

UCLA

UCLA Electronic Theses and Dissertations

Title

Sequential Learning and Inference of Causal Graphical Models

Permalink

<https://escholarship.org/uc/item/2qm1r70x>

ISBN

9798277445419

Author

Zhao, Yijia

Publication Date

2026-03-19

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Sequential Learning and Inference of Causal Graphical Models

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Yijia Zhao

2026

© Copyright by

Yijia Zhao

2026

ABSTRACT OF THE DISSERTATION

Sequential Learning and Inference of Causal Graphical Models

by

Yijia Zhao

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2026

Professor Qing Zhou, Chair

Many problems involve systems in which causal relationships among variables are represented by causal graphs. Because experimental data are often costly, learning procedures must combine statistical inference with sequential decision making. This dissertation develops methods for learning and inference of causal graphical models under sequential data collection, integrating observational and interventional information to support both optimal decision making and causal structure discovery. The analysis is conducted under linear structural equation models with Gaussian errors, which provide a direct link between causal effects and regression estimation.

The first part of the dissertation (Chapters 2 and 3) studies sequential identification of the intervention with the largest causal effect on a reward variable in an unknown graph. Interventions are treated as actions, and the objective is to determine which intervention produces the greatest effect on the reward. We develop upper-confidence-bound-based algorithms that exploit potential causal relationships among variables to improve learning efficiency relative to standard bandit methods. Under causal sufficiency, observational data are incorporated to construct valid confidence bounds for causal effects, leading to improved regret guarantees.

The framework is then extended to settings with latent confounders, where some causal effects may not be identifiable through adjustment. The resulting method distinguishes identifiable and non-identifiable interventions and provides regret bounds that reflect this distinction. Theoretical analysis establishes finite-sample regret guarantees, and simulation studies demonstrate improved learning efficiency compared with existing methods.

The second part of the dissertation (Chapter 4) studies adaptive causal structure learning. Observational data determine a directed acyclic graph only up to its Markov equivalence class, and interventions are required to resolve the remaining uncertainty. We propose a sequential experimental design procedure that selects interventions based on uncertainty in edge orientation and updates the graph using statistical evidence. Theoretical analysis shows that the proposed procedure recovers the true graph almost surely, and numerical experiments demonstrate that it requires fewer interventions than sequential-random and complete-random baseline strategies.

The dissertation of Yijia Zhao is approved.

Guido F. Montufar

Guang Cheng

Yingnian Wu

Qing Zhou, Committee Chair

University of California, Los Angeles

2026

To my family

TABLE OF CONTENTS

1	Introduction	1
2	Causal bandit on unknown graphs in a fully observed setting	4
2.1	Introduction	4
2.2	Problem formulation	7
2.3	The BA-UCB algorithm	10
2.3.1	The main idea	10
2.3.2	Identification of backdoor adjustment sets	11
2.3.3	Algorithm description	13
2.4	Regret analysis	15
2.4.1	Main results	16
2.4.2	Implications	21
2.4.3	Proof outline	22
2.5	Numerical experiments	24
2.5.1	Simulation settings	24
2.5.2	Simulation results of BA-UCB	25
2.5.3	Comparison with other algorithms	30
2.6	Discussion	33
2.7	Appendix	34
2.7.1	Some general bounds	34
2.7.2	When the underlying DAG is given	37
2.7.3	When the underlying DAG is unknown	48

3	Causal bandit on unknown graphs in presence of confounders	60
3.1	Introduction	60
3.2	Preliminaries	61
3.3	BA-UCB with latent confounders	63
3.4	Regret analysis	65
3.5	Numerical experiments	67
3.6	Discussion	69
3.7	Appendix	70
3.7.1	Proof of Lemma 3.1	70
3.7.2	Proof of Theorem 3.1	72
3.7.3	Proof of Theorem 3.2	77
4	Adaptive sequential intervention for learning causal DAGs	80
4.1	Introduction	80
4.2	Problem formulation	82
4.3	The Seq-UCB algorithm	83
4.3.1	The main idea	83
4.3.2	Algorithm description	86
4.4	Theoretical analysis of Seq-UCB	88
4.5	Numerical experiments	91
4.5.1	Simulation settings	91
4.5.2	True DAG identification rate comparisons	92
4.5.3	Early stopping and intervention time comparisons	94
4.6	Discussion	95

4.7	Appendix	96
4.7.1	Proof of Lemma 4.1	96
4.7.2	Proof of Lemma 4.2	97
4.7.3	Proof of Theorem 4.1	99
5	Summary and future directions	101

LIST OF FIGURES

2.1	An example DAG to illustrate Assumption 2.2.	17
2.2	Comparison of empirical cumulative regrets over 3000 time steps between the BA-UCB and standard UCB algorithms when DAG is given for $p = 10, 20, 30$. . .	27
2.3	Comparison of empirical cumulative regret over $T = 5000$ time steps for BA-UCB on DAGs of $p = 10, 20, 30$. BAUCB-lm uses estimates from the single linear regression and BAUCB uses weighted average.	28
2.4	Estimation accuracy of the optimal arm over time. The left panel shows the absolute difference between reward estimates based on observational and experimental data, while the right panel reports the absolute difference between the estimated reward and the true value. The box-plots summarize results across 100 randomly generated DAGs with $p = 10$	29
2.5	Average identification rate of a correct adjustment set for the optimal arm at each time step. Solid curves are smoothing averages with a 100-step window for better visualization.	30
2.6	Comparison of empirical cumulative regret over $T = 5000$ time steps among the BA-UCB, BBB-UCB, CN-UCB, and UCB algorithms on DAGs with $p = 10$. . .	31
2.7	Comparison of empirical cumulative regret over 5000 time steps among BA-UCB, CN-UCB, and UCB algorithms for $p = 20, 30, 50$. The top panel reports the cases where the optimal arm is not a parent of the reward variable and the lower panel reports the cases where the optimal arm is a parent of the reward variable. . . .	32
3.1	Comparison of empirical cumulative regrets over 5000 time steps between the BA-UCB and standard UCB algorithms on graphs with latent confounders. The number of observed variables $p = 9, 18, 27$, corresponding to a total of $N = 10, 20, 30$ variables, respectively.	68

3.2	Average identification rate of a correct adjustment set for the optimal arm at each time step in the presence of latent confounders. Solid curves are smoothing averages with a 100-step window for better visualization.	69
4.1	Average true DAG identification rate over time on 100 randomly simulated DAGs with $p = 20$, maximum in-degree 2, and between 10 and 15 undirected edges in the initial CPDAG.	93
4.2	Average times of interventions over time across all runs on 100 randomly simulated DAGs with $p = 20$	94
4.3	Average number of wrongly oriented edges over time on 100 randomly simulated DAGs with $p = 20$, maximum in-degree 2, and between 10 and 15 undirected edges in the initial CPDAG after 50 rounds.	94
4.4	Average number of times each node is selected for intervention across three graph structures.	95

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Qing Zhou, for his patience and detailed guidance throughout my Ph.D. journey. His careful mentorship, insightful feedback, and consistent support have been invaluable to my growth as a researcher. I especially appreciate the time and attention he devoted to discussing ideas, reading drafts, and helping me refine both technical details and presentation. His emphasis on clarity, rigor, and independence has had a lasting influence on me.

I am also sincerely grateful to my doctoral committee members, Professor Guang Cheng, Professor Guido Montúfar, and Professor Yingnian Wu, for their time, support, and thoughtful comments. Through their questions and suggestions, I was encouraged to consider my work from broader perspectives, which strengthened both the content and presentation.

Beyond my academic work, I owe my heartfelt thanks to my parents for their constant support and understanding, and to my friends for their companionship and care, which made this journey more joyful and less lonely. I am also grateful to my cat, Daves, for the comfort and calm he brought me during stressful periods, and to my favorite singer, Charlie Zhou, whose music provided emotional support and accompanied me through many long days and nights.

VITA

2014-2018	B.S., Pure and Applied Mathematics, Tsinghua University
2018-2020	M.S., Statistics, The University of Chicago
2021-2026	Teaching Assistant, Department of Statistics, University of California, Los Angeles

CHAPTER 1

Introduction

Many problems in science, engineering, and economics involve systems in which variables are connected through underlying causal mechanisms. These mechanisms are commonly represented by causal graphical models, in which nodes denote variables and directed edges represent direct causal effects. Learning in such systems requires not only statistical estimation from data, but also evaluation of the effects of interventions. In practice, experiments are often costly and only finite samples can be collected, so learning procedures must allocate limited experimental resources while accounting for uncertainty. This dissertation develops statistical methods for learning and inference in causal graphical models under sequential data collection.

A fundamental framework for sequential decision making is the multi-armed bandit (MAB) problem, which has been extensively studied in statistics, operations research, and machine learning [1–5]. At each round, a learner selects an arm based on past actions and observed rewards, with the goal of identifying the arm with the highest expected reward while minimizing cumulative regret over time. The central challenge is to balance exploration, which gathers information about uncertain actions, and exploitation, which selects actions believed to yield high reward. This problem has broad applications in online recommendation systems [6], adaptive clinical trials [7, 8], and dynamic pricing [9], among others.

Under the classic setting of the MAB problem [1, 2], the action set $\mathcal{A}$ consists of K independent arms and each arm $a \in \mathcal{A}$ corresponds to a real-valued distribution of the reward

variable. It is a simplified setting in which the entire system is regarded as a black box, relying solely on reward feedback to guide learning. Various algorithms have been developed in this setting, among which the upper confidence bound (UCB) algorithm [1] and Thompson sampling [10] are two classical examples. However, real-world decision-making processes often involve additional complexity, motivating numerous variants of the MAB problem. For example, the contextual bandit problem incorporates observed contextual features before each decision, allowing personalization and adaptive learning [11]; the combinatorial bandit focuses on actions formed by combinations of basic arms, leading to exponentially large action spaces [12]; and the budgeted bandit associates each action with a cost and requires the player to operate under a fixed budget [13].

When actions correspond to interventions in a causal system, the standard bandit formulation can be enriched by exploiting structural relationships among variables. In a causal bandit [14] setting, interventions affect the reward through a causal graph. Observational data and structural constraints may provide information about multiple interventions simultaneously, allowing learning to proceed more efficiently than in unstructured bandits. However, in many realistic settings the causal graph is not fully known, observational data may be limited, and experimental samples are costly. This leads to two distinct but related statistical problems.

The first problem concerns optimal decision making in unknown causal systems. The objective is to identify the intervention node with the largest causal effect on the reward while interacting sequentially with the system. Full recovery of the causal graph is not required in this problem. Instead, structural information is used to reduce uncertainty in estimating causal effects and to eliminate suboptimal interventions more quickly. Chapters 2 and 3 address this problem. Chapter 2 considers Gaussian causal directed acyclic graphs (DAGs) under causal sufficiency and develops a method that combines observational and experimental data to construct upper confidence bounds for causal effects, yielding improved regret guarantees relative to standard bandit approaches. Chapter 3 extends the framework

to settings with latent confounders, where the observed system is represented by an acyclic directed mixed graph and valid adjustment sets may not exist for all interventions. The resulting analysis distinguishes between identifiable and non-identifiable interventions and provides regret bounds that reflect this distinction.

The second problem concerns structural learning. Rather than focusing on reward maximization, the objective is to discover the causal relationships among variables. Observational data typically identify a directed acyclic graph only up to its Markov equivalence class, leaving certain edge orientations unresolved. Interventions provide additional information to distinguish among Markov equivalent DAGs. Because only finitely many samples can be collected under each intervention, experiments must be selected adaptively to reduce uncertainty about the graph as efficiently as possible. Chapter 4 develops an adaptive intervention procedure for recovering the true causal graph within its equivalence class with high probability, using confidence-bound-based criteria to guide sequential interventions.

Throughout the dissertation, data are assumed to arise from a linear structural equation model with additive Gaussian noise. Each variable is modeled as a linear function of its direct causes, and interventions modify the structural equation of the intervened variable while leaving the rest of the system unchanged. Under this model, causal effects correspond to regression coefficients and can be estimated from observational or interventional samples. This connection enables uncertainty quantification and supports sequential decision procedures based on confidence bounds.

CHAPTER 2

Causal bandit on unknown graphs in a fully observed setting

2.1 Introduction

A major source of complexity in real-world decision-making stems from the causal dependency among variables. In many domains, such as healthcare, social sciences, and economics, arms and rewards are connected through underlying causal mechanisms, typically represented by a causal directed acyclic graph (DAG). In this causal bandit (CB) framework [14], a causal DAG specifies the joint distribution of all variables in the system, including the reward. Interventions on variables correspond to arms, allowing the causal bandit problem to be cast naturally as a multi-armed bandit task. Unlike the standard MAB setting, however, leveraging the additional causal structure can enable substantially more efficient identification of the optimal intervention.

Most existing work on causal bandit relies on full or partial knowledge of the underlying causal graph to achieve improvements over standard MAB algorithms. For instance, Lattimore et al. [14], Sen et al. [15], and Yabe et al. [16] study the causal bandit problem assuming complete knowledge of the DAG and absence of unobserved confounders. Other approaches further impose structural restrictions while requiring the graph to be provided, such as the budgeted bandit under no-backdoor graphs [17] and the linear causal bandit [18]. More recent work in this category includes causal bandits for linear structural equation models [19] and causal bandits with general causal models and soft interventions [20], both of which

continue to assume a known causal graph. However, in practice, the underlying causal DAG is rarely known, which severely limits the applicability of these methods.

A few recent causal bandit methods have been developed without assuming the underlying causal graph is provided. For discrete variables, the optimal intervention is always among the parents of the reward variable [21]. A natural approach is therefore to first identify the parent set of the reward variable and then apply MAB algorithms with a reduced action set consisting of only the identified parents. Many recent works follow this strategy, including the central node - upper confidence bound (CN-UCB) algorithm [22] and the randomized parent search (RAPS) algorithm [23]. When the variables have continuous domains, the problem is significantly more challenging, since the optimal intervention is not necessarily among the parents of the reward. De Kroon, Mooij, and Belgrave [24] tackles the problem for discrete and Gaussian models. However, their method relies on the existence of a set of non-intervenable variables that d-separate the interventions and the reward variable, and no theoretical guarantee on the cumulative regret was provided. Huang and Zhou [25] propose a Bayesian framework that utilizes joint inferences from experimental and observational data without additional assumptions on the underlying graph structure. But the Bayesian framework requires updating posterior probabilities over all possible parent sets, which is extremely computationally intensive and becomes impractical even for moderately sized action set. Yan and Tajer [26] propose a graph-agnostic linear causal bandit (GA-LCB) algorithm for linear structural equation models with unknown graph structure and soft interventions, assuming no latent confounders. Their approach emphasizes learning the graph structure and model coefficients prior to reward evaluation, which is not required for identifying the intervention with the largest expected reward and leads to additional computational complexity.

To motivate our work, we revisit the farming example constructed by Lattimore, Lattimore, and Reid [14]. Suppose a farmer seeks to maximize crop yield and knows that the outcome is affected by temperature, moisture level, and soil nutrient levels, but the causal relationships among these factors are unknown. Due to limited time and budget, she can

intervene on at most one factor per season: for example, using artificial heating or shading to control temperature, irrigation or rain covers to control moisture, or adjusting fertilizer composition to modify nutrient levels. Each intervention is costly and yields limited experimental data per season. In contrast, historical observational data, collected without interventions, may be readily available. An important practical question is whether such observational data can be leveraged to help identify the variable with the largest causal impact on crop yield, thereby reducing the need for costly experimentation. Motivated by this example, our objective is to identify the variable with the largest causal effect on the reward, and we therefore consider the set of arms to be a collection of atomic interventions. In such continuous domains, the optimal intervention is not necessarily associated with a parent of the reward node, making it unclear which aspects of the underlying causal structure should be learned and how such information can be used to reduce the action space. Further challenges include how to efficiently integrate observational data into the sequential decision-making process to aid in identifying the most influential variable, and how to accommodate potential latent confounders.

To address these challenges, we develop the Backdoor Adjustment Upper Confidence Bound (BA-UCB) algorithm, which identifies the optimal intervention without requiring prior knowledge of the causal graph or imposing restrictive structural assumptions. We focus on settings in which all observed variables, including the reward, follow a Gaussian DAG model. Rather than attempting to recover the full DAG, BA-UCB exploits both observational and experimental data to identify candidate adjustment sets and integrates information from the two data sources to construct upper confidence bounds for each arm within a sequential decision-making framework.

We establish finite-time upper bounds on the cumulative regret of the BA-UCB algorithm. Specifically, suppose that m experimental samples and n observational samples are collected at each round. In the absence of latent confounders, and assuming a sufficiently large amount of prior observational data, we show that the worst-case cumulative regret is bounded by

$O(\sqrt{T \log T / (m + n)})$, where T denotes the total number of rounds. In contrast, the regret bound for the standard MAB algorithm is $O(\sqrt{KT \log T})$ [27], where $K = |\mathcal{A}|$ is the number of arms. Even if $m + n$ experimental samples are generated per round, the standard MAB regret bound remains at least $O(\sqrt{KT \log T / (m + n)})$. One sees that BA-UCB substantially weakens the dependence of cumulative regret on K and effectively incorporates observational data, which are typically much less costly to obtain than experimental data. Finally, through extensive simulation studies, we demonstrate that BA-UCB achieves lower cumulative regret than the standard UCB algorithm [1] and the CN-UCB algorithm [22], while being more computationally efficient than the Bayesian backdoor bandit algorithm [25].

The remainder of this chapter is organized as follows. Section 2.2 introduces the causal bandit framework under Gaussian DAG models without latent confounders. Section 2.3 presents the BA-UCB algorithm, and Section 2.4 develops its finite-sample regret analysis. Section 2.5 reports empirical results based on simulated data. Section 2.6 concludes with a summary of the main findings and directions for future work. Proofs of the theoretical results are provided in Section 2.7.

2.2 Problem formulation

In a causal graphical model [28], the causal relationships among a set of random variables $\mathcal{X} = \{X_1, \dots, X_p\}$ is represented by a DAG $\mathcal{G}$ over p nodes $V = \{1, \dots, p\}$, each corresponding to one random variable. A directed edge $i \rightarrow j$ in the causal DAG $\mathcal{G}$ means that i is a direct cause of j , in which case i is called a parent of j and j is called a child of i . The parent set of j , denoted as $\text{pa}(j)$, is the set of nodes that have an edge directed into node j , i.e. $\text{pa}(j) = \{i \in V : i \rightarrow j\}$. Similarly, the child set of j , denoted as $\text{ch}(j)$, is the set of all nodes that have an edge directed from node j . Each child-parent relationship in $\mathcal{G}$ is modeled by a function $X_j = f_j(X_{\text{pa}(j)}, \varepsilon_j)$, $j = 1, \dots, p$, where $\varepsilon_1, \dots, \varepsilon_p$ are exogenous errors.

Through Section 2.5, we focus on Gaussian DAGs whose functional relationships are

given by a linear structural equation model (SEM),

$$X_j = \sum_{i \in \text{pa}(j)} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \dots, p, \quad (2.1)$$

where $\varepsilon_j \stackrel{\text{indep.}}{\sim} \mathcal{N}(0, \omega_j^2)$ with $\omega_j^2 > 0$ for all $j \in V$ and $\beta_{ij} \neq 0$ for all $i \in \text{pa}(j)$. Without loss of generality, the reward variable is designated as $Y = X_p$ (noting that the nodes are not sorted). The causal effect of X_i on Y is defined via an atomic intervention that sets the variable X_i to a fixed value x_i , denoted as $do(X_i = x_i)$. In general, the causal effect γ_i of X_i on Y is determined by the mapping $x_i \mapsto \mathbb{P}[Y|do(X_i = x_i)]$. For Gaussian DAGs,

$$\gamma_i := \frac{\partial}{\partial x_i} \mathbb{E}(Y|do(X_i = x_i)) \quad (2.2)$$

and by linearity $\mathbb{E}(Y|do(X_i = x_i)) = \gamma_i x_i$. Thus, from any intervention $do(X_i = x_i)$ on X_i with $x_i \neq 0$, we can determine the causal effect $\gamma_i = \mathbb{E}(Y|do(X_i = x_i))/x_i$. Without loss of generality, we only consider $do(X_i = \pm 1)$ as the arms in our causal bandit problem, so that the reward magnitude across arms is on the same scale.

Treating atomic interventions as arms, we formulate the problem of finding the variable with the largest causal effect under the framework of a multi-armed bandit problem. The arm set is

$$\mathcal{A} = \{do(X_i = x_i) : x_i = -1 \text{ or } 1, i = 1, \dots, p-1\}. \quad (2.3)$$

Since $\mathbb{E}[Y|do(X_i = -1)] = -\mathbb{E}[Y|do(X_i = 1)]$ under the linear SEM (2.1), in most practical applications we only need to choose x_i to be either -1 or 1 . Consequently, the size of the arm set can be effectively reduced to $K = |\mathcal{A}| = p-1$. Note that our algorithm and theoretical results apply to the case with both $do(X_i = 1)$ and $do(X_i = -1)$ as separate arms.

For simplicity, let arm i correspond to the intervention $do(X_i = x_i)$ and identify the action set $\mathcal{A}$ with $\bar{V} := \{1, \dots, p-1\}$. For any arm $i \in \mathcal{A}$, the expected reward is defined as $\mu_i = \mathbb{E}[Y|do(X_i = x_i)]$, which has the same magnitude as the causal effect γ_i since $|x_i| = 1$

by design. The optimal reward is denoted by $\mu^* = \max_{i \in \mathcal{A}} \mu_i$ corresponding to the optimal arm $i^* = \operatorname{argmax}_{i \in \mathcal{A}} \mu_i$. The goal is to find the optimal arm while minimizing the expected loss in this process. Let $A_t \in \mathcal{A}$ denote the arm chosen at round t . The expected loss is measured by the cumulative regret up to round T ,

$$\mathcal{R}_T := \mu^* T - \sum_{t=1}^T \mathbb{E}(\mu_{A_t}) = \sum_{i \in \mathcal{A}} \Delta_i \mathbb{E}(n_i(T)), \quad (2.4)$$

where $\Delta_i = \mu^* - \mu_i$ and $n_i(t) = \sum_{s=1}^t 1\{A_s = i\}$ denotes the number of times that arm i has been chosen up to round t . A sequential decision-making policy that minimizes the cumulative regret has to handle the trade-off between exploration and exploitation.

Many algorithms have been proposed for the standard MAB problem with theoretical guarantees on cumulative regret, among which upper confidence bound (UCB) methods and Thompson sampling are two of the most widely studied. For normally distributed rewards, the UCB decision rule [1] takes the form

$$A_{t+1} \in \operatorname{argmax}_{i \in \mathcal{A}} \left\{ \hat{\mu}_i(t) + \hat{\sigma}_i(t) c \sqrt{\log t} \right\}, \quad (2.5)$$

where $\hat{\mu}_i(t)$ is an estimator of the expected reward for arm i , $\hat{\sigma}_i^2(t)$ is an estimator of the variance of $\hat{\mu}_i(t)$, and $c > 0$ is a tuning constant. When the standard MAB framework is applied to the causal bandit setting, only experimental data are collected at each round, and $\hat{\mu}_i(t)$ and $\hat{\sigma}_i(t)$ correspond, respectively, to the sample mean of Y under the intervention $\{do(X_i = x_i)\}$ and its associated standard error, computed from data available up to time t . This approach treats each intervention arm in isolation and does not exploit the causal relationships among the variables. In contrast, the BA-UCB method developed in the next section explicitly incorporates causal structure by leveraging observational data collected prior to or alongside the bandit process, leading to more efficient estimation and improved decision-making.

2.3 The BA-UCB algorithm

2.3.1 The main idea

Let $\beta_i(Y \sim X_i + X_S)$ denote the regression coefficient of X_i when regressing Y on (X_i, X_S) , i.e., $\mathbb{E}(Y|X_i, X_S) = \beta_i X_i + \beta_S^\top X_S$. For Gaussian DAGs, if $X_S \subseteq \mathcal{X} \setminus \{X_i, Y\}$ satisfies the backdoor criterion relative to (X_i, Y) [28], the causal effect γ_i of X_i on Y can be identified via backdoor adjustment,

$$\gamma_i = \beta_i(Y \sim X_i + X_S). \quad (2.6)$$

Therefore, the causal effect of X_i on Y can be estimated using not only the experimental data under intervention on X_i but also observational data as long as a backdoor adjustment set S is provided.

In many practical applications, a large amount of observational data is often available, while generating experimental data is typically much more costly. Leveraging both sources of data has at least two major benefits: (i) Observational data can be used with experimental data to identify valid backdoor adjustment sets, enabling causal identification using observational data alone; (ii) the abundance of observational data can reduce the variance of causal effect estimates, thereby improving accuracy. This motivates our work to integrate observational and experimental evidence in causal bandit settings.

In Section 2.3.2, we propose a method to identify backdoor adjustment sets for each node assuming the DAG is unknown. Once a backdoor adjustment set is identified, we can obtain an estimate of the causal effect μ_i and its variance σ_i^2 from either experimental or observational data. The two estimates can then be combined to improve estimation accuracy. Suppose that we have collected prior observational data of size $n_0 \geq 0$ before the sequential decision process, and will collect n observational data and m experimental data at each round. Note that $n_0 = 0$ if prior data is not available. Denote by $n_i(t) = \sum_{s=1}^t 1\{A_s = i\}$ the number of rounds at which intervention $do(X_i)$ is performed, $\mathcal{D}_i(t)$ the experimental data

generated under $do(X_i)$, and $\mathcal{D}_o(t)$ the observational data collected, all up to round t . Then, $N_i(t) = mn_i(t)$ is the size of $\mathcal{D}_i(t)$ and $N_o(t) = n_0 + nt$ is the size of $\mathcal{D}_o(t)$.

2.3.2 Identification of backdoor adjustment sets

It is straightforward to obtain a consistent estimator of μ_i and its variance from $\mathcal{D}_i(t)$ as follows:

$$\hat{\mu}_{i,\text{int}}(t) = \frac{1}{n_i(t)} \sum_{s=1}^t \bar{Y}_s 1\{A_s = i\}, \quad \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t)) = \frac{s_{i,t}^2}{N_i(t)}, \quad (2.7)$$

where $\bar{Y}_t$ is the mean of rewards collected under intervention at round t and $s_{i,t}^2$ is the sample variance of Y generated under $do(X_i = x_i)$ up to round t . From the property of Gaussian DAGs, the intervention distribution is $Y|do(X_i = x_i) \sim \mathcal{N}(\mu_i, \tilde{\sigma}_i^2)$ for some $\tilde{\sigma}_i^2 > 0$, and therefore $\hat{\mu}_{i,\text{int}}(t) \sim \mathcal{N}(\mu_i, \tilde{\sigma}_i^2/N_i(t))$ is an unbiased estimator of μ_i .

Estimates of causal effects μ_i can also be obtained from observational data by backdoor adjustment (2.6). For any subset S of nodes, let $\mathbf{X}_{i,S} = [\mathbf{X}_i \mid \mathbf{X}_S]$ denote a matrix of dimension $N_o(t) \times (s+1)$, where $s = |S|$, consisting of the columns corresponding to X_i and X_S in the observational dataset $\mathcal{D}_o(t)$. Denote by $\hat{\beta} = [\hat{\beta}_i, \hat{\beta}_S]$ the least-squares coefficient vector for the linear regression $Y \sim X_i + X_S$ using $\mathcal{D}_o(t)$. Construct the following estimates of the reward $\mu_i = \mathbb{E}(Y \mid do(X_i = x_i))$ and its variance,

$$\hat{\mu}_{i,\text{obs}}(t, S) = \hat{\beta}_i \cdot x_i, \quad \widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t, S)) = \frac{\rho_i(t, S) \|\mathbf{Y} - \mathbf{X}_i \hat{\beta}_i - \mathbf{X}_S \hat{\beta}_S\|^2}{N_o(t)(N_o(t) - s - 1)}, \quad (2.8)$$

where $\rho_i(t, S)$ is the $(1, 1)$ -th element of the matrix $(\mathbf{X}_{i,S}^\top \mathbf{X}_{i,S} / N_o(t))^{-1}$. It is easy to see that

$$\hat{\mu}_{i,\text{obs}}(t, S) \mid \mathbf{X}_{i,S} \sim \mathcal{N}(\beta_i x_i, \rho_i(t, S) \zeta_{i,S}^2 / N_o(t)),$$

where $\beta_i = \beta_i(Y \sim X_i + X_S)$ and $\zeta_{i,S}^2 = \text{Var}(Y \mid X_i, X_S)$. To simplify notation, let

$$\mu_i(S) = \beta_i x_i, \quad (2.9)$$

where $x_i \in \{-1, 1\}$ is the chosen intervention level in $do(X_i = x_i)$. If X_S satisfies the backdoor criterion relative to (X_i, Y) , then $\mu_i(S) = \mu_i$ and $\hat{\mu}_{i,\text{obs}}(t, S)$ is an unbiased estimator of μ_i with variance approximated by (2.8). It is a natural idea to combine the two estimates in (2.7) and (2.8) to improve the statistical efficiency in the upper confidence bounds for μ_i in a bandit algorithm.

Next, we discuss how to make a data-driven decision on whether a subset S is a valid adjustment set. Let $\mathcal{B}_i := \{S \subseteq \bar{V} \setminus \{i\} : \mu_i(S) = \mu_i\}$ be the collection of sets that yield the true causal effect of X_i on Y , which we call valid backdoor adjustment sets. Since experimental data and observational data are independent,

$$\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t) \mid \mathbf{X}_{i,S} \sim \mathcal{N}\left(\mu_i(S) - \mu_i, \frac{\rho_i(t, S)\zeta_{i,S}^2}{N_o(t)} + \frac{\tilde{\sigma}_i^2}{N_i(t)}\right).$$

As t increases, this Gaussian distribution will concentrate around $\mu_i(S) - \mu_i$, which equals 0 if S is a valid backdoor adjustment set relative to X_i and Y . We assume that for any set S , either $|\mu_i(S) - \mu_i| \geq 2\delta$ for some $\delta > 0$ or $\mu_i(S) = \mu_i$. Under this assumption, if $S \notin \mathcal{B}_i$, the difference $\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t)$ will deviate from 0 significantly as t becomes large. This allows us to distinguish valid sets $S \in \mathcal{B}_i$ from invalid ones $S' \notin \mathcal{B}_i$ by checking whether the difference between the two estimates is close to zero.

One might choose $S^* = \operatorname{argmin}_S |\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t)|$ at each time step t as a candidate adjustment set relative to X_i and Y . This idea turns out to be too greedy and leads to sub-optimal choices for a bandit problem. To encourage more exploration, we instead construct confidence intervals $CI_i(t, S)$ at round t for $\mu_i(S) - \mu_i$ as

$$\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t) \pm c_2 \sqrt{\log t} \sqrt{\widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t, S)) + \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))}, \quad (2.10)$$

where $c_2 > 0$ is some constant. We choose sets S whose corresponding confidence intervals cover 0 as our candidate backdoor adjustment sets. If such sets do not exist, we choose the set whose confidence interval has the smallest distance to 0, where the distance for an interval

$[a, b]$ is defined as $d(0, [a, b]) = \min\{|x| : a \leq x \leq b\}$. Formally, we identify a collection of candidate backdoor adjustment sets relative to (X_i, Y) at time step t as

$$\hat{\mathcal{B}}_i(t) = \operatorname{argmin}_S d(0, CI_i(t, S)). \quad (2.11)$$

Note that $\hat{\mathcal{B}}_i(t)$ may consist of multiple sets whose confidence intervals cover 0. Without any restriction, the choice of the set S can be large especially when the number of nodes p is large. Let M_i be the regression neighborhood of node i , i.e., the support of the regression of X_i on X_{-i} , and let $\mathcal{N}_i$ denote the collection of all subsets of M_i . For a Gaussian DAG formulated by (2.1), the regression neighborhood M_i contains the parents of node i . Since the parent set $\text{pa}(i) \subseteq M_i$ always satisfies the backdoor criterion, we restrict $S \in \mathcal{N}_i$ in the minimization (2.11) to reduce the number of candidate sets.

