Bandit Algorithms: Fairness, Welfare, and Applications in Causal Inference

A THESIS
SUBMITTED FOR THE DEGREE OF
Master of Technology (Research)
IN THE
Faculty of Engineering

BY

Ayush Sawarni

Computer Science and Automation
Indian Institute of Science
Bangalore – 560 012 (INDIA)

November, 2023

Declaration of Originality

I, **Ayush Sawarni**, with SR No. **19776** hereby declare that the material presented in the thesis titled

Bandit Algorithms: Fairness, Welfare and Applications in Causal Inference

represents original work carried out by me in the **Department of Computer Science and Automation** at **Indian Institute of Science** during the years **2022 - 2023**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Advisor Signature

DEDICATED TO

Mummy, Papa and Boku

Acknowledgements

I would like to begin this note of thanks by expressing my deep gratitude towards my advisor, Prof. Siddharth Barman. He has not only taught me the art of research but has also helped me appreciate the beauty in simple and elegant ideas. I consider myself extremely fortunate to have him as a mentor at this pivotal stage in my career. He strikes the perfect balance between being a mentor and an advisor. I can firmly state that my approach to every new research problem is profoundly influenced by what I have learned from him. While many can explain a proof, there are few who can artfully teach how to devise one – a skill I've observed him demonstrate even in his courses. I aspire to emulate these invaluable skills in my academic career going forward.

I extend my sincere thanks to Prof. Arindam Khan for his invaluable guidance and support. His dedication to teaching is truly amazing, and I have gained much from his courses, including those I experienced through online videos. Additionally, I am grateful for the thoughtful advice he provided about my future plans, always making time for me despite his busy schedule.

I am grateful to Dr. Gaurav Sinha and Dr. Soumyabrata Pal for their incredible collaboration and the opportunity to work with industry labs. Special thanks to Gaurav for his invaluable mentorship and guidance in shaping my career path, as well as his insightful discussions on the philosophy of research.

I thank the CSA department for their role in my growth, particularly Professors Chiranjib Bhattacharya, Gugan Thoppe, Shalabh Bhatnagar, and Anand Louis for their exceptional courses. My appreciation also goes to the CSA office staff, especially Kushael Madam, Padmavathi Madam, and Shubha Madam, for their indispensable support with administrative tasks.

Finally, I am fortunate to have collaborated with brilliant students. My heartfelt thanks to Arnab Maiti for his enthusiasm and motivation in our joint project, which stands out as a highlight of my time at IISc. I also cherish my productive collaboration with Nirjhar Das towards the end of my Masters and thank him for the enriching technical discussions.

A special mention goes to my labmates. I thank Shraddha for some of the most memorable conversations I've had on campus which, somewhat mysteriously, always took an unexpected turn from academic/philosophical questions to the existential question of "whats the Marathi name of that food?". I'm grateful to Debajyoti for all the Table Tennis sessions; he was an amazing captain, steering our team through victories, albeit from the comfort of his home. A big thanks to Manisha, a passionate pictionary player and the official Mentos supplier of the lab, ensuring our lab never faced a mint shortage crisis. Finally, I want to thank Kiran Shiragur for being an amazing mentor, the lab clown, and an inspirational researcher. Our philosophical discussions on research has helped me tremendouly in evaluating my own work. I thank Siddharth (an unofficial member of the lab), whose endless, mind-numbing debates and talent for countering every statement I made kept my mind sharp.

I'm grateful for the wonderful friendships I've formed at the department. I thank Aditya(s), Atasi, Arka, KVN, Rishikesh, and Rahul for being a fantastic peer group. I treasure our discussions over coffee, fish bowl sessions, and during theory lunches. I thank Anand, Shravani, and Vishakha for introducing me to amazing food and restaurants, and for being there whenever I needed support and advice. Sharing a passion for mangoes with them was a delightful bonus.

Outside the department, my friends from S-Block - Ayush, Keshav, Shreya, and Sneha - played a crucial role in ensuring I maintained a balance, preventing burnout, and encouraging occasional breaks. I also want to extend my gratitude to my undergrad friends from BITS Pilani - Rishabh, Ankur, Rhythm, Aakanksha, and Abhinav - who generously hosted me during my visits to California.

Finally, my deepest gratitude goes to my parents and my brother for their unwavering support and sacrifices. I am who I am today largely because of them. Their selflessness inspires me daily, and I can't imagine where I would be without their constant encouragement.

Abstract

This thesis explores different aspects of regret in online learning and its applications. We introduce Nash regret, which measures the difference between the optimal action choices and the algorithm's performance in terms of the Nash social welfare function. By providing bounds on Nash regret, we establish principled fairness guarantees for online learning algorithms. We investigate different online learning settings and derive tight bounds on Nash regret.

In the first part, we focus on the classic multi-armed bandit (MAB) framework and develop an algorithm that achieves a tight Nash regret bound. Specifically, given a horizon of play T, our algorithm achieves a Nash regret of $O\left(\sqrt{\frac{k \log T}{T}}\right)$, where k represents the number of arms in the MAB instance. The lower bound on average regret applies to Nash regret as well, making our guarantee essentially tight. Additionally, we propose an anytime algorithm with a Nash regret guarantee of $O\left(\sqrt{\frac{k \log T}{T}} \log T\right)$.

In the second part, we study the stochastic linear bandits problem with non-negative, ν-sub Poisson rewards. We present an algorithm that achieves a Nash regret bound of $O\left(\sqrt{\frac{d\nu}{T}} \log(T|\mathcal{X}|)\right)$, where $\mathcal{X}$ denotes the set of arms in ambient dimension d and T represents the number of rounds. Furthermore, for linear bandit instances with potentially infinite arm sets, we derive a Nash regret upper bound of $O\left(\frac{d^{5/4}\nu^{1/2}}{\sqrt{T}} \log(T)\right)$. Our algorithm builds upon the successive elimination method and incorporates novel techniques such as tailored concentration bounds and sampling via the John ellipsoid in conjunction with the Kiefer-Wolfowitz optimal design.

In the third part, we investigate Nash regret in the context of online concave optimization and the Expert's problem, assuming adversarially chosen reward functions. Our algorithm achieves Nash Regret of $O\left(\frac{\log N}{T}\right)$ for the Expert's problem where N is the number of experts. We provide a lower bound for this setting that is essentially tight with respect to the upper bound. Additionally, for online concave optimization, we provide a Nash regret guarantee of $O\left(\frac{d \log T}{T}\right)$, where d denotes the ambient dimension.

In the final part of this thesis, we focus on the causal bandit problem, which involves identifying near-optimal interventions in a causal graph. Previous works have provided a bound of

$\widetilde{O}(N/\sqrt{T})$ for simple regret for causal graphs with N vertices, constant in-degree, and Bernoulli random variables. In this thesis, we present a new approach for exploration using covering interventions. This allows us to achieve a significant improvement and provide a tighter simple regret guarantee of $\widetilde{O}(\sqrt{N/T})$. Furthermore, we extend our algorithm to handle the most general case of causal graphs with unobservable variables.

Publications based on this Thesis

1. **Fairness and Welfare Quantification for Regret in Multi-armed Bandits**
   Joint work with Siddharth Barman, Arindam Khan, and Arnab Maiti.
   Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2023)

2. **Learning Good Interventions in Causal Graphs via Covering**
   Joint work with Siddharth Barman, Rahul Madhavan, and Gaurav Sinha.
   Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI 2023)

3. **Nash Regret Bounds for Linear Bandits**
   Joint work with Siddharth Barman, Soumyabrata Pal.
   To appear in the Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS 2023)

Contents

Chapter 1

Introduction

Multi-armed bandits (MAB) is a mathematical framework for making sequential decisions in the face of uncertainty. In this framework, we have a set of arms (possible actions) with unknown means and a time horizon T. For T rounds, the online algorithm sequentially selects an arm and receives a reward drawn independently from the arm-specific distribution. Here, ex-ante, the optimal solution would be to select, in every round, the arm with the maximum mean. However, since the statistical properties of the arms are unknown, the algorithm accrues—in the T rounds—expected rewards that are not necessarily optimal. The overarching aim of the framework is to maximize the algorithm's performance (in some metric of choice) or, equivalently, minimize *regret*—which is a notion of loss defined over the bandit instance. As the algorithm gains more information about the arms, it faces an inherent tradeoff of whether to *explore*: pull arms that haven't been pulled enough times or to *exploit*: greedily choose the best arm based on the available information. This tradeoff is fundamental to numerous applications of MAB, some of which are detailed below

- *Medical Trials:* The MAB framework was introduced by William R. Thompson [Tho33] as a means to make medical trials less cruel. The idea was to dynamically allocate treatments based on the evolving knowledge of drug effectiveness, thus reducing harm. Since then, there have been numerous studies [VB22, TM17, VBW15, Git79] exploring the applications of bandit algorithms in medical trials. In particular, when modeled as a MAB problem, in each round $t \in \{1, \ldots, T\}$, the decision maker administers one of the candidate drugs to the tth patient. The reward received in each round indicates the efficacy of the drug administered in that round.

- *Applications in Economics:* Bandit algorithms have found extensive applications in various domains in economics, including dynamic pricing [DB15], where the objective is to determine the optimal prices for products based on customer preferences and market dynamics. Additionally, bandit algorithms have been explored in the context of auction design [BS10], where the online algorithm selects reserve prices to maximize revenue. Furthermore, bandit algorithms have been studied in the design of crowdsourcing platforms [SV14], aiming to dynamically match tasks with workers to maximize the number of completed tasks.

- *Applications to the Web:* The MAB framework naturally encompasses several web-based applications, including online advertisement [SBF17, PACJ07, LZ07], web search [RKJ08], and recommendation systems [BCS14, LKG16]. In these applications, during each round $t \in 1, \ldots, T$, a user t visits a website, and based on user information and web content, the algorithm selects an action to maximize revenue or user engagement.

Notably, the bandit algorithm induces value for a population of agents in the mentioned applications. This perspective encourages studying these algorithms from a welfarist point of view. In section 1.1, we explore the concept of regret in relation to welfarist considerations in more depth. Apart from the mentioned applications, recent works have also studied bandit algorithms for identifying good interventions in causal graphs. We provide a detailed discussion of this application in section 1.2.

1.1 A Welfarist Perspective on the MAB Framework

Ensuring a fair distribution of resources among agents is a central objective in economics and social choice theory [Mou04]. A key tool for evaluating the quality of allocations is the use of welfare functions, which map a set of positive real numbers to a single positive real number. In economics, several well-studied welfare functions include Egalitarian Welfare (EW), Utilitarian Welfare (UW), and Nash Social Welfare (NSW). EW assigns the minimum value from the set of values, UW calculates the arithmetic mean, and NSW computes the geometric mean of the values.

In the context of bandit problems, the primary metrics used to assess performance are average regret and simple regret [LS20, S$^+$19]. Average regret compares the optimal mean reward to the average (arithmetic mean) of the expected rewards obtained by the algorithm over the rounds. It quantifies the algorithm's performance as the arithmetic mean of the expected rewards accumulated throughout the T rounds. On the other hand, simple regret evaluates the algorithm's performance based solely on its expected reward in the Tth round, disregarding previous rounds. From a welfarist perspective, average regret corresponds to maximizing Social Welfare (SW), while simple regret corresponds to maximizing Egalitarian Welfare (EW). It is important to note that maximizing social welfare (or minimizing average regret) does not necessarily guarantee fairness. Even if the initial agents are treated unfairly, the social welfare can still be high. In contrast, simple regret captures the fairness guarantees only after excluding an initial set of agents.

To incorporate fairness and welfare considerations into the Multi-Armed Bandit (MAB) framework, we adopt a principled approach from mathematical economics. We employ a welfare function that is justified by axioms to quantify the algorithm's performance. Specifically, we utilize the Nash Social Welfare (NSW). NSW satisfies fundamental axioms such as symmetry, independence of unconcerned agents, scale invariance, and the Pigou-Dalton transfer principle [Mou04]. The Pigou-Dalton principle ensures that NSW increases under a policy change that transfers reward δ from an agent with a higher value to an agent with a lower value. This principle favours a more balanced distribution of rewards, hence, promoting fairness. However, if the relative increase in value for the lower-valued agent is significantly smaller than the decrease for the higher-valued agent, NSW would not favor such a transfer, thereby accommodating efficiency. The NSW strikes a balance between fairness and economic efficiency. It is positioned between egalitarian and social welfare, as the geometric mean is at least as large as the minimum reward and at most equal to the arithmetic mean (according to the AM-GM inequality).

In the current framework, each round of the algorithm corresponds to a distinct agent, and the algorithm's performance is measured as the geometric mean of its value induced over the T rounds. This introduces the concept of Nash regret, which captures the difference between the optimal action choices and the geometric mean of the expected rewards resulting from the algorithm's choices. On a meta-level, the notion of Nash regret can be loosely stated as

$$\mathrm{NR}_T = \texttt{OPT} - \left(\prod_{t=1}^{T} r_t\right)^{\frac{1}{T}}$$

where r_t denotes the value induced by an algorithm's decision in round t. In section 1.3, we provide more precise definitions of Nash regret pertaining different problems in online learning.

In the next section, we introduce an important application of the MAB framework within the domain of Causal Inference, namely the causal bandit problem.

1.2 Learning in Causal Bayesian Networks

Statistical inference involves understanding the characteristics of a high dimensional-distribution using observational data. Although probability distributions can effectively capture the relationships between these variables, they may fall short in providing accurate predictions when the outcomes result from interventions on specific variable, that is, when certain variables are set to specific values through an external intervention. Causal Bayesian Networks (CBNs) have emerged as a prominent paradigm for modeling such problems [Pea09]. Their recent applications span diverse domains such as language modeling [Sev20], medicine [KEG17, CKD$^+$15, LWB$^+$18], robotics [YN12], and computational advertising [BPQC$^+$13].

A Causal Bayesian Network (CBN) consists of a directed acyclic graph, known as a causal graph, indicating the direction of causation among random variables, with each vertex in the graph denoting a random variable. In this graph, the set of directed edges represents the causal relationships between these variables, and each variable (vertex) in the graph is determined by a function of its parent vertices. Additionally, a variable without parents, known as an exogenous variable, is an independent random variable following some distribution.

Despite the longstanding research focus on CBNs, studying online learning in the context of CBNs has only recently gained attention. The causal bandits model addresses fundamental questions at the intersection of online learning and CBNs. Introduced by Lattimore et al. [LLR16], causal bandits combine concepts from CBNs and multi-armed bandits (MABs) to establish a framework for learning good interventions in CBNs. The problem is defined by a causal graph, a set of interventions, and a designated reward node within the causal graph. Amongst the given set of interventions, the one that maximizes the expected value of the reward node is labeled as the best intervention. The online learning algorithm aims to use as few interventional and observational samples as possible to determine the best intervention.

We now present a stylized application of the causal bandit problem, similar to the one outlined in [MNS22]. Consider a scenario where a policy-maker is tasked with determining the most effective set of precautionary measures to mitigate the spread of a disease. The set of options available to the policy maker such as policing for social distancing, opening new vaccination centers, mandating the wearing of face masks, creating new remote-work policies, etc., are typically large, and the policymaker can choose multiple policies to implement as long as they fit the budget/time constraints. The task at hand is to find the policy most effective in mitigating the spread of the disease. Importantly, leveraging the domain expertise of health professionals, the policymaker can access an underlying causal graph. This graph represents causal relationships and helps to make informed decisions on when and if to enforce a specific measure. Each feasible combination of policies corresponds to an intervention in the causal graph, and the total number of possible choices can be exponential. When studied as a causal bandit problem, the decision maker gets to exploit the causal structure to identify the best intervention with fewer data samples.

1.3 Problem Definitions

1.3.1 Stochastic Multi-armed Bandits

In the classical (stochastic) multi-armed bandit problem, an online algorithm (decision maker) has access to samples from k (unknown) distributions that are defined on the interval $[0, 1]$. These distributions are referred to as arms, denoted by $i \in 1, \ldots, k$. The algorithm must repeatedly select (pull) an arm in each round, and this process continues for a total of $T \geq 1$ rounds. Each pull of arm i results in an independent and identically distributed (i.i.d.) reward from the ith distribution. Let $\mu_i \in [0, 1]$ denote the unknown mean of the ith arm, and we define μ^* as the maximum mean, i.e., $\mu^* \coloneqq \max_{i \in [k]} \mu_i$. Given a specific bandit instance and an algorithm, the random variable $I_t \in [k]$ denotes the arm pulled in round $t \in 1, \ldots, T$. Importantly, I_t depends on the observations made prior to round t.

We consider settings where rewards are distributed among a population of T agents. Specifically, for each agent $t \in 1, \ldots, T$, the expected reward received is $\mathbb{E}[r_t]$. Therefore, the algorithm induces rewards $\{\mathbb{E}[\mu_{I_t}]\}_{t=1}^{T}$ for all T agents. The algorithm's performance can be quantified by applying a welfare function to these induced rewards. We focus on Nash social welfare, which, in the context of stochastic Multi-armed Bandits, corresponds to the geometric mean of the agents' expected rewards: $\left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}$. The objective is to obtain guarantees on Nash regret which is defined as the difference between the optimal arm μ^* and the NSW of expected rewards. That is,

$$\mathrm{NR}_T \coloneqq \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}.$$

1.3.2 Online Learning with Full-feedback

The problem is defined over a set of actions $\mathcal{X} \subset \mathbb{R}^d$ represented as d-dimensional vectors, along with a sequence of concave reward functions $f_1, f_2, \ldots, f_T$. Each function $f_t : \mathcal{X} \to \mathbb{R}^+$ denotes the reward associated with the t-th round. In each round, the algorithm selects an action $x_t \in \mathcal{X}$ and receives a reward $r_t = f_t(x_t)$. The algorithm's performance is evaluated based on the Nash Social Welfare (NSW) across the rounds, and the Nash regret is defined with respect to the best fixed action in hindsight:

$$\mathrm{NR}_T = \max_{x \in \mathcal{X}} \left(\prod_{t=1}^{T} f_t(x)\right)^{\frac{1}{T}} - \left(\prod_{t=1}^{T} f_t(x_t)\right)^{\frac{1}{T}} \tag{1.1}$$

1.3.3 Linear Bandits

The stochastic linear bandit problem involves an online algorithm making sequential decisions over a time horizon of T rounds. The algorithm is provided with a set of arms $\mathcal{X} \subset \mathbb{R}^d$, where each arm represents a d-dimensional vector. Each arm $x \in \mathcal{X}$ is associated with a stochastic reward $r_x \in \mathbb{R}_+$. The key assumption in this problem is that the expected reward r_x is a linear function of the arm $x \in \mathbb{R}^d$, represented by an unknown parameter vector $\theta^* \in \mathbb{R}^d$. The expected value of the reward for arm $x \in \mathcal{X}$ is given by $\mathbb{E}[r_x] = \langle x, \theta^* \rangle$.

The online algorithm, which can be randomized, proceeds by selecting an arm X_t in each round $t \in [T]$ and observing the corresponding stochastic reward $r_{X_t} > 0$. Similar to the stochastic MAB problem the objective is to quantify the welfare induced by the algorithm in terms of the Nash Regret, defined as

$$\mathrm{NR}_T \coloneqq \max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \left(\prod_{t=1}^{T} \mathbb{E}[\langle X_t, \theta^* \rangle] \right)^{1/T}.$$

1.3.4 Causal Bandits

A causal bandit problem revolves around a causal Bayesian network (CBN) that consists of a directed acyclic graph $\mathcal{G} = (\mathcal{V}, E)$, representing the causal relationships among N random variables. In this graph, the vertices $\mathcal{V}$ represent the variables, and the set of directed edges E represents the causal relationships between these variables. Each variable is a function of its parents in the graph.

We consider the random variables in $\mathcal{V}$ to be Bernoulli variables. In the causal bandit problem, one variable $V_N \in \mathcal{V}$ is designated as the reward variable, and the objective is to optimize the expected value of this reward variable. The optimization is performed over a predefined set $\mathcal{A}$ consisting of interventions in the causal graph. These interventions, denoted as do() operations, fix the values of certain variables without considering their parents. Specifically, in an intervention $A = \mathrm{do}(S = s)$, the value of each variable i in the set $S \subseteq \mathcal{V}$ is fixed according to the corresponding binary assignment $s \in 0, 1^{|S|}$. Under this intervention, the unaffected variables (in $\mathcal{V} \setminus S$) follow the remaining causal relations. We denote the expected value of the reward variable under intervention A as $\mu(A)$.

The objective here is to perform exploratory interventions over a given number of rounds T (time horizon). At the end of this time horizon, the learner aims to identify a nearly optimal intervention from the target set $\mathcal{A}$. In other words, the main goal of the causal bandit problem is to find an intervention $A_T \in \mathcal{A}$ at the end of T rounds that maximizes the expected value of the reward variable V_N. The performance of the algorithm is measured in terms of simple regret, which is defined as

$$R_T = \max_{A \in \mathcal{A}} \mu(A) - \mathbb{E}\left[\mu(A_T)\right].$$

1.4 Overview of the Thesis

This thesis focuses on investigating the aforementioned problems, with each chapter dedicated to studying a specific problem. The contributions of each chapter are outlined below:

- Chapter 2 addresses the problem of Nash regret in stochastic multi-armed bandits. We introduce a novel algorithm that achieves near-optimal Nash regret, with only logarithmic factors of deviation. Additionally, we extend this algorithm to a variant that does not require prior knowledge of the time horizon.

- Chapter 3 delves into the problem of Nash regret in linear bandits with sub-Poisson rewards. We propose two algorithms: one that is dependent on the arm set, and another that is independent of the size of the arm set.

- Chapter 4 investigates the scenario of full feedback with adversarial rewards. Specifically, we examine the expert's problem and online concave optimization. In addition, we establish a tight lower bound up to logarithmic factors.

