathbf{U}$ and $\mathbf{A}$ that satisfy

$$\Big\|\mathbf{U}^\top\mathbf{U} - \mathbf{A}\Big\| \leq \gamma. \quad (149)$$

Then we have,

$$\sigma_{\max}(\mathbf{U}) \leq \sqrt{\sigma_{\max}(\mathbf{A}) + \gamma}, \quad (150)$$

$$\text{and} \quad \sigma_{\min}(\mathbf{U}) \geq \sqrt{\max\{0, \sigma_{\min}(\mathbf{A}) - \gamma\}}. \quad (151)$$

Now that we have bounds on the probability of error events, we define the union error event $\mathcal{E}_\cup$ as

$$\mathcal{E}_\cup \triangleq \{\exists \ (i, t, n) : i \in [N], t \in [T], n \in [t], \ \mathcal{E}_{i,n}(t)$$
$$\text{or } \mathcal{E}^*_{i,n}(t) \text{ or } \widetilde{\mathcal{E}}_{i,n}(t) \text{ or } \widetilde{\mathcal{E}}^*_{i,n}(t)\}. \quad (152)$$

By taking a union bound and using Lemma 5 we have

$$\mathbb{P}(\mathcal{E}_\cup) \leq \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{n=1}^{t}\Big(\mathbb{P}\big(\mathcal{E}_{i,n}(t)\big) + \mathbb{P}\big(\mathcal{E}^*_{i,n}(t)\big) \quad (153)$$

$$+ \mathbb{P}\big(\widetilde{\mathcal{E}}_{i,n}(t)\big) + \mathbb{P}\big(\widetilde{\mathcal{E}}^*_{i,n}(t)\big)\Big) \quad (154)$$

$$\leq 2NT(T+1)d \exp\Big(-\frac{3\alpha^2}{16}\Big). \quad (155)$$

b) Bounding $\mathbb{E}[\mathbb{1}\{\mathcal{E}_\cup\}\sum_{t=1}^{T}\lambda(t)]$:

Since $\lambda_{\min}\big(\mathbf{V}_{i,a(t)}(t)\big) \geq 1$, we have the following unconditional upper bound.

$$\lambda(t) = \max_{i \in [N]} \frac{\sqrt{\lambda_{\max}(\widetilde{\mathbf{V}}_{i,a(t)}(t))}}{\lambda_{\min}(\mathbf{V}_{i,a(t)}(t))} \quad (156)$$

$$\leq \max_{i \in [N]} \sqrt{\lambda_{\max}(\widetilde{\mathbf{V}}_{i,a(t)}(t))} \quad (157)$$

$$= \max_{i \in [N]} \sqrt{\lambda_{\max}\Big(\sum_{s=1}^{t}\mathbb{1}\{i \in a(s)\}w_i^2(s)\mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s) + \mathbf{I}_N\Big)} \quad (158)$$

$$\leq \max_{i \in [N]} \sqrt{1 + \frac{1}{C^2}\sum_{s=1}^{t}\mathbb{1}\{i \in a(s)\}\lambda_{\max}\Big(\mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s)\Big)} \quad (159)$$

$$\leq \max_{i \in [N]} \sqrt{1 + \frac{1}{C^2}\sum_{s=1}^{t}\lambda_{\max}\Big(\mathbf{X}_{\mathsf{pa}(i)}(s)\mathbf{X}_{\mathsf{pa}(i)}^\top(s)\Big)} \quad (160)$$

$$= \max_{i \in [N]} \sqrt{1 + \frac{1}{C^2}\sum_{s=1}^{t}\big\|\mathbf{X}_{\mathsf{pa}(i)}(s)\big\|^2} \quad (161)$$

$$\le \sqrt{\frac{m^2}{C^2}t + 1}, \tag{162}$$

where (162) follows from the fact that $\|\mathbf{X}\| \le m$. Hence, we have

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup\} \sum_{t=1}^{T} \lambda(t)\right] \tag{163}$$