2.3.3 Algorithm description

In this section, we describe the backdoor adjustment - upper confidence bound (BA-UCB) algorithm for causal bandit. Unlike most existing methods, our algorithm does not require prior knowledge or exploration of the underlying DAG. The main idea is to use estimates from both experimental data and observational data to identify backdoor adjustment sets alongside the decision process and then construct upper confidence bounds from both sources.

For a basic MAB problem, upper confidence bounds are of the form $\hat{\mu}_i(t) + \hat{\sigma}_i(t)c\sqrt{\log t}$, where $\hat{\mu}_i(t)$ and $\hat{\sigma}_i(t)$ are the mean and standard deviation of the sample mean of Y respectively. Since μ_i and σ_i^2 can be estimated from both experimental data as in (2.7) and observational data as in (2.8), we consider using weighted average of the two estimates to

construct UCBs, i.e.,

$$\begin{aligned}\hat{\mu}_i(t, S) &= \frac{N_o(t)\hat{\mu}_{i,\text{obs}}(t, S) + N_i(t)\hat{\mu}_{i,\text{int}}(t)}{N_o(t) + N_i(t)}, \\ \hat{\sigma}_i^2(t, S) &= \frac{N_o^2(t)\widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t, S)) + N_i^2(t)\widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))}{(N_o(t) + N_i(t))^2}.\end{aligned}\tag{2.12}$$

This effectively combines the two types of data to achieve higher estimation accuracy.

As discussed in Section 2.3.2, we restrict $S \in \mathcal{N}_i$ to find candidate backdoor adjustment sets relative to X_i and Y . We use observational data $\mathcal{D}_o(t)$ to estimate the regression neighborhood M_i of each node $i \in \bar{V}$. Given the estimated regression neighborhood $\widehat{M}_i$, we identify candidate backdoor adjustment sets $S \subseteq \widehat{M}_i$ as in (2.11), sample one set S from the candidate sets if $|\widehat{\mathcal{B}}_i(t)| \geq 2$, and construct the weighted version of UCBs to choose the intervention for the next round. The procedure is summarized in Algorithm 2.1. In practice, estimated regression neighborhoods and candidate backdoor adjustment sets will not vary significantly unless we collect enough new data. Therefore, to reduce computation, we update them every κ iterations.

Algorithm 2.1 BA-UCB for Gaussian DAGs

Input: Observational data $\mathcal{D}_o(0)$ of size n_0 , parameters c, c_3, n, m, κ , rounds T .

- 1: **for** $t = 0, \dots, T - 1$ **do**
- 2: **if** there is a node i which has been intervened less than $\lceil \frac{c_3 \log t + 1}{m} \rceil$ times **then**
- 3: Choose intervention node $A_{t+1} = i$;
- 4: **else**
- 5: **for** $i = 1, \dots, p - 1$ **do**
- 6: **if** $t \equiv 0 \pmod{\kappa}$ **then**
- 7: Estimate the regression neighborhood $\widehat{M}_i$ from $\mathcal{D}_o(t)$;
- 8: Identify candidate adjustment sets $\widehat{\mathcal{B}}_i = \text{argmin}_{S \subseteq \widehat{M}_i} d(0, CI_i(t, S))$;
- 9: Sample set S from $\widehat{\mathcal{B}}_i$, calculate $\hat{\mu}_{i,\text{obs}}(t, S)$ and $\hat{\sigma}_{i,\text{obs}}(t, S)$ as in (2.8);
- 10: Calculate estimates $\hat{\mu}_i(t, S)$, $\hat{\sigma}_i^2(t, S)$ as in (2.12);
- 11: $A_{t+1} = \text{argmax}_i \{\hat{\mu}_i(t, S) + \hat{\sigma}_i(t, S)c\sqrt{\log t}\}$;
- 12: Generate m data points under intervention A_{t+1} and n observational data.
- 13: Update $\hat{\mu}_{i,\text{int}}(t + 1)$ and $\widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t + 1))$ as in (2.7).

Output: Final intervention A_T and estimated reward $\hat{\mu}^*$.

Remark 2.1. According to Rule 2 of *do* Calculus by Pearl [28], if X_S blocks all backdoor paths from X to Y , we have $P(Y|do(X = x), X_S) = P(Y|X = x, X_S)$, i.e., conditioning on the backdoor adjustment set S , the intervention distribution $[Y|do(X = x), X_S]$ is identical to the conditional distribution $[Y|X = x, X_S]$. Hence, given a backdoor adjustment set S for node i , we can perform a single regression $Y \sim X_i + X_S$ on both experimental data under $do(X_i)$ and observational data to estimate μ_i and σ_i^2 . Thus, in Algorithm 2.1 Line 9, we could run one linear regression $Y \sim X_i + X_S$ using both experimental data $\mathcal{D}_i(t)$ and observational data $\mathcal{D}_o(t)$. In Line 10, we calculate $\hat{\mu}_i(t, S)$ and $\hat{\sigma}_i^2(t, S)$, as in (2.8) but with the combined data. This is another way to combine the two types of data to generate estimates and construct UCBs. However, this single-regression approach exhibits inferior empirical performance, as shown by the simulation results in Figure 2.3 (Section 2.5.2). The underlying issue is that when S is not a valid adjustment set, the estimator of μ_i from the regression $Y \sim X_i + X_S$ can be severely biased, particularly when applied to experimental data. Consequently, the single-regression estimator based on the pooled dataset $\mathcal{D}_i(t) \cup \mathcal{D}_o(t)$ could be very different from both the observational estimator $\hat{\mu}_{i,\text{obs}}(t, S)$ and the interventional estimator $\hat{\mu}_{i,\text{int}}(t)$, even though these two estimators are themselves close. In contrast, the weighted-average estimator in (2.12) combines the two estimates in a controlled manner, yielding a more robust estimate $\hat{\mu}_i(t)$ and superior performance in practice.

2.4 Regret analysis

In this section, we provide theoretical analysis on the cumulative regret $\mathcal{R}_T$ (2.4) of our proposed BA-UCB algorithm. Since using the same data to identify backdoor adjustment sets in Line 8 and to construct combined estimates in Line 10 of Algorithm 2.1 complicates the analysis, we consider sample splitting to simplify the proof. More specifically, we collect $2n_0$ observational data at the beginning of Algorithm 2.1, $2m$ experimental data and $2n$ observational data at each round in Line 12, and use half of the data to identify candidate

backdoor adjustment sets in Line 8 and the other half to calculate estimates in Line 10. All theoretical results in this section are established under this sample-splitting version of Algorithm 2.1. In practice, the algorithm works well without sample splitting, as demonstrated by the numerical experiments in Section 2.5.

2.4.1 Main results

The underlying model assumptions for Algorithm 2.1 are stated formally in Assumption 2.1.

Assumption 2.1. *The random vector $X = (X_1, \dots, X_p)$ satisfies the linear SEM (2.1) associated with a DAG $\mathcal{G}$.*

As explained in Section 2.3.2, we also assume a gap between $\mu_i(S)$ for $S \notin \mathcal{B}_i$ and the true causal effect μ_i , which is formally stated as Assumption 2.2.

Assumption 2.2 (Identifiability). *There exists $\delta > 0$, such that either $|\mu_i(S) - \mu_i| \geq 2\delta$ or $\mu_i(S) = \mu_i$, $\forall S \in \mathcal{N}_i$ and $i \in \{1, \dots, p-1\}$.*

Remark 2.2. Since we do not need to recover the entire graph, faithfulness assumption is not required in our work. Note that backdoor adjustment relies only on the recursive factorization of the joint distribution according to a DAG, which is implied by the SEM in (2.1).

If S satisfies the backdoor criterion, then $\mu_i(S)$ identifies the causal effect of X_i on Y ; in particular, the parent set $\text{pa}(i)$ of X_i always satisfies the backdoor criterion, and thus $\mu_i(\text{pa}(i))$ is the true causal effect. Under Assumption 2.2, for any S' that is not a valid backdoor adjustment set, there is a gap such that $|\mu_i(S') - \mu_i(\text{pa}(i))| \geq 2\delta$. Using Figure 2.1 as an example, suppose we want to estimate the causal effect of X_2 on $Y (= X_5)$, i.e. $\gamma_2 = \beta_{23}\beta_{34}\beta_{45}$. In this example, the regression neighborhood M_2 of X_2 consists of its parent X_1 and child X_3 , and thus $\mathcal{N}_2 = \{\emptyset, \{X_1\}, \{X_3\}, \{X_1, X_3\}\}$. Among these subsets, only $\{X_1\}$ satisfies the backdoor criterion relative to (X_2, Y) , i.e., $\mathcal{B}_2 = \{\{X_1\}\}$. From the SEM (2.1), the conditional expectations of Y given X_2 and different sets in $\mathcal{N}_2$ can be derived

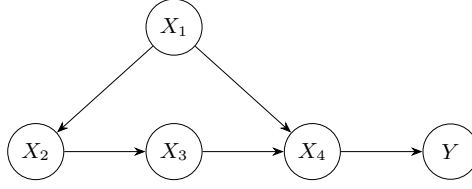


Figure 2.1: An example DAG to illustrate Assumption 2.2.

as follows:

$$\begin{aligned}
\mathbb{E}[Y \mid X_2] &= \beta_{45} \left(\frac{\beta_{12}\beta_{14}\omega_1^2}{\beta_{12}^2\omega_1^2 + \omega_2^2} + \beta_{34}\beta_{23} \right) X_2, \\
\mathbb{E}[Y \mid X_1, X_2] &= (\beta_{45}\beta_{14}) X_1 + (\beta_{45}\beta_{34}\beta_{23}) X_2, \\
\mathbb{E}[Y \mid X_2, X_3] &= \beta_{45} \left(\frac{\beta_{12}\beta_{14}\omega_1^2}{\beta_{12}^2\omega_1^2 + \omega_2^2} X_2 + \beta_{34}X_3 \right), \\
\mathbb{E}[Y \mid X_1, X_2, X_3] &= \beta_{45} (\beta_{14}X_1 + \beta_{34}X_3).
\end{aligned}$$

Then the minimum difference between the true causal effect $\gamma_2 = \beta_{23}\beta_{34}\beta_{45}$ and $\mu_2(S)$, $S \in \mathcal{N}_2 \setminus \mathcal{B}_2$ is

$$\min \left\{ \left| \frac{\beta_{12}\beta_{14}\beta_{45}\omega_1^2}{\beta_{12}^2\omega_1^2 + \omega_2^2} \right|, \left| \beta_{45} \left(\frac{\beta_{12}\beta_{14}\omega_1^2}{\beta_{12}^2\omega_1^2 + \omega_2^2} - \beta_{34}\beta_{23} \right) \right|, |\beta_{23}\beta_{34}\beta_{45}| \right\}.$$

Since the parameters β_{ij} are non-zero for directed edges, this difference becomes 0 only when $\beta_{12}(\beta_{14} - \beta_{23}\beta_{34}\beta_{12})\omega_1^2 = \beta_{23}\beta_{34}\omega_2^2$, a set of measure zero in the parameter space. This shows the rationale behind the gap 2δ in Assumption 2.2.

To make sure our algorithm works properly, we also need a mild assumption on the number of observational data for estimation of regression neighborhoods and $\hat{\beta}_i$ in (2.8) from backdoor adjustment sets. This is stated in Assumption 2.3.

Assumption 2.3. $n_0 + nK \geq p$.

Recall that $K = p - 1$ in our setting. This assumption is easily satisfied when $n_0, n \geq 1$ or $n_0 \geq p$. Let s_0 denote the maximum of in- and out-degrees among all nodes except the

reward variable. In fact, if we use Lasso-based neighborhood regression and restrict the search of candidate backdoor adjustment sets to subsets with cardinality $\leq s_0$ for computational efficiency, then Assumption 2.3 can be relaxed to $n_0 + nK \geq s_0 + 2$.

Recall a few relevant notations introduced earlier in this chapter. The joint distribution of X defined by the linear SEM (2.1) is $\mathcal{N}(0, \Sigma)$, where $\Sigma = (I - B)^{-\top} \Omega (I - B)^{-1}$, $\Omega = \text{diag}(\omega_1^2, \dots, \omega_p^2)$ and $B = (\beta_{ij})_{p \times p}$. The distribution of the reward variable Y under intervention is

$$Y|do(X_i = x_i) \sim \mathcal{N}(\mu_i, \tilde{\sigma}_i^2),$$

and the conditional distributions of $Y|X_i, X_S$ and $X_i|X_{-i}$ are

$$\begin{aligned} Y|X_i, X_S &\sim \mathcal{N}(\mu_i(S)X_i + \beta_S^\top X_S, \zeta_{i,S}^2), \\ X_i|X_{-i} &\sim \mathcal{N}(\sum_{j:j \neq i} \bar{\beta}_{ij}X_j, \bar{\sigma}_i^2). \end{aligned}$$

Let $\gamma_{\min}, \gamma_{\max}$, and tr denote the smallest eigenvalue, the largest eigenvalue, and the trace of a matrix, respectively. Let $s_1 = \max_i |\mathcal{M}_i|$ denote the maximum size of the regression neighborhood, and $s_2 = \max_i |\mathcal{N}_i \setminus \mathcal{B}_i|$ be the maximum number of subsets in the regression neighborhood that do not yield the true causal effect of X_i on Y across all i . Define a few constants:

$$\begin{aligned} \tilde{\eta}^2 &= \max_i \tilde{\eta}_i^2, \text{ where } \tilde{\eta}_i^2 = \max_{S \in \mathcal{B}_i} \left\{ \frac{9\zeta_{i,S}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}, \\ \tilde{\phi}^2 &= \max_i \max_{S \in \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S}^2}{\gamma_{\min}(\Sigma)\tilde{\sigma}_i^2}, \frac{25\tilde{\sigma}_i^2\gamma_{\max}(\Sigma)}{9\zeta_{i,S}^2} \right\}, \\ \psi^2 &= \max_i \max_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S'}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}, \\ \alpha &= \min_i \frac{\sqrt{\gamma_{\min}(\Sigma)}\bar{\beta}_i}{12\sqrt{3}\bar{\sigma}_i}, \text{ where } \bar{\beta}_i = \min_{j:\bar{\beta}_{ij} \neq 0} |\bar{\beta}_{ij}|, \\ \ell_0 &= \lceil \frac{9\text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil. \end{aligned} \tag{2.13}$$

Now we establish a case-dependent upper bound on the cumulative regret of the BA-UCB

algorithm. Here, “case-dependent” refers to a bound that explicitly depends on the gaps $\Delta_i = \mu^* - \mu_i$.

Theorem 2.1. *Under Assumptions 2.1, 2.2 and 2.3, the cumulative regret $\mathcal{R}_T$ of Algorithm 2.1, with parameters $c \geq 4\sqrt{2}\tilde{\phi}$ and $c_3 \geq \max\{64, 32\psi^2/\delta^2\}$, after T rounds is at most*

$$\log T \left(\frac{20c^2}{n+m} \sum_{i:\mu_i < \mu^*} \frac{\tilde{\eta}_i^2}{\Delta_i} + \frac{c_3 \sum_{i=1}^K \Delta_i}{m} \right) + C_3 \sum_{i=1}^K \Delta_i, \quad (2.14)$$

where $C_3 > 0$ is a constant that does not depend on T .

See Section 2.7.3.1 for a proof of this theorem. In general, C_3 depends on Σ , n_0 and n , and its exact expression is

$$\begin{aligned} C_3 = & \pi^2 \left(s_2 + \frac{4}{3} + \frac{1}{3} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + \ell_0 + \left(\frac{c_3}{n} \right)^2 + 1 \\ & + \frac{4K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{36(s_1 + s_2 + K + 3)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}. \end{aligned}$$

Theorem 2.1 shows the advantage of incorporating observational data into the UCB framework. The standard UCB algorithm has an upper bound of a similar form, such as in Theorem 4 of Auer et al. [1], which is

$$\frac{\log T}{M} \left(256 \sum_{i:\mu_i < \mu^*} \frac{\tilde{\sigma}_i^2}{\Delta_i} + 8 \sum_{i=1}^K \Delta_i \right) + \text{const} \cdot \sum_{i=1}^K \Delta_i$$

assuming M data points are generated under each round of intervention. Setting $n+m = M$ in the bound (2.14) will result in the same order of the leading term $O(\log T/M)$, showing that observational data, while much less costly, is as useful as experimental data in terms of reducing the regret for the BA-UCB algorithm. Increasing n , the size of the observational data in each round, directly reduces this dominating term and thus lower the bound. Moreover, the benefit of having an initial observational dataset is reflected in the constant C_3 , which also lowers the cumulative regret.

In the causal bandit literature, regret analyses typically fall into two categories: gap-dependent bounds that scale with Δ_i , and gap-independent bounds that hold uniformly over all problem instances. Since our algorithm is developed under Assumption 2.1, some dependence on the underlying distribution is unavoidable. Nevertheless, in line with the standard bandit setting, we also derive an upper bound that does not depend explicitly on the gaps Δ_i .

Specifically, assuming all rewards are bounded, i.e., $\mu_i \in [-1, 1]$ without loss of generality, and that the constructed upper confidence bounds are clipped to $[-1, 1]$ accordingly, we further establish a Δ -independent upper bound on the cumulative regret. While this bound does not involve Δ_i , it may still depend on other problem-specific parameters induced by the underlying distribution (e.g., through Σ).

Theorem 2.2. *Assume $\mu_i \in [-1, 1]$, $\forall i \in \{1, \dots, p-1\}$. Then under Assumptions 2.1, 2.2 and 2.3, the cumulative regret $\mathcal{R}_T$ of Algorithm 2.1, with parameters $c \geq 4\tilde{\phi}$ and $c_3 \geq \max\{32, 32\psi^2/\delta^2\}$, after T rounds is at most*

$$2c\tilde{\eta}\sqrt{\log T} \min \left\{ \frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m}, \frac{\sqrt{n_0 + nT} - \sqrt{n_0}}{n} 1_{\{n \geq 1\}} \right\} + \frac{2Kc_3 \log T}{m} + C_4, \quad (2.15)$$

where C_4 is a constant that does not depend on T .

See Section 2.7.3.2 for a proof of this theorem. In general, C_4 depends on Σ , n_0 , m , and n , and the exact expression is

$$C_4 = \pi^2 \left(2s_2 + \frac{7}{3} + \frac{1}{6} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + 2\ell_0 + 2 \left(\frac{c_3}{n} \right)^2 + 2K + \frac{8K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{72(s_1 + s_2 + K + 3)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}.$$

2.4.2 Implications

To better understand how observational data can benefit the sequential decision process, we take a closer look at (2.15). Since T is usually much larger than K , the first term in (2.15) is the leading term. If n_0 is large relative to T , more specifically $n_0 \geq T(n+m)$, we have

$$\begin{aligned} \frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m} &= \frac{\sqrt{K^2n_0}}{n+m} \left(\sqrt{1 + \frac{(n+m)T}{Kn_0}} - 1 \right) \\ &\leq \frac{T}{2\sqrt{n_0}} \leq \frac{1}{2} \sqrt{\frac{T}{n+m}}, \end{aligned}$$

where the first inequality is due to $(1+x)^{1/2} \leq 1+x/2$ for $x > 0$. In this case, the leading term of the cumulative regret is $O(\sqrt{T \log T/(n+m)})$, which does not depend on K and decreases with the total number $2(m+n)$ of experimental and observational data collected at each round. The cumulative regret of a standard UCB algorithm that generates $2(n+m)$ experimental data at each round is $O(\sqrt{KT \log T/(n+m)})$. This shows that by using a large prior observational data with backdoor adjustment, we reduce the regret substantially by a factor of $O(\sqrt{K})$ and remove its dependency on the action set size K . Moreover, the value of observational data generated at each round is comparable to that of experimental data in terms of minimizing the cumulative regret.

If n_0 is small relative to T , since $\sqrt{x+a} - \sqrt{x} \leq \sqrt{a}$ when $a > 0$, the leading term in (2.15) is bounded by the minimum between $O(\sqrt{KT \log T/(n+m)})$ and $O(\sqrt{T \log T/n})$. In general, it is cheaper to generate observational data than experimental data and thus m/n is often chosen to be small, say $m/n \leq 1$. Therefore, for many problems, we expect to have $K > 1 + m/n$ and in this case the leading term of (2.15) is $O(\sqrt{T \log T/n})$, which does not depend on K either. If the action set is large, i.e., $K \gg 1 + m/n$, then the BA-UCB bound $O(\sqrt{T \log T/n}) + O(K \log T/m)$ will be much smaller than the bound of the standard UCB $O(\sqrt{KT \log T/(n+m)})$ discussed above, under the typical setting that $T \gg K^2 \log T$.

In summary, as long as we collect enough observational data with either a large initial n_0

or during the bandit process ($n \geq 1$), we can relax the dependence of the upper bound (2.15) for the cumulative regret on the number of arms K , and substantially reduce the order of the cumulative regret bound compared to the traditional MAB algorithms.

The regret analysis demonstrates the advantage of the BA-UCB algorithm using observational data for the causal bandit problem. With backdoor adjustment, observational data can be used to estimate the causal effects of all arms simultaneously, which substantially relaxes the dependence of the cumulative regret on the number of arms. Note that, even with a sufficiently large amount of observational data, one can at most recover the equivalence class of a DAG, represented by a completed partially directed acyclic graph (CPDAG). In the Gaussian case, however, one cannot reduce the action set given a CPDAG to a small subset of the nodes in general. Thus, our approach is fundamentally different from the existing works [22, 24] that use prior data or knowledge to reduce the action set before a sequential bandit algorithm.

2.4.3 Proof outline

Complete proofs of Theorem 2.1 and 2.2 can be found in Section 2.7. Here, we give an outline of the main steps of the proofs.

We first establish case-dependent and Δ -independent upper bounds on the cumulative regret assuming the underlying causal DAG is given. When the graph structure is given, a valid adjustment set of each node i is its parent set $\text{pa}(i)$, which is used to generate estimates of causal effects from observational data in our BA-UCB algorithm. Then the central component is to show the concentration of the estimated mean $\hat{\mu}_i(t, \text{pa}(i))$ and variance $\hat{\sigma}_i^2(t, \text{pa}(i))$ in (2.12) around their expectations. By applying certain tail bounds, we control the deviation probabilities of these estimators after a sufficiently large number of rounds (i.e., t is large). Given the gaps Δ_i between the optimal and suboptimal arms, together with the concentration properties of the estimators, we can upper bound the probability $\mathbb{P}(A_t = i)$ for each sub-optimal arm i for a large enough t , leading to a finite upper bound

on the cumulative regret.

When the DAG structure is unknown, the regret analysis proceeds as above once valid backdoor adjustment sets have been identified. Conditional on a valid adjustment set, the cumulative regret bound coincides with that in the known-DAG scenario. Let $S(i, t)$ denote the sampled adjustment set for node i at time step t in Algorithm 2.1 Line 9. The main additional challenge lies in bounding the probability that the algorithm fails to identify a valid adjustment set at a given round, i.e., $\mathbb{P}(S(i, t) \notin \mathcal{B}_i)$. Such an error may arise from inaccurate estimation of the regression neighborhood in Line 7 and incorrect identification of candidate backdoor adjustment sets in Line 8. And we can control the error by bounding: (i) $\mathbb{P}(\widehat{M}_i \neq M_i)$, (ii) $\mathbb{P}(0 \notin CI_i(t, \text{pa}(i)))$, and (iii) $\mathbb{P}(0 \in CI_i(t, S') \text{ for some } S' \in \mathcal{N}_i \setminus \mathcal{B}_i)$.

Under the Gaussian linear SEM (2.1), we show that for each i , the estimated neighbor set $\widehat{M}_i$ satisfies $\mathbb{P}(\widehat{M}_i \neq M_i) = O(t^{-2})$, provided that the observational sample size $N_o(t)$ grows at least on the order of $\log t$. This result follows from Gaussian concentration inequalities for the t -statistics in linear regression and a union bound across all variables. Once M_i is recovered, the probability that $S(i, t)$ in Line 9 does not belong to the family of valid backdoor adjustment sets $\mathcal{B}_i$ can be upper bounded using the concentration properties of the difference $\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t)$ in (2.10). More specifically, we show that

$$\begin{aligned} \mathbb{P}(0 \in CI_i(t, S)) &\leq O(t^{-2}), \quad \forall S \in \mathcal{N}_i \setminus \mathcal{B}_i, \\ \mathbb{P}(0 \notin CI_i(t, S)) &\leq O(t^{-2}), \quad \forall S \in \mathcal{B}_i, \end{aligned}$$

provided that the sample sizes $N_i(t)$ and $N_o(t)$ both grow at least on the order of $\log t$. Combining the error bounds of the three events, we arrive at $\mathbb{P}(S(i, t) \notin \mathcal{B}_i) = O(t^{-2})$, and hence $\sum_{t=1}^{\infty} \mathbb{P}(S(i, t) \notin \mathcal{B}_i)$ is finite. Therefore, the increase in the cumulative regret due to these errors is bounded by a constant, and the dominant term in the regret bound remains the same as in the DAG-known case.

2.5 Numerical experiments

In this section, we evaluate the empirical performance of the BA-UCB algorithm (without sample splitting) through simulation. In our main experiments, we compare BA-UCB with three competing methods: the standard UCB, the central node (CN) - UCB [22], and the Bayesian backdoor bandit (BBB)- UCB [25].

2.5.1 Simulation settings

We simulate DAGs with the number of nodes $p \in \{10, 20, 30, 50\}$. To generate a Gaussian DAG, we randomly generate a DAG using the function `random.graph` in the `bnlearn` package, sample coefficients β_{ij} from $[-1, -0.5] \cup [0.5, 1]$ for $i \in \text{pa}(j)$ and standard deviations ω_i from $[\sqrt{0.2}, \sqrt{0.5}]$, and normalize each variable to have unit variance. We set maximum in-degree to be 3 and choose the last variable in the topological sort as the reward variable Y to minimize the number of trivial arms. In each simulation, we randomly generated 100 Gaussian DAGs, ran each bandit algorithm 5 times on each DAG with $T = 5000$ time steps, and recorded the average cumulative regrets.

Since the average execution time of BBB-UCB was 8.2 seconds for each iteration on a Gaussian DAG with 10 nodes, it is prohibitive to run it for 5000 time steps on 100 Gaussian DAGs. So for $p = 10$, we directly used the Gaussian DAGs and the corresponding simulation results provided by Huang and Zhou [25], available at <https://github.com/jirehhuang/bcb>. Since the computation time increases exponentially with the number of nodes, we exclude BBB-UCB from our comparison for $p \geq 20$.

The CN-UCB algorithm was proposed for discrete network settings, aiming to first identify the parents of the reward variable and then restrict the action set of UCB to the identified parent nodes. To mimic the mechanism under Gaussian settings, we first apply the PC algorithm [29] to identify neighbors of the reward variable Y using observational data, where a neighbor of X_i could be a parent ($X_i \rightarrow Y$) or be connected via an undirected edge ($X_i - Y$),

as the PC algorithm in general learns a partially directed acyclic graph. Then, we apply the UCB algorithm with these identified neighbors as the arms. The significance level of the PC algorithm for conditional independence test was set to $\alpha = 0.1$, which is slightly higher than the common cutoff of 0.05, so that true neighbors are less likely to be missed. Each node was intervened on at least 5 times at the beginning to allow the UCB algorithm to accumulate enough data. Note that, unlike discrete settings, the optimal intervention with the largest causal effect may not be a parent node of the reward under Gaussian settings. Consequently, the performance of this procedure depends on whether the optimal arm is in $\text{pa}(Y)$.

For our BA-UCB algorithm, we chose $\tau = 1/\sqrt{N_o(t)}$ as the threshold for p-values when estimating the regression neighborhood through linear regression. After several trials, we set $c = c_2 = 1/\sqrt{2}$ for all settings, where c and c_2 are the constants in (2.5) and (2.10), respectively. To reduce computational cost, we only considered $\{S \in \hat{\mathcal{N}}_i : |S| \leq 3\}$ in the identification of candidate backdoor adjustment sets in Algorithm 2.1. We estimated the regression neighborhood M_i , calculated $\hat{\mathcal{B}}_i$, and updated the estimates from observational data every κ steps, where $\kappa = 20, 50, 100, 200$ for $p = 10, 20, 30, 50$, respectively, so that the computation time of 5000 iterations was under 5 minutes for all settings. The numerical results were all generated on a macOS system with M1 chip, 8-core CPU, and 16GB memory.

2.5.2 Simulation results of BA-UCB

In this subsection, we present simulation results to evaluate different aspects of the BA-UCB algorithm. We begin with the setting where the true graph is given, which highlights the benefit of information sharing through observational data. We then compare two estimation approaches, i.e., the single regression versus weighted average (see Remark 2.1), and show that the empirical performance of the weighted average approach is better. Finally, we assess whether observational and experimental estimates are consistent and whether the combined estimate converges to the true reward, and validate the design of our algorithm by showing

the accuracy in identifying a true adjustment set.

2.5.2.1 Results when the true DAG is given

To highlight the benefit of having observational data, we examined the special case in which the underlying graph is given, for example through domain knowledge. In this setting, we ran our BA-UCB algorithm using the parent set of a node as the backdoor adjustment set and compared it to the standard UCB algorithm. For a fair comparison, BA-UCB collected $n = 5$ observational samples and $m = 5$ experimental samples per round, starting with no prior data, whereas UCB collected $m = 10$ experimental samples per round. Each variable was intervened upon at least once in both algorithms to ensure proper initialization. We recorded the median cumulative regret across five runs of each generated graph and reported the average over 100 random DAGs in Figure 2.2.

As shown in Figure 2.2, BA-UCB effectively avoids being trapped in sub-optimal arms and achieves much smaller empirical cumulative regret very quickly, while using the same total number of collected samples as the standard UCB algorithm. Moreover, the performance gap widens as the graph size p increases, highlighting the efficiency of information sharing through observational data. Here, information sharing refers to the fact that, via valid backdoor adjustment, observational data can simultaneously inform the estimation of causal effects across all arms. As the number of arms grows, this information sharing becomes increasingly valuable, resulting in higher efficiency gains for BA-UCB in larger graphs.

2.5.2.2 Single regression vs. weighted average estimates

Hereafter, we assume the underlying DAG is unknown for all numerical results. We compare BA-UCB with estimates constructed by the single regression approach described in Remark 2.1 and by the weighted average as in (2.12). The two approaches only differ in the way to combine observational and experimental data for constructing upper confidence

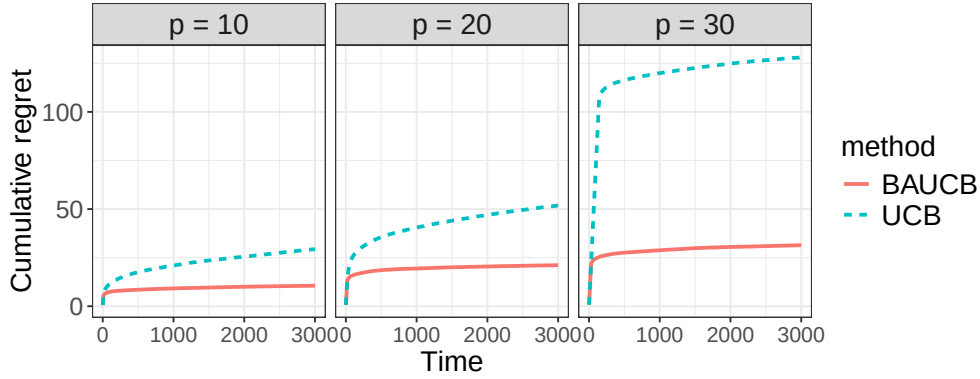


Figure 2.2: Comparison of empirical cumulative regrets over 3000 time steps between the BA-UCB and standard UCB algorithms when DAG is given for $p = 10, 20, 30$.

bounds. Given a backdoor adjustment set, the single regression approach runs one linear regression with observational data and experimental data combined together, in contrast to the weighted average of separate estimates from the two sources of data as in Line 10 of Algorithm 2.1.