- Chapter 5 explores the causal bandit problem and presents an algorithm for identifying optimal interventions in a causal graph. Our algorithm surpasses existing state-of-the-art approaches by introducing an exploration model that allows the algorithm to intervene beyond the given set of arms. Additionally, we extend the algorithm to provide simple regret guarantees for causal graphs with unobserved variables, a previously unsolved problem in the most general case of causal graphs.

Chapter 2

Nash Regret Bounds for Stochastic MAB

We extend the notion of regret with a welfarist perspective. Focussing on the classic multi-armed bandit (MAB) framework, this chapter quantifies the performance of bandit algorithms by applying a fundamental welfare function, namely the Nash social welfare (NSW) function. This corresponds to equating algorithm's performance to the geometric mean of its expected rewards and leads us to the study of *Nash regret*, defined as the difference between the—a priori unknown—optimal mean (among the arms) and the algorithm's performance.

Recall that the MAB framework provides an encapsulating abstraction for settings that entail sequential decision making under uncertainty. In this framework, a decision maker (online algorithm) has sample access to k distributions (arms), which are a priori unknown. For T rounds, the online algorithm sequentially selects an arm and receives a reward drawn independently from the arm-specific distribution. Here, ex ante, the optimal solution would be to select, in every round, the arm with the maximum mean. However, since the statistical properties of the arms are unknown, the algorithm accrues—in the T rounds—expected rewards that are not necessarily the optimal. The construct of regret captures this sub-optimality and, hence, serves as a key performance metric for algorithms. A bit more formally, regret is defined as the difference between the optimal mean (among the arms) and the algorithm's performance.

2.1 Results and Techniques

We develop an algorithm that achieves Nash regret of $O\left(\sqrt{\frac{k \log T}{T}}\right)$; here, k denotes the number of arms in the bandit instance and T is the given horizon of play (Theorem 2.1 and Theorem

2.2). Note that, for any algorithm, the Nash regret is at least as much as its average regret.[1] Therefore, the known $\Omega\left(\sqrt{\frac{k}{T}}\right)$ lower bound on average regret [ACBFS02] holds for Nash regret as well. This observation implies that, up to a log factor, our guarantee matches the best-possible bound, even for average regret.

We also show that the standard upper confidence bound (UCB) algorithm [LS20] does not achieve any meaningful guarantee for Nash regret (Section 2.8). This barrier further highlights that Nash regret is a more challenging benchmark than average regret. In fact, it is not obvious if one can obtain any nontrivial guarantee for Nash regret by directly invoking upper bounds known for average (cumulative) regret. For instance, a reduction from Nash regret minimization to average regret minimization, by taking logs of the rewards (i.e., by converting the geometric mean to the arithmetic mean of logarithms), faces the following hurdles: (i) for rewards that are bounded between 0 and 1, the log values can be in an arbitrarily large range, and (ii) an additive bound for the logarithms translates back to only a multiplicative guarantee for the underlying rewards.

Our algorithm (Algorithm 1) builds upon the UCB template with interesting technical insights; see Section 2.3 for a detailed description. The two distinctive features of the algorithm are: (i) it performs uniform exploration for a judiciously chosen number of initial rounds and then (ii) it adds a novel (arm-specific) confidence width term to each arm's empirical mean and selects an arm for which this sum is maximum (see equation (2.2)). Notably, the confidence width includes the empirical mean as well. These modifications enable us to go beyond standard regret analysis.[2]

The above-mentioned algorithmic result focusses on settings in which the horizon of play (i.e., the number of rounds) T is given as input. Extending this result, we also establish a Nash regret guarantee for T-oblivious settings. In particular, we develop an anytime algorithm with a Nash regret of $O\left(\sqrt{\frac{k \log T}{T}} \log T\right)$ (Theorem 2.3). This extension entails an interesting use of empirical estimates to identify an appropriate round at which the algorithm can switch out of uniform exploration.

2.1.1 Additional Related Work and Application

Given that learning algorithms are increasingly being used to guide socially sensitive decisions, there has been a surge of research aimed at achieving fairness in MAB contexts; see, e.g.,

[1] This follows from the AM-GM inequality: The average regret is equal to the difference between the optimal mean, μ^*, and the arithmetic mean of expected rewards. The arithmetic mean is at least the geometric mean, which in turn is considered in Nash regret.

[2] Note that the regret decomposition lemma [LS20], a mainstay of regret analysis, is not directly applicable for Nash regret.

[JKMR16a, CKSV19, PGNN20, BBLB20a] and references therein. This thread of research has predominantly focused on achieving fairness for the arms. By contrast, the current work establishes fairness (and welfare) guarantees across time.

In addition, [HMS21a] considers a multi-agent setting: each arm pull generates a (possibly distinct) reward among N agents. The goal in [HMS21a] is to find a distribution (over the arms) that is fair for the N agents. This objective differs from identifying an arm with a high mean reward, since, for each arm, the rewards can vary across the agents. On the other hand, our work conforms to the classic MAB setup and considers fairness across rounds; each round $t \in [T]$ represents a distinct agent.

The significance of Nash social welfare and its axiomatic foundations [KN79, NJ50] in fair division settings are well established; see [Mou04] for a textbook treatment. Specifically in the context of allocating divisible goods, NSW is known to uphold other important fairness and efficiency criteria [Var74]. In fact, NSW corresponds to the objective function considered in the well-studied convex program of Eisenberg and Gale [EG59]. NSW is an object of active investigation in discrete fair division literature as well; see, e.g., [CKM$^+$19].

While the focus of the current chapter is to develop provable algorithmic guarantees for Nash regret, we provide here an example to highlight the applicability of this fairness metric: Consider the use of MAB methods for displaying ad impressions [SBF17]. In this application, different ad configurations correspond to different arms, and the online users are the T agents. In round $t \in [T]$, the tth user visits the website and is shown an ad configuration (i.e., a chosen arm). The reward, for every user $t \in [T]$, is stochastic and based on the selected arm (i.e., the selected ad configuration). In this application, maximizing Nash welfare of rewards is a meaningful objective, since it qualitatively supports an online algorithm that is fair across the T agents. Indeed, Nash regret would dissuade sacrificing the experience of an initial set of users for overall utilitarian benefits.

2.2 Notation and Preliminaries

We study the classic (stochastic) multi-armed bandit problem. Here, an online algorithm (decision maker) has sample access to k (unknown) distributions, that are supported on $[0, 1]$. The distributions are referred to as arms $i \in \{1, \ldots, k\}$. The algorithm must iteratively select (pull) an arm per round and this process continues for $T \geq 1$ rounds overall. Successive pulls of an arm i yield i.i.d. rewards from the ith distribution. We will, throughout, write $\mu_i \in [0, 1]$ to denote the (a priori unknown) mean of the the ith arm and let μ^* be maximum mean, $\mu^* := \max_{i \in [k]} \ \mu_i$. Furthermore, given a bandit instance and an algorithm, the random variable $I_t \in [k]$ denotes the arm pulled in round $t \in \{1, \ldots, T\}$. Note that I_t depends on the draws observed before round t.

We address settings in which the rewards are distributed across a population of T agents. Specifically, for each agent $t \in \{1, \ldots, T\}$, the expected reward received is $\mathbb{E}[\mu_{I_t}]$ and, hence, the algorithm induces rewards $\{\mathbb{E}[\mu_{I_t}]\}_{t=1}^{T}$ across all the T agents. Notably, one can quantify the algorithm's performance by applying a welfare function on these induced rewards. Our focus is on Nash social welfare, which, in the current context, is equal to the geometric mean of the agents' expected rewards: $\left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}$. Here, the overarching aim of achieving a Nash social welfare as high as possible is quantitatively captured by considering *Nash regret*, NR_T; this metric is defined as

$$\mathrm{NR}_T \coloneqq \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T} \tag{2.1}$$

Note that the optimal value of Nash social welfare is μ^*, and our objective is to minimize Nash regret.

Furthermore, the standard notion of average (cumulative) regret is obtained by assessing the algorithm's performance as the induced social welfare $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$. Specifically, we write R_T to denote the average regret, $\mathrm{R}_T \coloneqq \mu^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$. The AM-GM inequality implies that Nash regret, NR_T, is a more challenging benchmark than R_T; indeed, for our algorithm, the Nash regret is $O\left(\sqrt{\frac{k \log T}{T}}\right)$ and the same guarantee holds for the algorithm's average regret as well.

2.3 The Nash Confidence Bound Algorithm

Our algorithm (Algorithm 1) consists of two phases. Phase I performs uniform exploration for $\widetilde{T} \coloneqq 16\sqrt{\frac{kT \log T}{\log k}}$ rounds. In Phase II, each arm is assigned a value (see equation (2.2)) and the algorithm pulls the arm with the highest current value. Based on the observed reward, the values are updated and this phase continues for all the remaining rounds.

We refer to the arm-specific values as the *Nash confidence bounds*, NCB_i-s. For each arm $i \in [k]$, we obtain NCB_i by adding a 'confidence width' to the empirical mean of arm i; in particular, NCB_i depends on the number of times arm i has been sampled so far and rewards experienced for i. Formally, for any round t and arm $i \in [k]$, let $n_i \geq 1$ denote the number of times arm i has been pulled before this round.[1] Also, for each $1 \leq s \leq n_i$, random variable $X_{i,s}$ be the observed reward when arm i is pulled the sth time. At this point, arm i has empirical

[1] Note that n_i is a random variable.

mean $\widehat{\mu}_i := \frac{1}{n_i}\sum_{s=1}^{n_i} X_{i,s}$ and we define the Nash confidence bound as

$$\mathrm{NCB}_i := \widehat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}} \tag{2.2}$$

It is relevant to observe that, in contrast to standard UCB (see, e.g., [LS20]), here the confidence width includes the empirical mean (i.e., the additive term has $\widehat{\mu}_i$ under the square-root). This is an important modification that enables us to go beyond standard regret analysis. Furthermore, we note that the Nash regret guarantee of Algorithm 1 can be improved by a factor of $\sqrt{\log k}$ (see Theorem 2.1 and Theorem 2.2). The initial focus on Algorithm 1 enables us to highlight the key technical insights for Nash regret. The improved guarantee is detailed in Section 2.4.1.

Algorithm 1 Nash Confidence Bound Algorithm

Input: Number of arms k and horizon of play T.

```
1: Initialize empirical means $\widehat{\mu}_i = 0$ and counts $n_i = 0$ for all arms $i \in [k]$. Also, set $\widetilde{T} := 16\sqrt{\frac{kT\log T}{\log k}}$.
   Phase I
2: for $t = 1$ to $\widetilde{T}$ do
3:     Select $I_t$ uniformly at random from $\{1, 2, \ldots, k\}$.
4:     Pull arm $I_t$ and observe reward $X_t$.
5:     For arm $I_t$, increment the count $n_{I_t}$ (by one) and update the empirical mean $\widehat{\mu}_{I_t}$.
6: end for
   Phase II
7: for $t = (\widetilde{T} + 1)$ to $T$ do
8:     Pull the arm $I_t$ with the highest Nash confidence bound, i.e., $I_t = \operatorname{argmax}_{i \in [k]} \ \mathrm{NCB}_i$.
9:     Observe reward $X_t$ and update $\widehat{\mu}_{I_t}$.
10:    Update the Nash confidence bound for $I_t$ (see equation (2.2)).
11: end for
```

The following theorem is the main result of this section and it establishes that Algorithm 1 achieves a tight—up to log factors—guarantee for Nash regret.

Theorem 2.1. *For any bandit instance with k arms and given any (moderately large) time horizon T, the Nash regret of Algorithm 1 satisfies*

$$\mathrm{NR}_T = O\left(\sqrt{\frac{k \log k \ \log T}{T}}\right).$$

2.3.1 Regret Analysis

We first define a "good" event G and show that it holds with high probability (Lemma 2.1); our Nash regret analysis is based on conditioning on G. In particular, we will first define three sub-events G_1, G_2, G_3 and set $G \coloneqq G_1 \cap G_2 \cap G_3$. For specifying these events, write $\widehat{\mu}_{i,s}$ to denote the empirical mean of arm i's rewards, based on the first s samples (of i).

G_1: Every arm $i \in [k]$ is sampled at least $\frac{\widetilde{T}}{2k}$ times in Phase I,[1] i.e., for each arm i we have $n_i \geq \frac{\widetilde{T}}{2k}$ at the end of the first phase in Algorithm 1.

G_2: For all arms $i \in [k]$, with $\mu_i > \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, and all sample counts $\frac{\widetilde{T}}{2k} \leq s \leq T$ we have $|\mu_i - \widehat{\mu}_{i,s}| \leq 3\sqrt{\frac{\mu_i \log T}{s}}$.

G_3: For all arms $j \in [k]$, with $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, and all $\frac{\widetilde{T}}{2k} \leq s \leq T$, we have $\widehat{\mu}_{j,s} \leq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}}$.

Here,[2] all the events are expressed as in the canonical bandit model (see, e.g., [LS20, Chapter 4]). In particular, for events G_2 and G_3, one considers a $k \times T$ reward table that populates T independent samples for each arm $i \in [k]$. All the empirical means are obtained by considering the relevant entries from the table; see section 2.7 for a more detailed description of the associated probability space. Also note that, conceptually, the algorithm gets to see the (i, s)th entry in the table only when it samples arm i the sth time.

The lemma below lower bounds the probability of event G; its proof is deferred to section 2.5.1.

Lemma 2.1. $\mathbb{P}\{G\} \geq \left(1 - \frac{4}{T}\right)$.

Next, we state a useful numeric inequality; for completeness, we provide its proof in section 2.5.2.

Claim 2.1. *For all reals $x \in \left[0, \frac{1}{2}\right]$ and all $a \in [0, 1]$, we have $(1 - x)^a \geq 1 - 2ax$.*

Now, we will show that the following key guarantees (events) hold under the good event G:

- Lemma 2.2: The Nash confidence bound of the optimal arm i^* is at least its true mean, μ^*, throughout Phase II.

[1] Recall that $\widetilde{T} \coloneqq 16\sqrt{\frac{kT \log T}{\log k}}$.

[2] Note that if, for all arms $i \in [k]$, the means $\mu_i \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, then, by convention, $\mathbb{P}\{G_2\} = 1$. Similarly, if all the means are sufficiently large, then $\mathbb{P}\{G_3\} = 1$.

- Lemma 2.3: Arms j with sufficiently small means (in particular, $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$) are never pulled in Phase II.

- Lemma 2.4: Arms i that are pulled many times in Phase II have means μ_i close to the optimal μ^*. Hence, such arms i do not significantly increase the Nash regret.

The proofs of these three lemmas are deferred to Section 2.5.3. In these results, we will address bandit instances wherein the optimal mean $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$. Note that in the complementary case (wherein $\mu^* < \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$) the Nash regret directly satisfies the bound stated in Theorem 2.1.

Lemma 2.2. *Let $\mathrm{NCB}_{i^*,t}$ be the Nash confidence bound of the optimal arm i^* at round t. Assume that the good event G holds and also $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$. Then, for all rounds $t > \widetilde{T}$ (i.e., for all rounds in Phase* II*), we have $\mathrm{NCB}_{i^*,t} \geq \mu^*$.*

Lemma 2.3. *Consider a bandit instance with optimal mean $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ and assume that the good event G holds. Then, any arm j, with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, is never pulled in all of Phase* II*.*

Lemma 2.4. *Consider a bandit instance with optimal mean $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ and assume that the good event G holds. Then, for any arm i that is pulled at least once in Phase* II *we have*

$$\mu_i \geq \mu^* - 8\sqrt{\frac{\mu^* \log T}{T_i - 1}},$$

where T_i is the total number of times that arm i is pulled in the algorithm.

2.3.2 Proof of Theorem 2.1

For bandit instances in which the optimal mean $\mu^* \leq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$, the theorem holds directly; specifically, the Nash regret $\mathrm{NR}_T = \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T} \leq \mu^*$. Therefore, in the remainder of the proof we will address instances wherein $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$.

The Nash social welfare of the algorithm satisfies[1] $\left(\prod_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} = \left(\prod_{t=1}^{\widetilde{T}} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \left(\prod_{t=\widetilde{T}+1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)$ In this product, the two terms account for the rewards accrued in the two phases, respectively. Next, we will lower bound these two terms.

[1] Recall that $\widetilde{T} \coloneqq 16\sqrt{\frac{kT \log T}{\log k}}$.

Phase I: In each round of the first phase, the algorithm selects an arm uniformly at random. Hence, $\mathbb{E}[\mu_{I_t}] \geq \frac{\mu^*}{k}$, for each round $t \leq \widetilde{T}$. Therefore, for Phase I we have

$$\left(\prod_{t=1}^{\widetilde{T}} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \geq \left(\frac{\mu^*}{k}\right)^{\frac{\widetilde{T}}{T}} = (\mu^*)^{\frac{\widetilde{T}}{T}} \left(\frac{1}{k}\right)^{\frac{16\sqrt{k}\log T}{\sqrt{T}\log k}} = (\mu^*)^{\frac{\widetilde{T}}{T}} \left(\frac{1}{2}\right)^{\frac{16\sqrt{k}\log T \log k}{\sqrt{T}\log k}}$$

$$= (\mu^*)^{\frac{\widetilde{T}}{T}} \left(1 - \frac{1}{2}\right)^{\frac{16\sqrt{k\log k}\log T}{\sqrt{T}}} \geq (\mu^*)^{\frac{\widetilde{T}}{T}} \left(1 - \frac{16\sqrt{k\log k}\log T}{\sqrt{T}}\right) \tag{2.3}$$

To obtain the last inequality we note that the exponent $\frac{16\sqrt{k\log k}\log T}{\sqrt{T}} < 1$ (for appropriately large T) and, hence, the inequality follows from Claim 2.1.

Phase II: For the second phase, the product of the expected rewards can be bounded as follows

$$\left(\prod_{t=\widetilde{T}+1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq \mathbb{E}\left[\left(\prod_{t=\widetilde{T}+1}^{T} \mu_{I_t}\right)^{\frac{1}{T}}\right] \geq \mathbb{E}\left[\left(\prod_{t=\widetilde{T}+1}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \middle| \, G\right] \mathbb{P}\{G\} \tag{2.4}$$

Here, the first inequality follows from the multivariate Jensen's inequality and the second one is obtained by conditioning on the good event G. To bound the expected value in the right-hand-side of inequality (2.4), we consider the arms that are pulled at least once in Phase II. In particular, with reindexing, let $\{1, 2, \ldots, \ell\}$ denote the set of all arms that are pulled at least once in the second phase. Also, let $m_i \geq 1$ denote the number of times arm $i \in [\ell]$ is pulled in Phase II and note that $\sum_{i=1}^{\ell} m_i = T - \widetilde{T}$. Furthermore, let T_i denote the total number of times any arm i is pulled in the algorithm. Indeed, $(T_i - m_i)$ is the number of times arm $i \in [\ell]$ is pulled in Phase I. With this notation, the expected value in the right-hand-side of inequality (2.4) can be expressed as $\mathbb{E}\left[\left(\prod_{t=\widetilde{T}}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \middle| \, G\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell} \mu_i^{\frac{m_i}{T}}\right) \middle| \, G\right]$. Moreover, since we are conditioning on the good event G, Lemma 2.4 applies to each arm $i \in [\ell]$. Hence,

$$\mathbb{E}\left[\left(\prod_{t=\widetilde{T}}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \middle| \, G\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell} \mu_i^{\frac{m_i}{T}}\right) \middle| \, G\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell} \left(\mu^* - 8\sqrt{\frac{\mu^*\log T}{T_i - 1}}\right)^{\frac{m_i}{T}} \middle| \, G\right]$$

(Lemma 2.4)

$$= (\mu^*)^{1-\frac{\widetilde{T}}{T}} \mathbb{E}\left[\prod_{i=1}^{\ell} \left(1 - 8\sqrt{\frac{\log T}{\mu^*(T_i - 1)}}\right)^{\frac{m_i}{T}} \middle| \, G\right] \tag{2.5}$$

For the last equality, we use $\sum_{i=1}^{\ell} m_i = T - \widetilde{T}$. Now, recall that, under event G, each arm is pulled at least $\frac{\widetilde{T}}{2k} = \frac{8}{k}\sqrt{\frac{kT\log T}{\log k}}$ times in Phase I. Hence, $T_i > \frac{\widetilde{T}}{2k}$ for each arm $i \in [\ell]$. Furthermore, since $\mu^* \geq \frac{32\sqrt{k\log k \log T}}{\sqrt{T}}$, we have $8\sqrt{\frac{\log T}{\mu^*(T_i-1)}} \leq 8\sqrt{\frac{1}{256}} = \frac{1}{2}$ for each $i \in [\ell]$. Therefore, we can apply Claim 2.1 to reduce the expected value in inequality (2.5) as follows

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-8\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right)^{\frac{m_i}{T}} \,\middle|\, G\right] &\geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16m_i}{T}\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right) \,\middle|\, G\right]\\
&\geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right) \,\middle|\, G\right] \qquad \text{(since } T_i \geq m_i+1\text{)}
\end{aligned}$$

We can further simplify the above inequality by noting that $(1-x)(1-y) \geq 1-x-y$, for all $x, y \geq 0$.