$$\overset{(157)}{\le} \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup\} \sum_{t=1}^{T} \sqrt{\lambda_{\max}\big(\widetilde{\mathbf{V}}_{N,a(t)}(t)\big)}\right] \tag{164}$$

$$\overset{(162)}{\le} \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup\} \sum_{t=1}^{T} \sqrt{\frac{m^2}{C^2}t + 1}\right] \tag{165}$$

$$= \mathbb{E}\big[\mathbb{1}\{\mathcal{E}_\cup\}\big] \sum_{t=1}^{T} \sqrt{\frac{m^2}{C^2}t + 1} \tag{166}$$

$$= \mathbb{P}(\mathcal{E}_\cup) \underbrace{\sum_{t=1}^{T} \sqrt{\frac{m^2}{C^2}t + 1}}_{A_1}. \tag{167}$$

Furthermore, A_1 is bounded as

$$\sum_{t=1}^{T} \sqrt{\frac{m^2}{C^2}t + 1} \le \frac{m}{C}\sqrt{T} + 1 + \sum_{t=1}^{T-1}\left(\frac{m}{C}\sqrt{t} + 1\right) \tag{168}$$

$$\le \frac{m}{C}\sqrt{T} + T + \int_{t=1}^{T} \frac{m}{C}\sqrt{t}\,dt \tag{169}$$

$$= \frac{m}{C}\sqrt{T} + T + \frac{2m}{3C}\big(T^{3/2} - 1\big). \tag{170}$$

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

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup\} \sum_{t=1}^{T} \sqrt{m^2 t + 1}\right] \tag{171}$$

$$\overset{(167)}{\le} \mathbb{P}(\mathcal{E}_\cup) \sum_{t=1}^{T} \sqrt{m^2 t + 1} \tag{172}$$

$$\overset{(155)}{\le} \underbrace{\frac{NT(T+1)d}{\exp\big(\log(dNT^{5/2}(T+1))\big)}}_{=T^{-3/2}} \sum_{t=1}^{T} \sqrt{m^2 t + 1} \tag{173}$$

$$\overset{(170)}{\le} T^{-3/2}\left(\frac{m}{C}\sqrt{T} + T + \frac{2m}{3C}\big(T^{3/2} - 1\big)\right) \tag{174}$$

$$< \frac{m}{CT} + \frac{2m}{3C} + 1. \tag{175}$$

c) Bounding $\mathbb{E}[\mathbb{1}\{\mathcal{E}_\cup^c\} \sum_{t=1}^{T} \lambda(t)]$:

Considering the event $\mathcal{E}_\cup^c$, which encompasses all the events $\{\mathcal{E}_{i,n}^c(t), \mathcal{E}_{i,n}^{*\ c}(t), \widetilde{\mathcal{E}}_{i,n}^c(t), \widetilde{\mathcal{E}}_{i,n}^{*\ c}(t): i \in [N], t \in [T], n \in [t]\}$ being hold. Therefore, we can use the following bounds on the singular values

$$\sigma_{\min}\big(\mathbf{U}_{i,a(t)}(t)\big) \ge \sqrt{\max\{0, N_{i,a(t)}(t)\kappa'_{\min} - \gamma'_n\}}, \tag{176}$$

$$\sigma_{\max}\big(\widetilde{\mathbf{U}}_{i,a(t)}(t)\big) \le \sqrt{N_{i,a(t)}(t)\tilde{\kappa}_{\max} + \tilde{\gamma}_n}. \tag{177}$$