In the simulations, both approaches used $m = n = 5$, a minimum intervention time of 1 for each arm, and 20 initial experimental samples per arm. The number of initial observational samples was set to $n_0 = 180, 380, 580$ for $p = 10, 20, 30$, respectively. For each generated graph, we reported the median cumulative regret over five runs and averaged the results across 100 random DAGs.

As shown in Figure 2.3, the weighted-average approach consistently achieves lower cumulative regret than the single-regression approach across all graph sizes considered. In BA-UCB, adjustment sets are identified based on the criterion that the observational estimate $\hat{\mu}_{i,\text{obs}}(t)$ obtained using a candidate set as in (2.8) is close to the corresponding interventional estimate $\hat{\mu}_{i,\text{int}}(t)$ as in (2.7). However, this does not guarantee that the identified set is a valid backdoor adjustment set, especially at the early stage of the algorithm. When an invalid adjustment set is selected, it is still possible for both $\hat{\mu}_{i,\text{obs}}(t)$ and $\hat{\mu}_{i,\text{int}}(t)$ to remain close to each other, and to the true causal effect μ_i , so that the weighted-average estimator remains accurate.

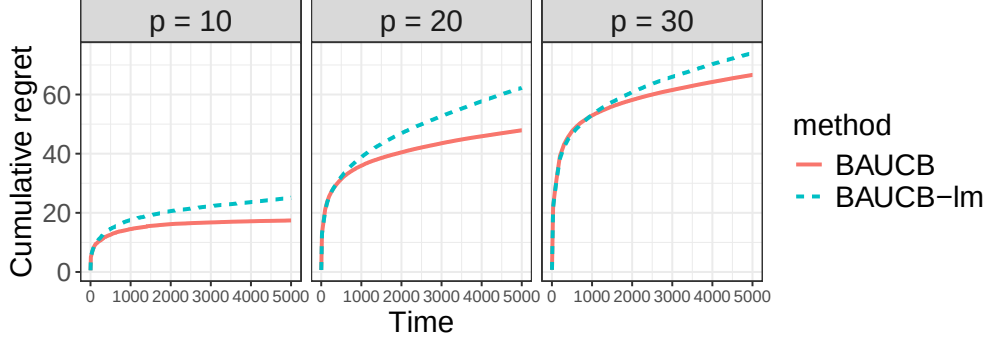


Figure 2.3: Comparison of empirical cumulative regret over $T = 5000$ time steps for BA-UCB on DAGs of $p = 10, 20, 30$. BAUCB-lm uses estimates from the single linear regression and BAUCB uses weighted average.

In contrast, the single-regression approach relies directly on the identified adjustment set when estimating the causal effect. If the selected adjustment set S is invalid, the regression $Y \sim X_i + X_S$ fitted on the combined data $\mathcal{D}_o(t) \cup \mathcal{D}_i(t)$ can yield estimates that deviate substantially from both $\hat{\mu}_{i,\text{obs}}(t)$ and $\hat{\mu}_{i,\text{int}}(t)$, leading to inferior empirical performance. By decoupling experimental estimate $\hat{\mu}_{i,\text{int}}(t)$ from adjustment set selection, the weighted-average approach is therefore more robust to adjustment set mis-specification. Consequently, for the subsequent results, we only focus on the weighted average approach.

2.5.2.3 Consistency of optimal arm estimation

To assess whether BA-UCB behaves as designed, we examined the consistency of the estimated reward for the optimal arm by analyzing the results from the simulations in Section 2.5.2.2. Specifically, we evaluated whether the observational estimate $\hat{\mu}_{\text{obs}}^*(t)$ and the interventional estimate $\hat{\mu}_{\text{int}}^*(t)$, defined in (2.8) and (2.7), converge over time, and whether the combined estimator $\hat{\mu}^*(t)$ in (2.12) converges to the true optimal reward μ^* as t increases.

Figure 2.4 summarizes estimation accuracy for the optimal arm. In both panels, the median error decreases steadily over time, indicating that observational and experimental estimates become increasingly aligned and that the estimated reward converges toward the true reward value. This behavior is consistent with improved identification of backdoor ad-

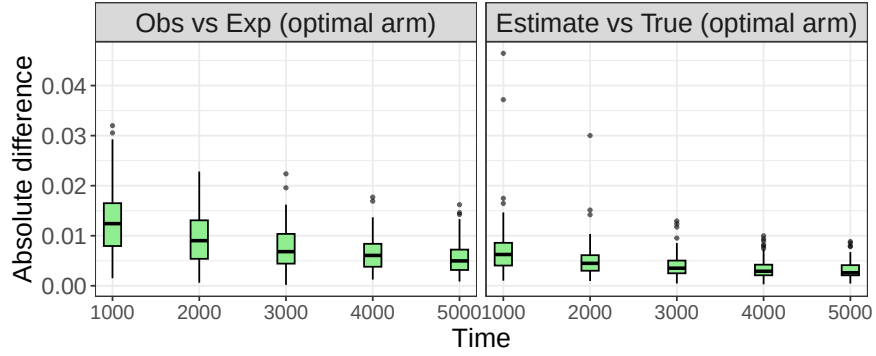


Figure 2.4: Estimation accuracy of the optimal arm over time. The left panel shows the absolute difference between reward estimates based on observational and experimental data, while the right panel reports the absolute difference between the estimated reward and the true value. The box-plots summarize results across 100 randomly generated DAGs with $p = 10$.

justment sets and leads to diminishing simple regret as the algorithm progresses. Outliers observed at early stages are expected, as adjustment set identification is noisy when limited data are available. Outliers observed at later stages can be attributed to the simulation design, which enforces only a small minimum number of experimental samples. Specifically, each arm is required to have at least 25 experimental samples, consisting of 20 initial experimental samples and at least one intervention round (Line 2 in Algorithm 2.1), which contributes $m = 5$ additional samples. Increasing the minimum intervention time is expected to enable more reliable identification of backdoor adjustment sets and more accurate estimation of the optimal reward.

To evaluate the accuracy of adjustment set identification, we recorded the frequency of the sampled adjustment set for the optimal arm being a true adjustment set over 5 independent runs and 100 graphs at each time step. Figure 2.5 reports this average identification rate over time. An identified adjustment set is regarded as correct if $|\mu_i(S) - \mu_i| < 0.01$, where $\mu_i(S)$ is defined in (2.9). Figure 2.5 shows a clear increasing trend over time for all graph sizes p , demonstrating that BA-UCB progressively improves its identification of valid adjustment sets as more data are collected. Nevertheless, because BA-UCB chooses an adjustment set even when the confidence interval $CI_i(t, S)$ in (2.10) does not cover 0, incorrect adjustment sets may still be selected, especially at early stages.

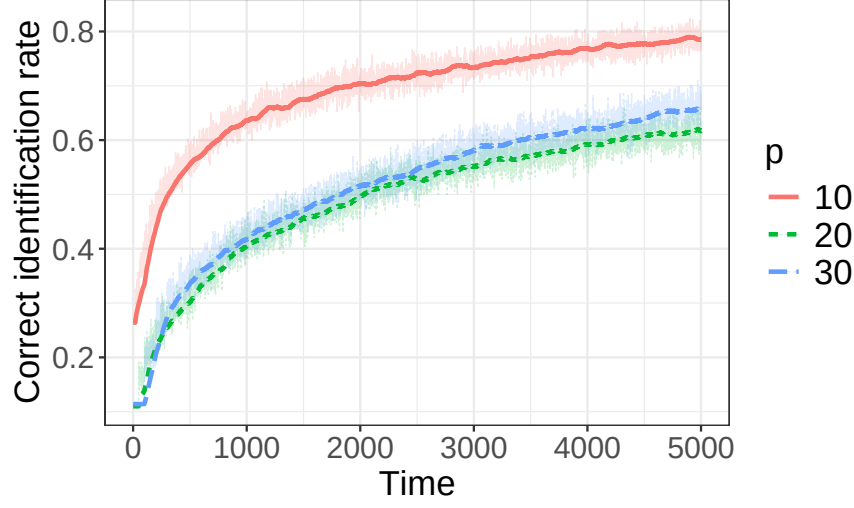


Figure 2.5: Average identification rate of a correct adjustment set for the optimal arm at each time step. Solid curves are smoothing averages with a 100-step window for better visualization.

2.5.3 Comparison with other algorithms

In this subsection, we compare our BA-UCB algorithm with existing causal bandit algorithms, including BBB-UCB, CN-UCB, and the standard UCB. We first present results for small graphs with $p = 10$ to illustrate the differences in performance and computational cost across methods. We then move to larger graphs with $p = 20, 30, 50$, excluding BBB-UCB due to its high computational cost. These comparisons demonstrate that BA-UCB attains consistently lower regret than UCB and CN-UCB, while offering substantially greater scalability than BBB-UCB.

Recall that n_0 is the number of initial observational data, n is the number of observational data collected at each round, and m is the number of experimental data collected at each round. In the comparison on small graphs with $p = 10$, the results of the BBB-UCB algorithm were generated with $n_0 = 320$ and $m = 1$. We chose $n_0 = 320$ and $n = m = 1$ for our BA-UCB algorithm, under which our algorithm took much less time than the BBB-UCB algorithm for $T = 5000$ steps. For the CN-UCB algorithm, we set $n_0 = 320$ and $m = 2$. For the UCB algorithm, we collected 4 experimental data from each arm at the

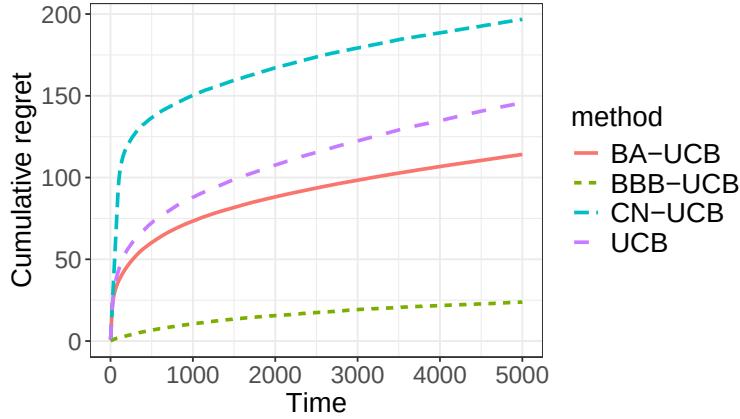


Figure 2.6: Comparison of empirical cumulative regret over $T = 5000$ time steps among the BA-UCB, BBB-UCB, CN-UCB, and UCB algorithms on DAGs with $p = 10$.

beginning, i.e., 36 initial experimental data in total, before incurring regrets, and set $m = 2$ for each round. Assuming the cost of collecting experimental data is around 10 times the cost of observational data, the initial cost of the UCB algorithm was comparable to that of the other three algorithms. Note that $m = 2$ experimental data were simulated at each round of CN-UCB and UCB, while we only generated $m = 1$ experimental data and $n = 1$ observational data for our BA-UCB algorithm. While these three algorithms collected the same number of data points at each round, the comparison setting here is in favor of the CN-UCB and the UCB algorithms because experimental data are much more informative and also more expensive than observational data.

As shown in Figure 2.6, our BA-UCB algorithm achieved a much lower average cumulative regret than the standard UCB algorithm and the CN-UCB algorithm. CN-UCB is seen to have poor performance because for 22 of the 100 Gaussian DAGs, the optimal arm does not correspond to intervention on a parent node of the reward variable. BBB-UCB achieved excellent performance at the cost of significantly higher computation time and resources. More specifically, BBB-UCB required exact computation of the parent set posterior probabilities at each time step, of which the complexity is $O(p \cdot 3^p)$. This is not feasible even for a moderate $p \geq 20$.

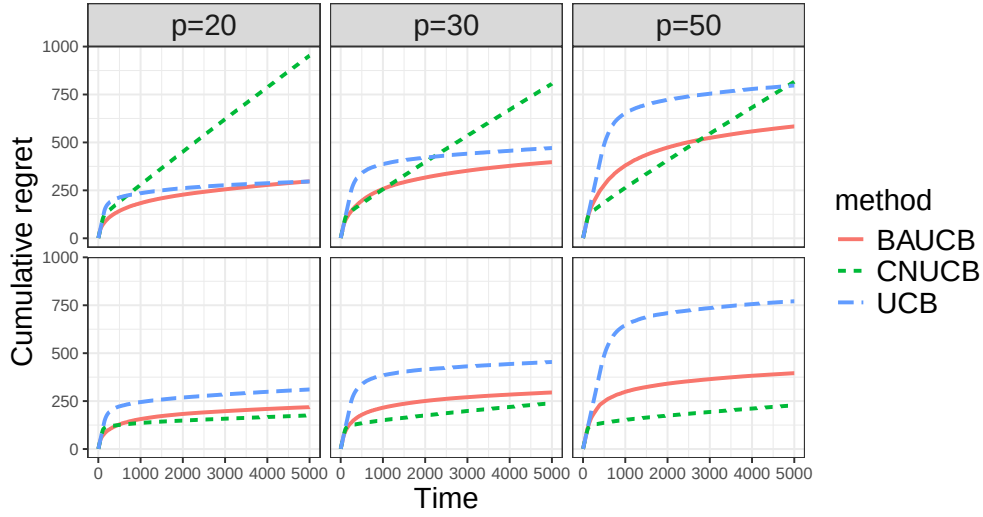


Figure 2.7: Comparison of empirical cumulative regret over 5000 time steps among BA-UCB, CN-UCB, and UCB algorithms for $p = 20, 30, 50$. The top panel reports the cases where the optimal arm is not a parent of the reward variable and the lower panel reports the cases where the optimal arm is a parent of the reward variable.

Since the BBB-UCB algorithm does not scale well with p , we excluded it from our next comparison for $p = 20, 30, 50$. The parameters were set as $n_0 = 100$ and $m = n = 1$ for BA-UCB; $n_0 = 5100$ and $m = 1$ for CN-UCB; $m = 2$ for UCB after a burning period of the first 50 steps. Note that the total number of data generated was the same for all three algorithms over $T = 5000$ steps. Since observational data are much cheaper to generate, BA-UCB and CN-UCB would be much less costly than the UCB algorithm in practical applications. Given that the performance of CN-UCB strongly depends on whether or not the optimal intervention is on a parent node of the reward variable, we report the results for these two different scenarios separately in Figure 2.7.

It is seen from Figure 2.7 that the BA-UCB algorithm outperforms uniformly the UCB algorithm across all scenarios. The cumulative regrets for both algorithms, as well as the difference between the two, increase as p increases. The CN-UCB algorithm is identical to the standard UCB algorithm but with a reduced set of nodes for intervention, which is usually much smaller than p . When the optimal arm indeed corresponds to intervention on a parent of the reward variable, it is not surprising that CN-UCB outperforms our method as

shown in the lower panel of the figure. However, in practice, one cannot guarantee that this is always the case. As reported in the top panel, the CN-UCB algorithm shows substantially inferior performance when the optimal arm is not an intervention on a parent node, and in such cases the cumulative regret increases linearly.

2.6 Discussion

In this chapter, we studied the causal bandit problem under the setting of Gaussian DAGs. Without assuming any prior knowledge about the underlying DAG structure, we proposed the BA-UCB algorithm which identifies candidate adjustment sets using estimates from both observational and experimental data. These candidate sets are used to construct weighted upper confidence bounds for making sequential decisions. The reason for updating adjustment sets along with more interventions is that the structure of a DAG cannot be fully recovered even with a sufficiently large number of observational data due to Markov equivalence. Therefore, our approach is fundamentally different from other existing works that use prior data or knowledge to reduce the action set before a sequential bandit algorithm.

Our worst case upper bound on the cumulative regret is proven to be of the order $O(\sqrt{T \log T / (n + m)})$ with a large number of initial observations, which does not depend on the size of action set and decreases with the number of data collected at each round. Empirical study shows that the cumulative regret of our algorithm is smaller than the UCB and the CN-UCB algorithms, especially when the optimal arm is not a parent of the reward. Moreover, it is much more scalable with the graph size than the BBB-UCB algorithm.

2.7 Appendix

2.7.1 Some general bounds

Definition 2.1. If each $x_i \in \mathbb{R}^p$ is drawn independently from $\mathcal{N}(0, \Sigma)$, then the random matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, with $x_i^\top$ as its i^{th} row, is said to be drawn from the Σ -Gaussian ensemble.

Lemma 2.1. Let $\mathbf{X} \in \mathbb{R}^{N_o(t) \times p}$ be drawn according to the Σ -Gaussian ensemble with $\Sigma > 0$.

For any $i \in V = \{1, \dots, p\}$ and any $S \subseteq V \setminus \{i\}$, let $\mathbf{X}_{i,S} = [\mathbf{X}_i \mid \mathbf{X}_S]$ and $\rho_i(t, S) = \left[\left(\frac{\mathbf{X}_{i,S}^\top \mathbf{X}_{i,S}}{N_o(t)} \right)^{-1} \right]_{i,i}$, then

$$\left(\frac{\sqrt{N_o(t)}}{\sigma_{\max}(\mathbf{X})} \right)^2 \leq \rho_i(t, S) \leq \left(\frac{\sqrt{N_o(t)}}{\sigma_{\min}(\mathbf{X})} \right)^2.$$

Further, when $N_o(t) \geq \max \left\{ \frac{9 \operatorname{tr}(\Sigma)}{\gamma_{\min}(\Sigma)}, p \right\}$,

$$\begin{aligned} \mathbb{P} \left(\rho_i(t, S) \geq \frac{9}{\gamma_{\min}(\Sigma)}, \text{ for some } i \in V \text{ and } S \subseteq V \setminus \{i\} \right) &\leq \mathbb{P} \left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N_o(t)}} \leq \frac{1}{3} \gamma_{\min}(\sqrt{\Sigma}) \right) \\ &\leq \exp \left(-\frac{N_o(t)}{18} \right), \\ \mathbb{P} \left(\rho_i(t, S) \leq \frac{9}{25 \gamma_{\max}(\Sigma)}, \text{ for some } i \in V \text{ and } S \subseteq V \setminus \{i\} \right) &\leq \mathbb{P} \left(\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N_o(t)}} \geq \frac{5}{3} \gamma_{\max}(\sqrt{\Sigma}) \right) \\ &\leq \exp \left(-\frac{N_o(t)}{18} \right). \end{aligned} \tag{2.16}$$

Proof. From Cauchy interlacing theorem, for all $i \in V$ and all $S \subseteq V \setminus \{i\}$,

$$\begin{aligned} \rho_i(t, S) &= \left[\left(\frac{\mathbf{X}_{i,S}^\top \mathbf{X}_{i,S}}{N_o(t)} \right)^{-1} \right]_{i,i} \leq \gamma_{\max} \left[\left(\frac{\mathbf{X}_{i,S}^\top \mathbf{X}_{i,S}}{N_o(t)} \right)^{-1} \right] = \left[\gamma_{\min} \left(\frac{\mathbf{X}_{i,S}^\top \mathbf{X}_{i,S}}{N_o(t)} \right) \right]^{-1} \\ &\leq \left[\gamma_{\min} \left(\frac{\mathbf{X}^\top \mathbf{X}}{N_o(t)} \right) \right]^{-1} = \left(\frac{\sqrt{N_o(t)}}{\sigma_{\min}(\mathbf{X})} \right)^2. \end{aligned}$$

Theorem 6.1 in [30] states that: Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be drawn according to the Σ -Gaussian ensemble. Then for all $\delta > 0$, the maximum singular value $\sigma_{\max}(\mathbf{X})$ satisfies the upper deviation inequality

$$\mathbb{P} \left[\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\Sigma})(1 + \delta) + \sqrt{\frac{\text{tr}(\Sigma)}{n}} \right] \leq e^{-n\delta^2/2}.$$

Moreover, for $n \geq d$, the minimum singular value $\sigma_{\min}(\mathbf{X})$ satisfies the analogous lower deviation inequality

$$\mathbb{P} \left[\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\Sigma})(1 - \delta) - \sqrt{\frac{\text{tr}(\Sigma)}{n}} \right] \leq e^{-n\delta^2/2}.$$

Since $N_o(t) \geq \frac{9 \text{tr}(\Sigma)}{\gamma_{\min}(\Sigma)}$, choose $\delta = \frac{2}{3} - \sqrt{\frac{\text{tr}(\Sigma)}{N_o(t)\gamma_{\min}(\Sigma)}} \geq \frac{1}{3}$, then

$$\mathbb{P} \left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N_o(t)}} \leq \frac{1}{3} \gamma_{\min}(\sqrt{\Sigma}) \right) \leq \exp \left(-\frac{N_o(t)\delta^2}{2} \right) \leq \exp \left(-\frac{N_o(t)}{18} \right).$$

Similarly, for all $i \in V$ and $S \subseteq V \setminus \{i\}$,

$$\rho_i(t, S) \geq \left(\frac{\sqrt{N_o(t)}}{\sigma_{\max}(\mathbf{X})} \right)^2,$$

and

$$\mathbb{P} \left(\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N_o(t)}} \geq \gamma_{\max}(\sqrt{\Sigma})(1 + \delta) + \sqrt{\frac{\text{tr}(\Sigma)}{N_o(t)}} \right) \leq \exp(-N_o(t)\delta^2/2).$$

Let $\delta = \frac{2}{3} - \sqrt{\frac{\text{tr}(\Sigma)}{N_o(t)\gamma_{\max}(\Sigma)}} \geq \frac{1}{3}$, then

$$\mathbb{P} \left(\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N_o(t)}} \geq \frac{5}{3} \gamma_{\max}(\sqrt{\Sigma}) \right) \leq \exp \left(-\frac{N_o(t)}{18} \right).$$

□

Remark 2.3. Lemma 2.1 applies to not only $\rho_i(t, S)$. In fact, when $\mathbf{X}$ is drawn from a Σ -Gaussian ensemble with $\Sigma > 0$, we can similarly prove the following results:

$$\begin{aligned} \mathbb{P} \left(\left[\left(\frac{\mathbf{X}_S^\top \mathbf{X}_S}{N_o(t)} \right)^{-1} \right]_{j,j} \geq \frac{9}{\gamma_{\min}(\Sigma)}, \text{ for some } S \subseteq V \text{ and } j \in S \right) &\leq \exp \left(-\frac{N_o(t)}{18} \right), \\ \mathbb{P} \left(\left[\left(\frac{\mathbf{X}_S^\top \mathbf{X}_S}{N_o(t)} \right)^{-1} \right]_{j,j} \leq \frac{9}{25\gamma_{\max}(\Sigma)}, \text{ for some } S \subseteq V \text{ and } j \in S \right) &\leq \exp \left(-\frac{N_o(t)}{18} \right). \end{aligned}$$

Lemma 2.2. *If $X \sim \chi_n^2$, then*

$$\mathbb{P}(X \geq 5n/2) \leq \exp(-n/4),$$

$$\mathbb{P}(X \leq n/3) \leq \exp(-n/9).$$

Proof. From Lemma 1 in [31],

$$\mathbb{P}(\chi_n^2 - n \geq 2\sqrt{n}\sqrt{x} + 2x) \leq e^{-x},$$

$$\mathbb{P}(\chi_n^2 - n \leq -2\sqrt{n}\sqrt{x}) \leq e^{-x}.$$

Let $x = \frac{n}{4}$ and $x = \frac{n}{9}$ respectively, then $\mathbb{P}(\chi_n^2 \geq 5n/2) \leq e^{-n/4}$ and $\mathbb{P}(\chi_n^2 \leq n/3) \leq e^{-n/9}$. □

Lemma 2.3. *If $X \sim t_n$, then for any $\delta > 0$,*

$$\mathbb{P}(|X| \geq \delta) \leq 2e^{-\delta^2/4} + e^{-n/16}.$$

Proof.

$$\mathbb{P}(|X| \geq \delta) = \mathbb{P} \left(\frac{|\mathcal{N}(0, 1)|}{\sqrt{\chi_n^2/n}} \geq \delta \right) \leq \mathbb{P}(|\mathcal{N}(0, 2)| \geq \delta) + \mathbb{P}(\chi_n^2 \leq n/2).$$

From Lemma 1 in [31], $\mathbb{P}(\chi_n^2 - n \leq -2\sqrt{n}\sqrt{x}) \leq e^{-x}$. Let $x = n/16$, then

$$\mathbb{P}(\chi_n^2 \leq n/2) \leq \exp(-n/16).$$

Combined with Gaussian tail probability bound, we have

$$\mathbb{P}(|X| \geq \delta) \leq 2\exp(-\delta^2/4) + \exp(-n/16).$$

□

2.7.2 When the underlying DAG is given

When the DAG is given, the parent set of each node X_i satisfies the backdoor criterion relative to (X_i, Y) . So Algorithm 2.1 can be reduced to the algorithm below.

Algorithm 2.2 Combined-UCB for Gaussian Causal Bandits (DAG given)

Input: DAG $\mathcal{G}$, Observational data $\mathbf{D}_0$, parameters m, n, c, c_3 , rounds T .

for $t = 1, \dots, T$ **do**

if There is a node i which has been intervened less than $\lceil \frac{c_3 \log t}{n+m} \rceil$ times **then**

 Choose intervention node $A_t = i$;

else

 Calculate $\hat{\mu}_i(t-1, \text{pa}(i))$ and $\hat{\sigma}_i(t-1, \text{pa}(i))$ for each $i \in \{1, \dots, K\}$;

 Choose intervention node $A_t \in \underset{i \in \{1, \dots, K\}}{\text{argmax}} \left\{ \hat{\mu}_i(t-1, \text{pa}(i)) + \hat{\sigma}_i(t-1, \text{pa}(i))c\sqrt{\log(t-1)} \right\}$.

 Sample m data points under intervention A_{t+1} and n observational data.

Output: Optimal intervention A_T and estimated reward $\hat{\mu}^*$.

Define the following constants:

$$\begin{aligned} \phi^2 &= \max_{1 \leq i \leq K} \max \left\{ \frac{9\zeta_i^2}{\gamma_{\min}(\Sigma)\tilde{\sigma}_i^2}, \frac{25\tilde{\sigma}_i^2\gamma_{\max}(\Sigma)}{9\zeta_i^2} \right\}, \text{ where } \zeta_i^2 = \zeta_{i, \text{pa}(i)}^2, \\ \eta^2 &= \max_i \eta_i^2, \text{ where } \eta_i^2 = \max \left\{ \frac{9\zeta_i^2}{\gamma_{\min}(\Sigma)}, \sigma_i^2 \right\}, \\ \ell_0 &= \lceil \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil. \end{aligned}$$

The theoretical results on the cumulative regret are summarized in the following theorems.

2.7.2.1 Case-dependent bound

Theorem 2.3. *Under Assumptions 2.1 and 2.3, the cumulative regret $\mathcal{R}_T$ of Algorithm 2.2 with parameters $c \geq 4\sqrt{2}\phi$ and $c_3 \geq 64$, after T rounds is at most*

$$\log T \left(\frac{20c^2}{n+m} \sum_{i: \mu_i < \mu^*} \frac{\eta_i^2}{\Delta_i} + \frac{c_3 \sum_{i=1}^K \Delta_i}{n+m} \right) + C_1 \left(\sum_{j=1}^K \Delta_j \right), \quad (2.17)$$

where $C_1 = \ell_0 + 1 + \frac{\pi^2}{3} \left(1 + \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + \frac{36}{n} \exp \left(-\frac{n_0 + n\ell_0}{18} \right)$.

Proof. The key idea of the proof follows the proof of Theorem 4 in Auer, Cesa-Bianchi, and Fischer [1]. Denote the optimal intervention node as a^* , with parent set pa^* . Let $c_t = c\sqrt{\log t}$, $n^*(t)$ be the times of round that the optimal intervention is picked, $\hat{\mu}_u^*(t)$ be the estimated mean rewards $\hat{\mu}^*(t, \text{pa}^*)$ for the optimal intervention when $n^*(t) = u$, $\hat{\mu}_{i,u_i}(t)$ be the estimated mean rewards $\hat{\mu}_i(t, \text{pa}_i)$ for any arm i when $n_i(t) = u_i$, similarly for $\hat{\sigma}$. Then we have

$$\begin{aligned} n_i(T) &= \sum_{t=1}^T \mathbb{1} \{A_t = i\} \leq \ell + \sum_{t=\ell+1}^T \mathbb{1} \{A_t = i, n_i(t-1) \geq \ell\} \\ &\leq \ell + \sum_{t=\ell}^{T-1} \mathbb{1} \{ \hat{\mu}^*(t) + \hat{\sigma}^*(t)c_t \leq \hat{\mu}_i(t) + \hat{\sigma}_i(t)c_t, n_i(t-1) \geq \ell \} \\ &\leq \ell + \sum_{t=\ell}^{T-1} \mathbb{1} \left\{ \min_{1 \leq u \leq t} \hat{\mu}_u^*(t) + \hat{\sigma}_u^*(t)c_t \leq \max_{\ell \leq u_i \leq t} \hat{\mu}_{i,u_i}(t) + \hat{\sigma}_{i,u_i}(t)c_t \right\} \\ &\leq \ell + \sum_{t=\ell}^{T-1} \sum_{u=1}^t \sum_{u_i=\ell}^t \mathbb{1} \{ \hat{\mu}_u^*(t) + \hat{\sigma}_u^*(t)c_t \leq \hat{\mu}_{i,u_i}(t) + \hat{\sigma}_{i,u_i}(t)c_t \}. \end{aligned}$$

When $\hat{\mu}_u^*(t) + \hat{\sigma}_u^*(t)c_t \leq \hat{\mu}_{i,u_i}(t) + \hat{\sigma}_{i,u_i}(t)c_t$, i.e.,

$$\begin{aligned}
0 &\leq \hat{\mu}_{i,u_i}(t) + \hat{\sigma}_{i,u_i}(t)c_t - (\hat{\mu}_u^*(t) + \hat{\sigma}_u^*(t)c_t) \\
&= \hat{\mu}_{i,u_i}(t) - \mu_i + \mu_i - \hat{\sigma}_{i,u_i}(t)c_t + 2\hat{\sigma}_{i,u_i}(t)c_t + \mu^* - \mu^* - \hat{\mu}_u^*(t) - \hat{\sigma}_u^*(t)c_t \\
&= (\hat{\mu}_{i,u_i}(t) - \mu_i - \hat{\sigma}_{i,u_i}(t)c_t) + (\mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t - \mu^*) + (\mu^* - \hat{\mu}_u^*(t) - \hat{\sigma}_u^*(t)c_t),
\end{aligned}$$

then at least one of the following three events must happen

$$\begin{aligned}
\hat{\mu}_{i,u_i}(t) - \mu_i - \hat{\sigma}_{i,u_i}(t)c_t &\geq 0, \\
\mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t - \mu^* &\geq 0, \\
\mu^* - \hat{\mu}_u^*(t) - \hat{\sigma}_u^*(t)c_t &\geq 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
n_i(T) &\leq \ell + \sum_{t=\ell}^{T-1} \sum_{u=1}^t \sum_{u_i=\ell}^t \left[\mathbb{1} \{ \hat{\mu}_u^*(t) \leq \mu^* - \hat{\sigma}_u^*(t)c_t \} \right. \\
&\quad \left. + \mathbb{1} \{ \hat{\mu}_{i,u_i}(t) \geq \mu_i + \hat{\sigma}_{i,u_i}(t)c_t \} + \mathbb{1} \{ \mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t \} \right].
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[n_i(T)] &\leq \ell + \sum_{t=\ell}^{T-1} \sum_{u=1}^t \sum_{u_i=\ell}^t \left[\mathbb{P}(\hat{\mu}_u^*(t) \leq \mu^* - \hat{\sigma}_u^*(t)c_t) \right. \\
&\quad \left. + \mathbb{P}(\hat{\mu}_{i,u_i}(t) \geq \mu_i + \hat{\sigma}_{i,u_i}(t)c_t) + \mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t) \right].
\end{aligned} \tag{2.18}$$

In what follows, we bound the three parts in (2.18), which provide an upper bound on $\mathbb{E}[n_i(T)]$ for $i \neq a^*$. Then this translates into a bound on $\mathcal{R}_T = \sum_i \Delta_i \mathbb{E}[n_i(T)]$. Denote $s = |\text{pa}(i)|$, $s_0 = \max_i |\text{pa}(i)|$.