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right) \,\middle|\, G\right] &\geq \mathbb{E}\left[1-\sum_{i=1}^{\ell}\left(\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right) \,\middle|\, G\right]\\
&= 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sum_{i=1}^{\ell}\sqrt{m_i} \,\middle|\, G\right]\\
&\geq 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell}\sqrt{\sum_{i=1}^{\ell} m_i} \,\middle|\, G\right]\\
&\qquad\qquad \text{(Cauchy-Schwarz inequality)}\\
&\geq 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell}\,T \,\middle|\, G\right] \qquad \text{(since } \textstyle\sum_i m_i \leq T\text{)}\\
&= 1-\left(16\sqrt{\frac{\log T}{\mu^* T}}\right)\mathbb{E}\left[\sqrt{\ell} \,\middle|\, G\right]\\
&\geq 1-\left(16\sqrt{\frac{k\log T}{\mu^* T}}\right) \qquad (2.6)
\end{aligned}$$

Here, the final inequality holds since $\ell \leq k$. Using (2.6), along with inequalities (2.4), and (2.5), we obtain for Phase II:

$$\left(\prod_{t=\widetilde{T}+1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq (\mu^*)^{1-\frac{\widetilde{T}}{T}}\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{G\} \qquad (2.7)$$

Inequalities (2.7) and (2.3) provide relevant bounds for Phase II and Phase I, respectively. Hence, for the Nash social welfare of the algorithm we have

$$\begin{aligned}
\left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq \mu^*\left(1-\frac{16\sqrt{k\log k\log T}}{\sqrt{T}}\right)\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{G\} \\
&\geq \mu^*\left(1-\frac{16\sqrt{k\log k\log T}}{\sqrt{T}}\right)\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\left(1-\frac{4}{T}\right) \quad \text{(via Lemma 2.1)} \\
&\geq \mu^*\left(1-\frac{32\sqrt{k\log k\log T}}{\sqrt{\mu^* T}}\right)\left(1-\frac{4}{T}\right) \\
&\geq \mu^* - \frac{32\sqrt{\mu^* k\log k\log T}}{\sqrt{T}} - \frac{4\mu^*}{T} \\
&\geq \mu^* - \frac{32\sqrt{k\log k\log T}}{\sqrt{T}} - \frac{4}{T} \qquad \text{(since } \mu^* \leq 1\text{)}
\end{aligned}$$

Therefore, the Nash regret of the algorithm satisfies $\text{NR}_T = \mu^* - \left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \leq \frac{32\sqrt{k\log k\log T}}{\sqrt{T}} + \frac{4}{T}$. Overall, we get that $\text{NR}_T = O\left(\sqrt{\frac{k\log k\log T}{T}}\right)$. The theorem stands proved.

Remark 1. Algorithm 1 is different from standard UCB, in terms of both design and analysis. For instance, here the empirical means appear in the confidence width and impact the concentration bounds utilized in the analysis.

Remark 2. As mentioned previously, the Nash regret guarantee obtained in Theorem 2.1 can be improved by a factor of $\sqrt{\log k}$. To highlight the key technical insights, in Algorithm 1 we fixed the number of rounds in Phase I (to $\widetilde{T}$). However, with an adaptive approach, one can obtain a Nash regret of $O\left(\sqrt{\frac{k\log T}{T}}\right)$, as stated in Theorem 2.2 (Section 2.4.2). A description of the modified algorithm (Algorithm 2) and the proof of Theorem 2.2 appear in Section 2.4.1.

2.4 Improved and Anytime Guarantees for Nash Regret

This section provides an improved (over Theorem 2.1) Nash regret guarantee for settings in which the horizon of play T is known in advance. Furthermore, here we also develop a Nash regret minimization algorithm for settings in which the horizon of play T is not known in advance. This anytime algorithm (Algorithm 3 in Section 2.4.3) builds upon the standard doubling trick. The algorithm starts with a guess for the time horizon, i.e., a *window* of length $W \in \mathbb{Z}_+$. Then, for W rounds it either (i) performs uniform exploration, with probability $\frac{1}{W^2}$, or (ii) invokes Algorithm 2 as a subroutine (with the remaining probability $\left(1-\frac{1}{W^2}\right)$). This execution for W rounds completes one *epoch* of Algorithm 3. In the subsequent epochs, the algorithm doubles the window length and repeats the same procedure till the end of the time horizon, i.e., till a stop signal is received.

In Section 2.4.1, we detail Algorithm 2, which is called as a subroutine in our anytime algorithm and it takes as input a (guess) window length W. We will also prove that if Algorithm 2 is in fact executed with the actual horizon of play (i.e., executed with $W = T$), then it achieves a Nash regret of $O\left(\sqrt{\frac{k \log T}{T}}\right)$ (Theorem 2.2); this provides the improvement mentioned above.

2.4.1 Modified Nash Confidence Bound Algorithm

Algorithm 2 consists of two phases:

- In Phase *1*, the algorithm performs uniform exploration until the sum of rewards for any arm i exceeds a certain threshold.[1] Specifically, with n_i denoting the number of times an arm i has been pulled so far and $X_{i,s}$ denoting the reward observed for arm i when it is pulled the sth time, the exploration continues as long as $\max_i \sum_{s=1}^{n_i} X_{i,s} \leq 420c^2 \log W$; here c is an absolute constant. Note that this stopping criterion is equivalent to $n_i \, \widehat{\mu}_i \leq 420c^2 \log W$, where $\widehat{\mu}_i$ is the empirical mean for arm i.

- In Phase *2*, the algorithm associates with each arm i the following Nash confined bound value, $\overline{\text{NCB}}_i$, and selects the arm for which that value is the maximized.[2]

$$\overline{\text{NCB}}_i := \widehat{\mu}_i + 2c\sqrt{\frac{2\widehat{\mu}_i \log W}{n_i}} \tag{2.8}$$

Recall that Algorithm 2 is called as a subroutine by our anytime algorithm (Algorithm 3) with a time window (guess) W as input. For the purposes of analysis (see Section 2.4.3 for details), it suffices to obtain guarantees for Algorithm 2 when W is at least $\sqrt{T}$. Hence, this section analyzes the algorithm with the assumption[3] that $\sqrt{T} \leq W \leq T$.

We first define a "good" event E and show that it holds with high probability; our analysis is based on conditioning on E. In particular, we will first define three sub-events E_1, E_2, E_3 and set $E := E_1 \cap E_2 \cap E_3$. For specifying these events, write $\widehat{\mu}_{i,s}$ to denote the empirical mean

[1] Note that this is in contrast to Algorithm 1, in which uniform exploration was performed for a fixed number of rounds.

[2] $\overline{\text{NCB}}_i$ differs from NCB_i (see equation (2.2)) in terms of constants. Specifically, the parameter c is an absolute constant and is fixed in the algorithm.

[3] We will also assume that the optimal mean μ^* is sufficiently greater than $\frac{1}{\sqrt{T}}$. In the complementary case, the stated Nash regret bound follows directly.

Algorithm 2 Modified NCB

Input: Number of arms k and time window W.

```
Initialize empirical means $\widehat{\mu}_i = 0$ and counts $n_i = 0$ for all $i \in [k]$.
Initialize round index $t = 1$ and set parameter $c = 3$.
Phase 1
while $\max_i n_i \, \widehat{\mu}_i \leq 420c^2 \log W$ and $t \leq W$ do
    Select $I_t$ uniformly at random from $[k]$. Pull arm $I_t$ and observe reward $X_t$.
    For arm $I_t$, increment the count $n_{I_t}$ (by one) and update the empirical mean $\widehat{\mu}_{I_t}$.
    Update $t \leftarrow t + 1$.
end while
Phase 2
while $t \leq W$ do
    Pull the arm $I_t$ with the highest Nash confidence bound, i.e., $I_t = \text{argmax}_{i \in [k]} \ \overline{\text{NCB}}_i$.
    Observe reward $X_t$ and update the Nash confidence bound for $I_t$ (see equation (2.8)).
    Update $t \leftarrow t + 1$.
end while
```

of arm i's rewards, based on the first s samples (of i). Also, define

$$S := \frac{c^2 \log T}{\mu^*} \tag{2.9}$$

E_1: For any number of rounds $r \geq 128kS$ and any arm $i \in [k]$, during the first r rounds of uniform sampling, arm i is sampled at least $\frac{r}{2k}$ times and at most $\frac{3r}{2k}$ times.

E_2: For all arms $i \in [k]$, with $\mu_i > \frac{\mu^*}{64}$, and all sample counts $64S \leq s \leq T$ we have $|\mu_i - \widehat{\mu}_{i,s}| \leq c\sqrt{\frac{\mu_i \log T}{s}}$.

E_3: For all arms $j \in [k]$, with $\mu_j \leq \frac{\mu^*}{64}$, and all sample counts $64S \leq s \leq T$, we have $\widehat{\mu}_{j,s} < \frac{\mu^*}{32}$.

Note that these events address a single execution of Algorithm 2 and are expressed in the canonical bandit model [LS20]. Furthermore, they are expressed using the overall horizon of play T. This, in particular, ensures that, irrespective of W, they are well specified.

We first obtain a probability bound for the event E; the proof of the following lemma is deferred to section 2.6.1.

Lemma 2.5. $\mathbb{P}\{E\} \geq 1 - \frac{4}{T}$.

The next lemma shows that, under event E, the total observed reward for any arm is low until certain number of samples. In the final analysis, this result will enable us to bound (under event E) the number of rounds in Phase *1* of Algorithm 2. The proofs of Lemmas 2.6, 2.7, and 2.8 appear in section 2.6.2.

Lemma 2.6. *Under the event E, for any arm i and any sample count $n \leq 192S$, we have $n\ \widehat{\mu}_{i,n} < 210c^2 \log T$.*

Recall that i^* denotes the optimal arm, i.e., $i^* = \arg\max_{i\in[k]} \mu_i$. The following lemma shows that, under event E and after certain number of samples, the total observed reward for i^* is sufficiently large.

Lemma 2.7. *Under the event E, for any sample count $n \geq 484S$, we have $n\ \widehat{\mu}_{i^*,n} \geq 462c^2 \log T$.*

The lemma below will help us in analyzing the termination of the first while-loop (Line 3) of Algorithm 2. Also, recall that in this section we analyze Algorithm 2 with the assumption that $\sqrt{T} \leq W \leq T$.

Lemma 2.8. *Assume that $\sqrt{T} \leq W \leq T$. Also, let random variable τ denote the number of rounds of uniform sampling at which the sum of observed rewards for any arm exceeds $420c^2 \log W$ (i.e., only after τ rounds of uniform sampling we have $\max_i\ n_i\widehat{\mu}_i > 420c^2 \log W$). Then, under event E, the following bounds hold*

$$128\ kS \leq \tau \leq 968\ kS.$$

As mentioned previously, the events E_1, E_2, and E_3 are defined under the canonical bandit model. Hence, Lemmas 2.6, 2.7, and 2.8 also conform to this setup.

Next, we will show that the following key guarantees (events) hold under the good event E:

- Lemma 2.9: The Nash confidence bound of the optimal arm i^* is at least its true mean, μ^*, throughout Phase *2* of Algorithm 2.
- Lemma 2.10: Arms j with sufficiently small means (in particular, $\mu_j \leq \frac{\mu^*}{64}$) are never pulled in Phase *2*.
- Lemma 2.11: Arms i that are pulled many times in Phase *2* have means μ_i close to the optimal μ^*. Hence, such arms i do not significantly increase the Nash regret.

Lemma 2.9. *Let $\overline{\mathrm{NCB}}_{i^*,t}$ be the Nash confidence bound of the optimal arm i^* at any round t in Phase 2. Assume that the good event E holds and $\sqrt{T} \leq W \leq T$. Then, we have $\overline{\mathrm{NCB}}_{i^*,t} \geq \mu^*$.*

Proof. Fix any round t in Phase *2* and write n_{i^*} to denote the number of times the optimal arm i^* has been pulled before that round. Also, let $\widehat{\mu}^*$ denote the empirical mean of arm i^* at round t. Hence, by definition, at this round the Nash confidence bound $\overline{\mathrm{NCB}}_{i^*,t} := \widehat{\mu}^* + 2c\sqrt{\frac{2\widehat{\mu}^* \log W}{n_{i^*}}}$.

Since event E holds, Lemma 2.8 implies that Algorithm 2 must have executed at least $128kS$ rounds in Phase *1* (before switching to Phase *2*): the termination condition of the first while-loop (Line 3) is realized only after $128kS$ rounds.

This lower bound on uniform exploration and event E_1 give us $n_{i^*} \geq 64S$. Therefore, the product $n_{i^*}\mu^* \geq 64c^2 \log T$. This inequality enables us to express the empirical mean of the optimal arm as follows

$$\begin{aligned}
\widehat{\mu}^* &\geq \mu^* - c\sqrt{\frac{\mu^* \log T}{n_{i^*}}} && \text{(since } n_{i^*} \geq 64S \text{ and event } E_2 \text{ holds)}\\
&= \mu^* - c\mu^* \sqrt{\frac{\log T}{\mu^* n_{i^*}}}\\
&\geq \mu^* - c\mu^* \sqrt{\frac{1}{64c^2}} && \text{(since } \mu^*\ n_{i^*} \geq 64c^2 \log T\text{)}\\
&= \frac{7}{8}\mu^*.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\overline{\text{NCB}}_{i^*,t} &= \widehat{\mu}^* + 2c\sqrt{\frac{2\widehat{\mu}^* \log W}{n_{i^*}}}\\
&\geq \widehat{\mu}^* + 2c\sqrt{\frac{\widehat{\mu}^* \log T}{n_{i^*}}} && \text{(since } 2\log W \geq \log T\text{)}\\
&\geq \mu^* - c\sqrt{\frac{\mu^* \log T}{n_{i^*}}} + 2c\sqrt{\frac{\widehat{\mu}^* \log T}{n_{i^*}}} && \text{(due to the event } E_2\text{)}\\
&\geq \mu^* - c\sqrt{\frac{\mu^* \log T}{n_{i^*}}} + 2c\sqrt{\frac{7\mu^* \log T}{8n_{i^*}}} && \text{(since } \widehat{\mu}^* \geq \tfrac{7}{8}\mu^*\text{)}\\
&\geq \mu^*
\end{aligned}$$

The lemma stands proved. □

Lemma 2.10. *Assume that the good event E holds and $\sqrt{T} \leq W \leq T$. Then, any arm j, with mean $\mu_j \leq \frac{\mu^*}{64}$, is never pulled in all of Phase 2.*

Proof. Fix any arm j with mean $\mu_j \leq \frac{\mu^*}{64}$. Let r_j denote the number of times arm j is pulled in Phase *1*.

We will fist show that $r_j \geq 64S$. Since event E holds, Lemma 2.8 ensures that Algorithm 2 must have executed at least $128kS$ rounds in Phase *1* (before switching to Phase *2*). This lower bound on uniform exploration and event E_1 give us $r_j \geq 64S$.

Furthermore, event E_3 and the fact that $r_j \geq 64S$ imply that (throughout Phase *2*) the empirical mean of arm j satisfies $\widehat{\mu}_j \leq \frac{\mu^*}{32}$.

For any round t in Phase *2*, write $\overline{\text{NCB}}_{j,t}$ to denote the Nash confidence bound of arm j at round t. Below we show that the $\overline{\text{NCB}}_{j,t}$ is strictly less than $\overline{\text{NCB}}_{i^*,t}$ and, hence, arm j is not even pulled once in all of Phase *2*.

$$\begin{aligned}
\overline{\text{NCB}}_{j,t} &= \widehat{\mu}_j + 2c\sqrt{\frac{2\widehat{\mu}_j \log W}{r_j}} \\
&\leq \widehat{\mu}_j + 2c\sqrt{\frac{2\widehat{\mu}_j \log T}{r_j}} && \text{(since } \log W \leq \log T\text{)} \\
&\leq \frac{\mu^*}{32} + 2c\sqrt{\frac{\mu^* \log T}{16 r_j}} && \text{(since } \widehat{\mu}_j \leq \tfrac{\mu^*}{32}\text{)} \\
&\leq \frac{\mu^*}{32} + \frac{c}{2}\sqrt{\frac{\mu^* \log T}{64S}} && \text{(since } r_j \geq 64S\text{)} \\
&\leq \frac{\mu^*}{32} + \frac{\mu^*}{16} \\
&= \frac{3}{32}\mu^* \\
&< \overline{\text{NCB}}_{i^*,t} && \text{(via Lemma 2.9)}
\end{aligned}$$

This completes the proof of the lemma. □

Lemma 2.11. *Assume that the good event E holds and $\sqrt{T} \leq W \leq T$. Then, for any arm i that is pulled at least once in Phase 2 we have $\mu_i \geq \mu^* - 4c\sqrt{\frac{\mu^* \log T}{T_i - 1}}$, where T_i is the total number of times that arm i is pulled in Algorithm 2.*

Proof. Fix any arm i that is pulled at least once in Phase *2*. When arm i was pulled the T_ith time during Phase *2*, it must have had the maximum Nash confidence bound value; in particular, at that round $\overline{\text{NCB}}_i \geq \overline{\text{NCB}}_{i^*} \geq \mu^*$; the last inequality follows from Lemma 2.9. Therefore, we have

$$\widehat{\mu}_i + 2c\sqrt{\frac{2\widehat{\mu}_i \log T}{T_i - 1}} \geq \mu^* \tag{2.10}$$

Here, $\widehat{\mu}_i$ denotes the empirical mean of arm i at this point.

As argued in the proof of Lemmas 2.9 and 2.10, event E ensures that any arm that is pulled in Phase *2* is sampled at least $64S$ times in Phase *1*. Hence, in particular, we have $T_i > 64S$. In addition, since arm i is pulled at least once in Phase *2*, Lemma 2.10 implies that $\mu_i > \frac{\mu^*}{64}$.

Now, complementing inequality (2.10), we will now upper bound the empirical mean $\widehat{\mu}_i$ in terms of μ^*. Specifically,

$$\begin{aligned}
\widehat{\mu}_i &\le \mu_i + c\sqrt{\frac{\mu_i \log T}{T_i - 1}} && \left(\text{since } \mu_i > \tfrac{\mu^*}{64} \text{ and event } E_2 \text{ holds}\right)\\
&\le \mu^* + c\sqrt{\frac{\mu^* \log T}{64S}} && (\text{since } T_i > 64S \text{ and } \mu_i \le \mu^*)\\
&= \mu^* + \frac{\mu^*}{8} && \left(\text{since } S = \tfrac{c^2 \log T}{\mu^*}\right)\\
&= \frac{9}{8}\mu^* && (2.11)
\end{aligned}$$

Inequalities (2.10) and (2.11) give us

$$\begin{aligned}
\mu^* &\le \widehat{\mu}_i + 2c\sqrt{\frac{9\mu^* \log T}{4(T_i - 1)}}\\
&\le \mu_i + c\sqrt{\frac{\mu_i \log T}{T_i - 1}} + 3c\sqrt{\frac{\mu^* \log T}{T_i - 1}} && (\text{via event } E_2)\\
&\le \mu_i + c\sqrt{\frac{\mu^* \log T}{T_i - 1}} + 3c\sqrt{\frac{\mu^* \log T}{T_i - 1}} && (\text{since } \mu_i \le \mu^*)\\
&\le \mu_i + 4c\sqrt{\frac{\mu^* \log T}{T_i - 1}}.
\end{aligned}$$

This completes the proof of the lemma. □

Using the above-mentioned lemmas, we will now establish an essential bound on the Nash social welfare of Algorithm 2.

Lemma 2.12. *Consider a bandit instance with optimal mean* $\mu^* \ge \frac{512\sqrt{k \log T}}{\sqrt{T}}$ *and assume that* $\sqrt{T} \le W \le T$. *Then, for any* $w \le W$, *we have*

$$\left(\prod_{t=1}^{w} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \ge \left(\mu^*\right)^{\frac{w}{T}} \left(1 - 1000c\sqrt{\frac{k \log T}{\mu^* T}}\right).$$

Proof. First, for the expected rewards $\mathbb{E}\left[\mu_{I_t}\right]$ (of Algorithm 2), we will derive a lower bound that holds for all rounds t. In Algorithm 2, for any round t (i.e., $t \leq W$), write p_t to denote the probability that the algorithm is in Phase *1* and, hence, with probability $(1-p_t)$ the algorithm is in Phase *2*. That is, with probability p_t the algorithm is selecting an arm uniformly at random and receiving an expected reward of at least $\frac{\mu^*}{k}$. Complementarily, if the algorithm is in Phase *2* at round t, then its expected reward is at least $\frac{\mu^*}{64}$ (Lemma 2.10). These observations give us

$$\begin{aligned} \mathbb{E}[\mu_{I_t}] &\geq \mathbb{E}\left[\mu_{I_t} | E\right] \; \mathbb{P}\{E\} \\ &\geq \mathbb{E}\left[\mu_{I_t} | E\right] \left(1 - \frac{4}{T}\right) && \text{(Lemma 2.5)} \\ &= \left(1 - \frac{4}{T}\right) \left(p_t \frac{\mu^*}{k} + (1-p_t)\frac{\mu^*}{64}\right) \\ &\geq \left(1 - \frac{4}{T}\right) \left(\frac{\mu^*}{64k}\right) && (2.12) \end{aligned}$$

Towards a case analysis, define threshold $\overline{T} := 968 \; kS$. We observe that Lemma 2.8 ensures that by the $\overline{T}$th round Algorithm 2 would have completed Phase *1*; in particular, the termination condition of the first while-loop (Line 3) in the algorithm would be met by the $\overline{T}$th round. Also, note that, under the lemma assumption on μ^* and for an appropriately large T we have

$$\frac{\overline{T} \; \log(64k)}{T} = \frac{968 \; kS \; \log(64k)}{T} = \frac{968 \; kc^2 \log T \; \log(64k)}{\mu^* T} \leq \frac{968 c^2 \log(64k) \sqrt{k \log T}}{512\sqrt{T}} \leq 1 \tag{2.13}$$

We establish the lemma considering two complementary and exhaustive cases based on the given round index w:
Case 1: $w \leq \overline{T}$, and
Case 2: $w > \overline{T}$.