Thus, the targeted sum can be upper-bounded by

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup^c\} \sum_{t=1}^{T} \lambda(t)\right] \tag{178}$$

$$= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup^c\} \sum_{t=1}^{T} \max_{i\in[N]} \frac{\sqrt{\lambda_{\max}\big(\widetilde{\mathbf{V}}_{i,a(t)}(t)\big)}}{\lambda_{\min}\big(\mathbf{V}_{i,a(t)}(t)\big)}\right] \tag{179}$$

$$= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup^c\} \sum_{t=1}^{T} \max_{i\in[N]} \frac{\sqrt{\sigma_{\max}^2\big(\widetilde{\mathbf{U}}_{i,a(t)}(t)\big) + 1}}{\sigma_{\min}^2\big(\mathbf{U}_{i,a(t)}(t)\big) + 1}\right] \tag{180}$$

$$\le \mathbb{E}\sum_{t=1}^{T} \max_{i\in[N]}\left[\frac{\sqrt{N_{i,a(t)}(t)\tilde{\kappa}_{\max} + \tilde{\gamma}_n + 1}}{\max\{0, N_{i,a(t)}(t)\kappa'_{\min} - \gamma'_n\} + 1}\right]. \tag{181}$$

It is worth noting that the term in the summation has a critical point, and we bound the two regions separately. To initiate this process, we define the function $h(x)$ as

$$h(x) \triangleq \frac{\sqrt{x\tilde{\kappa}_{\max} + \tilde{\gamma}_n + 1}}{\max\{0, x\kappa'_{\min} - \gamma'_n\} + 1}, \quad x > 0. \tag{182}$$

In order to analyze the behavior of the function h, we introduce $\tau \triangleq \frac{\alpha^2 m^6}{\kappa_{\min}^2}$ as the critical point. Note that when $x \le \tau$, we have $x\kappa'_{\min} < \gamma_n$. In this case, $h(x)$ is equal to

$$h(x) = \sqrt{x\tilde{\kappa}_{\max} + \tilde{\gamma}_n + 1}, \tag{183}$$

which is an increasing function over the region. To upper bound the h function when $x > \tau$, we define the g function when $x > \tau$ as follows.

$$g(x) \triangleq \frac{\sqrt{x\kappa_{\max} + \alpha m^2\sqrt{x}}}{x\kappa_{\min}/m - \alpha m^2\sqrt{x}} + \frac{C}{x\kappa_{\min}/m - \alpha m^2\sqrt{x}}. \tag{184}$$

Then when $x > \tau$, we establish the following relation.

$$h(x) = \frac{\sqrt{x\tilde{\kappa}_{\max} + \tilde{\gamma}_n + 1}}{x\kappa'_{\min} - \gamma'_n + 1} = \frac{\sqrt{x\kappa_{\max} + \gamma_n + C^2}}{x\kappa_{\min}/m - \gamma_n + C} \tag{185}$$

$$< \frac{\sqrt{x\kappa_{\max} + \gamma_n}}{x\kappa_{\min}/m - \gamma_n} = g(x), \tag{186}$$

where the equality holds due to the definitions in (140)-(142) and the inequality holds due to the fact that $\frac{\sqrt{x+a^2}}{y+a} < \frac{\sqrt{x}}{y} + \frac{a}{y}$ when $a > 0$. Moreover, when $x > \tau$, we have $\gamma_n = \alpha m^2\sqrt{n}$. Based on this, we can use the g function to upper bound h function when $x > \tau$. However, since g tends to infinity at τ, we begin bounding from a larger constant, specifically 4τ. However, for the summation when bounding $x \le 4\tau$, we need to consider the following monotonicity.

Lemma 7: $h(x)$ and $g(x)$ are both decreasing functions when $x > \tau$, where τ is defined as $\frac{\alpha^2 m^6}{\kappa_{\min}^2}$.

Proof: See [39, Appendix E]. ∎

Now we are ready to bound the last term

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_\cup^c\} \sum_{t=1}^{T} \lambda(t)\right] \le \mathbb{E}\sum_{t=1}^{T} \max_{i\in[N]} h\big(N_{i,a(t)}(t)\big). \tag{187}$$

We define the set of time indices at which the chosen actions are under-explored as

$$\mathcal{K}_h \triangleq \big\{t \in [T] \mid \exists\ i \in [N]: N_{i,a(t)}(t) \le 4\tau\big\}. \tag{188}$$

It can be readily verified that $|\mathcal{K}_h| \leq 8N\tau$. Furthermore, when $x \in \mathcal{K}_h$, we have

$$h(x) \leq h(\tau) \leq \frac{1}{C}\sqrt{\kappa_{\max}\tau + \alpha m^2 \sqrt{\tau}} + 1, \ x \leq \tau. \quad (189)$$

Then we can upper bound the summation when $\mathcal{K}_h$ occurs as follows.