(a) Bound on $\mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t)$.

For the last part in (2.18), denote $s = |\text{pa}(i)|$, then

$$\mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t) = \mathbb{P}\left(\hat{\sigma}_{i,u_i}^2(t) > \frac{\Delta_i^2}{4c^2 \log(t)}\right),$$

where

$$\hat{\sigma}_{i,u_i}^2(t) = \frac{(n_0 + nt)^2 \widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t)) + (mu_i)^2 \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))}{(n_0 + nt + mu_i)^2},$$

where $\widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))$ and $\widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t))$ are defined in equations (2.7) and (2.8) respectively.

From Gaussian sample variance,

$$\widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t)) = \frac{s_{i,t}^2}{N_i(t)} \sim \frac{\tilde{\sigma}_i^2}{mu_i(mu_i - 1)} \chi_{mu_i-1}^2.$$

From Gaussian linear regression,

$$\begin{aligned} \widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t)) &= \frac{1}{n_0 + nt} \rho_i(t, \text{pa}(i)) \hat{\zeta}_i^2(t), \\ \hat{\zeta}_i^2(t) &= \frac{\|\mathbf{Y} - \mathbf{X}_i \hat{\beta}_i - \mathbf{X}_{\text{pa}(i)} \hat{\beta}_{\text{pa}(i)}\|^2}{N_o(t) - s - 1} \Bigg| \mathbf{X}_{i,\text{pa}(i)} \sim \frac{\zeta_i^2}{n_0 + nt - s - 1} \chi_{n_0+nt-s-1}^2. \end{aligned}$$

Given the independence of observational data and interventional data, we have

$$\hat{\sigma}_{i,u_i}^2(t) \mid \mathbf{X}_{i,\text{pa}(i)} \sim \frac{\left[\frac{(n_0+nt)\rho_i(t,\text{pa}(i))}{n_0+nt-s-1} \zeta_i^2 \chi_{n_0+nt-s-1}^2 + \frac{mu_i}{mu_i-1} \tilde{\sigma}_i^2 \chi_{mu_i-1}^2 \right]}{(n_0 + nt + mu_i)^2}.$$

Since $t \geq \ell$, $u_i \geq \ell$, and let $\ell \geq \ell_0 = \lceil \frac{9 \text{tr}(\Sigma)}{n \gamma_{\min}(\Sigma)} \rceil$ and $\ell \geq 2$, we have $\frac{n_0+nt}{n_0+nt-s-1} \leq 2$ and $\frac{mu_i}{mu_i-1} \leq 2$. If we define $\eta_i^2 = \max\{\rho_i(t, \text{pa}(i)) \zeta_i^2, \tilde{\sigma}_i^2\}$, then we have

$$\frac{(n_0 + nt)\rho_i(t, \text{pa}(i))}{n_0 + nt - s - 1} \zeta_i^2 \chi_{n_0+nt-s-1}^2 + \frac{mu_i}{mu_i - 1} \tilde{\sigma}_i^2 \chi_{mu_i-1}^2 \leq 2\eta_i^2 \chi_{n_0+nt+mu_i-s-2}^2.$$

Therefore,

$$\mathbb{P} \left(\hat{\sigma}_{i,u_i}^2(t) > \frac{\Delta_i^2}{4c^2 \log t} \middle| \mathbf{X}_{i,\text{pa}(i)} \right) \leq \mathbb{P} \left(\chi_{n_0+nt+mu_i-s-2}^2 > \frac{(n_0 + nt + mu_i)^2 \Delta_i^2}{8c^2 \eta_i^2 \log t} \right).$$

According to Lemma 2.2, when $\frac{(n_0+nt+mu_i)^2 \Delta_i^2}{8c^2 \eta_i^2 \log t} \geq \frac{5}{2}(n_0 + nt + mu_i - s - 2)$, i.e., $u_i \geq \ell \geq \frac{20c^2 \eta_i^2 \log t}{(m+n)\Delta_i^2}$, we have

$$\begin{aligned} \mathbb{P} \left(\chi_{n_0+nt+mu_i-s-2}^2 > \frac{(n_0 + nt + mu_i)^2 \Delta_i^2}{8c^2 \eta_i^2 \log t} \right) &\leq \exp(-(n_0 + nt + mu_i - s - 2)/4) \\ &\leq \exp(-(n_0 - s - 2 + (m+n)u_i)/4). \end{aligned}$$

Let $u_i \geq \frac{16 \log t}{n+m}$, then

$$\mathbb{P} \left(\chi_{n_0+nt+mu_i-s-2}^2 > \frac{(n_0 + nt + mu_i)^2 \Delta_i^2}{8c^2 \eta_i^2 \log t} \right) \leq t^{-4} \exp(-(n_0 - s - 2)/4).$$

Therefore

$$\mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t | \mathbf{X}_{i,\text{pa}(i)}) \leq t^{-4} \exp(-(n_0 - s_0 - 2)/4). \quad (2.19)$$

(b) Bounds on $\mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i}c_t)$ and $\mathbb{P}(\hat{\mu}_u^*(t) \leq \mu^* - \hat{\sigma}_u^*(t)c_t)$.

Define $\phi_i^2 = \min\{\rho_i(t, \text{pa}(i))\zeta_i^2, \tilde{\sigma}_i^2\}$, then

$$\frac{(n_0 + nt)\rho_i(t, \text{pa}(i))}{n_0 + nt - s - 1} \zeta_i^2 \chi_{n_0+nt-s-1}^2 + \frac{mu_i}{mu_i - 1} \tilde{\sigma}_i^2 \chi_{mu_i-1}^2 \geq \phi_i^2 \chi_{n_0+nt+mu_i-s-2}^2.$$

Below ignoring t in the notation for simplicity, we have

$$\begin{aligned} \hat{\mu}_{i,u_i} &= \frac{N_o \hat{\mu}_{i,\text{obs}} + N_i \hat{\mu}_{i,\text{int}}}{N_o + N_i} \middle| \mathbf{X}_{i,\text{pa}(i)} \sim \mathcal{N} \left(\mu_i, \frac{N_i \tilde{\sigma}_i^2 + N_o \zeta_i^2 \rho_i(\text{pa}(i))}{(N_i + N_o)^2} \right), \\ \hat{\sigma}_{i,u_i}^2 \middle| \mathbf{X}_{i,\text{pa}(i)} &\sim \frac{\left[\frac{N_o \rho_i(\text{pa}(i))}{N_o - s - 1} \zeta_i^2 \chi_{N_o - s - 1}^2 + \frac{N_i}{N_i - 1} \tilde{\sigma}_i^2 \chi_{N_i - 1}^2 \right]}{(N_o + N_i)^2}. \end{aligned}$$

Therefore,

$$\frac{\hat{\mu}_{i,u_i} - \mu_i}{\hat{\sigma}_{i,u_i}} \Big| \mathbf{X}_{i,\text{pa}(i)} \sim \frac{\mathcal{N}(0, 1) \sqrt{N_i \tilde{\sigma}_i^2 + N_o \zeta_i^2 \rho_i(\text{pa}(i))}}{\sqrt{\frac{N_o \rho_i(\text{pa}(i))}{N_o - s - 1} \zeta_i^2 \chi_{N_o - s - 1}^2 + \frac{N_i}{N_i - 1} \tilde{\sigma}_i^2 \chi_{N_i - 1}^2}}.$$

Then

$$\begin{aligned} \mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i} c_t \mid \mathbf{X}_{i,\text{pa}(i)}) &= \mathbb{P}\left(\frac{\hat{\mu}_{i,u_i} - \mu_i}{\hat{\sigma}_{i,u_i}} \geq c_t \mid \mathbf{X}_{i,\text{pa}(i)}\right) \\ &\leq \mathbb{P}\left(\frac{\eta_i \mathcal{N}(0, 1) \sqrt{N_i + N_o}}{\sqrt{\phi_i^2 \chi_{N_o + N_i - s - 2}^2}} \geq c_t \mid \mathbf{X}_{i,\text{pa}(i)}\right) \\ &\leq \mathbb{P}\left(\frac{\eta_i \sqrt{N_i + N_o}}{\phi_i \sqrt{N_o + N_i - s - 2}} \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi_{N_o + N_i - s - 2}^2}{N_o + N_i - s - 2}}} \geq c_t \mid \mathbf{X}_{i,\text{pa}(i)}\right) \\ &\leq \mathbb{P}\left(\sqrt{2} \frac{\eta_i}{\phi_i} t_{N_o + N_i - s - 2} \geq c \sqrt{\log t} \mid \mathbf{X}_{i,\text{pa}(i)}\right) \\ &= \mathbb{P}\left(t_{N_o + N_i - s - 2} \geq \frac{c \phi_i \sqrt{\log t}}{\sqrt{2} \eta_i} \mid \mathbf{X}_{i,\text{pa}(i)}\right). \end{aligned}$$

According to Lemma 2.3, we have

$$\begin{aligned} \mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i} c_t \mid \mathbf{X}_{i,\text{pa}(i)}) &\leq \exp\left(-\frac{c^2 \phi_i^2 \log t}{8 \eta_i^2}\right) + \frac{1}{2} \exp\left(-\frac{n_0 + nt + mu_i - s - 2}{16}\right) \\ &\leq t^{-c^2 \phi_i^2 / 8 \eta_i^2} + \frac{1}{2} \exp\left(-\frac{n_0 + nt + mu_i - s - 2}{16}\right). \end{aligned} \tag{2.20}$$

Since $u_i \geq \frac{64 \log t}{n+m}$, $c \geq 4\sqrt{2} \eta_i / \phi_i$, then

$$\mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i} c_t \mid \mathbf{X}_{i,\text{pa}(i)}) \leq t^{-4} \left(1 + \frac{1}{2} \exp(-(n_0 - s_0 - 2)/16)\right). \tag{2.21}$$

Set minimum intervention times as $\frac{64 \log t}{n+m}$, similarly for the optimal arm, when $c \geq$

$4\sqrt{2}\eta^*/\phi^*$, we have

$$\mathbb{P}(\hat{\mu}_u^*(t) \leq \mu^* - \hat{\sigma}_u^*(t)c_t \mid \mathbf{X}_{a^*, \text{pa}^*}) \leq t^{-4} \left(1 + \frac{1}{2 \exp((n_0 - s_0 - 2)/16)} \right). \quad (2.22)$$

Define $\phi^2 = \max_i \eta_i^2 / \phi_i^2 = \max_i \max\{\frac{\rho_i(t, \text{pa}(i))\zeta_i^2}{\tilde{\sigma}_i^2}, \frac{\tilde{\sigma}_i^2}{\rho_i(t, \text{pa}(i))\zeta_i^2}\}$, $\bar{V} = \{1, 2, \dots, p-1\}$. Conditioning on $\mathbf{X}_{i, \text{pa}(i)}$, $\rho_i(t, \text{pa}(i))$ is a constant. Plugging (2.19)(2.21)(2.22) into (2.18), then $\mathbb{E}[n_i(T) \mid \mathbf{X}_{\bar{V}}]$ can be bounded by

$$\mathbb{E}[n_i(T) \mid \mathbf{X}_{\bar{V}}] \leq \ell + \sum_{t=\ell+1}^T t^{-2} \left(2 + 2 \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right), \quad (2.23)$$

where ℓ satisfies:

$$\ell \geq \max \left\{ \left\lceil \frac{20c^2 \frac{\eta_i^2}{\Delta_i^2} \log t}{m+n} \right\rceil, \left\lceil \frac{c_3 \log t}{n+m} \right\rceil, \left\lceil \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \right\rceil, 2 \right\} \quad (2.24)$$

where $c \geq 4\sqrt{2}\phi$, $c_3 \geq 64$.

Averaging over $\mathbf{X}_{\bar{V}}$, $\rho_i(t, \text{pa}(i))$ and $\rho^*(t, \text{pa}^*)$ can be bounded by Lemma 2.1 in Section 2.7.1. In fact, for any $t \geq \lceil \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil = \ell_0$, we have

$$\begin{aligned} & \mathbb{P} \left(\exists i \in \bar{V}, \rho_i(t, \text{pa}(i)) \geq \frac{9}{\gamma_{\min}(\Sigma)} \text{ or } \rho_i(t, \text{pa}(i)) \leq \frac{9}{25\gamma_{\max}(\Sigma)} \right) \\ & \leq \mathbb{P} \left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N_o(t)}} \leq \frac{1}{3}\gamma_{\min}(\sqrt{\Sigma}) \text{ or } \frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N_o(t)}} \geq \frac{5}{3}\gamma_{\max}(\sqrt{\Sigma}) \right) \\ & \leq 2 \exp \left(-\frac{n_0 + nt}{18} \right). \end{aligned} \quad (2.25)$$

Therefore, we have

$$\begin{aligned} & \sum_{t=\ell}^T \mathbb{P} \left(\exists i \in \bar{V}, \rho_i(t, \text{pa}(i)) \geq \frac{9}{\gamma_{\min}(\Sigma)} \text{ or } \rho_i(t, \text{pa}(i)) \leq \frac{9}{25\gamma_{\max}(\Sigma)} \right) \\ & \leq 2 \sum_{t=\ell}^{\infty} \exp \left(-\frac{n_0 + nt}{18} \right) \leq 2 \int_{\ell}^{+\infty} \exp \left(-\frac{n_0 + nt}{18} \right) dt \leq \exp \left(-\frac{n_0 + n\ell}{18} \right) \frac{36}{n} \end{aligned} \quad (2.26)$$

Now re-define

$$\begin{aligned} \eta_i^2 &= \max \left\{ \frac{9\zeta_i^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}, \\ \phi^2 &= \max_i \max \left\{ \frac{9\zeta_i^2}{\gamma_{\min}(\Sigma)\tilde{\sigma}_i^2}, \frac{25\tilde{\sigma}_i^2\gamma_{\max}(\Sigma)}{9\zeta_i^2} \right\}, \end{aligned}$$

and let $c \geq 4\sqrt{2}\phi$, given the conditions on ℓ , the final upper bound of $\mathbb{E}[n_i(T)]$ is

$$\begin{aligned} \mathbb{E}[n_i(T)] &\leq \ell + \sum_{t=\ell}^T t^{-2} \left(2 + 2 \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + (2.26) \\ &\leq 20c^2 \frac{\eta_i^2}{\Delta_i^2} \frac{\log T}{n+m} + \frac{c_3 \log T}{n+m} + \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} + 2 + \frac{\pi^2}{3} \left(1 + \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) \\ &\quad + \frac{36}{n} \exp \left(-\frac{n_0 + n\ell_0}{18} \right). \end{aligned} \quad (2.27)$$

Then the upper bound of $\mathcal{R}_T$ can be expressed as

$$\log T \left(\frac{20c^2}{n+m} \sum_{i: \mu_i < \mu^*} \frac{\eta_i^2}{\Delta_i} + \frac{c_3 \sum_{i=1}^K \Delta_i}{n+m} \right) + C_1 \left(\sum_{j=1}^K \Delta_j \right), \quad (2.28)$$

where

$$C_1 = \ell_0 + 1 + \frac{\pi^2}{3} \left(1 + \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + \frac{36}{n} \exp \left(-\frac{n_0 + n\ell_0}{18} \right). \quad (2.29)$$

□

2.7.2.2 Δ -independent bound

Theorem 2.4. Assume $\mu_i \in [-1, 1], \forall i \in \{1, \dots, p-1\}$. Then under Assumptions 2.1 and 2.3, the cumulative regret of Algorithm 2.2, with parameters $c \geq 4\phi$ and $c_3 \geq 32$, after T rounds is at most

$$2c\eta\sqrt{\log T} \min \left\{ \frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m}, \frac{\sqrt{n_0 + nT} - \sqrt{n_0}}{n} \mathbb{1}_{\{n \geq 1\}} \right\} + \frac{2Kc_3 \log T}{n+m} + C_2 \quad (2.30)$$

where $C_2 = 2\ell_0 + \frac{\pi^2}{3} + 2K + \frac{\pi^2}{6} \exp\left(-\frac{n_0-s_0-2}{16}\right) + \exp\left(-\frac{n_0+n\ell_0}{18}\right) \frac{72}{n}$.

Proof. Let $\ell \geq K\lceil \frac{c_3 \log T}{n+m} \rceil$, then for any $t \geq \ell + 1$, the optimal arm A_t is the arm with the largest upper confidence bound. Let $U_t(i) = \hat{\mu}_i(t, S) + \hat{\sigma}_i(t, S)c\sqrt{\log t}$ be the constructed upper confidence bound of arm i after t rounds. A^* is the optimal arm, μ^* is the largest reward, and all notations with $\star$ mean the corresponding quantities for the optimal arm. Since $\mu_i \in [-1, 1]$ and $U_t(i) \in [-1, 1]$ by construction, then $|\mu_i - U_t(i)| \leq 2, \forall i, t$. Therefore,

$$\begin{aligned} \mathcal{R}_T &= \sum_{t=1}^T \mathbb{E}[\mu^* - \mu_{A_t}] \leq 2\ell + \sum_{t \geq \ell+1} \mathbb{E}[\mu^* - U_{t-1}(A_t) + U_{t-1}(A_t) - \mu_{A_t}] \\ &\leq 2\ell + \sum_{t \geq \ell+1} \mathbb{E}[\mu^* - U_{t-1}(A^*)] + \sum_{t \geq \ell+1} \mathbb{E}[U_{t-1}(A_t) - \mu_{A_t}] \\ &\leq 2\ell + 2 \sum_{t=\ell}^{T-1} \mathbb{P}(\mu^* > U_t(A^*)) + \sum_{t=\ell+1}^T \mathbb{E}[U_{t-1}(A_t) - \mu_{A_t}]. \end{aligned} \quad (2.31)$$

According to Lemma 2.1, for any $t \geq \ell_0 = \lceil \frac{9 \text{tr}(\Sigma)}{n \gamma_{\min}(\Sigma)} \rceil$ and any $S(i) \subseteq \mathcal{V} \setminus \{i\}$, we have

$$\begin{aligned} & \mathbb{P} \left(\exists i \in \bar{V}, \rho_i(t, S(i)) \leq \frac{9}{25 \gamma_{\max}(\Sigma)} \text{ or } \rho_i(t, S(i)) \geq \frac{9}{\gamma_{\min}(\Sigma)} \right) \\ & \leq \mathbb{P} \left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N_o(t)}} \leq \frac{1}{3} \gamma_{\min}(\sqrt{\Sigma}) \text{ or } \frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N_o(t)}} \geq \frac{5}{3} \gamma_{\max}(\sqrt{\Sigma}) \right) \\ & \leq 2 \exp \left(-\frac{n_0 + nt}{18} \right). \end{aligned}$$

Therefore, we have

$$\sum_{t=\ell}^T \mathbb{P} \left(\exists i \in \bar{V}, \rho_i(t, \text{pa}(i)) \leq \frac{9}{25 \gamma_{\max}(\Sigma)} \text{ or } \rho_i(t, \text{pa}(i)) \geq \frac{9}{\gamma_{\min}(\Sigma)} \right) \leq \exp \left(-\frac{n_0 + n\ell}{18} \right) \frac{36}{n}. \quad (2.32)$$

Note that the set is not restricted to $\text{pa}(i)$, (2.32) is satisfied for any set $S(i, t)$ that consists of nodes except X_i and Y .

When $c \geq 4\phi$ and $c_3 \geq 32$, and viewing $\rho_i(t, \text{pa}(i))$ as constants between $\frac{9}{25 \gamma_{\max}(\Sigma)}$ and $\frac{9}{\gamma_{\min}(\Sigma)}$, by conditioning on $\mathbf{X}_{i, \text{pa}(i)}$, from (2.20), we have

$$\mathbb{P}(\mu^* > U_t(A^*) \mid \mathbf{X}_{a^*, \text{pa}^*}) \leq t^{-2} \left(1 + \frac{1}{2} \exp(-(n_0 - s_0 - 2)/16) \right).$$

Then

$$\sum_{t=K}^{T-1} \mathbb{P}(\mu^* > U_t(A^*) \mid \mathbf{X}_{a^*, \text{pa}^*}) \leq \frac{\pi^2}{6} \left(1 + \frac{1}{2} \exp(-(n_0 - s_0 - 2)/16) \right). \quad (2.33)$$

Let $\eta^2 = \max_i \max \left\{ \frac{9\zeta_{i, \text{pa}(i)}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}$, $\ell \geq \ell_0 = \lceil \frac{9 \text{tr}(\Sigma)}{n \gamma_{\min}(\Sigma)} \rceil$, $\mathcal{T}_a = \{t : A_t = a, \ell + 1 \leq t \leq T\}$, then

$$\sum_{t=\ell+1}^T \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t}) \leq \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}_a} \mathbb{E}(U_{t-1}(a) - \mu_a).$$

For $t \geq \ell$, $\forall a \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{E}[U_t(a) - \mu_a | \mathbf{X}_{\overline{V}}] &= \mathbb{E}(\hat{\sigma}_a | \mathbf{X}_{\overline{V}}) c \sqrt{\log t} \leq c \sqrt{\log t} \sqrt{\mathbb{E}(\hat{\sigma}_a^2 | \mathbf{X}_{\overline{V}})} \\
&= c \sqrt{\log t} \frac{\sqrt{N_o(t) \zeta_i^2 \rho_i(t, \text{pa}(i)) + N_a(t) \tilde{\sigma}_i^2}}{N_o(t) + N_a(t)} \\
&\leq c \eta \sqrt{\frac{\log t}{N_o(t) + N_a(t)}}
\end{aligned}$$

Since $N_o(t) + N_a(t) = n_0 + nt + mn_a(t) \geq n_0 + (n + m)n_a(t)$,

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}_a} \sqrt{\frac{\log t}{N_o(t) + N_a(t)}} &\leq \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}_a} \sqrt{\frac{\log t}{n_0 + (n + m)n_a(t)}} \leq \sum_{a \in \mathcal{A}} \sum_{j=0}^{n_a(T)} \sqrt{\frac{\log T}{n_0 + (n + m)j}} \\
&\leq \sqrt{\log T} \sum_{a \in \mathcal{A}} \int_0^{n_a(T)} \frac{1}{\sqrt{n_0 + (n + m)j}} dj \\
&\leq \frac{2\sqrt{\log T}}{n + m} \sum_{a \in \mathcal{A}} \left(\sqrt{n_0 + (n + m)n_a(T)} - \sqrt{n_0} \right). \tag{2.34}
\end{aligned}$$

By Cauchy-Schwartz inequality,

$$\sum_{a \in \mathcal{A}} \sqrt{n_0 + (n + m)n_a(T)} \leq \sqrt{K \left(\sum_{a \in \mathcal{A}} (n_0 + (n + m)n_a(T)) \right)} = \sqrt{(n + m)KT + K^2 n_0}.$$

Since $N_o(t) + N_a(t) \geq N_o(t)$, when $n \geq 1$, we also have

$$\begin{aligned}
\sum_{t=\ell}^T \sqrt{\frac{\log t}{N_o(t) + N_a(t)}} &\leq \sum_{t=\ell}^T \sqrt{\frac{\log t}{n_0 + nt}} \leq \sqrt{\log T} \int_0^T \frac{1}{\sqrt{n_0 + nt}} dt \\
&= \frac{2\sqrt{\log T}}{n} \left(\sqrt{n_0 + nT} - \sqrt{n_0} \right). \tag{2.35}
\end{aligned}$$

In summary, we have

$$\begin{aligned}\mathcal{R}_T &\leq 2\ell + c\eta \min\{(2.34), (2.35)\} + 2 \times (2.32) + 2 \times (2.33) \\ &\leq 2\ell + 2c\eta\sqrt{\log T} \min\left\{\frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m}, \frac{\sqrt{n_0 + nT} - \sqrt{n_0}}{n}\right\} \\ &\quad + \frac{\pi^2}{6} \left(2 + \exp\left(-\frac{n_0 - s_0 - 2}{16}\right)\right) + \exp\left(-\frac{n_0 + n\ell}{18}\right) \frac{72}{n}.\end{aligned}$$

Since $\ell \geq K \lceil \frac{c_3 \log T}{n+m} \rceil$ and $\ell \geq \ell_0$,

$$\begin{aligned}\mathcal{R}_T &\leq 2c\eta\sqrt{\log T} \min\left\{\frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m}, \frac{\sqrt{n_0 + nT} - \sqrt{n_0}}{n} \mathbb{1}\{n \geq 1\}\right\} \\ &\quad + \frac{2Kc_3 \log T}{n+m} + C_2,\end{aligned}\tag{2.36}$$

where

$$C_2 = 2\ell_0 + \frac{\pi^2}{3} + 2K + \frac{\pi^2}{6} \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) + \exp\left(-\frac{n_0 + n\ell_0}{18}\right) \frac{72}{n}.\tag{2.37}$$

□

2.7.3 When the underlying DAG is unknown

When the underlying DAG is unknown, we take an additional step to estimate the backdoor adjustment sets. In this section, we modify the proof for the DAG given setting to get our final theoretical results.

We adopt a sample-splitting strategy to decouple adjustment set identification from parameter estimation used to construct upper confidence bounds and select arms. Let M_i denote the regression neighborhood of a node X_i , $\mathcal{N}_i$ denote the collection of all subsets of M_i , $\mathcal{B}_i$ denote the collection of valid adjustment sets for (X_i, Y) , and $\tilde{\mathbf{X}}$ denote the observational data used exclusively for identifying adjustment sets. Also, denote the sampled set S

in Algorithm 2.1 Line 9 at time step t as $S(i, t)$ for node X_i and $S^*(t)$ for the optimal arm.

For any $S \in \mathcal{N}_i$, let $s = |S|$, we have

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\mu}_i(t, S) - \hat{\mu}_{i,\text{int}}(t)) &= \widehat{\text{Var}}(\hat{\mu}_i(t, S)) + \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t)) \\
&= \frac{\hat{\zeta}_{i,S}^2 \rho_i(t, S)}{N_o(t)} + \frac{s_{i,t}^2}{N_i(t)} \Big| \tilde{\mathbf{X}}_{i,S} \\
&\sim \frac{\zeta_{i,S}^2 \rho_i(t, S)}{N_o(t)(N_o(t) - s - 1)} \chi_{N_o(t)-s-1}^2 + \frac{\tilde{\sigma}_i^2}{N_i(t)(N_i(t) - 1)} \chi_{N_i(t)-1}^2, \\
\hat{\mu}_i(t, S) - \hat{\mu}_{i,\text{int}}(t) \mid \tilde{\mathbf{X}}_{i,S} &\sim \mathcal{N}\left(\mu_i(S) - \mu_i, \frac{\tilde{\sigma}_i^2}{N_i(t)} + \frac{\zeta_{i,S}^2 \rho_i(t, S)}{N_o(t)}\right).
\end{aligned} \tag{2.38}$$

As displayed (2.10), the confidence intervals $CI_i(t, S)$ at round t for $\mu_i(S) - \mu_i$ are

$$\hat{\mu}_{i,\text{obs}}(t, S) - \hat{\mu}_{i,\text{int}}(t) \pm c_2 \sqrt{\log t} \sqrt{\widehat{\text{Var}}(\hat{\mu}_{i,\text{obs}}(t, S)) + \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))}. \tag{2.39}$$

We now prove the upper bounds in Theorem 2.1 and Theorem 2.2.

2.7.3.1 Proof of Theorem 2.1

We first introduce and prove some lemmas that form the essential building blocks of the proof of Theorem 2.1.

Lemma 2.4. *Under the same setting of Theorem 2.1 and Assumptions 2.1 and 2.3, with parameters $c_2 \geq 2\sqrt{3}$ and $c_3 \geq 18$, for any $i \in \bar{V}$ and t ,*

$$\mathbb{P}\left(0 \notin CI_i(t, \text{pa}(i)) \mid \tilde{\mathbf{X}}_{i,\text{pa}(i)}\right) \leq 3t^{-2} + \exp\left(-\frac{N_o(t) - s_0 - 1}{9}\right). \tag{2.40}$$

Proof. Ignoring t for simplicity in notation, we have

$$\mathbb{P}\left(0 \notin CI_i(\text{pa}(i)) \mid \tilde{\mathbf{X}}_{i,\text{pa}(i)}\right) = \mathbb{P}\left(\left|\frac{\hat{\mu}_i(\text{pa}(i)) - \hat{\mu}_{i,\text{int}}}{\sqrt{\widehat{\text{Var}}(\hat{\mu}_i(\text{pa}(i)) - \hat{\mu}_{i,\text{int}})}}\right| > c_2 \sqrt{\log t} \mid \tilde{\mathbf{X}}_{i,\text{pa}(i)}\right).$$

According to Lemma 2.2 and (2.38), for any $S \in \mathcal{N}_i$,

$$\begin{aligned} & \mathbb{P} \left(\widehat{\text{Var}} (\hat{\mu}_i(S) - \hat{\mu}_{i,\text{int}}) \leq \frac{\zeta_{i,S}^2 \rho_i(S)}{3N_o} + \frac{\tilde{\sigma}_i^2}{3N_i} \middle| \tilde{\mathbf{X}}_{i,S} \right) \\ & \leq \mathbb{P} \left(\chi_{N_o-s-1}^2 \leq \frac{N_o-s-1}{3} \right) + \mathbb{P} \left(\chi_{N_i-1}^2 \leq \frac{N_i-1}{3} \right) \\ & \leq \exp \left(-\frac{N_o-s-1}{9} \right) + \exp \left(-\frac{N_i-1}{9} \right). \end{aligned}$$

Since $N_i \geq c_3 \log t + 1$, and $s = |\text{pa}(i)| \leq s_0$, let $c_3 \geq 18$, then we have

$$\begin{aligned} \mathbb{P} \left(0 \notin CI_i(\text{pa}(i)) \middle| \tilde{\mathbf{X}}_{i,\text{pa}(i)} \right) & \leq \mathbb{P} \left(|\mathcal{N}(0, 3)| > c_2 \sqrt{\log t} \right) \\ & \quad + \mathbb{P} \left(\widehat{\text{Var}} (\hat{\mu}_i(\text{pa}(i)) - \hat{\mu}_{i,\text{int}}) \leq \frac{\zeta_{i,\text{pa}(i)}^2 \rho_i(\text{pa}(i))}{3N_o} + \frac{\tilde{\sigma}_i^2}{3N_i} \middle| \tilde{\mathbf{X}}_{i,\text{pa}(i)} \right) \\ & \leq 2 \exp \left(-\frac{c_2^2 \log t}{6} \right) + \exp \left(-\frac{N_o-s-1}{9} \right) + \exp \left(-\frac{N_i-1}{9} \right) \\ & \leq 2t^{-c_2^2/6} + \exp \left(-\frac{n_0 + nt - s_0 - 1}{9} \right) + t^{-2}. \end{aligned}$$

Let $c_2 \geq 2\sqrt{3}$. Then

$$\mathbb{P} \left(0 \notin CI(\text{pa}(i)) \middle| \tilde{\mathbf{X}}_{i,\text{pa}(i)} \right) \leq 3t^{-2} + \exp \left(-\frac{n_0 + nt - s_0 - 1}{9} \right).$$

□

Define constants $\psi^2 = \max_i \max_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \max \{ \rho_i(t, S') \zeta_{i,S'}^2, \tilde{\sigma}_i^2 \}$, $\ell_0 = \lceil \frac{9 \text{tr}(\Sigma)}{n \gamma_{\min}(\Sigma)} \rceil$, and $s_2 = \max_i |\mathcal{N}_i \setminus \mathcal{B}_i|$.