For *Case 1* ($w \leq \overline{T}$), using inequality (2.12) we obtain

$$\begin{aligned} \left(\prod_{t=1}^{w} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &\geq \left(1 - \frac{4}{T}\right)^{\frac{w}{T}} \left(\frac{\mu^*}{64k}\right)^{\frac{w}{T}} \\ &\geq \left(1 - \frac{4}{T}\right) (\mu^*)^{\frac{w}{T}} \left(\frac{1}{64k}\right)^{\frac{w}{T}} \\ &= \left(1 - \frac{4}{T}\right) (\mu^*)^{\frac{w}{T}} \left(\frac{1}{2}\right)^{\frac{w \;\; \log(64k)}{T}} \end{aligned}$$

$$\begin{aligned}
&\geq \left(1-\frac{4}{T}\right)(\mu^*)^{\frac{w}{T}}\left(\frac{1}{2}\right)^{\frac{\overline{T}\ \log(64k)}{T}} && \text{(since } w \leq \overline{T}\text{)}\\
&= \left(1-\frac{4}{T}\right)(\mu^*)^{\frac{w}{T}}\left(1-\frac{1}{2}\right)^{\frac{\overline{T}\ \log(64k)}{T}}\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-\frac{\overline{T}\log(64k)}{T}\right)\left(1-\frac{4}{T}\right) && \text{(via inequality (2.13) and Claim 2.1)}\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-\frac{\overline{T}\log(64k)}{T}-\frac{4}{T}\right)\\
&= (\mu^*)^{\frac{w}{T}}\left(1-\frac{968c^2k\log T\ \log(64k)}{\mu^*T}-\frac{4}{T}\right)\\
&= (\mu^*)^{\frac{w}{T}}\left(1-\frac{968c\sqrt{k\log T}}{\sqrt{\mu^*T}}\cdot\frac{c\log(64k)\sqrt{k\log T}}{\sqrt{\mu^*T}}-\frac{4}{T}\right)\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-1000c\sqrt{\frac{k\log T}{\mu^*T}}\right) && (2.14)
\end{aligned}$$

The last inequality follows from the fact that $\frac{c\log(64k)\sqrt{k\log T}}{\sqrt{\mu^*T}} \leq 1$ for an appropriately large T; recall that $\mu^* \geq \frac{512\sqrt{k\log T}}{\sqrt{T}}$.

For *Case 2* ($w > \overline{T}$), we partition the Nash social welfare into two terms:

$$\left(\prod_{t=1}^{w}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} = \left(\prod_{t=1}^{\overline{T}}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}}\left(\prod_{t=\overline{T}+1}^{w}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \tag{2.15}$$

In this product, the two terms account for the rewards accrued in rounds $t \leq \overline{T}$ and in rounds $\overline{T} < t \leq w$, respectively. We will now lower bound these two terms separately.

The first term in the right-hand side of equation (2.15) can be bounded as follows

$$\begin{aligned}
\left(\prod_{t=1}^{\overline{T}}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &\geq \left(1-\frac{4}{T}\right)^{\frac{\overline{T}}{T}}\left(\frac{\mu^*}{64k}\right)^{\frac{\overline{T}}{T}} && \text{(via inequality (2.12))}\\
&\geq \left(1-\frac{4}{T}\right)(\mu^*)^{\frac{\overline{T}}{T}}\left(\frac{1}{64k}\right)^{\frac{\overline{T}}{T}}\\
&= \left(1-\frac{4}{T}\right)(\mu^*)^{\frac{\overline{T}}{T}}\left(\frac{1}{2}\right)^{\frac{\overline{T}\log(64k)}{T}}
\end{aligned}$$

$$= \left(1 - \frac{4}{T}\right) (\mu^*)^{\frac{\overline{T}}{T}} \left(1 - \frac{1}{2}\right)^{\frac{\overline{T}\log(64k)}{T}}$$
$$\geq (\mu^*)^{\frac{\overline{T}}{T}} \left(1 - \frac{\overline{T}\log(64k)}{T}\right) \left(1 - \frac{4}{T}\right) \tag{2.16}$$

For establishing the last inequality we note that the exponent $\frac{\overline{T}\log(64k)}{T} \leq 1$ (see inequality (2.13)) and apply Claim 2.1.

For the second term in the right-hand side of equation (2.15), we have

$$\left(\prod_{t=\overline{T}+1}^{w} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq \mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{w} \mu_{I_t}\right)^{\frac{1}{T}}\right] \qquad \text{(Multivariate Jensen's inequality)}$$
$$\geq \mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{w} \mu_{I_t}\right)^{\frac{1}{T}} \middle| E\right] \mathbb{P}\{E\} \tag{2.17}$$

As mentioned previously, Lemma 2.8 ensures that by the $\overline{T}$th round Algorithm 2 would have completed Phase *1*. Hence, any round $t > \overline{T}$ falls under Phase *2*. Now, to bound the expected value in the right-hand-side of inequality (2.17), we consider the arms that are pulled at least once after the first $\overline{T}$ rounds. In particular, with reindexing, let $\{1, 2, \dots, \ell\}$ denote the set of all arms that are pulled at least once after the first $\overline{T}$ rounds; note that these ℓ arms are in fact pulled in Phase *2*. Also, let $m_i \geq 1$ denote the number of times arm $i \in [\ell]$ is pulled after the first $\overline{T}$ rounds and note that $\sum_{i=1}^{\ell} m_i = w - \overline{T}$. Furthermore, let T_i denote the total number of times any arm i is pulled in the algorithm. Indeed, $(T_i - m_i)$ is the number of times arm $i \in [\ell]$ is pulled during the first $\overline{T}$ rounds. With this notation, the expected value in the right-hand-side of inequality (2.17) can be expressed as $\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \middle| E\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell} \mu_i^{\frac{m_i}{T}}\right) \middle| E\right]$. Moreover, since we are conditioning on the good event E, Lemma 2.11 applies to each arm $i \in [\ell]$. Hence,

$$\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{w} \mu_{I_t}\right)^{\frac{1}{T}} \middle| E\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell} \mu_i^{\frac{m_i}{T}}\right) \middle| E\right]$$
$$\geq \mathbb{E}\left[\prod_{i=1}^{\ell} \left(\mu^* - 4c\sqrt{\frac{\mu^* \log T}{T_i - 1}}\right)^{\frac{m_i}{T}} \middle| E\right] \qquad \text{(Lemma 2.11)}$$

$$= (\mu^*)^{\frac{w-\overline{T}}{T}} \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - 4c\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right)^{\frac{m_i}{T}} \;\middle|\; E\right] \tag{2.18}$$

For the last equality, we use $\sum_{i=1}^{\ell} m_i = w - \overline{T}$. Now under the good event E, recall that each arm is pulled at least $64S$ times during the first $\overline{T}$ rounds. Hence, $T_i > 64S$, for each arm $i \in [\ell]$, and we have $4c\sqrt{\frac{\log T}{\mu^*(T_i-1)}} \leq 4c\sqrt{\frac{\log T}{64c^2 \log T}} = \frac{1}{2}$ for each $i \in [\ell]$. Therefore, we can apply Claim 2.1 to reduce the expected value in inequality (2.18) as follows

$$\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - 4c\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right)^{\frac{m_i}{T}} \;\middle|\; E\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{8c\; m_i}{T}\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right) \;\middle|\; E\right]$$

$$\geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{8c}{T}\sqrt{\frac{m_i \log T}{\mu^*}}\right) \;\middle|\; E\right] \qquad (\text{since } T_i \geq m_i + 1)$$

We can further simplify the above inequality by noting that $(1-x)(1-y) \geq 1 - x - y$ for all $x, y \geq 0$.

$$\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{8c}{T}\sqrt{\frac{m_i \log T}{\mu^*}}\right) \;\middle|\; E\right] \geq \mathbb{E}\left[1 - \sum_{i=1}^{\ell}\left(\frac{8c}{T}\sqrt{\frac{m_i \log T}{\mu^*}}\right) \;\middle|\; E\right]$$

$$= 1 - \left(\frac{8c}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sum_{i=1}^{\ell}\sqrt{m_i} \;\middle|\; E\right]$$

$$\geq 1 - \left(\frac{8c}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell}\,\sqrt{\sum_{i=1}^{\ell} m_i} \;\middle|\; E\right]$$

(Cauchy-Schwarz inequality)

$$\geq 1 - \left(\frac{8c}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell\, T} \;\middle|\; E\right] \qquad (\text{since } \textstyle\sum_i m_i \leq T)$$

$$= 1 - \left(8c\sqrt{\frac{\log T}{\mu^* T}}\right)\mathbb{E}\left[\sqrt{\ell} \;\middle|\; E\right]$$

$$\geq 1 - \left(8c\sqrt{\frac{k \log T}{\mu^* T}}\right) \qquad (\text{since } \ell \leq k)$$

Using this bound, along with inequalities (2.17), and (2.18), we obtain

$$\left(\prod_{t=\overline{T}+1}^{w}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq (\mu^*)^{\frac{w-\overline{T}}{T}}\left(1-8c\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{E\} \tag{2.19}$$

Inequalities (2.19) and (2.16) provide relevant bounds for the two terms in equation (2.15), respectively. Hence, for the Nash social welfare of the algorithm, we have

$$\begin{aligned}
\left(\prod_{t=1}^{w}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq (\mu^*)^{\frac{w}{T}}\left(1-\frac{\overline{T}\cdot\log(64k)}{T}-\frac{4}{T}\right)\left(1-8c\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{E\}\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-\frac{\overline{T}\cdot\log(64k)}{T}-\frac{4}{T}\right)\left(1-8c\sqrt{\frac{k\log T}{\mu^* T}}\right)\left(1-\frac{4}{T}\right)\\
&\qquad\qquad\text{(via Lemma 2.5)}\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-\frac{\overline{T}\cdot\log(64k)}{T}-8c\sqrt{\frac{k\log T}{\mu^* T}}-\frac{8}{T}\right)\\
&= (\mu^*)^{\frac{w}{T}}\left(1-\frac{968c^2\cdot k\log T\cdot\log(64k)}{\mu^* T}-8c\sqrt{\frac{k\log T}{\mu^* T}}-\frac{8}{T}\right)\\
&\geq (\mu^*)^{\frac{w}{T}}\left(1-1000c\sqrt{\frac{k\log T}{\mu^* T}}\right).
\end{aligned}$$

Here, the final inequality follows along the lines of the last step in the derivation of (2.14). The lemma stands proved. □

2.4.2 Improved Guarantee for Nash Regret

Algorithm 2 not only serves as a subroutine in our anytime algorithm (Algorithm 3 in Section 2.4.3), it also provides an improved (over Theorem 2.1) Nash regret guarantee for settings in which the horizon of play T is known in advance. In particular, invoking Algorithm 2 with $W = T$ we obtain Theorem 2.2 (stated next).

Theorem 2.2. *For any bandit instance with k arms and given any (moderately large) T, there exists an algorithm that achieves Nash regret*

$$\mathrm{NR}_T = O\left(\sqrt{\frac{k\log T}{T}}\right).$$

Proof. The stated Nash regret guarantee follows directly by applying Lemma 2.12 with $w = T$. Specifically, $\mathrm{NR}_T \leq \mu^* - (\mu^*)^{\frac{T}{T}} \left(1 - 1000c\sqrt{\frac{k \log T}{\mu^* T}}\right) = 1000c\sqrt{\frac{\mu^* k \log T}{T}} \leq 1000c\sqrt{\frac{k \log T}{T}}$. This completes the proof of the theorem. □

2.4.3 Anytime Algorithm

As mentioned previously, our anytime algorithm (Algorithm 3) builds upon the standard doubling trick. The algorithm starts with a guess for the time horizon, i.e., a window of length $W \in \mathbb{Z}_+$. Then, for W rounds it either (i) performs uniform exploration, with probability $\frac{1}{W^2}$, or (ii) invokes Algorithm 2 as a subroutine (with the remaining probability $\left(1 - \frac{1}{W^2}\right)$). This execution for W rounds completes one *epoch* of Algorithm 3. In the subsequent epochs, the algorithm doubles the window length and repeats the same procedure till the end of the time horizon, i.e., till a stop signal is received.

Algorithm 3 Anytime Algorithm for Nash Regret

```
Input: Number of arms $k$
Initialize $W = 1$.
while the MAB process continues do
    With probability $\frac{1}{W^2}$ set flag = UNIFORM , otherwise, with probability $\left(1 - \frac{1}{W^2}\right)$, set flag = NCB
    if flag = UNIFORM then
        for $t = 1$ to $W$ do
            Select $I_t$ uniformly at random from $[k]$. Pull arm $I_t$ and observe reward $X_t$.
        end for
    else if flag = NCB then
        Execute Modified NCB($k$, $W$).
    end if
    Update $W \leftarrow 2 \times W$.
end while
```

Algorithm 3 gives us Theorem 2.3 (stated next and proved in Section 2.4.4).

Theorem 2.3. *There exists an anytime algorithm that, at any (moderately large) round T, achieves a Nash regret*

$$\mathrm{NR}_T = O\left(\sqrt{\frac{k \log T}{T}} \log T\right).$$

We will, throughout, use h to denote an epoch index in Algorithm 3 and the corresponding window length as W_h. Note that $h = \log W_h + 1$. Also, let e denote the total number of epochs during the T rounds of Algorithm 3. We have $e \leq \log_2 T + 1$. Furthermore, write R_h to denote the number of rounds before the hth epoch begins, i.e., $R_h = \sum_{z=1}^{h-1} W_z$. Note that $R_1 = 0$ and the hth epoch starts at round $(R_h + 1)$. In addition, let h^* be the smallest value of h for which $W_h \geq \sqrt{T}$, i.e., h^* is the first epoch in which the window length is at least $\sqrt{T}$.

We next provide three claims connecting these constructs.

Claim 2.2. *$W_h = R_h + 1$.*

Proof. In Algorithm 3, the window size doubles after each epoch, i.e., $W_z = 2^{z-1}$ for all epochs $z \geq 1$. Therefore, $R_h = \sum_{z=1}^{h-1} W_z = \sum_{z=1}^{h-1} 2^{z-1} = 2^{h-1} - 1 = W_h - 1$. Hence, we have $W_h = R_h + 1$. □

The next claim notes that the window length in the last epoch, e, is at most the horizon of play T.

Claim 2.3. *$W_e \leq T$.*

Proof. The definition of R_h implies that each epoch h starts at round $R_h + 1$. Hence, the last epoch e, in particular, starts at round $R_e + 1$. Indeed, $R_e + 1 \leq T$ and, via Claim 2.2, we get that $W_e \leq T$. □

Recall that h^* denotes the smallest value of h for which $W_h \geq \sqrt{T}$.

Claim 2.4. *$R_{h^*} < 2\sqrt{T}$.*

Proof. By definition of h^*, we have $W_{h^*-1} < \sqrt{T}$. Also, note that $R_{h^*} = R_{h^*-1} + W_{h^*-1} = 2W_{h^*-1} - 1 < 2\sqrt{T}$; here, the last equality follows from Claim 2.2. □

The following lemma provides a bound on the Nash social welfare accumulated by Algorithm 3 in an epoch $h \geq h^*$.

Lemma 2.13. *In any MAB instance with mean $\mu^* \geq \frac{512\sqrt{k \log T}}{\sqrt{T}}$, the following inequality holds for each epoch $h \geq h^*$ and all rounds $r \in \{R_h + 1, R_h + 2, \ldots, R_{h+1}\}$:*

$$\left(\prod_{t=R_h+1}^{r} \mathbb{E}\left[\mu_{I_t}\right] \right)^{\frac{1}{T}} \geq (\mu^*)^{\frac{r-R_h}{T}} \left(1 - 1001c\sqrt{\frac{k \log T}{\mu^* T}} \right).$$

Proof. Fix any epoch $h \geq h^*$ and let F_h denote the event that Algorithm 3 executes the Modified NCB algorithm (Algorithm 2) in epoch h; see Line 9. Note that $\mathbb{P}\{F_h\} = 1 - \frac{1}{W_h^2}$. The definition of h^* and Claim 2.3 give us $\sqrt{T} \leq W_h \leq T$. Hence, we can apply Lemma 2.12 with event F_h to obtain:

$$\begin{aligned}
\left(\prod_{t=R_h+1}^{r} \mathbb{E}\left[\mu_{I_t}\right] \right)^{\frac{1}{T}} &\geq \left(\prod_{t=R_h+1}^{r} \mathbb{E}\left[\mu_{I_t} | F_h\right] \ \mathbb{P}\{F_h\} \right)^{\frac{1}{T}} \\
&\geq \left(\prod_{t=R_h+1}^{r} \mathbb{E}\left[\mu_{I_t} | F_h\right] \left(1 - \frac{1}{W_h^2}\right) \right)^{\frac{1}{T}} \\
&\geq \left(1 - \frac{1}{W_h^2}\right) \left(\prod_{t=R_h+1}^{r} \mathbb{E}\left[\mu_{I_t} | F_h\right] \right)^{\frac{1}{T}} \\
&\geq (\mu^*)^{\frac{r-R_h}{T}} \left(1 - \frac{1}{W_h^2}\right) \left(1 - 1000c\sqrt{\frac{k \log T}{\mu^* T}}\right) && \text{(Lemma 2.12)} \\
&\geq (\mu^*)^{\frac{r-R_h}{T}} \left(1 - \frac{1}{T}\right) \left(1 - 1000c\sqrt{\frac{k \log T}{\mu^* T}}\right) && \text{(since } W_h \geq \sqrt{T}\text{)} \\
&\geq (\mu^*)^{\frac{r-R_h}{T}} \left(1 - 1000c\sqrt{\frac{k \log T}{\mu^* T}} - \frac{1}{T}\right) \\
&\geq (\mu^*)^{\frac{r-R_h}{T}} \left(1 - 1001c\sqrt{\frac{k \log T}{\mu^* T}}\right).
\end{aligned}$$

The lemma stands proved □

2.4.4 Proof of Theorem 2.3

For establishing the theorem, we focus on MAB instances in which the optimal mean $\mu^* \geq \frac{512\sqrt{k \log T}}{\sqrt{T}}$; otherwise, the stated guarantees on the Nash regret directly holds.

Recall that h^* denotes the smallest value of h for which $W_h \geq \sqrt{T}$. We will bound the Nash social welfare accrued by Algorithm 3 by first considering the initial R_{h^*} rounds and then separately analyzing the remaining rounds.

For the first R_{h^*} rounds, note that for every epoch $g \leq h^*$ we have $W_g < \sqrt{T}$. Hence, for each such epoch g, Algorithm 3 executes uniform sampling with probability $\frac{1}{W_g^2} \geq \frac{1}{T}$; see Line 6. Therefore, for all rounds $t \leq R_{h^*}$, we have $\mathbb{E}\left[\mu_{I_t}\right] \geq \frac{\mu^*}{k}\frac{1}{T}$. This bound gives us

$$\left(\prod_{t=1}^{R_{h^*}} \mathbb{E}\left[\mu_{I_t}\right] \right)^{\frac{1}{T}} \geq \left(\frac{\mu^*}{kT} \right)^{\frac{R_{h^*}}{T}}$$

$$
\begin{aligned}
&= (\mu^*)^{\frac{R_{h^*}}{T}} \left(\frac{1}{2}\right)^{\frac{R_{h^*}\log(kT)}{T}} \\
&\geq (\mu^*)^{\frac{R_{h^*}}{T}} \left(1 - \frac{R_{h^*}\log(kT)}{T}\right) && \text{(via Claim 2.1)} \\
&\geq (\mu^*)^{\frac{R_{h^*}}{T}} \left(1 - \frac{2\log(kT)}{\sqrt{T}}\right) && (2.20)
\end{aligned}
$$

Here, the last inequality follows from Claim 2.4.

For the remaining $T - R_{h^*}$ rounds, we perform an epoch-wise analysis and invoke Lemma 2.13. Specifically,

$$
\begin{aligned}
\left(\prod_{t=R_{h^*}+1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &= \left(\prod_{h=h^*}^{e-1} \prod_{t=R_h+1}^{R_{(h+1)}} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \cdot \left(\prod_{t=R_e+1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \\
&= \prod_{h=h^*}^{e-1} \left(\prod_{t=R_h+1}^{R_h+W_h} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \cdot \left(\prod_{t=R_e+1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \\
&\geq \prod_{h=h^*}^{e-1} (\mu^*)^{\frac{W_h}{T}} \left(1 - 1001c\sqrt{\frac{k\log T}{\mu^* T}}\right) \cdot (\mu^*)^{\frac{T-R_e}{T}} \left(1 - 1001c\sqrt{\frac{k\log T}{\mu^* T}}\right) \\
& && \text{(via Lemma 2.13)} \\
&\geq (\mu^*)^{1-\frac{R_{h^*}}{T}} \prod_{j=1}^{e} \left(1 - 1001c\sqrt{\frac{k\log T}{\mu^* T}}\right) \\
&\geq (\mu^*)^{1-\frac{R_{h^*}}{T}} \left(1 - 1001c\sqrt{\frac{k\log T}{\mu^* T}}\right)^{\log(2T)} && \text{(since } e \leq \log T + 1\text{)} \\
&\geq (\mu^*)^{1-\frac{R_{h^*}}{T}} \left(1 - 1001c\frac{\sqrt{k\log T}\log(2T)}{\sqrt{\mu^* T}}\right) && (2.21)
\end{aligned}
$$

The last inequality follows from the fact that $(1-x)(1-y) \geq 1 - x - y$, for all $x, y \geq 0$.

Inequalities (2.20) and (2.21), give us an overall bound on the Nash social welfare of Algorithm 3:

$$
\begin{aligned}
\left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &= \left(\prod_{t=1}^{R_{h^*}} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \left(\prod_{t=R_{h^*}+1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \\
&\geq \mu^* \left(1 - \frac{2\log(kT)}{\sqrt{T}}\right) \left(1 - 1001c\frac{\sqrt{k\log T}\log(2T)}{\sqrt{\mu^* T}}\right)
\end{aligned}
$$

$$\geq \mu^* \left(1 - \frac{2\log(kT)}{\sqrt{T}} - 1001c\frac{\sqrt{k\log T}\log(2T)}{\sqrt{\mu^* T}}\right).$$

Therefore, the Nash regret of Algorithm 3 satisfies

$$\mathrm{NR}_T = \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \leq \frac{2\mu^*\log(kT)}{\sqrt{T}} + 1001c\frac{\sqrt{\mu^* k\log T}\log(2T)}{\sqrt{T}} \leq 1003c\frac{\sqrt{k\log T}\log(2T)}{\sqrt{T}}.$$

The theorem stands proved.