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in \mathcal{K}_h\} \max_{i \in [N]} h(N_{i,a(t)}(t))\right] \quad (190)$$

$$\leq 2\sum_{i=1}^{N}\sum_{n=1}^{4\tau} h(n) \quad (191)$$

$$\leq 8N\tau\left(\frac{1}{C}\sqrt{\kappa_{\max}\tau + \alpha m^2 \sqrt{\tau}} + 1\right). \quad (192)$$

Now we only need to bound the remaining part when $t \notin \mathcal{K}_h$

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in \mathcal{K}_h^c\} \max_{i \in [N]} h(N_{i,a(t)}(t))\right]. \quad (193)$$

Note that when $t \in \mathcal{K}_h^c$, we have $N_{i,a(t)}(t) > \tau$ for all $i \in [N]$, and we have

$$\max_{i \in [N]} h(N_{i,a(t)}(t)) \leq \max_{i \in [N]} g(N_{i,a(t)}(t)). \quad (194)$$

Furthermore, note that there might be multiple nodes that achieve $\max_{i \in [N]} g(N_{i,a(t)}(t))$. Without loss of generality, we select the solution with minimum index as $\arg\max$ (or $\arg\min$). We denote the node that achieves the maximum value of the function g as i_t.

$$i_t \triangleq \arg\max_{i \in [N]} g(N_{i,a(t)}(t)) = \arg\min_{i \in [N]} N_{i,a(t)}(t), \quad (195)$$

where the last equality is a result of the fact that g is a decreasing function when $x \geq \tau$. Note that i_t does not capture whether i belongs to $a(t)$ or not. To address this challenge, we define the sets of time indices where $i_t = i$ for each of these two cases as follows for $i \in [N]$.

$$\mathcal{S}_i \triangleq \{t \in [T] : t \notin \mathcal{K}_h, \ i_t = i, \ i \notin a(t)\}, \quad (196)$$
$$\text{and} \quad \mathcal{S}_i^* \triangleq \{t \in [T] : t \notin \mathcal{K}_h, \ i_t = i, \ i \in a(t)\}. \quad (197)$$

Denote the elements of $\mathcal{S}_i$ by $S_{i,1}, \ldots, S_{i,|\mathcal{S}_i|}$. Until time $S_{i,n}$, the event $\{i_t = i, i \notin a(t)\}$ occurs exactly n times outside $\mathcal{K}_h$ set. Similarly $\{i_t = i, i \in a(t)\}$ occurs n times outside $\mathcal{K}_h$ set until time $S_{i,n}^*$. Then

$$N_i(S_{i,n}) = \sum_{t=1}^{S_{i,n}} \mathbb{1}\{i \notin a(t)\} \geq \sum_{t=1}^{S_{i,n}} \mathbb{1}\{i_t = i, i \notin a(t)\} \quad (198)$$

$$= n + 4\tau, \quad (199)$$

$$N_i^*(S_{i,n}^*) = \sum_{t=1}^{S_{i,n}^*} \mathbb{1}\{i \in a(t)\} \geq \sum_{t=1}^{S_{i,n}^*} \mathbb{1}\{i_t = i, i \in a(t)\} \quad (200)$$

$$= n + 4\tau. \quad (201)$$

Using the above results and noting that g is a decreasing function, we obtain

$$\sum_{t=1}^{T} \mathbb{1}\{t \in \mathcal{K}_h^c\} \max_{i \in [N]} g(N_{i,a(t)}(t)) \quad (202)$$

$$= \sum_{t=1}^{T} \mathbb{1}\{t \in \mathcal{K}_h^c\} g(N_{i_t,a(t)}(t)) \quad (203)$$