Lemma 2.5. *Under the same settings of Theorem 2.1 and Assumptions 2.1, 2.2, and 2.3, with parameters $c_2 = 2\sqrt{3}$ and $c_3 \geq \max\{32\psi^2/\delta^2, 8\}$, for any $i \in \bar{V}$ and $t \geq (c_3/n)^2$,*

$$\sum_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P} \left(d(0, CI_i(t, S')) = 0 \middle| \tilde{\mathbf{X}}_{i,S'} \right) \leq s_2 \left(3t^{-2} + \exp \left(-\frac{N_o(t) - p}{4} \right) \right). \quad (2.41)$$

Proof. Ignoring t for simplicity in notation, let $s = |S'|$, we have

$$\begin{aligned} \mathbb{P} \left(d(0, CI_i(S')) = 0 \mid \tilde{\mathbf{X}}_{i,S'} \right) &= \mathbb{P} \left(\left| \frac{\hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}}}{\sqrt{\widehat{\text{Var}}(\hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}})}} \right| \leq c_2 \sqrt{\log t} \mid \tilde{\mathbf{X}}_{i,S'} \right) \\ &\leq \mathbb{P} \left(|\hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}}| \leq c_2 \sqrt{\log t} \sqrt{3 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \mid \tilde{\mathbf{X}}_{i,S'} \right) \\ &\quad + \mathbb{P} \left(\widehat{\text{Var}}(\hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}}) \geq 3 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right) \mid \tilde{\mathbf{X}}_{i,S'} \right) \end{aligned}$$

According to Lemma 2.2 and (2.38), with $c_3 \geq 8$, the second term on the RHS satisfies

$$\begin{aligned} &\mathbb{P} \left(\widehat{\text{Var}}(\hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}}) \geq \frac{3\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{3\tilde{\sigma}_i^2}{N_i} \mid \tilde{\mathbf{X}}_{i,S'} \right) \\ &\leq \mathbb{P}(\chi_{N_o-s-1}^2 \geq 3(N_o - s - 1)) + \mathbb{P}(\chi_{N_i-1}^2 \geq 3(N_i - 1)) \\ &\leq \exp\left(-\frac{N_o - s - 1}{4}\right) + \exp\left(-\frac{N_i - 1}{4}\right) \\ &\leq \exp\left(-\frac{N_o - p}{4}\right) + t^{-2} \end{aligned}$$

According to (2.38) and identifiability assumption $|\mu_i(S') - \mu_i| \geq 2\delta$, $\forall S' \in \mathcal{N}_i \setminus \mathcal{B}_i$, we have

$$\hat{\mu}_i(t, S') - \hat{\mu}_{i,\text{int}}(t) \mid \tilde{\mathbf{X}}_{i,S'} \sim \mathcal{N} \left(\mu_i(S') - \mu_i, \frac{\tilde{\sigma}_i^2}{N_i(t)} + \frac{\zeta_{i,S'}^2 \rho_i(t, S')}{N_o(t)} \right),$$

and

$$\begin{aligned} \left| \mathcal{N} \left(\mu_i(S') - \mu_i, \frac{\tilde{\sigma}_i^2}{N_i} + \frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o(t)} \right) \right| &\geq |\mu_i(S') - \mu_i| - \left| \mathcal{N} \left(0, \frac{\tilde{\sigma}_i^2}{N_i} + \frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o(t)} \right) \right| \\ &\geq 2\delta - \left| \mathcal{N} \left(0, \frac{\tilde{\sigma}_i^2}{N_i} + \frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o(t)} \right) \right|. \end{aligned}$$

Therefore, with $c_2 = 2\sqrt{3}$, the first term on the RHS satisfies

$$\begin{aligned}
& \mathbb{P} \left(\left| \hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}} \right| \leq c_2 \sqrt{\log t} \sqrt{3 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \middle| \tilde{\mathbf{X}}_{i,S'} \right) \\
& \leq \mathbb{P} \left(\left| \mathcal{N} \left(0, \frac{\tilde{\sigma}_i^2}{N_i} + \frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o(t)} \right) \right| \geq 2\delta - 6\sqrt{\log t} \sqrt{\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i}} \middle| \tilde{\mathbf{X}}_{i,S'} \right) \\
& \leq 2 \exp \left(- \frac{\left(2\delta - 6\sqrt{\log t} \sqrt{\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i}} \right)^2}{2 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \middle| \tilde{\mathbf{X}}_{i,S'} \right).
\end{aligned}$$

Since $t \geq (c_3/n)^2$, we have $\log t/(n_0 + nt) \leq 1/c_3$. Given $N_o = n_0 + nt$, $N_i \geq c_3 \log t + 1$, and $c_3 \geq 32\psi^2/\delta^2$, then

$$\sqrt{\log t \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \leq \sqrt{\frac{\zeta_{i,S'}^2 \rho_i(S') + \tilde{\sigma}_i^2}{c_3}} \leq \sqrt{\frac{2\psi^2}{c_3}} \leq \frac{\delta}{4}.$$

Therefore, according to Lemma 2.1,

$$\begin{aligned}
& \mathbb{P} \left(\left| \hat{\mu}_i(S') - \hat{\mu}_{i,\text{int}} \right| \leq c_2 \sqrt{\log t} \sqrt{3 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \middle| \tilde{\mathbf{X}}_{i,S'} \right) \\
& \leq 2 \exp \left(- \frac{\left(2\sqrt{\log t} \sqrt{\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i}} \right)^2}{2 \left(\frac{\zeta_{i,S'}^2 \rho_i(S')}{N_o} + \frac{\tilde{\sigma}_i^2}{N_i} \right)} \right) \leq 2t^{-2}.
\end{aligned}$$

Combining the upper bounds for the two terms, we have

$$\sum_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P} \left(d(0, CI_i(t, S')) = 0 \middle| \tilde{\mathbf{X}}_{i,S'} \right) \leq s_2 \left(3t^{-2} + \exp \left(- \frac{N_o(t) - p}{4} \right) \right).$$

□

By assumption, $X_i = \sum_{j:j \neq i} \bar{\beta}_{ij} X_j + \varepsilon_i$, where $\varepsilon \sim N(0, \bar{\sigma}_i^2)$. Define constants $s_1 = \max_i |M_i|$ and $\bar{\beta}_i = \min_{j:\beta_{ij} \neq 0} |\bar{\beta}_{ij}|$. Denote $\rho_{ij}(t) = [(\tilde{\mathbf{X}}_{-i}^\top \tilde{\mathbf{X}}_{-i} / N_o(t))^{-1}]_{jj}$.

Lemma 2.6. *Under the same settings of Theorem 2.1 and Assumptions 2.1 and 2.3, and the regression neighborhood of X_i at time t is determined by $\widehat{M}_i(t) = \{j \neq i : |\hat{\beta}_{ij}| / \sqrt{\widehat{\text{Var}}(\hat{\beta}_{ij})} > \bar{\alpha} \sqrt{N_o(t)}\}$ with parameter $\bar{\alpha} \leq \min_{i,j} \frac{\bar{\beta}_i}{2\sqrt{3\bar{\sigma}_i^2 \rho_{ij}(t)}}$, then for any $i \in \bar{V}$ and $N_o(t) \geq p$,*

$$\begin{aligned} \mathbb{P}\left(\widehat{M}_i(t) \neq M_i \mid \tilde{\mathbf{X}}_{-i}\right) &\leq (p-1) \left[2 \exp\left(-\frac{\bar{\alpha}^2 N_o(t)}{4}\right) + \exp\left(-\frac{N_o(t)-p}{16}\right) \right] \\ &\quad + s_1 \exp\left(-\frac{N_o(t)-p}{4}\right). \end{aligned} \quad (2.42)$$

Proof. Write the linear regression in matrix form: $\hat{\beta}_i = \text{argmin}_{\beta_i} \|\tilde{X}_i - \tilde{\mathbf{X}}_{-i} \beta_i\|$. Ignoring t in notations for simplicity, then

$$\begin{aligned} \hat{\beta}_{ij} \mid \tilde{\mathbf{X}}_{-i} &\sim \mathcal{N}\left(\bar{\beta}_{ij}, \bar{\sigma}_i^2 \frac{\rho_{ij}}{N_o}\right), \\ \hat{\sigma}_i^2 &= \frac{\|\mathbf{X}_i - \mathbf{X}_{-i} \hat{\beta}_i\|^2}{N_o - p} \mid \tilde{\mathbf{X}}_{-i} \sim \bar{\sigma}_i^2 \frac{\chi_{N_o-p}^2}{N_o - p}, \\ \frac{\hat{\beta}_{ij} - \bar{\beta}_{ij}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_{ij})}} &= \frac{\hat{\beta}_{ij} - \bar{\beta}_{ij}}{\sqrt{\hat{\sigma}_i^2 \frac{\rho_{ij}}{N_o}}} \mid \tilde{\mathbf{X}}_{-i} \sim t_{N_o-p}. \end{aligned}$$

Then $\forall j$ with $\bar{\beta}_{ij} \neq 0$,

$$\begin{aligned} \mathbb{P}(X_j \notin \widehat{M}_i \mid \tilde{\mathbf{X}}_{-i}) &= \mathbb{P}\left(\left|\frac{\bar{\beta}_{ij}}{\sqrt{\hat{\sigma}_i^2 \frac{\rho_{ij}}{N_o}}} + t_{N_o-p}\right| \leq \bar{\alpha} \sqrt{N_o} \mid \tilde{\mathbf{X}}_{-i}\right) \\ &\leq \mathbb{P}\left(|t_{N_o-p}| \geq \left(\frac{\bar{\beta}_i}{\sqrt{\hat{\sigma}_i^2 \rho_{ij}}} - \bar{\alpha}\right) \sqrt{N_o} \mid \tilde{\mathbf{X}}_{-i}\right). \end{aligned}$$

Let $\bar{\alpha} \leq \frac{\bar{\beta}_i}{2\sqrt{3\bar{\sigma}_i^2\rho_{ij}}}$, according to Lemma 2.2 and 2.3,

$$\begin{aligned}
& \mathbb{P}(X_j \notin \widehat{M}_i \mid \tilde{\mathbf{X}}_{-i}) \\
& \leq \mathbb{P} \left(|t_{N_o-p}| \geq \left(\frac{\bar{\beta}_i}{\sqrt{\widehat{\sigma}_i^2\rho_{ij}}} - \bar{\alpha} \right) \sqrt{N_o} \mid \tilde{\mathbf{X}}_{-i} \right) \\
& \leq \mathbb{P} \left(|t_{N_o-p}| \geq \left(\frac{\bar{\beta}_i}{\sqrt{3\bar{\sigma}_i^2\rho_{ij}}} - \bar{\alpha} \right) \sqrt{N_o} \mid \tilde{\mathbf{X}}_{-i} \right) + \mathbb{P}(\chi_{N_o-p}^2 > 3(N_o - p)) \\
& \leq 2 \exp \left(-\frac{\bar{\beta}_i^2 N_o}{48\bar{\sigma}_i^2\rho_{ij}} \right) + \exp \left(-\frac{N_o(t) - p}{16} \right) + \exp \left(-\frac{N_o(t) - p}{4} \right) \\
& \leq 2 \exp \left(-\frac{\bar{\alpha}^2 N_o}{4} \right) + \exp \left(-\frac{N_o(t) - p}{16} \right) + \exp \left(-\frac{N_o(t) - p}{4} \right).
\end{aligned}$$

Similarly, $\forall j$ with $\bar{\beta}_{ij} = 0$,

$$\mathbb{P}(X_j \in \widehat{M}_i \mid \tilde{\mathbf{X}}_{-i}) = \mathbb{P} \left(|t_{N_o-p}| \geq \bar{\alpha} \sqrt{N_o} \right) \leq 2 \exp \left(-\frac{\bar{\alpha}^2 N_o}{4} \right) + \exp \left(-\frac{N_o - p}{16} \right).$$

Therefore,

$$\begin{aligned}
\mathbb{P} \left(\widehat{M}_i \neq M_i \mid \tilde{\mathbf{X}}_{-i} \right) & \leq \sum_{j: \bar{\beta}_{ij} \neq 0} \mathbb{P} \left(X_j \notin \widehat{M}_i \mid \tilde{\mathbf{X}}_{-i} \right) + \sum_{j: \bar{\beta}_{ij} = 0} \mathbb{P} \left(X_j \in \widehat{M}_i \mid \tilde{\mathbf{X}}_{-i} \right) \\
& \leq (p-1) \left[2 \exp \left(-\frac{\bar{\alpha}^2 N_o}{4} \right) + \exp \left(-\frac{N_o - p}{16} \right) \right] \\
& \quad + s_1 \exp \left(-\frac{N_o(t) - p}{4} \right).
\end{aligned}$$

□

Proof. Let $S(i, t)$ denote the backdoor adjustment set sampled from $\widehat{\mathcal{B}}_i$ for node X_i at time

step t , then

$$\begin{aligned} \mathcal{R}_T &= \sum_{i=1}^K \Delta_i \mathbb{E}[n_i(T)] = \sum_{i=1}^K \Delta_i \sum_{t=1}^T \mathbb{P}(A_t = i) \\ &\leq \sum_{i=1}^K \Delta_i \sum_{t=0}^{T-1} \left[\mathbb{P}(A_{t+1} = i | S(i, t) \in \mathcal{B}_i, S^*(t) \in \mathcal{B}^*) + \mathbb{P}(S(i, t) \notin \mathcal{B}_i) + \mathbb{P}(S^*(t) \notin \mathcal{B}^*) \right] \end{aligned} \quad (2.43)$$

Define the following constants

$$\begin{aligned} \tilde{\eta}_i^2 &= \max_{S \in \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}, \\ \tilde{\phi}^2 &= \max_i \max_{S \in \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S}^2}{\gamma_{\min}(\Sigma)\tilde{\sigma}_i^2}, \frac{25\tilde{\sigma}_i^2\gamma_{\max}(\Sigma)}{9\zeta_{i,S}^2} \right\}. \end{aligned}$$

We have proved that under the correct backdoor adjustment sets, the upper bound of $\mathcal{R}_T$ can be expressed as in (2.28), i.e., the first term is bounded.

For any i, t , since

$$\{0 \in CI_i(t, \text{pa}(i))\} \cap \{d(0, CI_i(t, S')) > 0, \forall S' \in \mathcal{N}_i \setminus \mathcal{B}_i\} \cap \{\widehat{M}_i = M_i\} \implies S \in \mathcal{B}_i,$$

thus

$$\mathbb{P}(S(i, t) \notin \mathcal{B}_i) \leq \mathbb{P}(0 \notin CI_i(t, \text{pa}(i))) + \mathbb{P}(\widehat{M}_i \neq M_i) + \sum_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P}(d(0, CI_i(t, S')) = 0).$$

Re-define the following constants:

$$\begin{aligned} \psi^2 &= \max_i \max_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S'}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}, \\ \bar{\alpha} &= \min_i \frac{\sqrt{\gamma_{\min}(\Sigma)}\bar{\beta}_i}{6\sqrt{3}\tilde{\sigma}_i}, \quad \alpha = \bar{\alpha}/2, \quad s_0 = \max_{i, S \in \mathcal{B}_i} |S|. \end{aligned}$$

For any $t \geq \ell_0$, we have the following bound from Lemma 2.1,

$$\mathbb{P} \left(\exists i, j \in V, i \neq j, \rho_{ij}(t) \geq \frac{9}{\gamma_{\min}(\Sigma)} \right) \leq \mathbb{P} \left(\frac{\sigma_{\min}(\tilde{\mathbf{X}})}{\sqrt{N_o(t)}} \leq \frac{1}{3} \gamma_{\min}(\sqrt{\Sigma}) \right) \leq \exp \left(-\frac{N_o(t)}{18} \right). \quad (2.44)$$

Incorporating Lemma 2.4, 2.5, and 2.6 altogether and averaging over $\tilde{\mathbf{X}}_{-i}$, then we have that for any i and $t \geq \ell_0$,

$$\begin{aligned} & \mathbb{P}(S(i, t) \notin \mathcal{B}_i) \\ & \leq \mathbb{P} \left(0 \notin CI_i(\text{pa}(i)) \middle| \tilde{\mathbf{X}}_{-i} \right) + \mathbb{P} \left(\widehat{M}_i \neq M_i \middle| \tilde{\mathbf{X}}_{-i} \right) + \sum_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P} \left(d(0, CI_i(S')) = 0 \middle| \tilde{\mathbf{X}}_{-i} \right) \\ & \quad + \mathbb{P} \left(\frac{\sigma_{\min}(\tilde{\mathbf{X}})}{\sqrt{N_o(t)}} \leq \frac{1}{3} \gamma_{\min}(\sqrt{\Sigma}) \right) \\ & \leq 3t^{-2} + \exp \left(-\frac{N_o(t) - s_0 - 1}{9} \right) + s_2 \left(3t^{-2} + \exp \left(-\frac{N_o(t) - p}{4} \right) \right) + \exp \left(-\frac{N_o(t)}{18} \right) \\ & \quad + (p-1) \left(2 \exp \left(-\frac{\bar{\alpha}^2 N_o(t)}{4} \right) + \exp \left(-\frac{N_o(t) - p}{16} \right) \right) + s_1 \exp \left(-\frac{N_o(t) - p}{4} \right) \\ & \leq 3(s_2 + 1)t^{-2} + (s_1 + s_2 + p + 1) \exp \left(-\frac{n_0 + nt - p}{18} \right) + 2(p-1) \exp(-\alpha^2(n_0 + nt)). \end{aligned} \quad (2.45)$$

Therefore,

$$\begin{aligned} & \sum_{t=\ell}^T \mathbb{P}(S(i, t) \notin \mathcal{B}_i) \\ & \leq \frac{(s_2 + 1)\pi^2}{2} + \frac{18(s_1 + s_2 + p + 1)}{n} \exp \left(-\frac{n_0 + n\ell - p}{18} \right) + \frac{2(p-1)}{\alpha^2 n} \exp(-\alpha^2(n_0 + n\ell)). \end{aligned} \quad (2.46)$$

Let $c \geq 4\sqrt{2}\tilde{\phi}$, $c_3 \geq \max \left\{ 64, \frac{32\psi^2}{\delta^2} \right\}$. Aside from the condition in (2.24), ℓ also need to satisfy $\ell \geq \left(\frac{c_3}{n} \right)^2$. The upper bound still has the form as in (2.28), i.e.,

$$\log T \left(\frac{20c^2}{n+m} \sum_{i: \mu_i < \mu^*} \frac{\tilde{\eta}_i^2}{\Delta_i} + \frac{c_3 \sum_{i=1}^K \Delta_i}{m} \right) + C_3 \left(\sum_{j=1}^K \Delta_j \right).$$

Since $p = K + 1$,

$$\begin{aligned}
C_3 &= \left(\frac{c_3}{n}\right)^2 + C_1 + \sum_{t=\ell_0}^T \mathbb{P}(S(i, t) \notin \mathcal{B}_i) + \sum_{t=\ell_0}^T \mathbb{P}(S^*(t) \notin \mathcal{B}^*) \\
&= \left(\frac{c_3}{n}\right)^2 + (2.29) + 2 \times (2.46) \\
&= \ell_0 + 1 + \pi^2 \left(s_2 + \frac{4}{3} + \frac{1}{3} \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right) + \left(\frac{c_3}{n}\right)^2 \\
&\quad + \frac{4K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{36(s_1 + s_2 + K + 3)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}.
\end{aligned} \tag{2.47}$$

□

Remark 2.4. If $N_o(t)$ can be less than p , we can consider using Lasso-based neighborhood regression [30] instead, i.e.,

$$\hat{\beta}_i = \operatorname{argmin}_{\beta_i \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2N_o(t)} \|X_i - \mathbf{X}_{-i}\bar{\beta}_i\|^2 + \lambda_{N_o} \|\bar{\beta}_i\|_1 \right\}.$$

According to Theorem 11.2 in Wainwright [30], if the non-zero entries of the precision matrix are bounded below in absolute value, then the estimated edge set $\hat{E}$ equals the actual edge set E with high probability, where $N_o(t) \gtrsim 2s_0 \log p$, and $\lambda_{N_o} = c_0 \left\{ \frac{1}{\alpha_0} \sqrt{\frac{\log p}{N_o(t)}} + v \right\}$ for some $v \in (0, 1]$. More specifically, $\mathbb{P}(\hat{E} = E) \geq 1 - c_4 \exp(-c_5 N_o(t))$, where c_4 and $c_5 \leq 1$ are constants. Therefore, if we use Lasso based neighborhood regression, we have

$$\mathbb{P}(\widehat{M}_i \neq M_i) \leq c_4 \exp(-c_5 N_o(t)).$$

Given this result, with Lasso based neighborhood regression, we can relax the dependency of this upper bound on K in Lemma 2.6.

2.7.3.2 Proof of Theorem 2.2

Proof. We start with the inequality (2.31) in the proof of Theorem 2.4:

$$\mathcal{R}_T \leq 2\ell + 2 \sum_{t=\ell}^{T-1} \mathbb{P}(\mu^* > U_t(A^*)) + \sum_{t=\ell+1}^T \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t})$$

If DAG is unknown, we have the following bounds:

$$\mathbb{P}(\mu^* > U_t(A^*)) \leq \mathbb{P}(\mu^* > U_t(A^*) | S^*(t) \in \mathcal{B}^*) + \mathbb{P}(S^*(t) \notin \mathcal{B}^*),$$

$$\mathbb{E}(U_{t-1}(A_t) - \mu_{A_t}) \leq \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t} | S(A_t, t-1) \in \mathcal{B}_{A_t}) + 2\mathbb{P}(S(A_t, t-1) \notin \mathcal{B}_{A_t}).$$

In Section 2.7.3.1 (2.45), we have proved an upper bound for $\mathbb{P}(S(i, t) \notin \mathcal{B}_i)$, $\forall i, \forall t \geq \ell$. Combined with the upper bound proved under the DAG given setting in Section 2.7.2.2, denote $\tilde{\eta}^2 = \max_i \tilde{\eta}_i^2$, and let $c \geq 4\tilde{\phi}$, $c_3 \geq \max\left\{32, \frac{32\psi^2}{\delta^2}\right\}$, $\ell \geq \lceil \frac{9\text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil = \ell_0$, we have the following result

$$\begin{aligned} \mathcal{R}_T &\leq 2c\tilde{\eta}\sqrt{\log T} \min \left\{ \frac{\sqrt{(n+m)KT + K^2n_0} - \sqrt{K^2n_0}}{n+m}, \frac{\sqrt{n_0 + nT} - \sqrt{n_0}}{n} \mathbb{1}\{n \geq 1\} \right\} \\ &\quad + \frac{2Kc_3 \log T}{m} + C_4, \end{aligned}$$

where

$$\begin{aligned} C_4 &= 2\left(\frac{c_3}{n}\right)^2 + C_2 + 2 \sum_{t=\ell}^{T-1} \mathbb{P}(S^*(t) \notin \mathcal{B}^*) + 2 \sum_{t=\ell+1}^T \mathbb{P}(S(A_t, t-1) \notin \mathcal{B}_{A_t}) \\ &= 2\left(\frac{c_3}{n}\right)^2 + (2.37) + 4 \times (2.46) \\ &= 2\ell_0 + \left(2s_2 + \frac{7}{3}\right)\pi^2 + 2K + 2\left(\frac{c_3}{n}\right)^2 + \frac{\pi^2}{6} \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \\ &\quad + \frac{72(s_1 + p + s_2 + 2)}{n} \exp\left(-\frac{n_0 + n\ell_0 - p}{18}\right) + \frac{8(p-1)}{\alpha^2 n} \exp(-\alpha^2(n_0 + n\ell_0)), \end{aligned}$$

Given $K = p - 1$, we have

$$\begin{aligned}
C_4 = & \pi^2 \left(2s_2 + \frac{7}{3} + \frac{1}{6} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + 2\ell_0 + 2 \left(\frac{c_3}{n} \right)^2 + 2K \\
& + \frac{8K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{72(s_1 + s_2 + K + 3)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}
\end{aligned}$$

□

CHAPTER 3

Causal bandit on unknown graphs in presence of confounders

3.1 Introduction

In Chapter 2, we studied causal bandits under the assumption of causal sufficiency, where all common causes of the observed variables are themselves observed. Under this assumption, the data-generating mechanism can be represented by a directed acyclic graph, and valid backdoor adjustment sets may exist for estimating causal effects. Observational data can therefore be leveraged to reduce uncertainty in reward estimation and improve learning efficiency.

However, in many applications, some variables in the underlying causal system are unobserved, which may lead to latent confounding among the observed variables. When hidden confounding is present, the observed system is more appropriately represented by an acyclic directed mixed graph (ADMG) containing both directed and bidirected edges [32]. In this setting, valid adjustment sets may not exist for all intervention nodes, introducing additional methodological challenges.

Several works in the causal bandit literature allow for hidden confounders while assuming that the complete causal structure is known. For example, Lee and Bareinboim [21] and Maiti, Nair, and Sinha [33] analyze regret guarantees in the presence of confounding when the graph is provided. Structural knowledge allows the learner to determine which effects are identifiable and to design appropriate estimation strategies. However, in realistic appli-

cations, neither the directed edges nor the confounding relationships are typically known in advance.

In this chapter, we extend the BA-UCB algorithm to accommodate latent confounders. We model the observed variables using Gaussian ADMGs and develop a procedure that integrates observational and experimental data without assuming prior knowledge of the graph. Rather than attempting to recover the full mixed graph, our objective remains focused on identifying the intervention with the largest expected causal effect. The proposed approach determines, for each arm, whether a valid adjustment set exists and adapts related estimates accordingly.

When latent confounders are present, we show that, with sufficiently large amount of prior observational data, the cumulative regret of BA-UCB is bounded by $O(\sqrt{T \log T / (m + n)}) + O(\sqrt{K_1 T \log T / m})$, where K_1 denotes the number of variables for which a valid adjustment set does not exist or cannot be identified. When K_1 is small, the dependence of the regret on the total number of arms K is again substantially reduced, highlighting the benefit of incorporating observational data into sequential decision-making.

The remainder of this chapter is organized as follows. Section 3.2 introduces the Gaussian ADMG model and the necessary preliminaries. Section 3.3 presents a modified BA-UCB algorithm assuming the underlying causal model is an ADMG. Section 3.4 establishes regret bounds that distinguish between identifiable and non-identifiable arms. Section 3.5 reports simulation results demonstrating the empirical effectiveness of the proposed algorithm.

3.2 Preliminaries

If some variables in a causal DAG are unobserved, the marginal distribution over the observed variables $\mathcal{X} = \{X_1, \dots, X_p\}$ can be represented by an ADMG $\mathcal{G}$ over nodes $V = \{1, \dots, p\}$ that contains both directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges. Directed edges encode direct causal relations among the observed variables, while a bidirected edge $i \leftrightarrow j$ indicates the

existence of one or more latent confounders influencing both X_i and X_j . If $i \rightarrow j$, then i is a parent of j and j is a child of i . If there is a directed path from i to j or $i = j$, then i is an ancestor of j and j is a descendant of i . Non-descendants of a node i consist of every node that is not a descendant of i . For each node i in $\mathcal{G}$, the parent set $\text{pa}(i) = \{j \in V : j \rightarrow i\}$. Similarly, we use $\text{ch}(i)$, $\text{an}(i)$, $\text{de}(i)$, and $\text{nd}(i)$ to denote the set of children, ancestors, descendants, and non-descendants of the node i . The district of i , denoted as $\text{dis}(i)$, is the maximal set of nodes connected to the node i through a bidirected path, i.e., its bidirected-connected component. The parent set of a district D in the graph $\mathcal{G}$ is defined as $\text{pa}(D) = \{\cup_{i \in D} \text{pa}(i)\} \setminus D$, which consists of the out-of-district parents of all nodes in the district D .

We assume a linear SEM for the observed variables $\mathcal{X}$ in an ADMG:

$$X_j = \sum_{i \in \text{pa}(j)} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \dots, p, \quad (3.1)$$

where $\beta_{ij} \neq 0$ if $i \in \text{pa}(j)$ and each error variable $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$ with $\omega_j^2 > 0$. Different from the fully observed case in (2.1), the error term $\varepsilon_1, \dots, \varepsilon_p$ here are not mutually independent. Let $\Omega = (\omega_{ij}^2)_{p \times p} = \text{Cov}(\varepsilon)$ be the covariance matrix of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^\top$. Then the off-diagonal entry $\omega_{ij}^2 \neq 0$ if and only if there is a bidirected edge $i \leftrightarrow j$. In matrix form, the model can be written as

$$X = B^\top X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Omega), \quad (3.2)$$

where $X = (X_1, \dots, X_p)^\top$ and $B = (\beta_{ij})_{p \times p}$. This model implies that $X \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = (I - B)^{-\top} \Omega (I - B)^{-1}$. In this setting, assuming $\Sigma > 0$, the regression neighborhood M_i of node i is given by

$$M_i = \{j \in V_{-i} : (\Sigma^{-1})_{ij} \neq 0\} = \{j \in V_{-i} : X_i \not\perp\!\!\!\perp X_j | X_{V \setminus \{i, j\}}\}. \quad (3.3)$$

Let $\mathcal{N}_i$ denote the collection of all subsets of M_i .

3.3 BA-UCB with latent confounders

Our BA-UCB algorithm proposed in Chapter 2 can be naturally extended to settings where the underlying causal DAG includes unobserved confounders. A key advantage of BA-UCB is that it does not require structure learning of the entire causal graph. Instead, it leverages observational data to generate estimates when the causal effect of interest is identifiable via covariate adjustment, that is, when there exists at least one valid adjustment set, $\mathcal{B}_i = \{S : \mu_i(S) = \mu_i\} \neq \emptyset$. However, in the presence of latent confounders, such valid adjustment sets may not always exist. For example, if X and Y are connected by a bidirected edge, no observed variable can block the backdoor path $X \leftrightarrow Y$, and consequently, no valid adjustment set exists. In such cases, the original BA-UCB design, $\hat{\mathcal{B}}_i = \arg \min_S d(0, CI_i(t, S))$, will still select a subset even though there is no valid adjustment set, which may lead to biased causal effect estimates.

To address this issue, we modify Line 8 in Algorithm 2.1 to:

$$\hat{\mathcal{B}}_i = \{S \subseteq \widehat{M}_i : 0 \in CI_i(t, S)\}, \quad (3.4)$$

where $\widehat{M}_i$ is the estimated regression neighborhood of node i . That is, we only consider subsets S of $\widehat{M}_i$ whose confidence interval $CI_i(t, S)$ for the difference between observational and experimental causal effect estimates contains zero. If $\hat{\mathcal{B}} \neq \emptyset$, the algorithm proceeds as usual with adjustment set sampling and causal effect estimation (Lines 9 and 10). If $\hat{\mathcal{B}}_i = \emptyset$, we default to purely interventional estimation using only the estimates in (2.7). The modified procedure is summarized in Algorithm 3.1.