2.5 Missing Proofs from Section 2.3.1

2.5.1 Proof of Lemma 2.1

In this section we prove that the good event G holds with probability at least $\left(1 - \frac{4}{T}\right)$. Towards this, we first state two standard concentration inequalities, Lemmas 2.14 and 2.15; see, e.g., [DP09, Chapter 1.6].

Lemma 2.14 (Hoeffding's Inequality)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables distributed in $[0,1]$. Consider their average $\widehat{Y} \coloneqq \frac{Y_1+\ldots+Y_n}{n}$ and let $\nu = \mathbb{E}[\widehat{Y}]$ be its expected value. Then, for any $0 \leq \delta \leq 1$,*

$$\mathbb{P}\left\{\left|\widehat{Y} - \nu\right| \geq \delta\nu\right\} \leq 2\,\exp\left(-\frac{\delta^2}{3}\,n\nu\right).$$

Lemma 2.15 (Hoeffding Extension)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables distributed in $[0,1]$. Consider their average $\widehat{Y} \coloneqq \frac{Y_1+\ldots+Y_n}{n}$ and suppose $\mathbb{E}[\widehat{Y}] \leq \nu_H$. Then, for any $0 \leq \delta \leq 1$,*

$$\mathbb{P}\left\{\widehat{Y} \geq (1+\delta)\,\nu_H\right\} \leq \ \exp\left(-\frac{\delta^2}{3} n\nu_H\right).$$

Using these concentration bounds, we establish two corollaries for the empirical means of the arms.

Corollary 2.1. *Consider any arm i, with mean $\mu_i > \frac{6\sqrt{k\log k\log T}}{\sqrt{T}}$, and sample count n such that $\frac{\widetilde{T}}{2k} \leq n \leq T$. Let $\widehat{\mu}_i$ be the empirical mean of arm i's rewards, based on n independent* *draws. Then,*

$$\mathbb{P}\left\{|\mu_i - \widehat{\mu}_i| \geq 3\sqrt{\frac{\mu_i \log T}{n}}\right\} \leq \frac{2}{T^3}.$$

Proof. We apply Lemma 2.14 (Hoeffding's inequality), with Y_s as the sth independent sample from arm i and for all $1 \leq s \leq n$. Also, we instantiate the lemma with $\delta = 3\sqrt{\frac{\log T}{\mu_i n}}$ and note that $\delta < 1$, since $\mu_i > \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$ and $n \geq \frac{\widetilde{T}}{2k} = \frac{8}{k}\sqrt{\frac{kT \log T}{\log k}}$. Specifically,

$$\begin{aligned}
\mathbb{P}\left\{|\mu_i - \widehat{\mu}_i| \geq 3\sqrt{\frac{\mu_i \log T}{n}}\right\} &= \mathbb{P}\left\{|\mu_i - \widehat{\mu}_i| \geq 3\sqrt{\frac{\log T}{\mu_i n}}\ \mu_i\right\} \\
&\leq 2\ \exp\left(-\frac{9 \log T}{3\mu_i n}\ n\mu_i\right) && \text{(via Lemma 2.14)} \\
&= 2\ \exp\left(-3 \log T\right) \\
&= \frac{2}{T^3}.
\end{aligned}$$

□

Corollary 2.2. *Consider any arm j, with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, and sample count n, such that $\frac{\widetilde{T}}{2k} \leq n \leq T$. Let $\widehat{\mu}_j$ be the empirical mean of arm j's rewards, based on n independent draws. Then,*

$$\mathbb{P}\left\{\widehat{\mu}_j \geq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}}\right\} \leq \frac{1}{T^3}.$$

Proof. We invoke Lemma 2.15, with $\delta = 1/2$ and $\nu_H = \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, to obtain

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\mu}_j \geq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}}\right\} &= \mathbb{P}\left\{\widehat{\mu}_j \geq (1 + \frac{1}{2})\nu_H\right\} \\
&\leq \exp\left(-\frac{1}{12}\ n\ \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}\right) \\
&\leq \exp\left(-\frac{1}{12}\ \frac{8}{k}\sqrt{\frac{kT \log T}{\log k}}\ \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}\right) \\
& \qquad \text{(since } n \geq \tfrac{8}{k}\sqrt{\tfrac{kT \log T}{\log k}}\text{)} \\
&= \exp\left(-4 \log T\right) \\
&\leq \frac{1}{T^3}.
\end{aligned}$$

□

Along with Corollary 2.1 and 2.2, we will invoke the Chernoff Bound (stated next).

Lemma 2.16 (Chernoff Bound)**.** *Let $Z_1, \ldots, Z_n$ be independent Bernoulli random variables. Consider the sum $S = \sum_{r=1}^{n} Z_r$ and let $\nu = \mathbb{E}[S]$ be its expected value. Then, for any $\varepsilon \in [0,1]$, we have*

$$\mathbb{P}\left\{S \leq (1-\varepsilon)\nu\right\} \leq \exp\left(-\frac{\nu\varepsilon^2}{2}\right), \quad \textit{and}$$
$$\mathbb{P}\left\{S \geq (1+\varepsilon)\nu\right\} \leq \exp\left(-\frac{\nu\varepsilon^2}{3}\right).$$

We now prove Lemma 2.1 by bounding the probabilities of the three sub-events G_1, G_2, and G_3, respectively. Recall that $G = G_1 \cap G_2 \cap G_3$.

For G_1^c (i.e., the complement of G_1), we will invoke Lemma 3.11 for every arm and then apply the union bound. In particular, fix any arm i and write random variable Z_r to indicate whether arm i is selected in round r of Phase I, or not. That is, $Z_r = 1$ if arm i is picked in round r and, otherwise, $Z_r = 0$. Note that n_i, the number of times arm i is sampled in Phase I, satisfies $n_i = \sum_{r=1}^{\widetilde{T}} Z_r$. Now, using the fact that the algorithm selects an arm uniformly at random in every round of Phase I and setting $\varepsilon = 1/2$ along with $\nu = \frac{\widetilde{T}}{k} = 16\sqrt{\frac{T \log T}{k \log k}}$, Lemma 3.11 gives us

$$\mathbb{P}\left\{n_i < \frac{\widetilde{T}}{2k}\right\} \leq \exp\left(-\frac{16\sqrt{T}\log T}{8\sqrt{k}\log k}\right) \leq \frac{1}{T^2} \tag{2.22}$$

Here, the last inequality follows from the theorem assumption that T is sufficiently large; specifically, $T \geq (k \log k)^2$ suffices. Inequality (2.22) and the union bound give us

$$\mathbb{P}\left\{G_1^c\right\} \leq \frac{1}{T^2}k \leq \frac{1}{T} \tag{2.23}$$

Next, we address G_2^c. Note that the arms and counts considered in G_2, respectively, satisfy the assumption in Corollary 2.1. Hence, the corollary ensures that, for each arm i, with mean $\mu_i > \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$ and each count $s \geq \frac{\widetilde{T}}{2k}$, we have $\mathbb{P}\left\{|\mu_i - \widehat{\mu}_{i,s}| \geq 3\sqrt{\frac{\mu_i \log T}{s}}\right\} \leq \frac{2}{T^3}$. Therefore, applying the union bound we get

$$\mathbb{P}\{G_2^c\} \leq \frac{2}{T^3}\, kT \leq \frac{2}{T} \tag{2.24}$$

In addition, Corollary 2.2 provides a probability bound for G_3^c. The arms and counts considered in G_3 satisfy the requirements of Corollary 2.2. Therefore, for any arm j, with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$ and count $s \geq \frac{\widetilde{T}}{2k}$, the probability $\mathbb{P}\left\{\widehat{\mu}_j \geq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}}\right\} \leq \frac{1}{T^3}$. Again, an application of union bound gives us

$$\mathbb{P}\{G_3^c\} \leq \frac{1}{T^3}\, kT \leq \frac{1}{T} \tag{2.25}$$

Inequalities (2.23), (2.24), and (2.25) lead to desired bound

$$\mathbb{P}\left\{G\right\} = 1 - \mathbb{P}\left\{G^c\right\} \geq 1 - \mathbb{P}\left\{G_1^c\right\} - \mathbb{P}\left\{G_2^c\right\} - \mathbb{P}\left\{G_3^c\right\} \geq 1 - \frac{4}{T}.$$

2.5.2 Proof of Claim 2.1

This section restates and proves Claim 2.1.

Claim 2.1. *For all reals $x \in \left[0, \frac{1}{2}\right]$ and all $a \in [0,1]$, we have $(1-x)^a \geq 1 - 2ax$.*

Proof. The binomial theorem gives us

$$\begin{aligned}(1-x)^a &= 1 - ax + \frac{a(a-1)}{2!}x^2 - \frac{a(a-1)(a-2)}{3!}x^3 + \ldots \\ &= 1 - ax - ax\left(\frac{(1-a)}{2!}x + \frac{(1-a)(2-a)}{3!}x^2 + \frac{(1-a)(2-a)(3-a)}{4!}x^3 + \ldots\right)\end{aligned} \tag{2.26}$$

We can bound the multiplied term as follows

$$\begin{aligned}\frac{(1-a)}{2!}x + \frac{(1-a)(2-a)}{3!}x^2 + &\frac{(1-a)(2-a)(3-a)}{4!}x^3 + \ldots \\ &\leq \frac{1}{2!}x + \frac{1\cdot 2}{3!}x^2 + \frac{1\cdot 2\cdot 3}{4!}x^3 + \ldots && \text{(since } a \in (0,1)\text{)} \\ &= \frac{1}{2}x + \frac{1}{3}x^2 + \frac{1}{4}x^3 \ldots \\ &\leq x + x^2 + x^3 \ldots \\ &= \frac{x}{1-x} && \text{(since } x < 1\text{)}\end{aligned}$$

Hence, equation (2.26) reduces to

$$(1-x)^a \geq 1 - ax - ax\frac{x}{1-x}.$$

Furthermore, since $x \leq \frac{1}{2}$, the ratio $\frac{x}{1-x} \leq 1$. Therefore, we obtain $(1-x)^a > 1 - 2ax$. This completes the proof of the claim. □

2.5.3 Proofs of Lemma 2.2 and 2.3

Next, we restate and prove Lemma 2.2.

Lemma 2.2. *Let* $\mathrm{NCB}_{i^*,t}$ *be the Nash confidence bound of the optimal arm* i^* *at round* t. *Assume that the good event* G *holds and also* $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$. *Then, for all rounds* $t > \widetilde{T}$ *(i.e., for all rounds in Phase* II*), we have* $\mathrm{NCB}_{i^*,t} \geq \mu^*$.

Proof. Fix any round $t > \widetilde{T}$ and write n_{i^*} to denote the number of times the optimal arm i^* has been pulled before that round. Also, let $\widehat{\mu}^*$ denote the empirical mean of arm i^* at round t. Hence, by definition, at this round the Nash confidence bound $\mathrm{NCB}_{i^*,t} := \widehat{\mu}^* + 4\sqrt{\frac{\widehat{\mu}^* \log T}{n_{i^*}}}$. Note that under the good event G (in particular, under G_1) we have $n_{i^*} \geq \frac{\widetilde{T}}{2k} = \frac{8\sqrt{T \log T}}{\sqrt{k \log k}}$. This inequality and the assumption $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ imply

$$\mu^* \; n_{i^*} \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}} \frac{8\sqrt{T \log T}}{\sqrt{k \log k}} = 256 \log T \tag{2.27}$$

In addition, the event G (specifically, G_2) gives us

$$\begin{aligned}
\widehat{\mu}^* &\geq \mu^* - 3\sqrt{\frac{\mu^* \log T}{n_{i^*}}} \\
&= \mu^* - 3\mu^* \sqrt{\frac{\log T}{\mu^* n_{i^*}}} \\
&\geq \mu^* - 3\mu^* \sqrt{\frac{1}{256}} && \text{(via inequality (2.27))} \\
&= \frac{13}{16}\mu^* && (2.28)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{NCB}_{i^*,t} &= \widehat{\mu}^* + 4\sqrt{\frac{\widehat{\mu}^* \log T}{n_{i^*}}} \\
&\geq \mu^* - 3\sqrt{\frac{\mu^* \log T}{n_{i^*}}} + 4\sqrt{\frac{\widehat{\mu}^* \log T}{n_{i^*}}} && \text{(via event } G_2\text{)} \\
&\geq \mu^* - 3\sqrt{\frac{\mu^* \log T}{n_{i^*}}} + 4\sqrt{\frac{13\mu^* \log T}{16 n_{i^*}}} && \text{(via inequality (2.28))}
\end{aligned}$$

$$\geq \mu^* + 0.6\sqrt{\frac{\mu^* \log T}{n_{i^*}}}.$$

The lemma stands proved. □

We restate and prove Lemma 2.3 below.

Lemma 2.3. *Consider a bandit instance with optimal mean* $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ *and assume that the good event* G *holds. Then, any arm* j*, with mean* $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$*, is never pulled in all of Phase* II.

Proof. Fix any arm j with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$. Let r_j denote the number of times arm j is pulled in Phase I. Under event G (in particular, G_1) we have $r_j \geq \frac{\widetilde{T}}{2k}$. In such a case, the good event (in particular, G_3) additionally ensures that (throughout Phase II) the empirical mean of arm j satisfies: $\widehat{\mu}_j < \frac{9\sqrt{k \log k \log T}}{\sqrt{T}}$.

Furthermore, under the good event, $\mathrm{NCB}_{i^*,t} \geq \mu^*$ for all rounds t in Phase II (Lemma 2.2). For any round t in Phase II (i.e., for any $t \geq \widetilde{T} = 16\sqrt{\frac{kT \log T}{\log k}}$), write $\mathrm{NCB}_{j,t}$ to denote the Nash confidence bound of arm j at round t. Below we show that the $\mathrm{NCB}_{j,t}$ is strictly less than $\mathrm{NCB}_{i^*,t}$ and, hence, arm j is not even pulled once in all of Phase II.

$$\begin{aligned}
\mathrm{NCB}_{j,t} &= \widehat{\mu}_j + 4\sqrt{\frac{\widehat{\mu}_j \log T}{r_j}} \\
&\leq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}} + 4\sqrt{\frac{9\sqrt{k \log k \log T}}{\sqrt{T}} \cdot \frac{\log T}{r_j}} && \text{(via event } G_3\text{)} \\
&\leq \frac{9\sqrt{k \log k \log T}}{\sqrt{T}} + 4\sqrt{\frac{9\sqrt{k \log k \log T}}{\sqrt{T}} \cdot \frac{k \log T \sqrt{\log k}}{8\sqrt{Tk \log T}}} && \text{(since } r_j \geq \tfrac{\widetilde{T}}{2k} \text{ under } G_1\text{)} \\
&< \frac{32\sqrt{k \log k \log T}}{\sqrt{T}} \quad \leq \mu^* \quad \leq \mathrm{NCB}_{i^*,t} && \text{(via Lemma 2.2)}
\end{aligned}$$

This completes the proof of the lemma. □

Finally, we prove Lemma 2.4.

Lemma 2.4. *Consider a bandit instance with optimal mean* $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ *and assume that the good event* G *holds. Then, for any arm* i *that is pulled at least once in Phase* II *we have*

$$\mu_i \geq \mu^* - 8\sqrt{\frac{\mu^* \log T}{T_i - 1}},$$

where T_i *is the total number of times that arm* i *is pulled in the algorithm.*

Proof. Any arm j with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$ is never pulled in Phase II (Lemma 2.3). Hence, we focus on arms i with $\mu_i > \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$. Note that when arm i was pulled the T_ith time during Phase II, it must have had the maximum Nash confidence bound value; in particular, at that round $\mathrm{NCB}_i \geq \mathrm{NCB}_{i^*} \geq \mu^*$. Here, the last inequality follows from Lemma 2.2. Therefore, with $\widehat{\mu}_i$ denoting the empirical mean of arm i at this point, we have

$$\widehat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{T_i - 1}} \geq \mu^* \tag{2.29}$$

Complementing inequality (2.29), we will now upper bound the empirical mean $\widehat{\mu}_i$ in terms of μ^*. Note that, since arm i is pulled at least once in Phase II and event G_1 holds, we have $T_i > \frac{\widetilde{T}}{2k} = \frac{8\sqrt{T \log T}}{\sqrt{k \log k}}$. Using this fact and event G_2 we obtain

$$\begin{aligned}
\widehat{\mu}_i &\leq \mu_i + 3\sqrt{\frac{\mu_i \log T}{T_i - 1}} && \text{(since } G_2 \text{ holds)} \\
&\leq \mu^* + 3\sqrt{\frac{\mu^* \log T}{\frac{8\sqrt{T \log T}}{\sqrt{k \log k}}}} && \left(\text{since } T_i > \tfrac{8\sqrt{T \log T}}{\sqrt{k \log k}} \text{ and } \mu_i \leq \mu^*\right) \\
&= \mu^* + 3\sqrt{\frac{\mu^* \sqrt{k \log k \log T}}{8\sqrt{T}}} \\
&\leq \mu^* + 3\sqrt{\frac{\mu^* \mu^*}{256}} && \left(\text{since } \mu^* \geq \tfrac{32\sqrt{k \log k \log T}}{\sqrt{T}}\right) \\
&= \frac{19}{16}\mu^* && (2.30)
\end{aligned}$$

Inequalities (2.29) and (2.30) give us

$$\begin{aligned}
\mu^* &\leq \widehat{\mu}_i + 4\sqrt{\frac{19\mu^* \log T}{16(T_i - 1)}} \\
&\leq \mu_i + 3\sqrt{\frac{\mu_i \log T}{T_i - 1}} + 4\sqrt{\frac{19\mu^* \log T}{16(T_i - 1)}} && \text{(via event } G_2) \\
&\leq \mu_i + 3\sqrt{\frac{\mu^* \log T}{T_i - 1}} + 4\sqrt{\frac{19\mu^* \log T}{16(T_i - 1)}} && \text{(since } \mu_i \leq \mu^*) \\
&\leq \mu_i + 8\sqrt{\frac{\mu^* \log T}{T_i - 1}}.
\end{aligned}$$

This completes the proof of the lemma. □

2.6 Missing Proofs from Section 2.4.1

2.6.1 Proof of Lemma 2.5

This section shows that the good event E occurs with probability at least $\left(1 - \frac{4}{T}\right)$. Toward this, we will upper bound the probabilities of the complements of the events E_1, E_2, and E_3, respectively, and apply union bound to establish the lemma.

For E_1^c, we invoke the Chernoff bound (Lemma 3.11) with random variable $Z_{i,t}$ indicating whether the arm i is selected in round t of uniform sampling, or not. That is, $Z_{i,t} = 1$ if the arm i is picked in round t and, otherwise, $Z_{i,t} = 0$. Using the fact that the algorithm selects an arm uniformly at random in every round of Phase *1* and setting $\varepsilon = 1/2$ along with $\nu = \frac{r}{k} \geq 128S$, Lemma 3.11 along with union bound gives us[1]

$$\mathbb{P}\left\{E_1^c\right\} \leq 2 \cdot \exp\left(-\frac{128 \cdot c^2 \log T}{12\mu^*}\right) \cdot kT \leq \frac{1}{T} \tag{2.31}$$

Next, we address E_2^c. Note that for each arm i, with mean $\mu_i > \frac{\mu^*}{64}$, and for each count $s \geq 64S$, we have $c\sqrt{\frac{\log T}{\mu_i s}} < 1$. Hence, Lemma 2.14 gives us

$$\mathbb{P}\left\{|\mu_i - \widehat{\mu}_{i,s}| \geq c\sqrt{\frac{\mu_i \log T}{s}}\right\} = \mathbb{P}\left\{|\mu_i - \widehat{\mu}_{i,s}| \geq c\sqrt{\frac{\log T}{\mu_i \ s}}\mu_i\right\} \leq 2 \ \exp\left(-\frac{c^2 \log T}{3\mu_i \ s} s\mu_i\right) = \frac{2}{T^3}.$$

Therefore, via the union bound we obtain

$$\mathbb{P}\{E_2^c\} \leq \frac{2}{T^3} \ kT \leq \frac{2}{T} \tag{2.32}$$

Finally, we address E_3^c. Consider any arm j, with mean $\mu_j \leq \frac{\mu^*}{64}$ and any count $s \geq 64S$. Lemma 2.15 (applied with $\nu_H = \frac{\mu^*}{64}$ and $\delta = 1$) leads to

$$\mathbb{P}\left\{\widehat{\mu}_{j,s} \geq \frac{\mu^*}{32}\right\} \leq \exp\left(-\frac{1}{3} \ \frac{\mu^*}{64} \ s\right) = \exp\left(-\frac{c^2 \log T}{3}\right) = \frac{1}{T^3}.$$

Again, an application of union bound gives us

$$\mathbb{P}\{E_3^c\} \leq \frac{1}{T^3} \ kT \leq \frac{1}{T} \tag{2.33}$$

[1] Recall that $S = \frac{c^2 \log T}{\mu^*}$ and $\mu^* \leq 1$.

Inequalities (2.31), (2.32), and (2.33) establish the lemma:

$$\mathbb{P}\{E\} = 1 - \mathbb{P}\{E^c\} \geq 1 - \mathbb{P}\{E_1^c\} - \mathbb{P}\{E_2^c\} - \mathbb{P}\{E_3^c\} \geq 1 - \frac{4}{T}.$$

2.6.2 Proof of Supporting Lemmas

Here, we restate and prove Lemma 2.6.