$$= \sum_{i=1}^{N} \sum_{t:t \in \mathcal{S}_i} g\left(N_i(t)\right) + \sum_{i=1}^{N} \sum_{t:t \in \mathcal{S}_i^*} g\left(N_i^*(t)\right) \quad (204)$$

$$= \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i|} \underbrace{g(N_i(S_{i,n}))}_{\substack{(199) \\ \leq \ g(n+4\tau)}} + \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i^*|} \underbrace{g(N_i^*(S_{i,n}^*))}_{\substack{(201) \\ \leq \ g(n+4\tau)}} \quad (205)$$

$$\leq \sum_{i=1}^{N} \sum_{n=4\tau+1}^{|\mathcal{S}_i|+4\tau} g(n) + \sum_{i=1}^{N} \sum_{n=4\tau+1}^{|\mathcal{S}_i^*|+4\tau} g(n). \quad (206)$$

We bound the discrete sums through integrals and define

$$G_\tau(y) = \int_{x=4\tau}^{y} g(x)dx, \quad y \geq 4\tau. \quad (207)$$

Since $g(x)$ is a positive, non-increasing function, for any $k \in \mathbb{N}, k \geq 4\tau + 1$ we have

$$\sum_{n=4\tau+1}^{k} g(n) \leq \int_{x=4\tau}^{k} g(x)dx = G_\tau(k). \quad (208)$$

Then, the summation is upper bounded by

$$\sum_{t=1}^{T} \mathbb{1}\{t \in \mathcal{K}_h^c\} \max_{i \in [N]} g(N_{i_t,a(t)}(t)) \quad (209)$$

$$\leq \sum_{i=1}^{N} \sum_{n=4\tau+1}^{|\mathcal{S}_i|+4\tau} g(n) + \sum_{i=1}^{N} \sum_{n=4\tau+1}^{|\mathcal{S}_i^*|+4\tau} g(n) \quad (210)$$

$$\overset{(208)}{\leq} \sum_{i=1}^{N} G_\tau\left(|\mathcal{S}_i| + 4\tau\right) + \sum_{i=1}^{N} G_\tau\left(|\mathcal{S}_i^*| + 4\tau\right). \quad (211)$$

Since $g(x)$ is positive and decreasing, and $G(y)$ is defined as an integral of the g function with a positive first derivative and negative second derivative, it can be deduced that G is a concave function.

$$\sum_{i=1}^{N} G_\tau(|\mathcal{S}_i| + 4\tau) + \sum_{i=1}^{N} G_\tau\left(|\mathcal{S}_i^*| + 4\tau\right) \quad (212)$$

$$\leq 2N \times G_\tau\left(\frac{1}{2N}\sum_{i=1}^{N}|\mathcal{S}_i| + \frac{1}{2N}\sum_{i=1}^{N}|\mathcal{S}_i^*| + 4\tau\right) \quad (213)$$

$$\leq 2N \times G_\tau\left(\frac{T}{2N} + 4\tau\right). \quad (214)$$

Next, we proceed to establish an upper bound for the function G.

Lemma 8: The G function can be upper bounded as

$$G_\tau\left(\frac{T}{2N} + 4\tau\right)$$

$$\leq 2\frac{\sqrt{m\kappa_{\max}}}{\kappa_{\min}}\left(\sqrt{\frac{T}{2N}} + \sqrt{\tau}\log\left(\sqrt{\frac{T}{2N}} + \sqrt{\tau}\right)\right)$$

$$+ \frac{4}{\kappa_{\min}}\sqrt[4]{\frac{T}{2N}} + 2\sqrt{\frac{\alpha m^5}{\kappa_{\min}^3}}\log\left(\frac{\sqrt{\frac{1}{\tau}}\sqrt[4]{\frac{T}{2N}} + \sqrt[4]{4} + 1}{\sqrt{\frac{1}{\tau}}\sqrt[4]{\frac{T}{2N}} + \sqrt[4]{4} - 1}\right)$$

$$+ \frac{2m \log\left(\frac{\kappa_{\min}}{m}\sqrt{\frac{T}{2N}} + \alpha m^2\right)}{\kappa_{\min}} C. \qquad (215)$$