This modification ensures that BA-UCB remains theoretically valid and empirically robust in the presence of latent confounders. When valid adjustment sets exist in the regression neighborhood, the algorithm efficiently integrates observational data to improve estimation accuracy and sample efficiency. When no such sets exist, it automatically falls back to unbiased experimental estimation. As we will show in the subsequent sections, this extension

Algorithm 3.1 BA-UCB for Gaussian DAGs with confounders

Input: Observational data $\mathcal{D}_o(0)$ of size n_0 , parameters c, c_3, n, m, κ , rounds T .

```

1: for  $t = 0, \dots, T - 1$  do
2:   if there is a node  $i$  which has been intervened less than  $\lceil \frac{c_3 \log t + 1}{m} \rceil$  times then
3:     Choose intervention node  $A_{t+1} = i$ ;
4:   else
5:     for  $i = 1, \dots, p - 1$  do
6:       if  $t \bmod \kappa = 0$  then
7:         Estimate the regression neighborhood  $\widehat{M}_i$  from  $\mathcal{D}_o(t)$ ;
8:         Identify candidate adjustment sets  $\widehat{\mathcal{B}}_i = \{S \subseteq \widehat{M}_i : 0 \in CI_i(t, S)\}$ ;
9:         if  $\widehat{\mathcal{B}}_i \neq \emptyset$  then
10:          Sample set  $S$  from  $\widehat{\mathcal{B}}_i$ , calculate  $\hat{\mu}_{i,\text{obs}}(t, S)$  and  $\hat{\sigma}_{i,\text{obs}}(t, S)$  as in (2.8);
11:          Calculate estimates  $\hat{\mu}_i(t, S)$ ,  $\hat{\sigma}_i^2(t, S)$  as in (2.12);
12:        else
13:          Update  $\hat{\mu}_i(t) = \hat{\mu}_{i,\text{int}}(t)$  and  $\hat{\sigma}_i^2(t) = \widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t))$ ;
14:         $A_{t+1} = \text{argmax}_i \{\hat{\mu}_i(t) + \hat{\sigma}_i(t)c\sqrt{\log t}\}$ ;
15:   Generate  $m$  data points under intervention  $A_{t+1}$  and  $n$  observational data.
16:   Update  $\hat{\mu}_{i,\text{int}}(t + 1)$  and  $\widehat{\text{Var}}(\hat{\mu}_{i,\text{int}}(t + 1))$  as in (2.7).

```

Output: Final intervention A_T and estimated reward $\hat{\mu}^*$.

preserves strong theoretical guarantees while maintaining superior empirical performance, even when valid adjustment sets exist for only a subset of the variables.

Similar to the fully observed setting, the restriction of the search for adjustment sets to the regression neighborhood M_i of each variable is for computational efficiency. We next show that under certain condition an adjustment set is contained in the regression neighborhood. Define $\text{mb}(i) := \text{pa}(\text{dis}(i)) \cup (\text{dis}(i) \setminus \{i\})$ as the Markov blanket of i with respect to an ordering of an ADMG [32]. The following lemma gives a sufficient condition for $\text{mb}(i)$ to be a valid adjustment set for the causal effect of X_i on Y . If $j \in \text{mb}(i)$, then $X_i \not\perp\!\!\!\perp X_j | X_{-\{i,j\}}$ and thus $\text{mb}(i) \subseteq M_i$ by equality (3.3).

Lemma 3.1. *Assume the Gaussian ADMG model (3.1) with $\Sigma > 0$. If $\text{dis}(i) \cap \text{de}(i) = \{i\}$ and $Y \notin \text{mb}(i) \cup \{i\}$, then $\text{mb}(i)$ is a valid adjustment set for the causal effect of X_i on Y .*

The proof of Lemma 3.1 is provided in Section 3.7.1. In an ADMG, latent confounding

is represented by bidirected edges and induces districts. The condition $\text{dis}(i) \cap \text{de}(i) = \{i\}$ ensures that there are no directed paths from i to other nodes in its district. Under this condition, the node i is fixable, and the post-intervention distribution $p(x_{V \setminus \{i\}} \mid \text{do}(x_i))$ is identifiable [34]. Consequently, the Markov blanket $\text{mb}(i)$ provides a valid adjustment set for identifying the causal effect of X_i on Y as stated in Lemma 3.1.

3.4 Regret analysis

The underlying model assumptions for Algorithm 3.1 are stated formally in Assumption 3.1.

Assumption 3.1. *The random vector $X = (X_1, \dots, X_p)$ satisfies the linear SEM (3.1) associated with an ADMG $\mathcal{G}$ such that $\beta_{ij} \neq 0$ if and only if $i \in \text{pa}(j)$, $\omega_{ij} \neq 0$ if and only if $i \leftrightarrow j$ for $i \neq j$, and $\Omega > 0$.*

Let $\mathcal{I}_0 := \{i \in \bar{V} : \mu_i(S) = \mu_i \text{ for some } S \in \mathcal{N}_i\}$ be the set of nodes for which there is a valid adjustment set in its regression neighborhood for identifying its causal effect on Y , and let $\mathcal{I}_1 = \bar{V} \setminus \mathcal{I}_0$. Recall the constants in (2.13) in Section 2.4. Further define $\tilde{\sigma} = \max_{i \in \bar{V}} \tilde{\sigma}_i$. We now establish regret bounds for the BA-UCB algorithm under the ADMG model. The first result provides a case-dependent upper bound that depends on the reward gaps Δ_i , which separates the regrets from arms in $\mathcal{I}_0$ and $\mathcal{I}_1$.

Theorem 3.1. *Let $\mathcal{R}_T$ be the cumulative regret of Algorithm 3.1 with the modification (3.4), and $c \geq 4\sqrt{2}\tilde{\phi}$, $c_3 \geq \max\{64, 32\psi^2/\delta^2\}$. If Assumptions 3.1, 2.2 and 2.3 hold, then after T rounds the cumulative regret $\mathcal{R}_T$ is at most*

$$\frac{20c^2 \log T}{m+n} \sum_{i \in \mathcal{I}_0} \frac{\tilde{\eta}_i^2}{\Delta_i} + \frac{10c^2 \log T}{m} \sum_{i \in \mathcal{I}_1} \frac{\tilde{\sigma}_i^2}{\Delta_i} + \left(\frac{c_3 \log T}{m} + C_5 \right) \sum_i \Delta_i, \quad (3.5)$$

where C_5 is a constant independent of T .

See Section 3.7.2 for a proof of this theorem. In general, C_5 depends on the covariance matrix Σ , the initial observational sample size n_0 , and the per-round observational data size

n , and its exact expression is

$$C_5 = \pi^2 \left(s_2 + \frac{2}{3} + \frac{1}{3} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + \ell_0 + \left(\frac{c_3}{n} \right)^2 + 1 \\ + \frac{4K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{36(s_1 + s_2 + K + 2)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}.$$

This regret bound explicitly reflects the heterogeneity across arms in terms of causal effect identifiability. For arms in $\mathcal{I}_0$, our algorithm benefits from both observational and experimental data, reflected in the $(m + n)^{-1}$ scaling. For arms in $\mathcal{I}_1$, only experimental samples contribute to reducing uncertainty, leading to a larger regret term. This decomposition generalizes the result from the fully observed setting (Theorem 2.1): If the causal effects of all arms are identifiable, i.e., $\mathcal{I}_1 = \emptyset$, the bound (3.5) reduces to the form of Theorem 2.1; conversely, if none of them is identifiable, i.e., $\mathcal{I}_0 = \emptyset$, the regret scales as $O((\log T)/m)$, where m is the experimental sample size. This latter scaling is identical to that of the standard UCB.

Similar to Theorem 2.2, we also establish a Δ -independent upper bound that does not depend on the reward gaps Δ_i , under the additional assumption that all rewards are bounded.

Theorem 3.2. *Assume $\mu_i \in [-1, 1], \forall i \in \{1, \dots, p-1\}$. Then under Assumptions 3.1, 2.2 and 2.3, the cumulative regret $\mathcal{R}_T$ of Algorithm 3.1, with the modification (3.4) and parameters $c \geq 4\tilde{\phi}$ and $c_3 \geq \max\{32, 32\psi^2/\delta^2\}$, after T rounds is at most*

$$2c\sqrt{\log T} \left[\frac{\tilde{\eta} \left(\sqrt{|\mathcal{I}_0|(n+m)T + n_0|\mathcal{I}_0|^2} - \sqrt{n_0|\mathcal{I}_0|} \right)}{n+m} + \tilde{\sigma} \sqrt{\frac{|\mathcal{I}_1|T}{m}} \right] + \frac{2Kc_3 \log T}{m} + C_6, \quad (3.6)$$

where C_6 is a constant independent of T .

See Section 3.7.3 for a proof of this theorem. In general, C_6 depends on Σ , n_0 , m , and n ,

and the exact expression is

$$C_6 = \pi^2 \left(2s_2 + \frac{1}{2} + \frac{1}{6} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + 2\ell_0 + 2 \left(\frac{c_3}{n} \right)^2 + 2K \\ + \frac{8K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{72(s_1 + s_2 + K + 2)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}.$$

This regret bound captures the worst-case scenario across all problem instances and avoids dependence on the reward gaps. Similar to Theorem 3.1, it separates arms according to identifiability. As discussed in Section 2.4.2, if a sufficient amount of prior observational data is available, i.e., with a large initial sample size n_0 , the first term in the cumulative regret upper bound in (3.6)

$$\frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \left(\sqrt{|\mathcal{I}_0|(n+m)T + n_0|\mathcal{I}_0|^2} - \sqrt{n_0}|\mathcal{I}_0| \right) = O \left(\sqrt{T \log T / (n+m)} \right).$$

As such, it does not depend on the number of identifiable arms $|\mathcal{I}_0|$, in contrast to the standard MAB bound $O(\sqrt{|\mathcal{I}_0|T \log T / (n+m)})$. Consequently, the overall cumulative regret can be significantly reduced compared to traditional MAB algorithms that rely solely on experimental data, especially if the number of identifiable arms $|\mathcal{I}_0|$ is large.

3.5 Numerical experiments

We evaluate the empirical performance of the BA-UCB algorithm in comparison with the standard UCB algorithm when the underlying causal DAG includes latent confounders. DAGs were randomly generated with a total number of variables (including unobserved confounders) $N \in \{10, 20, 30\}$ and a maximum in- and out-degree of 3. In each graph, 10% of the variables were chosen as unobserved confounders, resulting in $p \in \{9, 18, 27\}$ observed variables. To control the complexity of the induced ADMG, confounders were selected such that they did not share any common child.

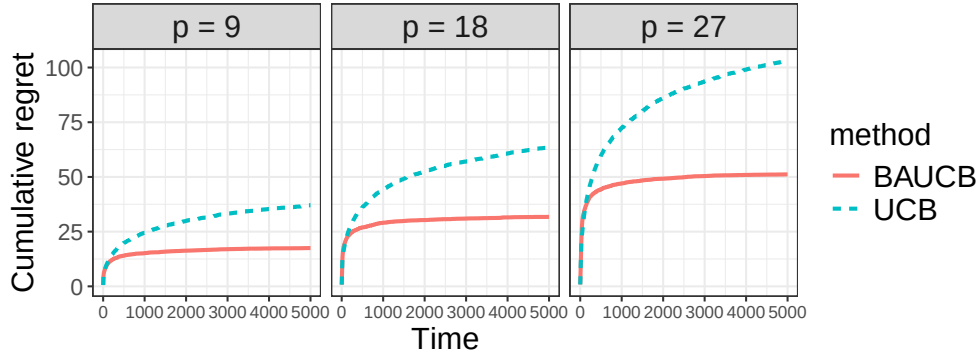


Figure 3.1: Comparison of empirical cumulative regrets over 5000 time steps between the BA-UCB and standard UCB algorithms on graphs with latent confounders. The number of observed variables $p = 9, 18, 27$, corresponding to a total of $N = 10, 20, 30$ variables, respectively.

For BA-UCB, the initial observational sample size was set to $n_0 = 160, 340$, and 520 for $p = 9, 18$, and 27 , respectively. We generated 20 experimental samples for each arm before incurring any regret, in order to satisfy the minimum intervention requirement in Algorithm 3.1 Line 2. Both observational and experimental sample sizes per round were set to $n = m = 5$. Other hyper-parameters were chosen as follows: $c = \sqrt{2}, \tau = 0.1/p$ (threshold for regression neighborhood identification), $l = 4$, and $c_2 = \frac{1}{2\sqrt{2}}$ (confidence interval threshold for identifying adjustment sets). For the standard UCB algorithm, 22 experimental samples were generated for each arm before incurring regret, and the per-round sample size was $m = 10$. Under these choices, the computational costs of the two algorithms were comparable. Each algorithm was executed five times on every generated graph.

Figure 3.1 summarizes the average results over 100 independently simulated graphs. For each graph, we computed the median cumulative regret over five runs, and then averaged the median cumulative regrets across the graphs. The results clearly demonstrate that BA-UCB consistently achieves lower cumulative regret than the standard UCB algorithm for all values of p , confirming its advantages in using observational data in the presence of latent confounders.

We also examined whether BA-UCB behaved as expected in terms of identifying adjust-

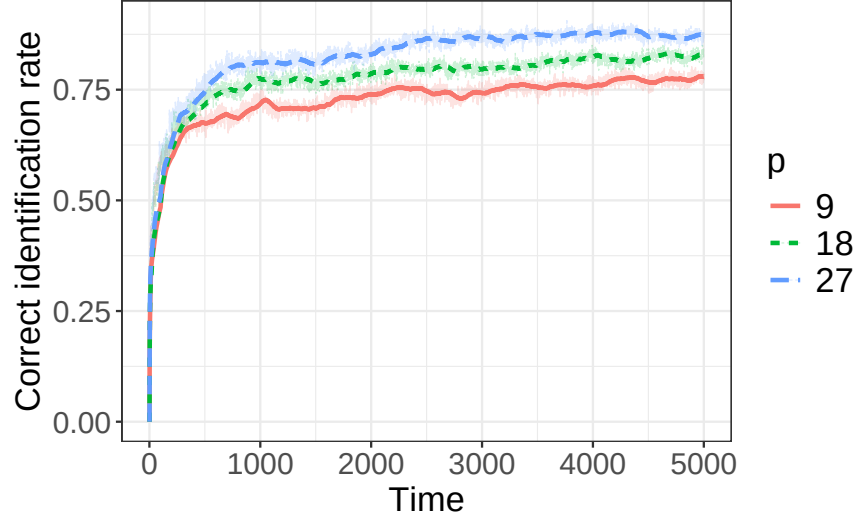


Figure 3.2: Average identification rate of a correct adjustment set for the optimal arm at each time step in the presence of latent confounders. Solid curves are smoothing averages with a 100-step window for better visualization.

ment sets. As the number of rounds increases, the algorithm accumulates more observational and experimental data, particularly for the optimal arm, leading to more reliable estimation of causal effects and more accurate identification of valid adjustment sets. In our numerical results, an identified adjustment set for the causal effect μ_i is regarded as correct (up to certain numerical precision) if $|\mu_i(S) - \mu_i| < 0.01$, where $\mu_i(S)$ is the theoretical value of the adjusted estimate of the causal effect as defined in (2.9). At each time step, we computed the frequency that the sampled adjustment set for the optimal arm is a true adjustment set, among 5 independent runs and 100 graphs. Figure 2.5 presents this identification rate over time for different graph sizes p , showing a steady improvement as BA-UCB accumulates more data. In particular, the rate of correct identification increased to more than 75% near $T = 5000$ for all graph sizes.

3.6 Discussion

In this chapter, we extended the BA-UCB algorithm to causal bandit problems in the presence of latent confounders. Unlike the setting studied in Chapter 2, valid backdoor ad-

justment sets may not exist for every intervention when latent confounders are present. The proposed extension determines, for each arm, whether a valid adjustment set can be identified and adapts its estimation mechanism accordingly. When a valid adjustment set is available, both observational and experimental data are incorporated into reward estimation. When adjustment is not feasible, the algorithm relies solely on experimental samples. This adaptive design ensures robustness to unobserved confounding while preserving the benefit of incorporating observational data whenever identification is possible.

Suppose that m experimental samples and n observational samples are collected at each round, and that a sufficiently large amount of prior observational data is available. We prove that the cumulative regret of BA-UCB after T rounds is bounded by $O(\sqrt{T \log T / (m + n)}) + O(\sqrt{K_1 T \log T / m})$, where K_1 denotes the number of variables for which a valid adjustment set does not exist or cannot be identified. When K_1 is small relative to the total number of arms K , the overall dependence on K is substantially reduced compared to classical bandit algorithms. This makes clear that observational data remain useful even under latent confounding, provided that adjustment is valid for a nontrivial subset of interventions.

The current analysis assumes Gaussian linear structural equation models. The Gaussian assumption ensures closed-form expressions for adjustment estimators and allows us to derive concentration bounds using standard tail inequalities. Extending the method to non-Gaussian noise or nonlinear structural equations would require new concentration results and possibly different estimates for causal effects. Whether similar regret rates can be obtained in those settings remains an open question.

3.7 Appendix

3.7.1 Proof of Lemma 3.1

Proof. Let $<$ be a topological order of the ADMG $\mathcal{G}$, and denote by $[v]$ the rank of v under $<$. Define $V_{[i]} := \{v \in V : [v] \leq [i]\}$ and let $\mathcal{G}_{[i]}$ be the induced subgraph of $\mathcal{G}$ on $V_{[i]}$. If

$\text{dis}(i) \cap \text{de}(i) = \{i\}$, then i is fixable in $\mathcal{G}(V)$ [34]. According to Definition 19 in Richardson et al. [34], for a fixable vertex i , we have

$$p(x_{V \setminus \{i\}} | do(x_i)) = \frac{p(x_V)}{p(x_i | x_{\text{mb}(i, \mathcal{G}_{[i]})})}, \quad (3.7)$$

where $\text{mb}(i, \mathcal{G}_{[i]})$ is the Markov blanket of the node i in the induced graph $\mathcal{G}_{[i]}$.

Since $\text{dis}(i) \cap \text{de}(i) = \{i\}$, the node i can be sorted as the last vertex in its district. Therefore, under this topological sorting, $\text{mb}(i, \mathcal{G}_{[i]}) = \text{mb}(i)$. Then we have

$$\begin{aligned} p(x_{V \setminus \{i\}} | do(x_i)) &= \frac{p(x_V)}{p(x_i | x_{\text{mb}(i)})} \\ &= \frac{p(x_V)}{p(x_i, x_{\text{mb}(i)}) / p(x_{\text{mb}(i)})} = \frac{p(x_V)}{p(x_i, x_{\text{mb}(i)})} p(x_{\text{mb}(i)}) \\ &= p(x_{V \setminus \{i, \text{mb}(i)\}} | x_i, x_{\text{mb}(i)}) p(x_{\text{mb}(i)}). \end{aligned}$$

Therefore

$$\begin{aligned} p(y | do(x_i)) &= \int p(x_{V \setminus \{i\}} | do(x_i)) dx_{\overline{V \setminus \{i\}}} \\ &= \int p(x_{V \setminus \{i, \text{mb}(i)\}} | x_i, x_{\text{mb}(i)}) p(x_{\text{mb}(i)}) dx_{\overline{V \setminus \{i\}}} \\ &= \int p(x_{\text{mb}(i)}) \left[\int p(x_{V \setminus \{i, \text{mb}(i)\}} | x_i, x_{\text{mb}(i)}) dx_{\overline{V \setminus \{i, \text{mb}(i)\}}} \right] dx_{\text{mb}(i)} \\ &= \int p(x_{\text{mb}(i)}) p(y | x_i, x_{\text{mb}(i)}) dx_{\text{mb}(i)}, \end{aligned}$$

which shows that $\text{mb}(i)$ is a valid adjustment set for $p(y | do(x_i))$. □

3.7.2 Proof of Theorem 3.1

Proof. The proof starts from

$$\mathcal{R}_T = \sum_{i=1}^K \Delta_i \mathbb{E}[n_i(T)] = \sum_{i=1}^K \Delta_i \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = i).$$

Similar to (2.43), $\mathbb{P}(A_{t+1} = i)$ can be upper bounded the conditional probability of $A_{t+1} = i$ given the sampled backdoor adjustment set is correct (none if not exists) for arm i and the optimal arm plus the probability of the sampled backdoor adjustment sets are wrong for arm i or the optimal arm.

As long as the identification of backdoor adjustment sets is correct, we can always decompose

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = i) &\leq \ell + \sum_{t=\ell}^{T-1} \sum_{u=1}^t \sum_{u_i=\ell}^t \left[\mathbb{P}(\hat{\mu}_u^*(t) \leq \mu^* - \hat{\sigma}_u^*(t)c_t) + \mathbb{P}(\hat{\mu}_{i,u_i}(t) \geq \mu_i + \hat{\sigma}_{i,u_i}(t)c_t) \right. \\ &\quad \left. + \mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t) \right]. \end{aligned}$$

Define $\tilde{\phi}^2 = \max_{i \in \mathcal{I}_0} \max_{S \in \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S}^2}{\gamma_{\min}(\Sigma)\tilde{\sigma}_i^2}, \frac{25\tilde{\sigma}_i^2\gamma_{\max}(\Sigma)}{9\zeta_{i,S}^2} \right\}$, $s_0 = \max_{i \in \mathcal{I}_0, S \in \mathcal{B}_i} |S|$.

If a node $i \in \mathcal{I}_0$, i.e., $\mathcal{B}_i \neq \emptyset$, as proved in (2.19) of Section 2.7.2.1, when $u_i \geq \ell \geq \frac{20c^2\tilde{\eta}_i^2 \log t}{(m+n)\Delta_i^2}$ and $u_i \geq \frac{16 \log t}{n+m}$, we have

$$\mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t \mid \mathbf{X}_{\bar{V}}) \leq t^{-4} \exp\left(-\frac{n_0 - s_0 - 2}{4}\right) = F_1, \quad (3.8)$$

and when $u_i \geq \frac{64 \log t}{n+m}$, $c \geq 4\sqrt{2}\tilde{\phi}$, as proved in (2.21) of Section 2.7.2.1, we have

$$\mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i}c_t \mid \mathbf{X}_{\bar{V}}) \leq t^{-4} \left(1 + \frac{1}{2} \exp\left(-\frac{n_0 - s_0 - 2}{16}\right)\right) = F_2. \quad (3.9)$$

If $i \in \mathcal{I}_1$, when $u_i \geq \ell \geq \frac{10c^2\tilde{\sigma}_i^2 \log t + 1}{m\Delta_i^2}$, $u_i \geq \frac{16 \log t + 1}{m}$, we have

$$\begin{aligned} \mathbb{P}(\mu^* < \mu_i + 2\hat{\sigma}_{i,u_i}(t)c_t | \mathbf{X}_{\bar{V}}) &= \mathbb{P}\left(\hat{\sigma}_{i,u_i}^2(t) > \frac{\Delta_i^2}{4c^2 \log(t)} \middle| \mathbf{X}_{\bar{V}}\right) \\ &\leq \mathbb{P}\left(\chi_{mu_i-1}^2 > \frac{(mu_i)(mu_i-1)\Delta_i^2}{4c^2\tilde{\sigma}_i^2 \log t}\right) \\ &\leq t^{-4} = F_3. \end{aligned} \quad (3.10)$$

And when $u_i \geq \frac{16 \log t + 1}{m}$, and $c \geq 4$, we have

$$\mathbb{P}(\hat{\mu}_{i,u_i} \geq \mu_i + \hat{\sigma}_{i,u_i}c_t) = \mathbb{P}(t_{mu_i-1} \geq c\sqrt{\log t}) \leq t^{-c^2/4} + \frac{1}{2}e^{-\frac{mu_i-1}{16}} \leq \frac{3}{2}t^{-4} = F_4. \quad (3.11)$$

Aside from the above bounds, for $t \geq \ell \geq \ell_0$, we also have

$$\mathbb{P}\left(\exists i \in \bar{V}, S \subseteq \bar{V} \setminus \{i\}, \rho_i(t, S) \geq \frac{9}{\gamma_{\min}(\Sigma)} \text{ or } \rho_i(t, S) \leq \frac{9}{25\gamma_{\max}(\Sigma)}\right) \leq 2 \exp\left(-\frac{N_o(t)}{18}\right) = F_5. \quad (3.12)$$

Now we continue to bound the probability of the sampled backdoor adjustment sets are wrong for arm i or the optimal arm. For any node $i \in \mathcal{I}_0$, $\mathcal{B}_i \neq \emptyset$, and

$$\begin{aligned} \mathbb{P}(S(i, t) \notin \mathcal{B}_i) &\leq \mathbb{P}(S(i, t) \notin \mathcal{B}_i \text{ and } \widehat{M}_i = M_i) + \mathbb{P}(\widehat{M}_i \neq M_i) \\ &\leq \mathbb{P}(\exists S \in \mathcal{N}_i \setminus \mathcal{B}_i, 0 \in CI_i(t, S)) + \mathbb{P}(\widehat{M}_i \neq M_i) \\ &\leq \sum_{S \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P}(0 \in CI_i(t, S)) + \mathbb{P}(\widehat{M}_i \neq M_i). \end{aligned}$$

For $i \in \mathcal{I}_1$, $\mathcal{B}_i = \emptyset$. The probability of incorrect adjustment is bounded by

$$\begin{aligned} \mathbb{P}(\widehat{\mathcal{B}}_i \neq \emptyset) &\leq \mathbb{P}(\widehat{\mathcal{B}}_i \neq \emptyset \text{ and } \widehat{M}_i = M_i) + \mathbb{P}(\widehat{M}_i \neq M_i) \\ &\leq \mathbb{P}(\exists S \in \mathcal{N}_i \setminus \mathcal{B}_i, 0 \in CI_i(t, S)) + \mathbb{P}(\widehat{M}_i \neq M_i) \\ &\leq \sum_{S \in \mathcal{N}_i \setminus \mathcal{B}_i} \mathbb{P}(0 \in CI_i(t, S)) + \mathbb{P}(\widehat{M}_i \neq M_i). \end{aligned}$$

Therefore, the upper bound of incorrect backdoor adjustment set sampled is the same for $i \in \mathcal{I}_0$ and $i \in \mathcal{I}_1$, i.e., define $\psi^2 = \max_i \max_{S' \in \mathcal{N}_i \setminus \mathcal{B}_i} \max \left\{ \frac{9\zeta_{i,S'}^2}{\gamma_{\min}(\Sigma)}, \tilde{\sigma}_i^2 \right\}$, when $N_i(t) \geq c_3 \log t + 1$, where $c_3 \geq 32\psi^2/\delta^2$, we have

$$\begin{aligned} & \mathbb{P}(\text{Incorrect } S(i, t)) \leq \text{RHS of } [(2.41) + (2.42) + (2.44)] \\ & = 3s_2t^{-2} + (s_1 + s_2 + K + 1) \exp \left(-\frac{n_0 + nt - K - 1}{18} \right) + 2K \exp(-\alpha^2(n_0 + nt)) \quad (3.13) \\ & = F_6. \end{aligned}$$

Now we can summarize all conditions as follows.

For $i \in \mathcal{I}_0$, we need $\ell \geq \frac{c_3 \log t + 1}{m}$, $\ell \geq \frac{20c^2\tilde{\eta}_i^2 \log t}{(m+n)\Delta_i^2}$, $\ell \geq \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)}$, $\ell \geq 2$, $\ell \geq \left(\frac{c_3}{n}\right)^2$, where $c \geq 4\sqrt{2}\tilde{\phi}$, $c_3 \geq \max\{64, 32\psi^2/\delta^2\}$, i.e.,

$$\ell \geq \frac{20c^2\tilde{\eta}_i^2 \log T}{(m+n)\Delta_i^2} + \frac{c_3 \log T}{m} + \lceil \frac{9 \text{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil + \left(\frac{c_3}{n}\right)^2 + 1.$$

If the optimal arm $a^* \in \mathcal{I}_0$, $\forall t \geq \ell$,

$$\begin{aligned} \mathbb{P}(A_{t+1} = i) & \leq \mathbb{P}(A_{t+1} = i | S(i, t) \in \mathcal{B}_i, S^*(t) \in \mathcal{B}^*) + \mathbb{P}(S(i, t) \notin \mathcal{B}_i) + \mathbb{P}(S^*(t) \notin \mathcal{B}^*) \\ & \leq t^2 (F_1 + 2F_2) + F_5 + 2F_6 \\ & = 2t^{-2} \left(1 + \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + F_5 + 2F_6. \end{aligned}$$

If the optimal arm $a^* \in \mathcal{I}_1$, $\forall t \geq \ell$,

$$\begin{aligned} \mathbb{P}(A_{t+1} = i) & \leq \mathbb{P}(A_{t+1} = i | S(i, t) \in \mathcal{B}_i, S^*(t) \in \mathcal{B}^*) + \mathbb{P}(S(i, t) \notin \mathcal{B}_i) + \mathbb{P}(S^*(t) \notin \mathcal{B}^*) \\ & \leq t^2 (F_1 + F_2 + F_4) + F_5 + 2F_6 \\ & = t^{-2} \left(\frac{5}{2} + \frac{3}{2} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) + F_5 + 2F_6. \end{aligned}$$

If $i \in \mathcal{I}_1$, we need $\ell \geq \frac{10c^2\tilde{\sigma}_i^2 \log t+1}{m\Delta_i^2}$, $\ell \geq \frac{c_3 \log t+1}{m}$, $\ell \geq \frac{9 \operatorname{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)}$, $\ell \geq \left(\frac{c_3}{n}\right)^2$, i.e.,

$$\ell \geq \frac{10c^2\tilde{\sigma}_i^2 \log T}{m\Delta_i^2} + \frac{c_3 \log T}{m} + \frac{9 \operatorname{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} + \left(\frac{c_3}{n}\right)^2 + 1.$$

If $a^* \in \mathcal{I}_0$, $\forall t \geq \ell$,

$$\begin{aligned} \mathbb{P}(A_{t+1} = i) &\leq \mathbb{P}\left(A_{t+1} = i | \hat{\mathcal{B}}_i = \emptyset, S^*(t) \in \mathcal{B}^*\right) + \mathbb{P}\left(\hat{\mathcal{B}}_i \neq \emptyset\right) + \mathbb{P}\left(S^*(t) \notin \mathcal{B}^*\right) \\ &\leq t^2 (F_2 + F_3 + F_4) + F_5 + 2F_6 \\ &= t^{-2} \left(\frac{7}{2} + \frac{1}{2} \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right) + F_5 + 2F_6. \end{aligned}$$

If $a^* \in \mathcal{I}_1$, $\forall t \geq \ell$,

$$\begin{aligned} \mathbb{P}(A_{t+1} = i) &\leq \mathbb{P}\left(A_{t+1} = i | \hat{\mathcal{B}}_i = \emptyset, \hat{B}^* = \emptyset\right) + \mathbb{P}\left(\hat{\mathcal{B}}_i \neq \emptyset\right) + \mathbb{P}\left(\hat{B}^* \neq \emptyset\right) \\ &\leq t^2 (F_3 + 2F_4) + F_5 + 2F_6 \\ &= 4t^{-2} + F_5 + 2F_6. \end{aligned}$$

Therefore, $\forall i$, $\forall t \geq \ell$, we have

$$\begin{aligned} \mathbb{P}(A_{t+1} = i) &\leq t^{-2} \left(4 + 2 \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right) + F_5 + 2F_6 \\ &= t^{-2} \left(6s_2 + 4 + 2 \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right) + 4K \exp(-\alpha^2(n_0 + nt)) \\ &\quad + 2 \exp\left(-\frac{n_0 + nt}{18}\right) + 2(s_1 + s_2 + K + 1) \exp\left(-\frac{n_0 + nt - K - 1}{18}\right) \\ &\leq t^{-2} \left(6s_2 + 4 + 2 \exp\left(-\frac{n_0 - s_0 - 2}{16}\right) \right) \\ &\quad + \frac{2(s_1 + s_2 + K + 2)}{\exp((n_0 + nt - K - 1)/18)} + 4K \exp(-\alpha^2(n_0 + nt)). \end{aligned}$$