Lemma 2.6. *Under the event E, for any arm i and any sample count $n \leq 192S$, we have $n\,\widehat{\mu}_{i,n} < 210c^2 \log T$.*

Proof. Write $N := 192S$. Note that, for any arm i, the product $n\,\widehat{\mu}_{i,n}$ is equal to the sum of the rewards observed for arm i in the first n samples. Therefore, for all $n \leq N$, we have

$$n\,\widehat{\mu}_{i,n} \leq N\,\widehat{\mu}_{i,N} \tag{2.34}$$

Using inequality (2.34), we first show that the lemma holds for arms j whose mean $\mu_j \leq \frac{\mu^*}{64}$. Note that for any such arm j, event E_3 gives us $\widehat{\mu}_{j,N} \leq \frac{\mu^*}{32}$. Therefore,

$$n\,\widehat{\mu}_{j,n} \underset{\text{via (2.34)}}{\leq} N\,\widehat{\mu}_{j,N} \leq 192S\frac{\mu^*}{32} = 6c^2 \log T.$$

Next, we complete the proof by proving the lemma for arms i whose mean $\mu_i \geq \frac{\mu^*}{64}$. For any such arm i, we have

$$\begin{aligned}
\widehat{\mu}_{i,N} &\leq \mu_i + c\sqrt{\frac{\mu_i \log T}{N}} && \text{(via event } E_2\text{)}\\
&\leq \mu^* + c\sqrt{\frac{\mu^* \log T}{N}} && \text{(since } \mu_i \leq \mu^*\text{)}\\
&= \mu^* + \frac{\mu^*}{\sqrt{192}} && \text{(since } N = 192S = \tfrac{192c^2 \log T}{\mu^*}\text{)}\\
&< \frac{210}{192}\mu^*
\end{aligned}$$

Hence, even for arms with high enough means we have $N\,\widehat{\mu}_{i,N} < \frac{210}{192}\mu^*\ 192S = 210c^2 \log T$. □

Lemma 2.7 is established next.

Lemma 2.7. *Under the event E, for any sample count $n \geq 484S$, we have $n\,\widehat{\mu}_{i^*,n} \geq 462c^2 \log T$.*

Proof. Write $M := 484S$ and note that, for all $n \geq M$, we have $n\ \widehat{\mu}_{i,n} \geq M\ \widehat{\mu}_{i,M}$. Furthermore,

$$\begin{aligned}
\widehat{\mu}_{i^*,M} &\geq \mu^* - c\sqrt{\frac{\mu^* \log T}{M}} && \text{(via event } E_2\text{)} \\
&= \mu^* - \frac{\mu^*}{\sqrt{484}} && \left(\text{since } M = 484S = \tfrac{484c^2 \log T}{\mu^*}\right) \\
&= \frac{21}{22}\mu^*
\end{aligned}$$

Hence, for any $n \geq M = 484S$, the total observed reward $n\ \widehat{\mu}_{i,n} \geq \frac{21}{22}\mu^*\ 484S = 462c^2 \log T$. □

Next, we restate and prove Lemma 2.8

Lemma 2.8. *Assume that $\sqrt{T} \leq W \leq T$. Also, let random variable τ denote the number of rounds of uniform sampling at which the sum of observed rewards for any arm exceeds $420c^2 \log W$ (i.e., only after τ rounds of uniform sampling we have $\max_i\ n_i \widehat{\mu}_i > 420c^2 \log W$). Then, under event E, the following bounds hold*

$$128\ kS \leq \tau \leq 968\ kS.$$

Proof. Write $t_1 := 128kS$ and note that event E (specifically, E_1) ensures that *at* t_1 rounds of uniform sampling, no arm has been sampled more than $192S$ times. This, in fact, implies that no arm gets sampled more than $192S$ times throughout the first t_1 rounds of uniform exploration. Hence, Lemma 2.6 implies that, till the round t_1, for every arm i the sum of observed rewards $n_i\ \widehat{\mu}_i$ is less than $210c^2 \log T \leq 420c^2 \log W$. Therefore, $\tau \geq 128kS$.

In addition, let $t_2 := 968kS$. Under event E, we have that each arm i is sampled at least $484S$ times by the t_2th round of uniform sampling. Therefore, Lemma 2.7 implies that, by round t_2 and for the optimal arm i^*, the sum of rewards $n_{i^*}\widehat{\mu}_{i^*}$ is at least $462c^2 \log T > 420c^2 \log W$. Hence, $\tau \leq 968kS$. This completes the proof of the lemma. □

2.7 Other Formulations of Nash Regret

This section compares Nash regret NR_T with two variants, $\text{NR}_T^{(0)}$ and $\text{NR}_T^{(1)}$, defined below. For completeness, we also detail the relevant aspects of the canonical bandit model (see, e.g., [LS20, Chapter 4]), which is utilized in analysis of MAB algorithms.

For a bandit instance with k arms and horizon of play T, the canonical model works with a $k \times T$ reward table $(Y_{i,s})_{i\in[k],s\in[T]}$ that populates T independent samples for each arm $i \in [k]$. That is, $Y_{i,s}$ is sth independent draw from the ith distribution. As before, for a given bandit algorithm, the random variable $I_t \in [k]$ denotes the arm pulled in round $t \in \{1, \ldots, T\}$. Recall that μ_{I_t} denotes the associated mean. Furthermore, in each round t, the reward observed, X_t, is obtained from the reward table: $X_t = Y_{I_t,t}$.

Therefore, the sample space $\Omega := \{1, 2, \ldots, k\}^T \times [0, 1]^{k\times T}$. Each element of the sample space is a tuple $\omega = ((I_1, I_2, \ldots, I_T), (Y_{i,s})_{i\in[k],s\in[T]})$, denoting the list of arms pulled at each time t and the rewards table. Precisely, let $\mathcal{F}_1 = \mathfrak{P}(\{1, 2, \ldots, k\}^T)$ be the power set of $\{1, 2, \ldots, k\}^T$, $\mathcal{F}_2 = \mathcal{B}(\mathbb{R}^{k\times T})$ be the Borel σ-algebra on $\mathbb{R}^{k\times T}$, and $\mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ be the product σ-algebra. The probability measure on $(\Omega, \mathcal{F})$ is induced by the bandit instance and the algorithm. The regret definitions and analysis are based on this probability measure.

Recall that Nash regret is defined as $\text{NR}_T := \mu^* - \left(\prod_{t=1}^T \mathbb{E}_{I_t}[\mu_{I_t}]\right)^{\frac{1}{T}}$. Towards a variant, one can first consider the difference between the optimal mean, μ^*, and geometric mean of the realized rewards:

$$\text{NR}_T^{(0)} := \mu^* - \mathbb{E}_{X_1,\ldots X_T}\left[\left(\prod_{t=1}^T X_t\right)^{\frac{1}{T}}\right].$$

Indeed, $\text{NR}_T^{(0)}$ is an unreasonable metric: Even if the algorithm pulls the optimal arm in every round, a single draw X_t can be equal to zero with high probability (in a general MAB instance). In such cases, $\text{NR}_T^{(0)}$ would be essentially as high as μ^*.

A second variant is obtained as follows:

$$\text{NR}_T^{(1)} := \mu^* - \mathbb{E}_{I_1,\ldots I_T}\left[\left(\prod_{t=1}^T \mu_{I_t}\right)^{\frac{1}{T}}\right].$$

While $\text{NR}_T^{(1)}$ upper bounds Nash regret NR_T (see Theorem 2.4 below), it does not conform to a per-agent ex ante assessment. We also note the regret guarantee we obtain for Phase II (of Algorithm 1) in fact holds for $\text{NR}_T^{(1)}$; see inequality (2.4) and the following analysis. From a technical point of view, we also note that it is unreasonable to expect bounds for $\text{NR}_T^{(1)}$ that hold through all the rounds: consider a bandit instance in which all, expect one of the arms (i.e., all except the optimal arm), have zero rewards. As soon as, in the initial (say k) rounds one of these arms get pulled, $\text{NR}_T^{(1)}$ becomes as high as μ^* and cannot be salvaged.

Theorem 2.4. *For any MAB instance and bandit algorithm we have* $\text{NR}_T^{(0)} \geq \text{NR}_T^{(1)} \geq \text{NR}_T$.

Proof. Since the geometric mean is a concave function, the multivariate form of Jensen's inequality gives us $\mathbb{E}\left[\left(\prod_{t=1}^T \mu_{I_t}\right)^{\frac{1}{T}}\right] \leq \left(\prod_{t=1}^T \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}}$. Therefore, $\text{NR}_T^{(1)} \geq \text{NR}_T$.

Next, we compare $\mathrm{NR}_T^{(0)}$ and $\mathrm{NR}_T^{(1)}$. We have

$$\begin{aligned}
\mathbb{E}\left[\left(\prod_{t=1}^{T} X_t\right)^{\frac{1}{T}}\right] &= \mathbb{E}_{\substack{I_1,\dots,I_T \\ X_1,\dots,X_T}}\left[\left(\prod_{t=1}^{T} X_t\right)^{\frac{1}{T}}\right] \\
&= \mathbb{E}_{I_1,\dots,I_T}\left[\mathbb{E}_{X_1,\dots,X_T}\left[\left(\prod_{t=1}^{T} X_t\right)^{\frac{1}{T}} \;\middle|\; (I_t)_t\right]\right] \\
&\leq \mathbb{E}_{I_1,\dots,I_T}\left[\left(\prod_{t=1}^{T} \mathbb{E}_{X_t}\left[X_t \mid (I_t)_t\right]\right)^{\frac{1}{T}}\right] \qquad \text{(Multivariate Jensen's inequality)} \\
&= \mathbb{E}_{I_1,\dots,I_T}\left[\left(\prod_{t=1}^{T} \mu_{I_t}\right)^{\frac{1}{T}}\right].
\end{aligned}$$

This last inequality gives us $\mathrm{NR}_T^{(0)} \geq \mathrm{NR}_T^{(1)}$. The theorem stands proved. □

2.8 Counterexample for the UCB algorithm

This section shows that, in general, the Nash regret of UCB does not decrease as a function of T. In particular, there exist MAB instances in which the UCB algorithm could incur Nash regret close to 1. Recall that, in each round t, the UCB algorithm pulls an arm

$$A_t = \operatorname*{argmax}_{i \in [k]} \left(\widehat{\mu_i} + \sqrt{\frac{2 \log T}{n_{i,t}}}\right),$$

here $\widehat{\mu_i}$ denotes the empirical mean of arm i's reward at round t and $n_{i,t}$ denotes the number of times arm i has been pulled before the tth round. Also, write $\mathrm{UCB}_{i,t}$ to denote the upper confidence bound associated with arm i, i.e.,

$$\mathrm{UCB}_{i,t} \coloneqq \widehat{\mu_i} + \sqrt{\frac{2 \log T}{n_{i,t}}}.$$

UCB algorithm follows an arbitrary, but consistent, tie breaking rule.

We will next detail an MAB instance that illustrates the high Nash regret of UCB. Consider an instance with two arms, $\texttt{arm}_1$ and $\texttt{arm}_2$, and time horizon T, such that $T > 25 \log T$. Also, let the means of the two arms be $\mu_1 = (2e)^{-T}$ and $\mu_2 = 1$, respectively. The rewards of both the arms follow a Bernoulli distribution.[1]

Write random variable $X_{i,t}$ to denote the reward observed for arm i in the tth round. For any given sequence of $X_{i,t}$-s (as in the canonical bandit model), the order of arm pulls in UCB is fixed. That is, for a given sequence of $X_{i,t}$-s, one can deterministically ascertain the arm that will be pulled in a particular round r–this can be done by comparing the values $\text{UCB}_{i,r}$-s and applying the tie-breaking rule.

Furthermore, write Z to denote the event wherein, for $\texttt{arm}_1$, the first T pulls yield a reward of 0. We have $\mathbb{P}\{Z\} = \left(1 - (2e)^{-T}\right)^T \geq 1 - (2e)^{-T}T$.

We will next prove that, under event Z, there exist at least $\log T$ rounds in which $\texttt{arm}_1$ is pulled. Assume, towards a contradiction, that $\texttt{arm}_1$ is pulled less than $\log T$ times. Since, for $\texttt{arm}_2$, we have $X_{2,t} = 1$ for all rounds t, under event Z the sequence of arm pulls in UCB is fixed. Now, consider the round s in which $\texttt{arm}_2$ is pulled the $\left(\frac{T}{2}+1\right)$th time. Then, it must be the case that $\text{UCB}_{1,s} \leq \text{UCB}_{2,s}$. However, $UCB_{1,s} \geq \sqrt{2}$ (given that $\texttt{arm}_1$ has been pulled less than $\log T$ many times) and $\text{UCB}_{2,s} = 1 + \sqrt{\frac{4\log T}{T}}$. This leads to a contradiction and, hence, shows that $\texttt{arm}_1$ is pulled at least $\log T$ times. Moreover, the rounds in which $\texttt{arm}_1$ is pulled are fixed under event Z.

Write $\mathcal{R}$ to denote the specific rounds in which $\texttt{arm}_1$ is pulled under the event Z. Note that $|\mathcal{R}| \geq \log T$. For any $r \in \mathcal{R}$ we have

$$\begin{aligned}
\mathbb{E}[\mu_{I_r}] &= \mathbb{E}[\mu_{I_r}|Z]\ \mathbb{P}\{Z\}\ +\ \mathbb{E}\left[\mu_{I_r}|Z^c\right]\ \mathbb{P}\{Z^c\} \\
&\leq \mathbb{E}[\mu_{I_r}|Z]\cdot 1\ + \mathbb{E}\left[\mu_{I_r}|Z^c\right]\ (2e)^{-T}T && \text{(since } \mathbb{P}\{Z^c\} \leq (2e)^{-T}T\text{)} \\
&\leq (2e)^{-T}\ + \mathbb{E}\left[\mu_{I_r}|Z^c\right]\ (2e)^{-T}T && \text{(since } I_r = \texttt{arm}_1 \text{ under } Z\text{)} \\
&\leq (2e)^{-T}\ + (2e)^{-T}T && \text{(since } \mathbb{E}[\mu_{I_r}|Z^c] \leq 1\text{)} \\
&\leq e^{-T} && (2.35)
\end{aligned}$$

Here, the last inequality follows from the fact that $2^T > T+1$. Also, observe that inequality (2.35) is a bound on the expected value at round r; it holds irrespective of whether Z holds or not.

Therefore, the Nash social welfare of UCB satisfies

$$\begin{aligned}
\left(\prod_{t=1}^{T}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &\leq \left(\prod_{r\in\mathcal{R}}\mathbb{E}[\mu_{I_r}]\right)^{\frac{1}{T}} && \text{(since } \mathbb{E}[\mu_{I_t}] \leq 1 \text{ for all } t\text{)} \\
&\leq \left(e^{-T}\right)^{\frac{|\mathcal{R}|}{T}} && \text{(via inequality (2.35))}
\end{aligned}$$

[1] Hence, $\texttt{arm}_2$ is a point mass.

$$
\begin{aligned}
&\leq \left(e^{-T}\right)^{\frac{\log T}{T}} && \text{(since } |\mathcal{R}| \geq \log T\text{)} \\
&= e^{-\log T} \\
&= \frac{1}{T}
\end{aligned}
$$

Hence, the Nash regret of UCB is at least $\left(1 - \frac{1}{T}\right)$.

2.9 Conclusion and Future Work

This chapter considers settings in which a bandit algorithm's expected rewards, $\{\mathbb{E}\left[\mu_{I_t}\right]\}_{t=1}^{T}$, correspond to values distributed among T agents. In this ex ante framework, we apply Nash social welfare (on the expected rewards) to evaluate the algorithm's performance and thereby formulate the notion of Nash regret. Notably, in cumulative regret, the algorithm is assessed by the social welfare it generates. That is, while cumulative regret captures a utilitarian objective, Nash regret provides an axiomatically-supported primitive for achieving both fairness and economic efficiency.

We establish an instance-independent (and essentially tight) upper bound for Nash regret. Obtaining a Nash regret bound that explicitly depends on the gap parameters, $\Delta_i := \mu^* - \mu$, is an interesting direction of future work. It would also be interesting to formulate regret under more general welfare functions. Specifically, one can consider the generalized-mean welfare [Mou04] which—in the current context and for parameter $p \in (-\infty, 1]$—evaluates to $\left(\frac{1}{T} \sum_t \mathbb{E}\left[\mu_{I_t}\right]^p\right)^{1/p}$. Generalized-means encompass various welfare functions, such as social welfare ($p = 1$), egalitarian welfare ($p \to -\infty$), and Nash social welfare ($p \to 0$). Hence, these means provide a systematic tradeoff between fairness and economic efficiency. Studying Nash regret in broader settings—such as contextual or linear bandits—is a meaningful research direction as well.

Chapter 3

Nash Regret Bounds for Linear Bandits

This chapter focuses on obtaining tight upper bounds for Nash regret in the stochastic linear bandits framework. Nash regret measures the performance of a bandit algorithm by quantifying the difference between the unknown optimal reward and the geometric mean of the expected rewards accumulated over the rounds. This formulation aligns with the well-studied Nash social welfare (NSW) function, which captures the collective welfare generated by the bandit algorithm.

We consider the stochastic linear bandits problem with a horizon of T rounds and a set of arms $\mathcal{X}$ in an ambient dimension of d. Specifically, we examine settings where the stochastic reward associated with each arm in $\mathcal{X}$ follows a non-negative, ν-sub Poisson distribution. In this context, we propose an algorithm that achieves a Nash regret upper bound of $O\left(\sqrt{\frac{d\nu}{\mathsf{T}}}\log(\mathsf{T}|\mathcal{X}|)\right)$. Furthermore, for linear bandit instances with an infinite set of arms $\mathcal{X}$, we derive a Nash regret bound of $O\left(\frac{d^{\frac{5}{4}}\nu^{\frac{1}{2}}}{\sqrt{\mathsf{T}}}\log(\mathsf{T})\right)$. It is worth noting that these results apply to scenarios with bounded, positive rewards, as bounded random variables are sub-Poisson.

Our proposed linear bandit algorithm builds upon the successive elimination method and incorporates novel technical insights. These insights include tailored concentration bounds, as well as the utilization of sampling via John ellipsoid in conjunction with the Kiefer-Wolfowitz optimal design.

3.1 Our Contributions and Techniques.

We consider the stochastic linear bandits setting with a set of arms $\mathcal{X}$ over a finite horizon of T rounds. Since we consider the welfarist viewpoint, we assume that the rewards across all the rounds are positive and, in particular, model the distribution of the arm rewards to be ν-sub Poisson, for parameter $\nu \in \mathbb{R}_+$. As mentioned previously, our goal is to minimize the Nash regret NR_{T}. We develop a novel algorithm LinNash that obtains essentially optimal Nash regret guarantees for this setting. Specifically, for a finite set of arms $\mathcal{X} \subset \mathbb{R}^d$, our algorithm LinNash achieves Nash regret $\mathrm{NR}_{\mathsf{T}} = O\left(\sqrt{\frac{d\nu}{\mathsf{T}}} \log(\mathsf{T}|\mathcal{X}|)\right)$. For infinite sets of arms, a modified version of LinNash achieves Nash regret $\mathrm{NR}_{\mathsf{T}} = O\left(\frac{d^{\frac{5}{4}}\nu^{\frac{1}{2}}}{\sqrt{\mathsf{T}}} \log(\mathsf{T})\right)$.

Note that an ostensible approach for minimizing Nash regret is to take the logarithm of the observed rewards and, then, solve the average regret problem. However, this approach has the following shortcomings: (i) Taking log implies the modified rewards can have a very large range possibly making the regret vacuous and (ii) This approach leads to a multiplicative guarantee and not an additive one. In a chapter 2 we studied Nash regret in the context of stochastic multi-armed bandits (with bounded rewards) and provided optimal guarantees. This chapter notably generalizes this result to linear bandits.

Recall that Nash regret is a strengthening of the average regret; the AM-GM inequality implies that, for any bandit algorithm, the Nash regret is never less than its average regret. Hence, in the linear bandits context, the known $\Omega\left(d/\sqrt{\mathsf{T}}\right)$ lower bound on average regret (see [LS20], Chapter 24) holds for Nash regret as well.[1] This observation implies that, up to a logarithmic factor, our upper bound on Nash regret is tight with respect to the number of rounds T. We also note that for instances in which the number of arms $|\mathcal{X}| = \omega(2^d)$, the Nash-regret dependence on d has a slight gap. Tightening this gap is an interesting direction of future work.

We note that bounded, positive random variables are sub Poisson (Lemma 3.1). Hence, our results hold for linear bandit instances wherein the stochastic rewards are bounded and positive. This observation also highlights the fact that the current work is a generalization of the result obtained in chapter 2. In addition, notice that, by definition, Poisson distributions are 1-sub Poisson. Hence, our guarantees further hold of rewards that are not necessarily sub Gaussian. Given the recent interest in obtaining regret guarantees beyond sub Gaussian rewards[MY16, AJK21], our study of sub Poisson rewards is interesting in its own right.

Our linear bandit algorithm, LinNash, has two parts. In the first part, we develop a novel approach of sampling arms such that in expectation the reward obtained is a linear function of the center of John Ellipsoid [How97]. Such a strategy ensures that the expected reward in any round of the first part is sufficiently large. The second part of LinNash runs in phases of exponentially increasing length. In each phase, we sample arms according to a distribution that is obtained as a solution of a concave optimization problem, known as D-optimal design. We construct confidence intervals at each phase and eliminate sub-optimal arms. A key novelty

[1] This lower bound on average regret is obtained for instances in which the set of arms $\mathcal{X}$ are the corners of a hypercube [LS20].

in our algorithm and analysis is the use of confidence widths that are estimate dependent. We define these widths considering multiplicative forms of concentration bounds and crucially utilize the sub Poisson property of the rewards. The tail bounds we propose might be of independent interest.