Proof: See [39, Appendix F] ∎

Combining the results in (175), (206), (214) and (215), let E_1 denote the accumulation of terms that exhibit at most logarithmic growth rates with respect to T and C.

$$E_1 = 4N\frac{\sqrt{m\kappa_{\max}}}{\kappa_{\min}}\sqrt{\tau}\log\left(\sqrt{\frac{T}{2N}} + \sqrt{\tau}\right) \qquad (216)$$

$$+ 4N\sqrt{\frac{\alpha m^5}{\kappa_{\min}^3}}\log\left(\frac{\sqrt{\frac{1}{\tau}}\sqrt[4]{\frac{T}{2N}} + \sqrt[4]{4} + 1}{\sqrt{\frac{1}{\tau}}\sqrt[4]{\frac{T}{2N}} + \sqrt[4]{4} - 1}\right) \qquad (217)$$

$$+ 8N\tau\left(\frac{1}{C}\sqrt{\kappa_{\max}\tau + \alpha m^2\sqrt{\tau}} + 1\right) \qquad (218)$$

$$+ \frac{m}{CT} + \frac{2m}{3C} + 1. \qquad (219)$$

Therefore, the final result for the bound is

$$\mathbb{E}\left[\sum_{t=1}^{T}\lambda(t)\right] \leq \frac{4\sqrt{m\kappa_{\max}}}{\kappa_{\min}}\sqrt{NT} + \frac{8}{\kappa_{\min}}\sqrt[4]{\frac{N^3 T}{2}} \qquad (220)$$

$$+ \frac{4Nm}{\kappa_{\min}}\log\left(\frac{\kappa_{\min}}{m}\sqrt{\frac{T}{2N}} + \alpha m^2\right)C + E_1. \qquad (221)$$

Plugging (221) into (128), we have

$$\mathbb{E}[R(T)] \leq 4(\beta_T + 1)^L d^{\frac{L-1}{2}}\mathbb{E}\left[\sum_{t=1}^{T}\lambda(t)\right] \qquad (222)$$

$$= \tilde{\mathcal{O}}\left(d^{L-\frac{1}{2}}\sqrt{NT} + d^{L-\frac{1}{2}}NC\right). \qquad (223)$$

∎

REFERENCES

[1] O. Ahmed et al., "CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–18.

[2] M. B. Badsha and A. Q. Fu, "Learning causal biological networks with the principle of mendelian Randomization," *Front. Genet.*, vol. 10, May 2019, Art. no. 454043.

[3] 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," *J. Med. Internet Res.*, vol. 22, no. 7, Jul. 2020, Art. no. e18477.

[4] Y. Zhao et al., "Mitigating targeting bias in content recommendation with causal bandits," in *Proc. ACM Conf. Recomm. Syst. Workshop Multi-Object. Recomm. Syst.*, 2022, pp. 1–9.

[5] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[6] J. Zhang, C. Squires, and C. Uhler, "Matching a desired causal state via shift interventions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–12.

[7] Y. Lu, A. Meisami, and A. Tewari, "Causal bandits with unknown graph structure," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–12.

[8] F. Lattimore, T. Lattimore, and M. D. Reid, "Causal bandits: Learning good interventions via causal inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.

[9] A. De Kroon, J. Mooij, and D. Belgrave, "Causal bandits without prior knowledge using separating sets," in *Proc. Conf. Causal Learn. Reason.*, 2022, pp. 1–21.

[10] S. Feng, N. Xiong, and W. Chen, "Combinatorial causal bandits without graph skeleton," 2023, *arXiv:2301.13392.*

[11] B. Bilodeau, L. Wang, and D. M. Roy, "Adaptively exploiting *d*-separators with causal bandits," in *Proc. 36th Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–12.