Then

$$\begin{aligned} \sum_{t \geq \ell} \mathbb{P}(A_{t+1} = i) &\leq \pi^2 \left(s_2 + \frac{2}{3} + \frac{1}{3} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) \\ &\quad + \frac{36(s_1 + s_2 + K + 2)}{n \exp((n_0 + n\ell - K - 1)/18)} + \frac{4K}{n\alpha_2^2 \exp(\alpha_2^2(n_0 + n\ell))}. \end{aligned}$$

Incorporating the condition for ℓ , we have that for $i \in \mathcal{I}_0$,

$$\mathbb{E}[n_i(T)] \leq \frac{20c^2 \tilde{\eta}_i^2 \log T}{(m+n)\Delta_i^2} + \frac{c_3 \log T}{m} + C_5,$$

and for $i \in \mathcal{I}_1$,

$$\mathbb{E}[n_i(T)] \leq \frac{10c^2 \tilde{\sigma}_i^2 \log T}{m\Delta_i^2} + \frac{c_3 \log T}{m} + C_5,$$

where

$$\begin{aligned} C_5 &= \ell_0 + \left(\frac{c_3}{n} \right)^2 + 1 + \pi^2 \left(s_2 + \frac{2}{3} + \frac{1}{3} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) \\ &\quad + \frac{36(s_1 + s_2 + K + 2)}{n \exp((n_0 + n\ell_0 - K - 1)/18)} + \frac{4K}{n\alpha_2^2 \exp(\alpha_2^2(n_0 + n\ell_0))}. \end{aligned}$$

And the final cumulative regret bound is

$$\mathcal{R}_T \leq \frac{20c^2 \log T}{m+n} \sum_{i \in \mathcal{I}_0} \frac{\tilde{\eta}_i^2}{\Delta_i} + \frac{10c^2 \log T}{m} \sum_{i \in \mathcal{I}_1} \frac{\tilde{\sigma}_i^2}{\Delta_i} + \left(\frac{c_3 \log T}{m} + C_5 \right) \left(\sum_i \Delta_i \right).$$

□

3.7.3 Proof of Theorem 3.2

Proof.

$$\mathcal{R}_T \leq 2\ell + 2 \sum_{t=\ell}^{T-1} \mathbb{P}(\mu^* > U_t(A^*)) + \sum_{t=\ell+1}^T \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t}),$$

where

$$\mathbb{P}(\mu^* > U_t(A^*)) \leq \mathbb{P}(\mu^* > U_t(A^*) | S^*(t) \in \mathcal{B}^*) + \mathbb{P}(\text{Incorrect } S^*(t)),$$

$$\mathbb{E}(U_{t-1}(A_t) - \mu_{A_t}) \leq \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t} | S(A_t, t-1) \in \mathcal{B}_{A_t}) + 2\mathbb{P}(\text{Incorrect } S(A_t, t-1)).$$

If $A^* \in \mathcal{I}_0$, then

$$\mathbb{P}(\mu^* > U_t(A^*) | S^*(t) \in \mathcal{B}^*) \leq F_2 \leq t^{-2} \left(1 + \frac{1}{2} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right).$$

If $A^* \in \mathcal{I}_1$, then

$$\mathbb{P}(\mu^* > U_t(A^*) | \widehat{\mathcal{B}}^* = \emptyset) \leq F_4 \leq \frac{3}{2} t^{-2}.$$

Therefore,

$$\begin{aligned} \sum_{t=\ell}^{T-1} \mathbb{P}(\mu^* > U_t(A^*) | S^*(t) \in \mathcal{B}^*) &\leq \sum_{t=\ell}^{T-1} t^{-2} \left(\frac{3}{2} + \frac{1}{2} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) \\ &\leq \frac{\pi^2}{6} \left(\frac{3}{2} + \frac{1}{2} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right). \end{aligned} \tag{3.14}$$

Let $\mathcal{T}_a = \{t : A_t = a, \ell + 1 \leq t \leq T\}$, then

$$\begin{aligned} \sum_{t=\ell+1}^T \mathbb{E}(U_{t-1}(A_t) - \mu_{A_t} | S(A_t, t-1) \in \mathcal{B}_{A_t}) &\leq \sum_{a \in \mathcal{I}_0} \sum_{t \in \mathcal{T}_a} \mathbb{E}(U_{t-1}(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a) \\ &\quad + \sum_{a \in \mathcal{I}_1} \sum_{t \in \mathcal{T}_a} \mathbb{E}(U_{t-1}(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a). \end{aligned}$$

If $a \in \mathcal{I}_0$, then

$$\mathbb{E}(U_{t-1}(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a) \leq c\tilde{\eta} \sqrt{\frac{\log t}{N_o(t) + N_a(t)}}.$$

According to (2.34) and Cauchy-Schwarz inequality,

$$\begin{aligned} &\sum_{a \in \mathcal{I}_0} \sum_{t \in \mathcal{T}_a} \mathbb{E}(U_{t-1}(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a) \\ &\leq \frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \sum_{a \in \mathcal{I}_0} \left(\sqrt{n_0 + (m+n)n_a(T)} - \sqrt{n_0} \right) \\ &\leq \frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \left(\sqrt{|\mathcal{I}_0| \left(\sum_{a \in \mathcal{I}_0} (n+m)n_a(T) \right) + n_0|\mathcal{I}_0|^2} - \sqrt{n_0}|\mathcal{I}_0| \right) \\ &= \frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \left(\sqrt{|\mathcal{I}_0|(n+m)n_{\mathcal{I}_0}(T) + n_0|\mathcal{I}_0|^2} - \sqrt{n_0}|\mathcal{I}_0| \right) \\ &\leq \frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \left(\sqrt{|\mathcal{I}_0|(n+m)T + n_0|\mathcal{I}_0|^2} - \sqrt{n_0}|\mathcal{I}_0| \right). \end{aligned}$$

If $a \in \mathcal{I}_1$, then

$$\mathbb{E}(U_t(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a) \leq c\tilde{\sigma} \sqrt{\frac{\log t}{N_a(t)}},$$

and

$$\begin{aligned}
\sum_{a \in \mathcal{I}_1} \sum_{t \in \mathcal{T}_a} \mathbb{E} (U_{t-1}(a) - \mu_a | S(a, t-1) \in \mathcal{B}_a) &\leq \frac{2c\tilde{\sigma}\sqrt{\log T}}{m} \sum_{a \in \mathcal{I}_1} \left(\sqrt{mn_a(T)} \right) \\
&\leq \frac{2c\tilde{\sigma}\sqrt{\log T}}{m} \sqrt{|\mathcal{I}_1| \left(\sum_{a \in \mathcal{I}_1} mn_a(T) \right)} \\
&= \frac{2c\tilde{\sigma}\sqrt{\log T}}{m} \sqrt{|\mathcal{I}_1| mn_{\mathcal{I}_1}(T)} \\
&\leq \frac{2c\tilde{\sigma}\sqrt{\log T}}{m} \sqrt{|\mathcal{I}_1| mT}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{t=\ell+1}^T \mathbb{E} (U_{t-1}(A_t) - \mu_{A_t} | S(A_t, t-1) \in \mathcal{B}_{A_t}) \\
&\leq \frac{2c\tilde{\eta}\sqrt{\log T}}{n+m} \left(\sqrt{|\mathcal{I}_0|(n+m)T + n_0|\mathcal{I}_0|^2} - \sqrt{n_0|\mathcal{I}_0|} \right) + \frac{2c\tilde{\sigma}\sqrt{\log T}}{m} \sqrt{|\mathcal{I}_1|mT}.
\end{aligned} \tag{3.15}$$

Since $\ell \geq \frac{9 \operatorname{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)}$, $\ell \geq \left(\frac{c_3}{n}\right)^2$, $\ell \geq K \lceil \frac{c_3 \log T + 1}{m} \rceil$ let $\ell_0 = \lceil \frac{9 \operatorname{tr}(\Sigma)}{n\gamma_{\min}(\Sigma)} \rceil$, the final cumulative regret bound is

$$\begin{aligned}
\mathcal{R}_T &\leq 2\ell_0 + 2 \left(\frac{c_3}{n} \right)^2 + \frac{2Kc_3 \log T}{m} + 2K + \text{RHS of } [2 \times (2.32) + 2 \times (3.14) + (3.15)] + \sum_{t \geq \ell_0} 4F_6 \\
&\leq 2c\sqrt{\log T} \left(\frac{\tilde{\eta} \left(\sqrt{|\mathcal{I}_0|(n+m)T + n_0|\mathcal{I}_0|^2} - \sqrt{n_0|\mathcal{I}_0|} \right)}{n+m} + \tilde{\sigma} \sqrt{\frac{|\mathcal{I}_1|T}{m}} \right) + \frac{2Kc_3 \log T}{m} + C_6,
\end{aligned}$$

where

$$\begin{aligned}
C_6 &= 2\ell_0 + 2 \left(\frac{c_3}{n} \right)^2 + 2K + \pi^2 \left(\frac{1}{2} + 2s_2 + \frac{1}{6} \exp \left(-\frac{n_0 - s_0 - 2}{16} \right) \right) \\
&\quad + \frac{8K}{n\alpha^2 \exp(\alpha^2(n_0 + n\ell_0))} + \frac{72(s_1 + s_2 + K + 2)}{n \exp((n_0 + n\ell_0 - K - 1)/18)}.
\end{aligned}$$

□

CHAPTER 4

Adaptive sequential intervention for learning causal DAGs

4.1 Introduction

Structural learning is a central problem in causal inference, whose objective is to determine causal relationships among variables in a system from data [28, 35, 36]. Causal relationships are commonly represented by directed acyclic graphs (DAGs), where vertices correspond to random variables and directed edges represent direct causal effects. When only observational data are available, the underlying DAG is generally not identifiable. In particular, observational distributions determine only the Markov equivalence class (MEC), a collection of graphs that encode the same conditional independence relations. Consequently, observational data identify the causal structure only up to its MEC, and additional information is required to orient the remaining edges. Interventional data, obtained by actively perturbing variables in the system, provide such information and enable further identification of the causal structure.

There has been increasing interest in designing interventions for learning causal graphs. Early work focuses on non-adaptive settings in which the intervention scheme is fixed prior to data collection and algorithm execution [37–40]. Subsequent research considers adaptive procedures that select interventions sequentially based on previously observed outcomes. Many existing adaptive methods [41, 42] assume access to conditional independence oracles under each intervention, which is equivalent to having an infinite number of samples from

each interventional distribution. Under this assumption, statistical uncertainty is ignored and the problem reduces to determining which interventions are sufficient for identification. In practice, however, experiments produce only finite samples, and edge orientations must be inferred statistically rather than deterministically.

Recent work begins to address adaptive causal discovery under finite samples. The algorithm of Greenewald et al. [43] operates in a finite-sample sequential framework but is restricted to tree-structured graphs. The track-and-stop algorithm of Elahi et al. [44] extends adaptive discovery to general graphs and provides sample complexity guarantees, but it requires repeated evaluation over candidate graph structures and becomes computationally expensive even for moderate graph sizes. These limitations motivate the development of methods that accommodate finite samples, apply to general graph structures, and remain computationally tractable.

Throughout this chapter we assume causal sufficiency, meaning that no latent confounders are present. Under this assumption, observational data can be used to recover the Markov equivalence class of the true DAG. A representative example is the PC algorithm [35], a constraint-based procedure that combines conditional independence tests with Meek’s orientation rules [45] to identify all causal directions determined by observational distributions. The resulting completed partially directed acyclic graph (CPDAG) contains directed edges whose orientations are identifiable from observational data and undirected edges whose directions require interventional information.

This chapter studies adaptive intervention design under finite samples. We formulate the problem in an adaptive learning framework where interventions are selected sequentially and information is accumulated over time. The goal is to identify the true DAG within its Markov equivalence class with high confidence while minimizing the number of interventional samples. We propose a computationally efficient greedy strategy that prioritizes interventions expected to orient the largest number of currently undirected edges based on statistical estimates from accumulated data. The resulting Seq-UCB algorithm is shown to consistently

recover the true DAG, and finite-time guarantees on the probability of incorrect identification are established. Numerical experiments on randomly generated graphs demonstrate that the adaptive intervention strategy requires substantially fewer samples than non-adaptive baselines while remaining computationally practical.

4.2 Problem formulation

In a causal graphical model, the causal relationships among a set of random variables $\mathcal{X} = \{X_1, \dots, X_p\}$ is represented by a DAG $\mathcal{G}$ over p nodes $V = \{1, \dots, p\}$, each corresponding to one random variable. A directed edge $i \rightarrow j$ in the causal DAG $\mathcal{G}$ means that i is a direct cause of j , in which case i is called a parent of j and j is called a child of i . The parent set of j , denoted as $\text{pa}(j)$, is the set of nodes that have an edge directed into node j , i.e. $\text{pa}(j) = \{i \in V : i \rightarrow j\}$. Similarly, the child set of j , denoted as $\text{ch}(j)$, is the set of all nodes that have an edge directed from node j . Each child-parent relationship in $\mathcal{G}$ is modeled by a function $X_j = f_j(X_{\text{pa}(j)}, \varepsilon_j)$, $j = 1, \dots, p$, where $\varepsilon_1, \dots, \varepsilon_p$ are exogenous errors.

A partially directed acyclic graph (PDAG) is a graph that contains both directed and undirected edges but does not have directed cycles. Due to Markov equivalent among DAGs, one in general cannot learn the full structure of a DAG even with infinite amount of observational data. The best one can estimate from purely observational data is its equivalence class, represented by a completed partially DAG (CPDAG). In the following sections, we assume the CPDAG $\mathcal{G}_0$ is given, or we have access to infinite observational data. Experimental data need to be collected and utilized in order to distinguish between DAGs in the same equivalence class. The goal is to identify the true DAG $\mathcal{G}$ from its Markov equivalence class.

In this chapter, we focus on Gaussian DAGs whose functional relationships are given by

a linear structural equation model (SEM),

$$X_j = \sum_{i \in \text{pa}(j)} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \dots, p, \quad (4.1)$$

where $\varepsilon_j \stackrel{\text{indep.}}{\sim} \mathcal{N}(0, \omega_j^2)$ with $\omega_j^2 > 0$ for all $j \in V$ and $\beta_{ij} \neq 0$ for all $i \in \text{pa}(j)$.

The causal effect of X_i on X_j is defined via an atomic intervention that sets the variable X_i to a fixed value x_i , denoted as $do(X_i = x_i)$. In general, the causal effect γ_{ij} of X_i on X_j is determined by the mapping $x_i \mapsto \mathbb{P}[X_j | do(X_i = x_i)]$. For Gaussian DAGs,

$$\gamma_{ij} := \frac{\partial}{\partial x_i} \mathbb{E}(Y | do(X_i = x_i)) = \beta_{ij}. \quad (4.2)$$

If the true direction is $i \rightarrow j$ in the true DAG, then $\gamma_{ij} \neq 0$ and $\gamma_{ji} = 0$. The intervention $do(X_i = x_i)$ removes all edges into X_i and consequently blocks all backdoor paths from X_i to X_j . Therefore, experimental data under $do(X_i = x_i)$ can be used directly to estimate γ_{ij} by running the linear regression $X_j \sim X_i$.

4.3 The Seq-UCB algorithm

4.3.1 The main idea

Suppose the CPDAG $\mathcal{G}_0$ is given, with the set of undirected edges

$$\mathcal{E}_0 = \{(i, j) \in E(\mathcal{G}_0) : i - j\}.$$

To identify the true DAG from its Markov equivalence class (MEC), it is sufficient to determine the direction of all undirected edges in $\mathcal{E}_0$. As explained in Section 4.2, the direction of an undirected edge $i - j$ can be identified as $i \rightarrow j$ if and only if the causal effects satisfy $\gamma_{ij} \neq 0$ and $\gamma_{ji} = 0$. Therefore, if sufficient experimental data are available to accurately

estimate both γ_{ij} and γ_{ji} , then true DAG can be recovered from its equivalence class. We consider an adaptive experimental design in which edge directions are determined sequentially, and the next intervention node is selected based on the currently accumulated data and estimates.

At each time step, to determine edge directions, we run linear regressions of X_j on X_i using the accumulated data $\mathcal{D}_i[t]$ collected under the intervention $do(X_i)$, for each $(i, j) \in \mathcal{E}_0$. The design is normalized so that $X_i^\top X_i = N_i(t)$, where $N_i(t)$ denotes the number of experimental data under intervention on node i up to time t . It yields estimates of the causal effects $\hat{\gamma}_{ij}(t)$, and their corresponding standard errors $\hat{se}_{ij}(t)$. By standard properties of least-squares estimators, the following distributions hold:

$$\begin{aligned}\hat{\gamma}_{ij}(t) &\sim \mathcal{N}\left(\gamma_{ij}, \frac{\sigma_{ij}^2}{N_i(t)}\right), \\ \hat{se}_{ij}^2(t) &\sim \frac{\sigma_{ij}^2}{N_i(t)(N_i(t) - 1)} \chi_{N_i(t)-1}^2, \\ \frac{\hat{\gamma}_{ij}(t)}{\hat{se}_{ij}(t)} &\sim \frac{\gamma_{ij}\sqrt{N_i(t)(N_i(t) - 1)}}{\sigma_{ij}\sqrt{\chi_{N_i(t)-1}^2}} + t_{N_i(t)-1}.\end{aligned}\tag{4.3}$$

Based on these distributions, we can conduct inference on the hypotheses $\gamma_{ij} = 0$ and $\gamma_{ji} = 0$ with high confidence to determine the direction of the edge $i - j$. Specifically, let

$$z_{ij}(t) = \frac{|\hat{\gamma}_{ij}(t)|}{\hat{se}_{ij}(t)}.$$

Given a threshold λ_t , we can orient edges as follows: orient $i \rightarrow j$ if $z_{ij}(t) \geq \lambda_t > z_{ji}(t)$; orient $j \rightarrow i$ if $z_{ji}(t) \geq \lambda_t > z_{ij}(t)$; otherwise, the edge $i - j$ remains undirected.

If the direction of an edge $i - j$ is determined, one can apply Meek's rule to orient as many additional edges as possible, updating the PDAG to $\mathcal{G}_t$. Let

$$\mathcal{E}_t = \{(i, j) \in E(\mathcal{G}_t) : i - j\}$$

denote the remaining undirected edges. Define

$$\mathcal{V}_t = \{i : (i, j) \in \mathcal{E}_t \text{ for some } j\}$$

as the set of nodes connected to at least one undirected edge in $\mathcal{G}_t$. Interventions can then be restricted to nodes in $\mathcal{V}_t$, since intervening on nodes outside this set cannot further reduce the remaining ambiguity.

To choose the intervention node at each round, the simplest strategy is to sample uniformly at random from $\mathcal{V}_t$. We refer to this baseline as the sequential random (Seq-Rand) algorithm. Another natural heuristic is to select the node incident to the largest number of undirected edges. However, this approach does not perform well in practice. (Give an example)

To overcome this issue, we propose a sequential upper confidence bound (Seq-UCB) strategy. The main idea is to evaluate, for each node $i \in \mathcal{V}_t$, the number of edges that are likely to become orientable if an intervention on i is performed, taking into account both the estimated causal effects and their associated uncertainty. For each node $i \in \mathcal{V}_t$, let $E_i(t)$ denote the set of edges that can potentially be oriented by intervening on node i at time step t , either directly or through Meek's rule. We next describe how the number of orientable edges $|E_i(t)|$ is computed.

Let $c_t = c\sqrt{\log t}$. Consider an edge $i - j \in \mathcal{E}_t$, which is undirected at time step t by itself. Below are two representative cases in which intervening on node i may resolve the edge direction.

1. $z_{ij}(t), z_{ji}(t) < \lambda_t$: In this case, if node i has a larger causal effect estimate or a larger standard error (equivalently, fewer interventions), then intervening on node i is more likely to push $z_{ij}(t+1)$ above the threshold λ_{t+1} . Specifically, if

$$|\hat{\gamma}_{ij}(t)| + c_t \hat{\text{se}}_{ij}(t) > |\hat{\gamma}_{ji}(t)| + c_t \hat{\text{se}}_{ji}(t),$$

then intervening on node i can help determine the orientation $i \rightarrow j$.

2. $z_{ij}(t), z_{ji}(t) \geq \lambda_t$: In this case, if node i has a smaller causal effect estimate or a larger standard error (equivalently, fewer interventions), then intervening on node i is more likely to invalidate the direction $i \rightarrow j$. Specifically, if

$$|\hat{\gamma}_{ij}(t)| - c_t \hat{s}e_{ij}(t) < |\hat{\gamma}_{ji}(t)| - c_t \hat{s}e_{ji}(t),$$

then intervening on node i can help determine the direction $j \rightarrow i$.

Define the upper and lower confidence bounds

$$\text{UCB}_{ij}(t) = |\hat{\gamma}_{ij}(t)| + c_t \hat{s}e_{ij}(t), \quad \text{LCB}_{ij}(t) = |\hat{\gamma}_{ij}(t)| - c_t \hat{s}e_{ij}(t). \quad (4.4)$$

The analysis above motivates the following counting rule: for an undirected edge $i - j$, if $z_{ij}(t) < \lambda_t$ and $\text{UCB}_{ij}(t) > \text{UCB}_{ji}(t)$, count $i \rightarrow j$ and all resulting edges implied by Meek's rule into $E_i(t)$; if $z_{ij}(t) \geq \lambda_t$ and $\text{LCB}_{ij}(t) < \text{LCB}_{ji}(t)$, count $j \rightarrow i$ and all resulting edges implied by Meek's rule into $E_i(t)$; otherwise, no edge is added into $E_i(t)$.

Once $|E_i(t)|$ has been computed for each node $i \in \mathcal{V}_t$, we select the node that can orient the largest number of undirected edges as the intervention node for the next round. This constitutes the greedy component of our proposed algorithm. To balance exploitation and exploration, we introduce a small probability δ_t , decreasing with t , under which the intervention node is sampled uniformly at random from $\mathcal{V}_0$.

4.3.2 Algorithm description

Based on the main idea introduced in Section 4.3.1, we propose the Seq-UCB algorithm, summarized in Algorithm 4.1. The algorithm proceeds iteratively by (i) orienting edges to update the current PDAG, (ii) selecting an intervention node that is expected to orient the largest number of currently undirected edges, and (iii) generating experimental data under

the selected intervention and updating causal effect estimates.

Algorithm 4.1 Causal graph learning via sequential experiments (Seq-UCB)

Input: CPDAG $\mathcal{G}_0$, parameters λ_t, δ_t, c .

- 1: Initialize $\{(\hat{\gamma}_{ij}(1), \hat{s}_{ij}(1)) : (i, j) \in \mathcal{E}_0\}$.
- 2: **for** $t = 1, \dots, T$ **do**
- 3: Orient edges in $\mathcal{E}_0$ based on $\{(\hat{\gamma}_{ij}(t), \hat{s}_{ij}(t))\}$ and update PDAG to $\mathcal{G}_t$.
- 4: For all $(i, j) \in \mathcal{E}_t$, calculate $\text{UCB}_{ij}(t)$ and $\text{LCB}_{ij}(t)$.
- 5: For all $i \in \mathcal{V}_t$, find orientable edges $E_i(t)$.
- 6: Update $\tilde{\mathcal{C}}_t = \text{argmax}_{i \in \mathcal{V}_t} |E_i(t)|$.
- 7: With probability $|\mathcal{V}_0|\delta_t$, sample $a_t \in \mathcal{V}_0$; with probability $1 - |\mathcal{V}_0|\delta_t$, sample $a_t \in \tilde{\mathcal{C}}_t$.
- 8: Generate experimental data under $do(X_{a_t})$ and add to $\mathcal{D}_{a_t}[t]$.
- 9: For $j \in \mathcal{N}_0(a_t)$, regress $X_j \sim X_{a_t}$ using $\mathcal{D}_{a_t}[t]$, update $\hat{\gamma}_{a_t j}(t+1), \hat{s}_{a_t j}(t+1)$.
- 10: For $i \neq a_t$ and $j \in \mathcal{N}_0(i)$, $\hat{\gamma}_{ij}(t+1) \leftarrow \hat{\gamma}_{ij}(t), \hat{s}_{ij}(t+1) \leftarrow \hat{s}_{ij}(t)$.

Output: Graph $\mathcal{G}_T$.

To update the PDAG at round t (Line 3 of Algorithm 4.1), the algorithm computes $z_{ij}(t) = |\hat{\gamma}_{ij}(t)/\hat{s}_{ij}(t)|$ for all undirected edges $(i, j) \in \mathcal{E}_0$. All ordered pairs are then ranked in descending order of $z_{ij}(t)$, and the algorithm iterates through this list. For each pair, if

$$z_{ij}(t) \geq \lambda_t > z_{ji}(t)$$

and orienting the edge as $i \rightarrow j$ does not contradict with any previously oriented edges, the edge is oriented accordingly. After each successful orientation, Meek's rules are applied to infer as many additional edge directions as possible. After a full pass, the resulting graph is the updated PDAG $\mathcal{G}_t$.

Given $\mathcal{G}_t$, the algorithm evaluates the informativeness of intervening on each node $i \in \mathcal{V}_t$, where $\mathcal{V}_t$ denotes the set of nodes connected to undirected edges in $\mathcal{G}_t$. The upper and lower confidence bounds defined in (4.4) are computed for all $(i, j) \in \mathcal{E}_t$ (Line 4 of Algorithm 4.1) and are used to count the number of orientable edges associated with intervening on node i (Line 5 of Algorithm 4.1). Specifically,

- If $z_{ij}(t) < \lambda_t$ and $\text{UCB}_{ij}(t) > \text{UCB}_{ji}(t)$, count $i \rightarrow j$ and additional edges by Meek's

rule to $E_i(t)$;

- If $z_{ij}(t) \geq \lambda_t$ and $\text{LCB}_{ij}(t) < \text{LCB}_{ji}(t)$, count $j \rightarrow i$ and additional edges by Meek's rule to $E_i(t)$.

In Line 7 of Algorithm 4.1, the algorithm selects the intervention node by balancing exploitation and exploration, controlled by the parameter δ_t . The exploration parameter δ_t is chosen as a decreasing sequence in t . At early stages, δ_t is relatively large, so the algorithm behaves nearly randomly and accumulates sufficient experimental data to obtain reliable estimates of causal effects. As t increases, δ_t becomes small, and the algorithm increasingly selects the node with the largest number of orientable edges. When all edges have been directed, i.e., $\mathcal{E}_t = \emptyset$ and $\tilde{\mathcal{C}}_t = \emptyset$, the algorithm may perform no intervention at certain rounds, and consequently no experimental data are generated. This behavior highlights the sample efficiency of the proposed algorithm. The choice of the tuning parameters λ_t and δ_t is discussed in Remark 4.1.

In practice, an early stopping rule can be incorporated into Algorithm 4.1. If the updated graph $\mathcal{G}_t$ contains no undirected edges and remains unchanged for a fixed number of consecutive rounds (e.g., 50), it provides empirical evidence that the true DAG has been identified, and the algorithm can be terminated.

4.4 Theoretical analysis of Seq-UCB

Our theoretical analysis focuses on linear Gaussian structural equation models. Let $\mathcal{G}$ denote the true DAG, and define its directed edge set as $\mathcal{E} = \{(i, j) : i \rightarrow j \in E(\mathcal{G})\}$. To ensure identifiability of edge directions from experimental data, we need the following faithfulness assumption.

Assumption 4.1. *The random vector $X = (X_1, \dots, X_p)$ satisfies the linear SEM (4.1) associated with a DAG $\mathcal{G}$.*

Assumption 4.1 guarantees that each directed edge in the true DAG induces a nonzero causal effect in the correct direction and no causal effect in the reverse direction, which allows edge orientations to be identified from interventional estimates. Under an intervention $do(X_i = x)$, the interventional distribution of X_j is given by

$$X_j \mid do(X_i = x) \sim \mathcal{N}(\gamma_{ij}x, \sigma_{ij}^2). \quad (4.5)$$

Define $N_i(t) = \sum_{u=1}^t \mathbb{1}\{A_u = i\}$ as the number of times node i is selected for intervention up to time t .

Lemma 4.1. *Let $u_t > 0$ be a deterministic sequence. Under Algorithm 4.1, the number of interventions on each node $i \in \mathcal{V}_0$ satisfies*

$$\mathbb{P}(N_i(t) < u_t) \leq \exp\left(-\frac{(\sum_{u=1}^t \delta_u - u_t)^2}{2t}\right), \quad \forall i \in \mathcal{V}_0. \quad (4.6)$$

The proof of Lemma 4.1 is provided in Section 4.7.1. By choosing the exploration sequence δ_t appropriately, the bound in (4.6) ensures that every node $i \in \mathcal{V}_0$ is intervened upon at least u_t times by round t with high probability.

Define

$$u_i(t) = \left\lceil \max\left\{48 \log t + 1, \max_{j: \gamma_{ij} \neq 0} \frac{10\sigma_{ij}^2 \lambda_t^2}{\gamma_{ij}^2}\right\} \right\rceil.$$

The quantity $u_i(t)$ represents a minimum number of interventions on node i required to reliably determine the orientations of its connected undirected edges.

The following lemma shows that, once this minimum intervention requirement is satisfied for all nodes in $\mathcal{V}_0$, the probability that the updated PDAG $\mathcal{G}_t$ differs from the true DAG $\mathcal{G}$ decays at the order of $O(t^{-3})$ at most. The proof of Lemma 4.2 is provided in Section 4.7.2.

Lemma 4.2. *Suppose that for all nodes $i \in \mathcal{V}_0$, $t > 0$, $N_i(t) \geq u_i(t)$ with $\lambda_t \geq 2\sqrt{3 \log t}$.*

Then under Assumption 4.1,

$$\mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) \leq 7|\mathcal{E}_0|t^{-3}. \quad (4.7)$$

Define $u_t = \max_{i \in \mathcal{V}_0} u_i(t)$. With appropriate choices of the tuning parameters δ_t and λ_t , the following theorem establishes that Algorithm 4.1 consistently recovers the true DAG. The proof of this result is provided in Section 4.7.3.

Theorem 4.1. *Suppose $\sum_{u=1}^t \delta_u - u_t$ has a higher order in t than $\sqrt{t \log t}$, with $\lambda_t \geq 2\sqrt{3 \log t}$. Then under Assumption 4.1, Algorithm 4.1 satisfies*

$$\sum_{t=1}^{\infty} \mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) < \infty,$$

and hence

$$\mathbb{P}\left(\lim_{t \rightarrow \infty} \mathcal{G}_t = \mathcal{G}\right) = 1.$$

Remark 4.1 (Choice of tuning parameters). Graph recovery consistency requires the threshold to satisfy $\lambda_t \geq 2\sqrt{3 \log t}$. Under this condition, the minimum intervention requirement u_t is dominated by the order of λ_t^2 . In addition, the exploration parameter δ_t must be chosen such that

$$\sum_{u=1}^t \delta_u - u_t \geq 2\sqrt{t \log t},$$

which ensures that each node receives a sufficient number of interventions with high probability. These conditions imply that λ_t can grow at any rate between $O(\sqrt{\log t})$ and $O(\sqrt{t})$, while $\sum_{u=1}^t \delta_u$ must lie between $O(\max\{\lambda_t^2, \sqrt{t \log t}\})$ and $O(t)$. A convenient parameterization that satisfies all theoretical requirements is to set

$$\lambda_t = O(t^\alpha), \quad \delta_t = O(t^{2\alpha-1}),$$

for some $\alpha \in (1/4, 1/2)$. This choice balances statistical accuracy in edge orientation with

sufficient exploration of intervention nodes.

In practice, the parameter α controls the trade-off between early exploration and aggressive exploitation. Larger values of α lead to more conservative edge orientation thresholds and increased random exploration in early rounds, while smaller values favor faster exploitation. We investigate the empirical impact of different choices of α on the performance of Algorithm 4.1 in Section 4.5.2.