Other Related Work: There has been a recent surge in interest to achieve fairness guarantees in the context of multi-armed bandits; see, e.g., [JKMR16b, CKM$^+$19, PGNN21, BBLB20b, HMS21b]. However, these works mostly consider fairness across arms and, in particular, impose fairness constraints that require each arm to be pulled a pre-specified fraction of times. By contrast, our work considers fairness across rounds.

Alternative Regret Formulations. In the current work, for the welfare computation, each agent t's value is considered as the expected reward in round t. One can formulate stronger notions of regret by, say, considering the expectation of the geometric mean of the rewards, rather than the geometric mean of the expectations. However, as discussed in chapter 2, it is not possible to obtain non-trivial guarantees for such reformulations in general: every arm must be pulled at least once. Hence, if one considers the realized rewards (and not their expectations), even a single pull of a zero-reward arm will make the geometric mean zero.

3.2 Problem Formulation and Preliminaries

We will write $[m]$ to denote the set $\{1, 2, \ldots, m\}$. For a matrix $\mathbf{X}$, let $\mathsf{Det}(\mathbf{X})$ to denote the determinant of $\mathbf{X}$. For any discrete probability distribution λ with sample space Ω, write $\mathsf{Supp}(\lambda) \triangleq \{x \in \Omega : \Pr_{X \sim \lambda}\{X = x\} > 0\}$ to denote the points for which the probability mass assigned by λ is positive. For a vector $\mathbf{a} \in \mathbb{R}^d$ and a positive definite matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$, we will denote $||a||_{\mathbf{V}} := \sqrt{a^T \mathbf{V} a}$. Finally, let $\mathcal{B} := \{x \in \mathbb{R}^d \mid ||x||_2 = 1\}$ be the d-dimensional unit ball.

We address the problem of stochastic linear bandits with a time horizon of $\mathsf{T} \in \mathbb{Z}_+$ rounds. Here, an online algorithm (decision maker) is given a set of arms $\mathcal{X} \subset \mathbb{R}^d$. Each arm corresponds to a d-dimensional vector. Furthermore, associated with each arm $x \in \mathcal{X}$, we have a stochastic reward $r_x \in \mathbb{R}_+$. In the linear bandits framework, the expected value of the reward r_x is modeled to be a linear function of $x \in \mathbb{R}^d$. In particular, there exists an unknown parameter vector $\theta^* \in \mathbb{R}^d$ such that, for each $x \in \mathcal{X}$, the associated reward's expected value $\mathbb{E}[r_x] = \langle x, \theta^* \rangle$. Given the focus on welfare contexts, we will, throughout, assume that the rewards are positive, $r_x > 0$, for all $x \in \mathcal{X}$.

The online algorithm (possibly randomized) must sequentially select an arm X_t in each round $t \in [\mathsf{T}]$ and, then, it observes the corresponding (stochastic) reward $r_{X_t} > 0$.[1] For notational convenience, we will write r_t to denote r_{X_t}. In particular, if in round t the selected arm $X_t = x$,

[1] Note that, for a randomized online algorithm, the selected arm X_t is a random variable.

then the expected reward is $\langle x, \theta^* \rangle$, i.e., $\mathbb{E}[r_t \mid X_t = x] = \langle x, \theta^* \rangle$. We will, throughout, use x^* to denote the optimal arm, $x^* = \text{argmax}_{x \in \mathcal{X}} \langle x, \theta^* \rangle$ and $\widehat{\theta}$ to denote estimator of θ^*.

In the stochastic linear bandits framework, our overarching objective is to minimize the Nash regret, defined as follows:

$$\text{NR}_{\mathsf{T}} := \max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \left(\prod_{t=1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^* \rangle] \right)^{1/\mathsf{T}} \tag{3.1}$$

Note that the definition of Nash regret is obtained by applying the Nash social welfare (geometric mean) onto ex ante rewards, $\mathbb{E}\left[\langle X_t, \theta^* \rangle\right]$,[1] accrued across the T rounds.

3.2.1 Sub-Poisson Rewards

In order to model the environment with positive rewards ($r_x > 0$), we assume that the rewards r_x associated with the arms $x \in \mathcal{X}$ are *ν-sub Poisson*, for some parameter $\nu > 0$. Formally, their moment-generating function satisfies the following bound

$$\mathbb{E}\left[e^{\lambda\ r_x}\right] \leq \exp\left(\nu^{-1} \mathbb{E}[r_x]\ \left(e^{\nu\lambda} - 1\right)\right) = \exp\left(\nu^{-1} \langle x, \theta^* \rangle \left(e^{\nu\lambda} - 1\right)\right) \text{ for all } \lambda \in \mathbb{R}. \tag{3.2}$$

Note that a Poisson random variable is 1-sub Poisson. To highlight the generality of ν-sub-Poisson distributions, we note that bounded, non-negative random variables are sub-Poisson (Lemma 3.1). Further, in Lemma 3.2, we establish a connection between non-negative sub-Gaussian and sub-Poisson random variables. .

Lemma 3.1. *Any non-negative random variable $X \in [0, \mathsf{B}]$ is B-sub-Poisson, i.e., if mean $\mathbb{E}[X] = \mu$, then for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[e^{\lambda X}] \leq \exp\left(B^{-1}\mu\left(e^{B\lambda} - 1\right)\right)$.*

Lemma 3.2. *Let X be a non-negative sub-Gaussian random variable X with mean $\mu = \mathbb{E}[X]$ and sub-Gaussian norm σ. Then, X is also $\left(\frac{\sigma^2}{\mu}\right)$-sub-Poisson.*

The proofs of Lemmas 3.1 and 3.2 appear in Appendix 3.6. Lemma 3.2 has useful instantiations. In particular, the lemma implies that the half-normal random variable, with variance of σ, is also a $(C\sigma)$-sub-Poisson, where C is a constant (independent of distribution parameters). Similarly, for other well-studied, positive sub-Gaussian random variables (including truncated and folded normal distributions), the sub-Poisson parameter is small.

Next, we discuss the necessary preliminaries for our algorithm and analysis.

[1] Here, the expectation is with respect to the random variable X_t.

3.2.2 Optimal Design.

Write $\Delta(\mathcal{X})$ to denote the probability simplex associated with the set of arms $\mathcal{X}$. Let $\lambda \in \Delta(\mathcal{X})$ be such a probability distribution over the arms, with λ_x denoting the probability of selecting arm x. The following optimization problem, defined over the set of arms $\mathcal{X}$, is well-known and is referred to as the G-optimal design problem.

$$\text{Minimize } g(\lambda) \triangleq \max_{x \in \mathcal{X}} ||x||^2_{\mathbf{U}(\lambda)^{-1}}, \text{ where } \lambda \in \Delta(\mathcal{X}) \text{ and } \mathbf{U}(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^T \tag{3.3}$$

The solution to (3.3) provides the optimal sequence of arm pulls (for a given budget of rounds) to minimize the confidence width of the estimated rewards for all arms $x \in \mathcal{X}$. The G-optimal design problem connects to the following optimization problem (known as D-optimal design problem):

$$\text{Maximize } f(\lambda) \triangleq \log \mathsf{Det}(\mathbf{U}(\lambda)), \text{ where } \lambda \in \Delta(\mathcal{X}) \text{ and } \mathbf{U}(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^T \tag{3.4}$$

The lemma below provides an important result of Kiefer and Wolfowitz [KW60].

Lemma 3.3 (Kiefer-Wolfowitz). *If the set $\mathcal{X}$ is compact and $\mathcal{X}$ spans $\mathbb{R}^d$, then there exists $\lambda^* \in \Delta(\mathcal{X})$ supported over at most $d(d+1)/2$ arms such that λ^* minimizes the objective in equation (3.3) with $g(\lambda^*) = d$. Furthermore, λ^* is also a maximizer of the D-optimal design objective, i.e., λ^* maximizes the function $f(\lambda) = \log \mathsf{Det}(\mathbf{U}(\lambda))$ subject to $\lambda \in \Delta(\mathcal{X})$.*

At several places in our algorithm, our goal is to find a probability distribution that minimizes the non-convex optimization problem (3.3). However, instead we will maximize the concave function $f(\lambda) = \log \mathsf{Det}(\mathbf{U}(\lambda))$ over $\lambda \in \Delta(\mathcal{X})$. The Frank-Wolfe algorithm, for instance, can be used to solve the D-optimal design problem (3.4) and compute λ^* efficiently ([LS20], Chapter 21). Lemma 3.3 ensures that this approach works, since the G-optimal and the D-optimal design problems have the same optimal solution $\lambda^* \in \Delta(\mathcal{X})$, which satisfies $\mathsf{Supp}(\lambda^*) \leq d(d+1)/2$.[1]

3.2.3 John Ellipsoid.

For any convex body $K \subset \mathbb{R}^d$, a John ellipsoid is an ellipsoid with maximal volume that can be inscribed within K. It is known that K itself is contained within the John Ellipsoid dilated

[1] Even though the two optimization problems (3.3) and (3.4) share the optimal solution, the optimal objective function values can be different.

Algorithm 4 `GenerateArmSequence` (Subroutine to generate Arm Sequence)

```
Input: Arm set $\mathcal{X}$ and sequence length $\widetilde{\mathsf{T}} \in \mathbb{Z}_+$.
Find the probability distribution $\lambda \in \Delta(\mathcal{X})$ by maximizing the following objective
    $\log \mathsf{Det}(\mathbf{U}(\lambda_0))$ subject to $\lambda_0 \in \Delta(\mathcal{X})$ and $\mathsf{Supp}(\lambda_0) \leq d(d+1)/2$     (3.5)
Initialize multiset $\mathcal{S} = \emptyset$ and set $\mathcal{A} = \mathsf{Supp}(\lambda)$. Also, initialize count $c_z = 0$, for each arm $z \in \mathcal{A}$.
Compute distribution $U$ as described in Section 3.3.1.
for $i = 1$ to $\widetilde{\mathsf{T}}$ do
    With probability $1/2$ set $\mathsf{flag}$ = SAMPLE-U, otherwise, set $\mathsf{flag}$ = D/G-OPT.
    if $\mathsf{flag}$ = SAMPLE-U or $\mathcal{A} = \emptyset$ then
        Sample an arm $\widehat{x}$ from the distribution $U$, and update multiset $\mathcal{S} \leftarrow \mathcal{S} \cup \{\widehat{x}\}$.
    else if $\mathsf{flag}$ = D/G-OPT then
        Pick the next arm $z$ in $\mathcal{A}$ (round robin).
        Update multiset $\mathcal{S} \leftarrow \mathcal{S} \cup \{z\}$ and increment count $c_z \leftarrow c_z + 1$.
        If $c_z \geq \lceil \lambda_z \widetilde{\mathsf{T}}/3 \rceil$, then update $\mathcal{A} \leftarrow \mathcal{A} \setminus \{z\}$.
    end if
end for
return multiset $\mathcal{S}$
```

by a factor of d. Formally,[1]

Lemma 3.4 ([GLS12]). *Let $K \subset \mathbb{R}^d$ be a convex body (i.e., a compact, convex set with a nonempty interior). Then, there exists an ellipsoid E (called the John ellipsoid) that satisfies $E \subseteq K \subseteq c + d(E - c)$. Here, $c \in \mathbb{R}^d$ denotes the center of E and $c + d(E - c)$ refers to the (dialated) set $\{c + d(x - c) : x \in E\}$.*

3.3 Our Algorithm LinNash and Main Results

In this section, we detail our algorithm LinNash (Algorithm 5), and establish an upper bound on the Nash regret achieved by this algorithm. Subsection 3.3.1 details Part I of LinNash and related analysis. Then, Subsection 3.3.2 presents and analyzes Part II of the algorithm. Using the lemmas from these two subsections, the regret bound for the algorithm is established in Subsection 3.3.3.

3.3.1 Part I: Sampling via John Ellipsoid and Kiefer-Wolfowitz Optimal Design

As mentioned previously, Nash regret is a more challenging objective than average regret: if in any round $t \in [\mathsf{T}]$, the expected[2] reward $\mathbb{E}[r_t]$ is zero (or very close to zero), then geometric mean

[1] The ellipsoid E considered in Lemma 3.4 is also the ellipsoid of maximal volume contained in K [GLS12].

[2] Here, the expectation is over randomness in algorithm and the reward noise.

Algorithm 5 LinNash (Nash Regret Algorithm for Finite Set of Arms)

```
Input: Arm set $\mathcal{X}$ and horizon of play $T$.
Initialize matrix $\mathbf{V} \leftarrow [0]_{d,d}$ and number of rounds $\widetilde{\mathsf{T}} = 3\sqrt{\mathsf{T} d\nu \log(\mathsf{T}|\mathcal{X}|)}$.
Part I
Generate arm sequence $\mathcal{S}$ for the first $\widetilde{\mathsf{T}}$ rounds using Algorithm 4.
for $t = 1$ to $\widetilde{T}$ do
    Pull the next arm $X_t$ from the sequence $\mathcal{S}$, observe corresponding reward $r_t$, and update $\mathbf{V} \leftarrow \mathbf{V} + X_t X_t^T$
end for
Set estimate $\widehat{\theta} := \mathbf{V}^{-1}\left(\sum_{t=1}^{\widetilde{\mathsf{T}}} r_t X_t\right)$
Compute confidence bounds $\mathrm{LNCB}(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3)$ and $\mathrm{UNCB}(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3)$, for all $x \in \mathcal{X}$ (see equation (3.7))
Set $\widetilde{\mathcal{X}} = \left\{x \in \mathcal{X} : \mathrm{UNCB}(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3) \geq \max_{z \in \mathcal{X}} \mathrm{LNCB}(z, \widehat{\theta}, \widetilde{\mathsf{T}}/3)\right\}$ and initialize $\mathsf{T}' = \frac{2}{3}\widetilde{\mathsf{T}}$
Part II
while end of time horizon $\mathsf{T}$ is reached do
    Initialize $V = [0]_{d,d}$ to be an all zeros $d \times d$ matrix and $s = [0]_d$ to be an all-zeros vector.
    // Beginning of new phase.
    Find the probability distribution $\lambda \in \Delta(\widetilde{\mathcal{X}})$ by maximizing the following objective

        $\log \mathsf{Det}(\mathbf{U}(\lambda_0))$ subject to $\lambda_0 \in \Delta(\widetilde{\mathcal{X}})$ and $\mathsf{Supp}(\lambda_0) \leq d(d+1)/2$.    (3.6)

    for each arm $a$ in $\mathsf{Supp}(\lambda)$ do
        Pull arm $a$ for the next $\lceil \lambda_a \, \mathsf{T}' \rceil$ rounds. Update $\mathbf{V} \leftarrow \mathbf{V} + \lceil \lambda_a \mathsf{T}' \rceil \cdot a a^T$.
        Observe $\lceil \lambda_a \, \mathsf{T}' \rceil$ corresponding rewards $z_1, z_2, \ldots$ and update $s \leftarrow s + (\sum_j z_j) a$.
    end for
    Set estimate $\widehat{\theta} = \mathbf{V}^{-1} s$ and compute $\mathrm{LNCB}(x, \widehat{\theta}, \mathsf{T}')$ and $\mathrm{UNCB}(x, \widehat{\theta}, \mathsf{T}')$, for all $x \in \mathcal{X}$ (see equation (3.7))
    Set $\widetilde{\mathcal{X}} = \left\{x \in \widetilde{\mathcal{X}} : \mathrm{UNCB}(x, \widehat{\theta}, \mathsf{T}') \geq \max_{z \in \mathcal{X}} \mathrm{LNCB}(z, \widehat{\theta}, \mathsf{T}')\right\}$.    // End of phase.
    Update $\mathsf{T}' \leftarrow 2\,\mathsf{T}'$.
end while
```

$(\prod_{t=1}^{\mathsf{T}} \mathbb{E}[r_{X_t}])^{1/\mathsf{T}}$ goes to zero, even if the expected rewards in the remaining rounds are large. Hence, we need to ensure that in every round $t \in [\mathsf{T}]$, specifically the rounds in the beginning of the algorithm, the expected rewards are bounded from below. In [BKMS22], this problem was tackled for stochastic multi-armed bandits (MAB) by directly sampling each arm uniformly at random in the initial rounds. Such a sampling ensured that, in each of those initial rounds, the expected reward is bounded from below by the average of the expected rewards. While such a uniform sampling strategy is reasonable for the MAB setting, it can be quite unsatisfactory in the current context of linear bandits. To see this, consider a linear bandit instance in which, all—except for one—arms in $\mathcal{X}$ are orthogonal to θ^*. Here, a uniform sampling strategy will lead to an expected reward of $\langle x^*, \theta^* \rangle / |\mathcal{X}|$, which can be arbitrarily small for large cardinality

$\mathcal{X}$.

To resolve this issue we propose a novel approach in the initial $\widetilde{\mathsf{T}} := 3\sqrt{\mathsf{T} d\nu \log(\mathsf{T}|\mathcal{X}|)}$ rounds. In particular, we consider the convex hull of the set of arms $\mathcal{X}$—denoted as $\mathrm{cvh}(\mathcal{X})$—and find the center $c \in \mathbb{R}^d$ of the John ellipsoid E for the convex hull $\mathrm{cvh}(\mathcal{X})$. Since $E \subseteq \mathrm{cvh}(\mathcal{X})$, the center c of the John ellipsoid is contained within $\mathrm{cvh}(\mathcal{X})$ as well. Furthermore, via Carathéodory's theorem [Eck93], we can conclude that the center c can be expressed as a convex combination of at most $(d+1)$ points in $\mathcal{X}$. Specifically, there exists a size-$(d+1)$ subset $\mathcal{Y} := \{y_1, \ldots, y_{d+1}\} \subseteq \mathcal{X}$ and convex coefficients $\alpha_1, \ldots, \alpha_{d+1} \in [0,1]$ such that $c = \sum_{i=1}^{d+1} \alpha_i y_i$ with $\sum_{i=1}^{d+1} \alpha_i = 1$. Therefore, the convex coefficients induce a distribution $U \in \Delta(\mathcal{X})$ of support size $d+1$ and with $\mathbb{E}_{x \sim U}[x] = c$.

Lemma 3.5 below asserts that sampling according to the distribution U leads to an expected reward that is sufficiently large. Hence, U is used in the subroutine `GenerateArmSequence` (Algorithm 4).

In particular, the purpose of the subroutine is to carefully construct a sequence (multiset) of arms $\mathcal{S}$, with size $|\mathcal{S}| = \widetilde{\mathsf{T}}$ and to be pulled in the initial $\widetilde{\mathsf{T}}$ rounds. The sequence $\mathcal{S}$ is constructed such that (i) upon pulling arms from $\mathcal{S}$, we have a sufficiently large expected reward in each pull, and (ii) we obtain an initial estimate of the inner product of the unknown parameter vector θ^* with all arms in $\mathcal{X}$. Here, objective (i) is achieved by considering the above-mentioned distribution U. Now, towards the objective (ii), we compute distribution $\lambda \in \Delta(\mathcal{X})$ by solving the optimization problem (also known as the D-optimal design problem) stated in equation (3.5).

We initialize sequence $\mathcal{S} = \emptyset$ and run the subroutine `GenerateArmSequence` for $\widetilde{\mathsf{T}}$ iterations. In each iteration (of the for-loop in Line 4), with probability $1/2$, we sample an arm according to the distribution U (Line 7) and include it in $\mathcal{S}$. Also, in each iteration, with remaining probability $1/2$, we consider the computed distribution λ and, in particular, pick arms z from the support of λ in a round-robin manner. We include such arms z in $\mathcal{S}$ while ensuring that, at the end of the subroutine, each such arm $z \in \mathsf{Supp}(\lambda)$ is included at least $\lceil \lambda_z \widetilde{\mathsf{T}}/3 \rceil$ times. We return the curated sequence of arms $\mathcal{S}$ at the end of the subroutine.

Our main algorithm LinNash (Algorithm 5) first calls subroutine `GenerateArmSequence` to generated the sequence $\mathcal{S}$. Then, the algorithm LinNash sequentially pulls the arms X_t from $\mathcal{S}$, for $1 \leq t \leq \widetilde{\mathsf{T}}$ rounds. For these initial $\widetilde{\mathsf{T}} = |\mathcal{S}|$ rounds, let r_t denote the noisy, observed rewards. Using these $\widetilde{\mathsf{T}}$ observed rewards, the algorithm computes the ordinary least squares (OLS) estimate $\widehat{\theta}$ (see Line 6 in Algorithm 5); in particular, $\widehat{\theta} := (\sum_{t=1}^{\widetilde{\mathsf{T}}} X_t X_t^T)^{-1} (\sum_{t=1}^{\widetilde{\mathsf{T}}} r_t X_t)$. The algorithm uses the OLS estimate $\widehat{\theta}$ to eliminate several low rewarding arms (in Lines 7 and 8 in Algorithm 5). This concludes Part I of the algorithm LinNash.

Before detailing Part II (in Subsection 3.3.2), we provide a lemma to be used in the analysis of Part I of LinNash.