[12] M. Konobeev, J. Etesami, and N. Kiyavash, "Causal bandits without graph learning," 2023, *arXiv:2301.11401.*

[13] A. Maiti, V. Nair, and G. Sinha, "A causal bandit approach to learning good atomic interventions in presence of unobserved confounders," in *Proc. 38th Conf. Uncertainty Artif. Intell.*, 2022, pp. 1328–1338.

[14] S. Feng and W. Chen, "Combinatorial causal bandits," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 7550–7558.

[15] B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer, "Causal bandits for linear structural equation models," *J. Mach. Learn. Res.*, vol. 24, no. 297, pp. 1–59, 2023.

[16] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, "Behind distribution shift: Mining driving forces of changes and causal arrows," in *Proc. IEEE Int. Conf. Data Min.*, 2017, pp. 913–918.

[17] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf, "Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1347–1353.

[18] K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour, "Causal discovery in the presence of measurement error: Identifiability conditions," 2017, *arXiv:1706.03768.*

[19] K. Zhang, J. Zhang, B. Huang, B. Schölkopf, and C. Glymour, "On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection," in *Proc. Conf. Uncertainty Artif. Intell.*, 2016, pp. 825–834.

[20] R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang, "Causal discovery in the presence of missing data," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2019, pp. 1762–1770.

[21] J. Li, D. Cai, and X. He, "Learning graph-level representation for drug discovery," 2017, *arXiv:1709.03741.*

[22] Y. Lu, A. Meisami, A. Tewari, and W. Yan, "Regret analysis of bandit problems with causal background knowledge," in *Proc. Conf. Uncertainty Artif. Intell.*, 2020, pp. 1–10.

[23] V. Nair, V. Patil, and G. Sinha, "Budgeted and non-budgeted causal bandits," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2021, pp. 1–10.

[24] S. Sussex, A. Makarova, and A. Krause, "Model-based causal Bayesian optimization," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–24.

[25] A. Ghosh, S. R. Chowdhury, and A. Gopalan, "Misspecified linear bandits," in *Proc. Conf. Artif. Intell.*, 2017, pp. 3761–3767.

[26] T. Lattimore, C. Szepesvari, and G. Weisz, "Learning with good feature representations in bandits and in RL with a generative model," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1–9.

[27] D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert, "Adapting to misspecification in contextual bandits," in *Proc. 34th Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–12.

[28] S. K. Krishnamurthy, V. Hadad, and S. Athey, "Adapting to misspecification in contextual bandits with offline regression oracles," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1-10.

[29] Y. Li, E. Y. Lou, and L. Shan, "Stochastic linear optimization with adversarial corruption," 2019, *arXiv:1909.02109.*

[30] I. Bogunovic, A. Losalka, A. Krause, and J. Scarlett, "Stochastic linear bandits robust to adversarial attacks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2021, pp. 1–10.

[31] C.-W. Lee, H. Luo, C.-Y. Wei, M. Zhang, and X. Zhang, "Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1–10.

[32] H. Zhao, D. Zhou, and Q. Gu, "Linear contextual bandits with adversarial corruptions," 2021, *arXiv:2110.12615.*

[33] C.-Y. Wei, C. Dann, and J. Zimmert, "A model selection approach for corruption robust reinforcement learning," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2022, pp. 1–54.

[34] J. He, D. Zhou, T. Zhang, and Q. Gu, "Nearly optimal algorithms for linear contextual bandits with adversarial corruptions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–12.

[35] Y. Russac, C. Vernade, and O. Cappé, "Weighted linear bandits for non-stationary environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–10.

[36] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1–9.

[37] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[38] Q. Ding, C.-J. Hsieh, and J. Sharpnack, "Robust stochastic linear contextual bandits under adversarial attacks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2022, pp. 7111–7123.

[39] Z. Yan, A. Mukherjee, B. Varıcı, and A. Tajer, "Robust causal bandits for linear models," 2024, *arXiv:2310.19794.*