4.5 Numerical experiments

4.5.1 Simulation settings

We compare the performance of three algorithms that differ only in how the intervention node A_t is selected at each time step. For all three methods, the procedure for orienting edges and updating the PDAG $\mathcal{G}_t$ is identical; in particular, the same significance-based orientation rule is used at every round. The purpose of the numerical experiments is to evaluate whether the proposed Seq-UCB algorithm can select intervention nodes more effectively and thereby recover the true DAG with fewer interventions.

The three algorithms considered are as follows:

- Completely random (Baseline 1): At each time step t , the intervention node a_t is sampled uniformly at random from the initial candidate set $\mathcal{V}_0$.
- Sequentially random (Seq-rand, Baseline 2): At each times step t , with probability $|\mathcal{V}_0|\delta_t$, the intervention node a_t is sampled uniformly from $\mathcal{V}_0$; with probability $1 - |\mathcal{V}_0|\delta_t$, it is sampled uniformly from the current candidate set $\mathcal{V}_t$.
- Sequential UCB (Seq-UCB): At each times step t , with probability $|\mathcal{V}_0|\delta_t$, the intervention node a_t is sampled uniformly from $\mathcal{V}_0$; with probability $1 - |\mathcal{V}_0|\delta_t$, it is sampled uniformly from $\tilde{\mathcal{C}}_t$, the set of nodes maximizing the number of orientable edges.

Following the discussion in Remark 4.1, we choose the tuning parameters as

$$\lambda_t = 0.5t^\alpha, \quad \delta_t = \frac{11\alpha}{20}t^{2\alpha-1},$$

so that $\sum_{u=1}^t \delta_u \approx 1.1t^{2\alpha}$, with $\alpha \in \{0.25, 0.3, 0.35\}$. For the Seq-UCB algorithm, an additional parameter $c = 1/\sqrt{2}$ is used in the construction of the confidence bounds. To improve stability in finite-sample settings, we impose a minimum intervention requirement of $2\sqrt{\log t}$ for all three algorithms. Although this condition is not required by the theoretical results, it helps ensure more reliable performance in practice by preventing insufficient sampling at early stages.

For all three algorithms, at each round we collect two experimental samples under $\text{do}(X_{a_t})$. The intervention value is drawn uniformly from $\{-1, 1\}$ for simplicity. The structural coefficients in this section are generated according to the linear SEM model in (4.1), where the edge weights are sampled uniformly from $\beta_{ij} \in [-1, -0.5] \cup [0.5, 1]$ and the noise variances are sampled from $\omega_{ij}^2 \in [0.5, 1]$.

4.5.2 True DAG identification rate comparisons

In this subsection, we compare the true DAG identification rates of the three algorithms described in Section 4.5.1. We generate 100 random DAGs with 20 nodes, maximum indegree equal to 2, and between 10 and 15 initially undirected edges. For each simulated graph, we run each algorithm independently 5 times for 1000 time steps. We consider three choices of the tuning parameter $\alpha \in \{0.25, 0.30, 0.35\}$, which determine the growth rate of λ_t , in order to examine how the threshold sequence affects empirical performance. At each time step, we compute the true DAG identification rate averaged over the five runs, and then further average this quantity across the 100 simulated graphs. The results are shown in Figure 4.1.

As illustrated in Figure 4.1, when λ_t grows relatively slowly (i.e., smaller α), the true DAG identification rates of Seq-UCB and Seq-Rand are very similar over time. This behavior

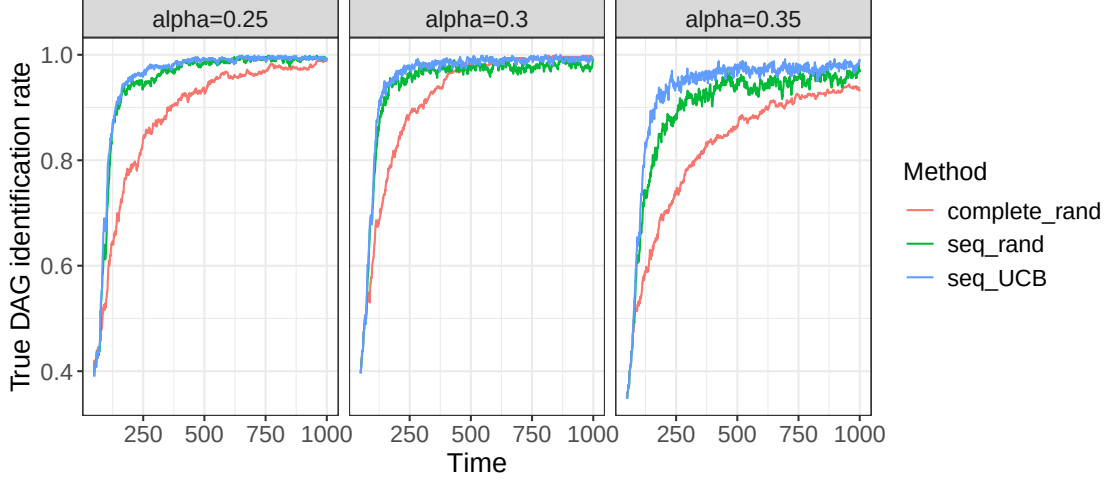


Figure 4.1: Average true DAG identification rate over time on 100 randomly simulated DAGs with $p = 20$, maximum in-degree 2, and between 10 and 15 undirected edges in the initial CPDAG.

can be attributed to the fact that, after a certain number of rounds, the number of undirected edges is typically small when λ_t is small. In such cases, the UCB-based intervention criterion provides limited advantage, since most edges have already been oriented and the graph has a fixed maximum degree. In contrast, when λ_t grows faster (i.e., larger α), the performance gap between Seq-UCB and Seq-Rand becomes more evident. Seq-UCB consistently achieves higher identification rates by selecting intervention nodes that are more informative in orienting the remaining undirected edges. In all settings, the completely random algorithm generates interventional samples at every round, whereas Seq-UCB and Seq-Rand require substantially fewer samples. As shown in Figure 4.2, Seq-UCB uses the smallest number of interventional samples while achieving higher true DAG identification rates.

Moreover, we compute the average number of wrongly oriented edges averaged over all runs and display the results in Figure 4.3. As illustrated in Figure 4.3, with small λ_t and δ_t , the average number of wrongly oriented edges decreases slower over time than with moderate λ_t and δ_t . This suggests that, when avoiding false positives is a primary concern, choosing sequences λ_t and δ_t with a relatively higher growth rate may be preferable in practice.

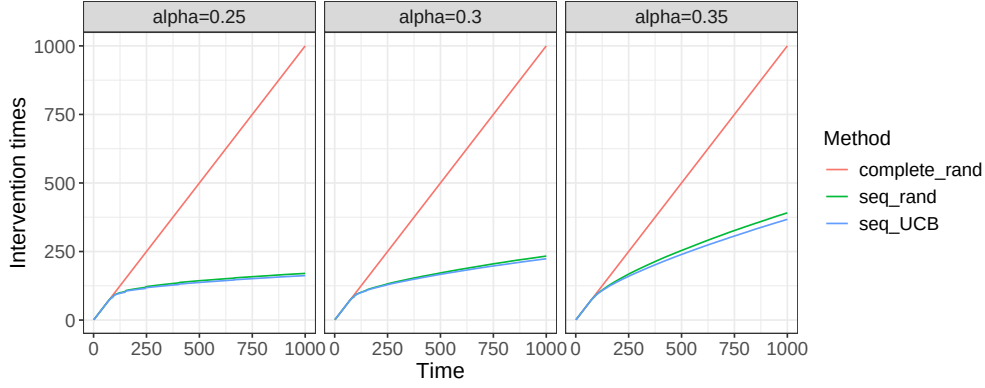


Figure 4.2: Average times of interventions over time across all runs on 100 randomly simulated DAGs with $p = 20$.

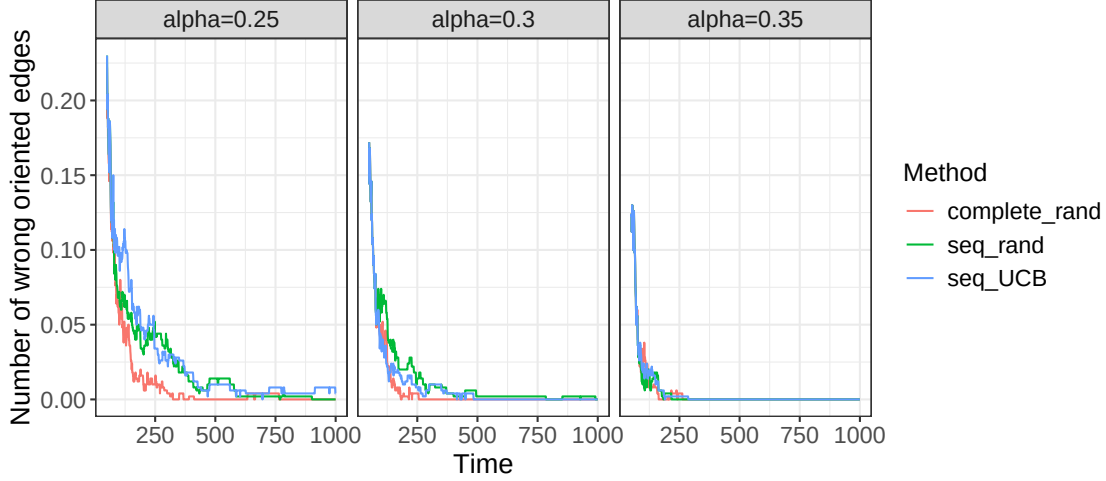


Figure 4.3: Average number of wrongly oriented edges over time on 100 randomly simulated DAGs with $p = 20$, maximum in-degree 2, and between 10 and 15 undirected edges in the initial CPDAG after 50 rounds.

4.5.3 Early stopping and intervention time comparisons

To illustrate differences in intervention-selection behavior, we compare Seq-Rand and Seq-UCB on several structured graphs with 15 nodes. We adopt an early stopping criterion under which the algorithm terminates once the estimated graph $\mathcal{G}_t$ is fully oriented and remains unchanged for 50 consecutive rounds (excluding rounds in which $\mathcal{G}_t$ contains undirected edges). The tuning parameters are set to $\lambda_t = t^{0.3}/2$ and $\delta_t = 0.165t^{-0.4}$, i.e., $\alpha = 0.3$.

For each graph type, we generate 30 independent sets of coefficients and run each algorithm 5 times per coefficient set. The reported intervention times are averaged over all

runs. The results are shown in Figure 4.4. As illustrated in the figure, Seq-UCB consistently requires fewer total interventions than Seq-Rand to recover the true DAG under the early stopping criterion. This advantage is particularly evident for the star-shaped graph, where node degrees are highly heterogeneous. In such settings, Seq-UCB more effectively prioritizes interventions on high-impact nodes, leading to faster orientation of undirected edges.

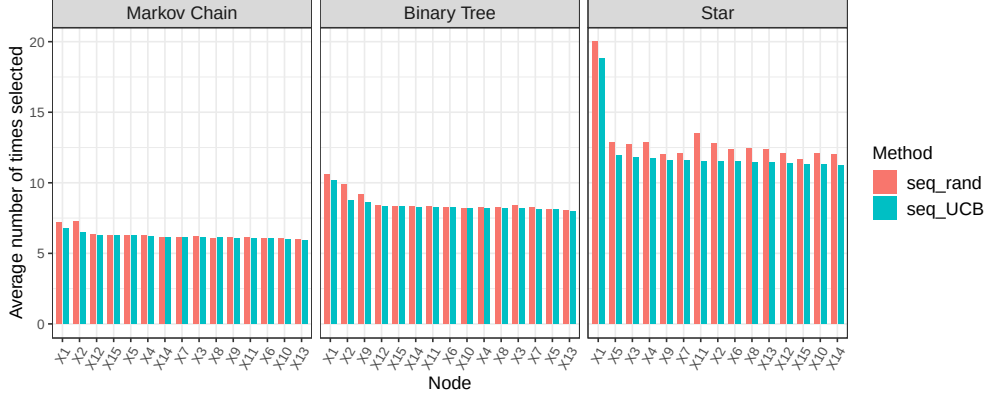


Figure 4.4: Average number of times each node is selected for intervention across three graph structures.

4.6 Discussion

This chapter develops a finite-sample adaptive framework for causal DAG structure learning, where interventions are selected sequentially and edge orientations are inferred statistically rather than deterministically through conditional independence oracles. The key idea is to update the current graph using two complementary sources of information. First, the method performs uncertainty-aware inference on edge directions using estimated causal effects together with their standard errors. Second, it propagates newly oriented edges through Meek’s rules to accelerate progress toward full identification. These components together yield the Seq-UCB algorithm, which prioritizes resolving the remaining undirected edges in the CPDAG as quickly as possible while remaining computationally practical under finite experimental samples.

Theoretical guarantees are established for linear structural equation models with Gaus-

sian noise. It shows that, with appropriate tuning of parameters, every candidate node receives sufficiently many interventions with high probability. Once this minimum intervention requirement is satisfied, the probability of incorrect graph identification decays at a rate no worse than $O(t^{-3})$, implying almost sure convergence to the true DAG. Simulation results demonstrate clear empirical gains from adaptive intervention selection. Seq-UCB achieves higher true-DAG identification rates with fewer experimental samples than sequential random baselines, especially for graphs with heterogeneous degree patterns, such as star-shaped graphs, where certain interventions are substantially more informative than others.

4.7 Appendix

4.7.1 Proof of Lemma 4.1

Proof. Define $\mathcal{F}_t = \sigma(A_1, \dots, A_t)$, which is the σ -algebra generated by the random variables $A_1, \dots, A_t$, representing all information available up to time t , i.e., $\{\mathcal{F}_t\}_{t \geq 1}$ denote the natural filtration generated by the action process $\{A_t\}$, where $\mathcal{F}_t$ contains all events measurable with respect to the history $(A_1, \dots, A_t)$.

Then $\forall i \in \mathcal{V}_0$, we have

$$\mathbb{P}(A_{t+1} = i | \mathcal{F}_t) \geq \delta_t, \forall t \geq 1.$$

Define

$$\begin{aligned} Z_u &= \mathbb{1} \{A_u = i\} \in \{0, 1\}, \\ N_i(t) &= \sum_{u=1}^t Z_u, \\ p_u &= \mathbb{E} [Z_u \mid \mathcal{F}_{u-1}] = \mathbb{P}(A_u = i \mid \mathcal{F}_{u-1}) \geq \delta_t, \\ M_t &= \sum_{u=1}^t (p_u - Z_u), \end{aligned}$$

then $\mathbb{E} [p_u - Z_u \mid \mathcal{F}_{u-1}] = 0$, so (M_t) is a martingale, and each increment satisfies $|(p_u - Z_u)| \leq 1$.

1. Now on the event $\{N_i(t) < u_t\}$,

$$M_t = \sum_{u=1}^t p_u - \sum_{u=1}^t Z_u \geq \sum_{u=1}^t \delta_u - u_t.$$

So

$$\{N_i(t) < u_t\} \subseteq \left\{ M_t \geq \sum_{u=1}^t \delta_u - u_t \right\}.$$

Then according to Azuma-Hoeffding, when $u_i(t) \leq \sum_{u=1}^t \delta_u$, we have

$$\mathbb{P}(N_i(t) < u_t) \leq \mathbb{P}(M_t \geq \sum_{u=1}^t \delta_u - u_t) \leq \exp \left(-\frac{(\sum_{u=1}^t \delta_u - u_t)^2}{2t} \right).$$

□

4.7.2 Proof of Lemma 4.2

Proof. According to (4.3),

$$\frac{\hat{\gamma}_{ij}(t)}{\hat{se}_{ij}(t)} \sim \frac{\gamma_{ij} \sqrt{N_i(t)(N_i(t) - 1)}}{\sigma_{ij} \sqrt{\chi_{N_i(t)-1}^2}} + t_{N_i(t)-1}.$$

Specially, if $\gamma_{ji} = 0$, then

$$\frac{\hat{\gamma}_{ji}(t)}{\hat{se}_{ji}(t)} \sim t_{N_j(t)-1}.$$

Our proof is based on the following tail properties of normal, t and χ^2 distributions:

$$\begin{aligned} \mathbb{P}(|\mathcal{N}(0, \sigma^2)| \geq \delta) &\leq 2e^{-\delta^2/(2\sigma^2)}, \\ \mathbb{P}(|t_r| \geq \delta) &\leq 2e^{-\delta^2/4} + e^{-r/16}, \\ \mathbb{P}(\chi_r^2 \geq 5r/2) &\leq e^{-r/4}. \end{aligned} \tag{4.8}$$

If $|t_{N_i(t)-1}| \leq \lambda_t$, we have

$$\frac{|\gamma_{ij}|\sqrt{N_i(t)(N_i(t)-1)}}{\sigma_{ij}\sqrt{\chi_{N_i(t)-1}^2}} - \lambda_t \leq z_{ij}(t) \leq \frac{|\gamma_{ij}|\sqrt{N_i(t)(N_i(t)-1)}}{\sigma_{ij}\sqrt{\chi_{N_i(t)-1}^2}} + \lambda_t.$$

Then $\forall (i, j) \in \mathcal{E}_0$ with $\gamma_{ij} \neq 0$, $\forall t$,

$$\begin{aligned} \mathbb{P}(z_{ij}(t) < \lambda_t) &\leq \mathbb{P}\left(\frac{|\gamma_{ij}|\sqrt{N_i(t)(N_i(t)-1)}}{\sigma_{ij}\sqrt{\chi_{N_i(t)-1}^2}} - \lambda_t < \lambda_t\right) + \mathbb{P}(|t_{N_i(t)-1}| \geq \lambda_t) \\ &\leq \mathbb{P}\left(\chi_{N_i(t)-1}^2 > \frac{\gamma_{ij}^2 N_i(t)(N_i(t)-1)}{4\sigma_{ij}^2 \lambda_t^2}\right) + 2\exp(-\lambda_t^2/4) \\ &\quad + \exp(-(N_i(t)-1)/16), \end{aligned} \tag{4.9}$$

$$\mathbb{P}(z_{ji}(t) \geq \lambda_t) = \mathbb{P}(|t_{N_j(t)-1}| \geq \lambda_t) \leq 2\exp(-\lambda_t^2/4) + \exp(-(N_j(t)-1)/16).$$

Since $N_i(t) \geq \frac{10\sigma_{ij}^2 \lambda_t^2}{\gamma_{ij}^2}$, i.e., $\frac{\gamma_{ij}^2 N_i(t)(N_i(t)-1)}{4\sigma_{ij}^2 \lambda_t^2} \geq \frac{5(N_i(t)-1)}{2}$, then

$$\mathbb{P}\left(\chi_{N_i(t)-1}^2 > \frac{\gamma_{ij}^2 N_i(t)(N_i(t)-1)}{4\sigma_{ij}^2 \lambda_t^2}\right) \leq \exp(-(N_i(t)-1)/4).$$

Since $N_i(t) \geq 48 \log t + 1$, $\forall i \in \mathcal{V}_0$ and $\lambda_t \geq 2\sqrt{3 \log t}$, then

$$\begin{aligned} \mathbb{P}(z_{ij}(t) < \lambda_t) + \mathbb{P}(z_{ji}(t) \geq \lambda_t) &\leq \exp(-(N_i(t) - 1)/4) + \exp(-(N_i(t) - 1)/16) \\ &\quad + \exp(-(N_j(t) - 1)/16) + 4 \exp(-\lambda_t^2/4) \\ &\leq 7t^{-3}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) &\leq \mathbb{P}(z_{ij}(t) \geq \lambda_t > z_{ji}(t) \text{ does not hold for some } (i, j) \in \mathcal{E}_0 \text{ with } \gamma_{ij} \neq 0) \\ &\leq \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} [\mathbb{P}(z_{ij}(t) < \lambda_t) + \mathbb{P}(z_{ji}(t) \geq \lambda_t)] \\ &\leq 7|\mathcal{E}_0|t^{-3}. \end{aligned}$$

□

4.7.3 Proof of Theorem 4.1

Proof. Since $\sum_{u=1}^t \delta_u - u_t$ has a higher order in t than $\sqrt{t \log t}$, after some $t \geq \ell_0$,

$$\sum_{u=1}^t \delta_u - u_i(t) \geq 2\sqrt{t \log t}, \quad \forall i \in \mathcal{V}_0,$$

where ℓ_0 is a constant determined by γ_{ij} , σ_{ij} . Then according to Lemma 4.1, $\forall (i, j) \in \mathcal{E}_0$,

$$\mathbb{P}(N_i(t) < u_i(t)) \leq \exp\left(-\frac{(\sum_{u=1}^t \delta_u - u_i(t))^2}{2t}\right) \leq t^{-2}.$$

Therefore, for any $t \geq \ell_0$,

$$\begin{aligned}
\mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) &\leq \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \mathbb{P}(z_{ij}(t) < \lambda_t) + \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \mathbb{P}(z_{ji}(t) \geq \lambda_t) \\
&\leq \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \sum_{u=0}^{t-1} \mathbb{P}(z_{ij}(t) < \lambda_t, N_i(t) = u) \\
&\quad + \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \sum_{u=0}^{t-1} \mathbb{P}(z_{ji}(t) \geq \lambda_t, N_j(t) = u) \\
&\leq \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \left(\sum_{u=u_i(t)}^{t-1} \mathbb{P}(z_{ij}(t) < \lambda_t, N_i(t) = u) + \mathbb{P}(N_i(t) < u_i(t)) \right) \\
&\quad + \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \left(\sum_{u=u_j(t)}^{t-1} \mathbb{P}(z_{ji}(t) < \lambda_t, N_j(t) = u) + \mathbb{P}(N_j(t) < u_j(t)) \right) \\
&\leq 7|\mathcal{E}_0|t^{-2} + \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \mathbb{P}(N_i(t) < u_i(t)) + \sum_{(i,j) \in \mathcal{E}_0: \gamma_{ij} \neq 0} \mathbb{P}(N_j(t) < u_j(t)) \\
&\leq 9|\mathcal{E}_0|t^{-2}.
\end{aligned}$$

Then

$$\sum_{t=1}^{\infty} \mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) \leq \ell_0 + \sum_{t=\ell_0+1}^{\infty} \mathbb{P}(\mathcal{G}_t \neq \mathcal{G}) \leq \ell_0 + 9|\mathcal{E}_0| \sum_{t=1}^{\infty} t^{-2} = \ell_0 + \frac{3\pi^2|\mathcal{E}_0|}{2}.$$

By Borel-Cantelli lemma,

$$\mathbb{P}\left(\lim_{t \rightarrow \infty} \mathcal{G}_t = \mathcal{G}\right) = 1.$$

□

CHAPTER 5

Summary and future directions

This dissertation studies sequential learning in causal graphical models under finite samples collected at each round. Two related learning objectives are investigated: identifying the intervention with the largest causal effect on a reward variable and learning the underlying causal graph structure. In both settings, the main challenge is how to allocate limited experimental interventions while exploiting information available from observational data and structural constraints.

Chapter 2 considers the problem of identifying the variable with the largest causal effect on the reward when the causal DAG is unknown. The proposed BA-UCB algorithm uses adjustment sets to combine observational and experimental samples for estimating causal effects. Through valid backdoor adjustment, observational data can simultaneously inform the estimation of causal effects across multiple arms. The algorithm constructs confidence bounds based on weighted estimates from both data sources, and the accumulation of observational data progressively reduces uncertainty across arms. As a result, the dependence of the cumulative regret on the number of arms is reduced once a sufficiently large amount of initial observational data is available.

Chapter 3 extends the BA-UCB algorithm to settings with latent confounders, where some causal effects may not be identifiable from observational data. The algorithm performs an identifiability check for each arm and adapts the estimates accordingly. Observational data are incorporated in constructing UCBs only when a valid adjustment set exists; otherwise, the estimates rely solely on experimental samples. The resulting cumulative regret

bound shows that, given a sufficiently large amount of initial observational data, performance depends on the number of non-identifiable interventions rather than the total number of arms. Therefore, as long as adjustment is valid for a nontrivial subset of interventions, observational data can still improve the efficiency of sequential learning through information sharing across identifiable arms.

Chapter 4 studies adaptive learning of the causal structure. The Seq-UCB algorithm selects intervention nodes that are most informative for orienting the remaining undirected edges under finite samples within a given CPDAG. The procedure combines statistical uncertainty in estimated causal effects with graphical propagation rules, so that each successful orientation can trigger additional orientations through Meek’s rules. Theoretical guarantees establish almost sure convergence to the true DAG, and simulations demonstrate that Seq-UCB effectively reduces the number of experimental samples required for structure recovery compared with both non-adaptive and randomly adaptive baselines.

Several directions naturally follow from our work.

First, for both learning objectives, the current theoretical analysis relies on linear structural equation models with Gaussian noise, which enables explicit estimators and concentration bounds. Extending the algorithms and analysis to broader settings is therefore important. One direction is to establish theoretical guarantees under more general error distributions, such as sub-Gaussian noise, by replacing Gaussian tail bounds with appropriate concentration inequalities. Another direction is to handle nonlinear structural equations by employing alternative causal effect estimators and deriving corresponding confidence bounds for sequential learning.

Second, the Seq-UCB algorithm assumes that the CPDAG is known or accurately estimated from observational data. In practice, observational structure learning is itself uncertain and may introduce orientation errors or extra edges that significantly affect intervention selection in our setting. Incorporating CPDAG estimation uncertainty into adaptive intervention design is nontrivial and represents an important direction for future work.

Third, this dissertation studies reward optimization and structure learning separately because the two tasks have different primary objectives. However, the two objectives are closely related. Better structural knowledge can improve the efficiency of identifying the best intervention, and data collected for reward optimization can also inform structure learning. An interesting research direction is to design adaptive policies that allocate experiments between reward estimation and structure identification according to a pre-specified trade-off between the two tasks, and to analyze how improved structural knowledge facilitates identification of the largest reward and reduces long-term regret.

Bibliography

- [1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47 (2002), pp. 235–256.
- [2] T.L Lai and Herbert Robbins. “Asymptotically efficient adaptive allocation rules”. In: *Advances in Applied Mathematics* 6.1 (1985), pp. 4–22. ISSN: 0196-8858.
- [3] Donald A Berry and Bert Fristedt. “Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)”. In: *London: Chapman and Hall* 5.71-87 (1985), pp. 7–7.
- [4] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. “Regret analysis of stochastic and non-stochastic multi-armed bandit problems”. In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122.
- [5] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [6] Lihong Li et al. “A contextual-bandit approach to personalized news article recommendation”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 661–670.
- [7] JA Bather. “On the allocation of treatments in sequential medical trials”. In: *International Statistical Review/Revue Internationale de Statistique* (1985), pp. 1–13.
- [8] Sofia S Villar, Jack Bowden, and James Wason. “Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges”. In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 30.2 (2015), p. 199.
- [9] Omar Besbes and Assaf Zeevi. “Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms”. In: *Operations research* 57.6 (2009), pp. 1407–1420.

- [10] William R Thompson. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. In: *Biometrika* 25.3-4 (1933), pp. 285–294.
- [11] John Langford and Tong Zhang. “The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information”. In: *Advances in Neural Information Processing Systems*. Vol. 20. 2007.
- [12] Nicolo Cesa-Bianchi and Gábor Lugosi. “Combinatorial bandits”. In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1404–1422.
- [13] Long Tran-Thanh et al. “Epsilon-first policies for budget-limited multi-armed bandits”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 24. 2010, pp. 1211–1216.
- [14] Finnian Lattimore, Tor Lattimore, and Mark D Reid. “Causal bandits: Learning good interventions via causal inference”. In: *Advances in neural information processing systems* 29 (2016).
- [15] Rajat Sen et al. “Identifying best interventions through online importance sampling”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3057–3066.
- [16] Akihiro Yabe et al. “Causal Bandits with Propagating Inference”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, 2018, pp. 5512–5520.
- [17] Vineet Nair, Vishakha Patil, and Gaurav Sinha. “Budgeted and non-budgeted causal bandits”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2017–2025.
- [18] Yangyi Lu et al. “Regret analysis of bandit problems with causal background knowledge”. In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 141–150.

- [19] Burak Varici et al. “Causal bandits for linear structural equation models”. In: *Journal of Machine Learning Research* 24.297 (2023), pp. 1–59.
- [20] Zirui Yan and Ali Tajer. “Linear causal bandits: Unknown graph and soft interventions”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 23939–23987.
- [21] Sanghack Lee and Elias Bareinboim. “Structural causal bandits: Where to intervene?”. In: *Advances in neural information processing systems* 31 (2018).
- [22] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. “Causal bandits with unknown graph structure”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24817–24828.
- [23] Mikhail Konobeev, Jalal Etesami, and Negar Kiyavash. “Causal Bandits without Graph Learning”. In: *Proceedings of the Fourth Conference on Causal Learning and Reasoning*. Vol. 275. PMLR, 2025, pp. 31–63.
- [24] Arnoud De Kroon, Joris Mooij, and Danielle Belgrave. “Causal bandits without prior knowledge using separating sets”. In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 407–427.
- [25] Jireh Huang and Qing Zhou. “Bayesian Causal Bandits with Backdoor Adjustment Prior”. In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856.
- [26] Zirui Yan and Ali Tajer. “Linear Causal Bandits: Unknown Graph and Soft Interventions”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 23939–23987.
- [27] Daniel Russo and Benjamin Van Roy. “Learning to optimize via posterior sampling”. In: *Mathematics of Operations Research* 39.4 (2014), pp. 1221–1243.
- [28] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

- [29] Peter Spirtes and Clark Glymour. “An algorithm for fast recovery of sparse causal graphs”. In: *Social science computer review* 9.1 (1991), pp. 62–72.
- [30] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.
- [31] Beatrice Laurent and Pascal Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of statistics* (2000), pp. 1302–1338.
- [32] Thomas S. Richardson. “Markov properties for acyclic directed mixed graphs”. In: *Scandinavian Journal of Statistics* 30.1 (2003), pp. 145–157.
- [33] Aurghya Maiti, Vineet Nair, and Gaurav Sinha. “A causal bandit approach to learning good atomic interventions in presence of unobserved confounders”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1328–1338.
- [34] Thomas S. Richardson et al. “Nested Markov properties for acyclic directed mixed graphs”. In: *The Annals of Statistics* 51.1 (2023), pp. 334–361.
- [35] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. 2nd ed. MIT Press, 2000.
- [36] Qing Zhou. *Latent Structure and Causality*. WORLD SCIENTIFIC, 2025. DOI: [10.1142/13773](https://doi.org/10.1142/13773).
- [37] Alain Hauser and Peter Bühlmann. “Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs”. In: *J. Mach. Learn. Res.* 13.1 (Aug. 2012), pp. 2409–2464. ISSN: 1532-4435.
- [38] Huining Hu, Zhentao Li, and Adrian Vetta. “Randomized Experimental Design for Causal Graph Discovery”. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.
- [39] Karthikeyan Shanmugam et al. “Learning causal graphs with small interventions”. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’15. MIT Press, 2015, pp. 3195–3203.

- [40] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. “Cost-optimal learning of causal graphs”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1875–1884.
- [41] Chandler Squires et al. “Active structure learning of causal DAGs via directed clique trees”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.
- [42] Davin Choo and Kirankumar Shiragur. “Adaptivity complexity for causal graph discovery”. In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. UAI ’23. Pittsburgh, PA, USA: JMLR.org, 2023.
- [43] Kristjan Greenewald et al. “Sample Efficient Active Learning of Causal Trees”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.
- [44] Muhammad Qasim Elahi et al. “Adaptive online experimental design for causal discovery”. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. Vienna, Austria: JMLR.org, 2024.
- [45] Christopher Meek. “Causal inference and causal explanation with background knowledge”. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 403–410. ISBN: 1558603859.