Lemma 3.5. *Let $c \in \mathbb{R}^d$ denote the center of a John ellipsoid for the convex hull* $\text{cvh}(\mathcal{X})$ *and let $U \in \Delta(\mathcal{X})$ be a distribution that satisfies $\mathbb{E}_{x\sim U}\ x = c$. Then, it holds that*

$$\mathbb{E}_{x\sim U}[\langle x, \theta^* \rangle] \geq \frac{\langle x^*, \theta^* \rangle}{(d+1)}.$$

Proof. Lemma 3.4 ensures that there exists a positive definite matrix $\mathbf{H}$ with the property that

$$\left\{ x \in \mathbb{R}^d : \sqrt{(x-c)^T \mathbf{H}(x-c)} \leq 1 \right\} \subseteq \text{cvh}(\mathcal{X}) \subseteq \left\{ x \in \mathbb{R}^d : \sqrt{(x-c)^T \mathbf{H}(x-c)} \leq d \right\}.$$

Now, write $y := c - \frac{x^*-c}{d}$ and note that

$$\sqrt{(y-c)^T \mathbf{H}(y-c)} = \sqrt{\frac{(x^*-c)^T \mathbf{H}(x^*-c)}{d^2}} \leq 1 \qquad \text{(since } x^* \in \text{cvh}(\mathcal{X})\text{)}$$

Therefore, $y \in \text{cvh}(\mathcal{X})$. Recall that, for all arms $x \in \mathcal{X}$, the associated reward (r_x) is non-negative and, hence, the rewards' expected value satisfies $\langle x, \theta^* \rangle \geq 0$. This inequality and the containment $y \in \text{cvh}(\mathcal{X})$ give us $\langle y, \theta^* \rangle \geq 0$. Substituting $y = c - \frac{x^*-c}{d}$ in the last inequality leads to $\langle c, \theta^* \rangle \geq \langle x^*, \theta^* \rangle/(d+1)$. Given that $\mathbb{E}_{x\sim U}\ [x] = c$, we obtain the desired inequality $\mathbb{E}_{x\sim U}\langle x, \theta^* \rangle = \langle c, \theta^* \rangle \geq \frac{\langle x^*, \theta^* \rangle}{(d+1)}$. □

Note that at each iteration of the subroutine `GenerateArmSequence`, with probability 1/2, we insert an arm into $\mathcal{S}$ that is sampled according to U. Using this observation and Lemma 3.5, we obtain that, for any round $t \in [\widetilde{\mathsf{T}}]$ and for the random arm X_t pulled from the sequence $\mathcal{S}$ according to our procedure, the observed reward r_{X_t} must satisfy $\mathbb{E}[r_{X_t}] \geq \frac{\langle x^*, \theta^* \rangle}{2(d+1)}$.[1]

Further, recall that in the subroutine `GenerateArmSequence`, we insert arms $x \in \mathsf{Supp}(\lambda)$ at least $\lceil \lambda_x \widetilde{\mathsf{T}}/3 \rceil$ times, where λ corresponds to the solution of D-optimal design problem defined in equation (3.5). Therefore, we can characterize the confidence widths of the estimated rewards for each arm in $\mathcal{X}$ computed using the least squares estimate $\widehat{\theta}$ computed in Line 7 in Algorithm 5.

Broadly speaking, we can show that all arms with low expected reward (less than a threshold) also have an estimated reward at most twice the true reward. On the other hand, high rewarding arms must have an estimated reward to be within a factor of 2 of the true reward.

[1] Here, the expectation is over both the randomness in including arm X_t in $\mathcal{S}$ and the noise in the reward.

Thus, based on certain high probability confidence bounds (equation (3.7)), we can eliminate arms in $\mathcal{X}$ with true expected reward less than some threshold, with high probability.

3.3.2 Part II: Phased Elimination via Estimate Dependent Confidence Widths

Note that while analyzing average regret via confidence bound algorithms, it is quite common to use, for each arm x, a confidence width (interval) that does not depend on x's estimated reward. This is a reasonable design choice for bounding average regret, since the regret incurred at each round is the sum of confidence intervals that grow smaller with the round index and, hence, this choice leads to a small average regret. However, for the analysis of the Nash regret, a confidence width that is independent of the estimated reward can be highly unsatisfactory: the confidence width might be larger than the optimal $\langle x^*, \theta^* \rangle$. This can in turn allow an arm with extremely low reward to be pulled leading to the geometric mean going to zero. In order to alleviate this issue, it is vital that our confidence intervals are reward dependent. This in turn, requires one to instantiate concentration bounds similar to the multiplicative version of the standard Chernoff bound. In general, multiplicative forms of concentration bounds are much stronger than the additive analogues [KQ21]. In prior work [BKMS22] on Nash regret for the stochastic multi-armed bandits setting, such concentration bounds were readily available through the multiplicative version of the Chernoff bound. However, in our context of linear bandits, the derivation of analogous concentration bounds (and the associated confidence widths) is quite novel and requires a careful use of the sub-Poisson property.

In particular, we use the following confidence bounds (with estimate dependent confidence widths) in our algorithm. We define the lower and upper confidence bounds considering any arm x, any least squares estimator ϕ (of θ^*), and t the number of observations used to compute the estimator ϕ. That is, for any triple $(x, \phi, t) \in \mathcal{X} \times \mathbb{R}^d \times [\mathsf{T}]$, we define Lower Nash Confidence Bound (LNCB) and Upper Nash Confidence Bound (UNCB) as follows:

$$\begin{aligned} \mathrm{LNCB}(x, \phi, t) &:= \langle x, \phi \rangle - 6\sqrt{\frac{\langle x, \phi \rangle \nu d \log(\mathsf{T}|\mathcal{X}|)}{t}} \\ \mathrm{UNCB}(x, \phi, t) &:= \langle x, \phi \rangle + 6\sqrt{\frac{\langle x, \phi \rangle \nu d \log(\mathsf{T}|\mathcal{X}|)}{t}}. \end{aligned} \tag{3.7}$$

As mentioned previously, the confidence widths in equation (3.7) are estimate dependent.

Next, we provide a high level overview of Part II in Algorithm 5. This part is inspired from the phased elimination algorithm for the average regret ([LS20], Chapter 21); a key distinction here is to use the Nash confidence bounds defined in (3.7). Part II in Algorithm 5 begins with the set of arms $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ obtained after an initial elimination of low rewarding arms in Part I . Subsequently, Part II runs in phases of exponentially increasing length and eliminates sub-optimal arms in every phase.

Suppose at the beginning of the ℓ^{th} phase, $\widetilde{\mathcal{X}}$ is the updated set of arms. We solve the D-optimal design problem (see (3.6)) corresponding to $\widetilde{\mathcal{X}}$ to obtain a distribution $\lambda \in \Delta(\widetilde{\mathcal{X}})$. For the next $O(d^2 + 2^\ell \widetilde{\mathsf{T}})$ rounds, we pull arms a in the support of λ (Line 13): each arm $a \in \mathsf{Supp}(\lambda)$ is pulled $\lceil \lambda_a \mathsf{T}' \rceil$ times where $\mathsf{T}' = O(2^\ell \widetilde{\mathsf{T}})$. Using the data covariance matrix and the observed noisy rewards, we recompute: (1) an improved estimate $\widehat{\theta}$ (of θ^*) and (2) improved confidence bounds for every surviving arm. Then, we eliminate arms based on the confidence bounds and update the set of surviving arms (Lines 16 and 17).

The following lemma provides the key concentration bound for the least squares estimate.

Lemma 3.6. *Let $x_1, x_2, \ldots, x_s \in \mathbb{R}^d$ be a fixed set of vectors and let $r_1, r_2, \ldots, r_s$ be independent ν-sub-Poisson random variables satisfying $\mathbb{E} r_s = \langle x_s, \theta^* \rangle$ for some unknown θ^*. Further, let matrix $\mathbf{V} = \sum_{j=1}^{s} x_j x_j^T$ and $\widehat{\theta} = \mathbf{V}^{-1}\left(\sum_j r_j x_j\right)$ be the least squares estimator of θ^*. Consider any $z \in \mathbb{R}^d$ with the property that $z^T \mathbf{V}^{-1} x_j \leq \gamma$ for all $j \in [s]$. Then, for any $\delta \in [0, 1]$ we have*

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{\delta^2 \langle z, \theta^*\rangle}{3\nu\gamma}\right) \quad \text{and} \tag{3.8}$$

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq (1-\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{\delta^2 \langle z, \theta^*\rangle}{2\nu\gamma}\right). \tag{3.9}$$

Lemma 3.6 is established in Appendix 3.7. Using this lemma, we can show that the optimal arm x^* is never eliminated with high probability.

Lemma 3.7. *Consider any bandit instance in which for the optimal arm $x^* \in \mathcal{X}$ we have $\langle x^*, \theta^* \rangle \geq 192\sqrt{\frac{d\ \nu}{T}} \log(\mathsf{T}|\mathcal{X}|)$. Then, with probability at least $\left(1 - \frac{4 \log \mathsf{T}}{\mathsf{T}}\right)$, the optimal arm x^* always exists in the surviving set $\widetilde{\mathcal{X}}$ in Part* I *and in every phase in Part* II *of Algorithm 5.*

Finally, using Lemmas 3.6 and 3.7, we show that, with high probability, in every phase of Part II all the surviving arms $x \in \widetilde{\mathcal{X}}$ have sufficiently high reward means.

Lemma 3.8. *Consider any phase ℓ in Part* II *of Algorithm 5 and let $\widetilde{\mathcal{X}}$ be the surviving set of arms at the beginning of that phase. Then, with $\widetilde{\mathsf{T}} = \sqrt{d\nu \mathsf{T} \log(\mathsf{T}\,|\mathcal{X}|)}$, we have*

$$\Pr\left\{\langle x, \theta^*\rangle \geq \langle x^*, \theta^*\rangle - 25\sqrt{\frac{3d\nu\langle x^*, \theta^*\rangle \log\left(\mathsf{T}|\mathcal{X}|\right)}{2^\ell \cdot \widetilde{\mathsf{T}}}} \ \text{for all } x \in \widetilde{\mathcal{X}}\right\} \geq 1 - \frac{4\log \mathsf{T}}{\mathsf{T}} \tag{3.10}$$

Here, ν is the sub-Poisson parameter of the stochastic rewards.

The proofs of the Lemmas 3.7 and 3.8 are deferred to Appendix 3.8.

3.3.3 Main Result

This section states and proves the Nash regret guarantee achieved by LinNash (Algorithm 5).

Theorem 3.1. *For any given stochastic linear bandits problem with (finite) set of arms $\mathcal{X} \subset \mathbb{R}^d$, time horizon $\mathsf{T} \in \mathbb{Z}_+$, and ν-sub-Poisson rewards, Algorithm 5 achieves Nash regret*

$$\mathrm{NR}_{\mathsf{T}} = O\left(\beta\sqrt{\frac{d\ \nu}{\mathsf{T}}}\log(\mathsf{T}|\mathcal{X}|)\right).$$

Here, $\beta = \max\{1,\ \langle x^*, \theta^*\rangle \log d\}$, with $x^* \in \mathcal{X}$ denoting the optimal arm and θ^* the (unknown) parameter vector.

Proof. We will assume, without loss of generality, that $\langle x^*, \theta^*\rangle \geq 192\sqrt{\frac{d\ \nu}{\mathsf{T}}}\log(\mathsf{T}|\mathcal{X}|)$, otherwise the stated Nash Regret bound directly holds (see equation (3.1)). Write E to denote the 'good' event identified in Lemma 3.8; the lemma ensures that $\mathbb{P}\{E\} \geq 1 - \frac{4\log\mathsf{T}}{\mathsf{T}}$. During Part I of Algorithm 5, the product of expected rewards, conditioned on E, satisfies

$$\begin{aligned}\prod_{t=1}^{\widetilde{\mathsf{T}}} \mathbb{E}[\langle X_t, \theta^*\rangle \mid E]^{\frac{1}{\mathsf{T}}} &\geq \left(\frac{\langle x^*, \theta^*\rangle}{2(d+1)}\right)^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}} && \text{(via Lemma 3.5)}\\ &= \langle x^*, \theta^*\rangle^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}}\left(1 - \frac{1}{2}\right)^{\frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}}}\\ &\geq \langle x^*, \theta^*\rangle^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}}\left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}}\right).\end{aligned}$$

For analyzing Part II, we will utilize Lemma 3.8. Write $\mathcal{B}_\ell$ to denote all the rounds t that belong to ℓ^{th} phase (in Part II). Also, let T'_ℓ denote the associated phase-length parameter, i.e., $\mathsf{T}'_\ell = 2^\ell\ \widetilde{\mathsf{T}}/3$. Note that in each phase ℓ (i.e., in the for-loop at Line 12 of Algorithm 5), every arm a in $\mathsf{Supp}(\lambda)$ (the support of D-optimal design) is pulled $\lceil \lambda_a \mathsf{T}'_\ell \rceil$ times. Given that $|\mathsf{Supp}(\lambda)| \leq d(d+1)/2$, we have $|\mathcal{B}_\ell| \leq \mathsf{T}'_\ell + \frac{d(d+1)}{2}$. By construction $\mathsf{T}'_\ell \geq \frac{d(d+1)}{2}$ and, hence, $|\mathcal{B}_\ell| \leq 2\mathsf{T}'_\ell$. Since the phase length parameter, T'_ℓ, doubles after each phase, the algorithm would have at most $\log\mathsf{T}$ phases. Hence, the product of expected rewards in Part II satisfies

$$\prod_{t=\widetilde{\mathsf{T}}+1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^*\rangle \mid E]^{\frac{1}{\mathsf{T}}} = \prod_{\mathcal{B}_\ell}\prod_{t\in\mathcal{B}_\ell} \mathbb{E}[\langle X_t, \theta^*\rangle \mid E]^{\frac{1}{\mathsf{T}}}$$

$$\geq \prod_{\mathcal{B}_\ell} \left(\langle x^*, \theta^* \rangle - 25 \sqrt{\frac{d\ \nu\ \langle x^*, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'_\ell}} \right)^{\frac{|\mathcal{B}_\ell|}{\mathsf{T}}} \qquad \text{(Lemma 3.8)}$$

$$\geq \langle x^*, \theta^* \rangle^{\frac{\mathsf{T} - \widetilde{\mathsf{T}}}{\mathsf{T}}} \prod_{\ell=1}^{\log \mathsf{T}} \left(1 - 25 \sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle \mathsf{T}'_\ell}} \right)^{\frac{|\mathcal{B}_\ell|}{\mathsf{T}}}$$

$$\geq \langle x^*, \theta^* \rangle^{\frac{\mathsf{T} - \widetilde{\mathsf{T}}}{\mathsf{T}}} \prod_{\ell=1}^{\log \mathsf{T}} \left(1 - 50 \frac{|\mathcal{B}_\ell|}{\mathsf{T}} \sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle \mathsf{T}'_\ell}} \right).$$

The last inequality follows from the fact that $(1-x)^r \geq (1-2rx)$, for any $r \in [0,1]$ and $x \in [0, 1/2]$. Note that the term $\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle \mathsf{T}'_\ell}} \leq 1/2$, since $\langle x^*, \theta^* \rangle \geq 192 \sqrt{\frac{d\nu}{\mathsf{T}}} \log(\mathsf{T}|\mathcal{X}|)$ along with $\mathsf{T}'_\ell \geq 2\sqrt{\mathsf{T} d\nu \log \mathsf{T}|\mathcal{X}|}$ and $\mathsf{T} \geq e^4$. We further simplify the expression as follows

$$\prod_{\ell=1}^{\log \mathsf{T}} \left(1 - 50 \frac{|\mathcal{B}_\ell|}{\mathsf{T}} \sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle \mathsf{T}'_\ell}} \right) \geq \prod_{\ell=1}^{\log \mathsf{T}} \left(1 - 100 \frac{\sqrt{\mathsf{T}'_\ell}}{\mathsf{T}} \sqrt{\frac{d\ \nu \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle}} \right)$$

$$\text{(since } |\mathcal{B}_\ell| \leq 2\mathsf{T}'_\ell\text{)}$$

$$\geq 1 - \frac{100}{\mathsf{T}} \sqrt{\frac{d\ \nu \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle}} \left(\sum_{\ell=1}^{\log \mathsf{T}} \sqrt{\mathsf{T}'_\ell} \right)$$

$$\text{(since } (1-a)(1-b) \geq 1 - a - b \quad \text{for } a, b \geq 0\text{)}$$

$$\geq 1 - \frac{100}{\mathsf{T}} \sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\langle x^*, \theta^* \rangle}} \left(\sqrt{\mathsf{T} \log \mathsf{T}} \right)$$

$$\text{(via Cauchy-Schwarz inequality)}$$

$$\geq 1 - 100 \sqrt{\frac{d\nu}{\mathsf{T} \langle x^*, \theta^* \rangle}} \log(\mathsf{T}|\mathcal{X}|).$$

Combining the lower bound for the expected rewards in the two parts we get

$$\prod_{t=1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^* \rangle]^{\frac{1}{\mathsf{T}}} \geq \prod_{t=1}^{\mathsf{T}} \left(\mathbb{E}[\langle X_t, \theta^* \rangle \mid E]\ \mathbb{P}\{E\} \right)^{\frac{1}{\mathsf{T}}}$$

$$\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} \right) \left(1 - 100 \sqrt{\frac{d\nu}{\mathsf{T} \langle x^*, \theta^* \rangle}} \log(\mathsf{T}|\mathcal{X}|) \right) \mathbb{P}\{E\}$$

$$\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} - 100 \sqrt{\frac{d\nu}{\mathsf{T} \langle x^*, \theta^* \rangle}} \log(\mathsf{T}|\mathcal{X}|) \right) \mathbb{P}\{E\}$$

$$\begin{aligned}
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} - 100\sqrt{\frac{d\nu}{\mathsf{T}\langle x^*, \theta^*\rangle}} \log\left(\mathsf{T}|\mathcal{X}|\right) \right) \left(1 - \frac{4\log \mathsf{T}}{\mathsf{T}} \right) \\
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))3\sqrt{\mathsf{T}d\nu\log(\mathsf{T}|\mathcal{X}|)}}{\mathsf{T}} - 100\sqrt{\frac{d\nu}{\mathsf{T}\langle x^*, \theta^*\rangle}} \log\left(\mathsf{T}|\mathcal{X}|\right) - \frac{4\log \mathsf{T}}{\mathsf{T}} \right) \\
&\geq \langle x^*, \theta^* \rangle - 100\sqrt{\frac{\langle x^*, \theta^*\rangle d\ \nu}{\mathsf{T}}} \log\left(\mathsf{T}|\mathcal{X}|\right) - 6\langle x^*, \theta^* \rangle \sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} \log(2(d+1)).
\end{aligned}$$

Therefore, the Nash Regret can be bounded as

$$\begin{aligned}
\mathrm{NR}_{\mathsf{T}} &= \langle x^*, \theta^* \rangle - \left(\prod_{t=1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^* \rangle] \right)^{1/\mathsf{T}} \\
&\leq 100\sqrt{\frac{\langle x^*, \theta^*\rangle d\ \nu}{\mathsf{T}}} \log\left(\mathsf{T}|\mathcal{X}|\right) + 6\sqrt{\frac{d\ \nu\ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} \log(2(d+1))\langle x^*, \theta^* \rangle \qquad (3.11) \\
&\leq \left(100\sqrt{\langle x^*, \theta^*\rangle} + 6\log(2(d+1))\langle x^*, \theta^* \rangle \right) \sqrt{\frac{d\nu}{\mathsf{T}}} \log\left(\mathsf{T}|\mathcal{X}|\right) \qquad (3.12)
\end{aligned}$$

Hence, with $\beta = \max\left\{1, \sqrt{\langle x^*, \theta^*\rangle}, \langle x^*, \theta^* \rangle \log d \right\} = \max\left\{1,\ \langle x^*, \theta^* \rangle \log d\right\}$, from equation (3.12) we obtain the desired bound on Nash regret $\mathrm{NR}_{\mathsf{T}} = O\left(\beta\sqrt{\frac{d\ \nu}{\mathsf{T}}} \log(\mathsf{T}|\mathcal{X}|)\right)$. The theorem stands proved. □

Note that, in Theorem 3.1, lower the value of the optimal expected reward, $\langle x^*, \theta^* \rangle$, stronger is the Nash regret guarantee. In particular, with a standard normalization assumption that $\langle x^*, \theta^* \rangle \leq 1$ and for 1-sub Poisson rewards, we obtain a Nash regret of $O\left(\sqrt{\frac{d}{\mathsf{T}}} \log(\mathsf{T}|\mathcal{X}|)\right)$. Also, observe that the regret guarantee provided in Theorem 3.1 depends logarithmically on the size of $\mathcal{X}$. Hence, the Nash regret is small even when $|\mathcal{X}|$ is polynomially large in d.

Computational Efficiency of LinNash. We note that Algorithm 5 (LinNash) executes in polynomial time. In particular, the algorithm calls the subroutine GenerateArmSequence in Part I for computing the John Ellipsoid. Given a set of arm vectors as input, this ellipsoid computation can be performed efficiently (see Chapter 3 in [Tod16]). In fact, for our purposes an approximate version of the John Ellipsoid suffices, and such an approximation can be found much faster [CCLY19]; specifically, in time $O(|\mathcal{X}|^2 d)$. Furthermore, the algorithm solves the D-optimal design problem, once in Part I and at most $O(\log \mathsf{T})$ times in Part II. The D-optimal design is a concave maximization problem, which can be efficiently solved using, say, the Frank-Wolfe algorithm with rank-1 updates. Each iteration takes $O(|\mathcal{X}|^2)$ time, and the total number of iterations is at most $O(d)$ (see, e.g., Chapter 21 of [LS20] and Chapter 3 in [Tod16]). Overall, we get that LinNash is a polynomial-time algorithm.

3.4 Extension of Algorithm LinNash for Infinite Arms

Algorithm 6 LinNash (Nash Confidence Bound Algorithm for Infinite Set of Arms)

```
Input: Arm set $\mathcal{X}$ and horizon of play $T$.
Initialize matrix $\mathbf{V} \leftarrow [0]_{d,d}$ and number of rounds $\widetilde{\mathsf{T}} = 3\sqrt{\mathsf{T}d^{2.5}\nu\log(\mathsf{T})}$.
Part I
Generate arm sequence $\mathcal{S}$ for the first $\widetilde{\mathsf{T}}$ rounds using Algorithm 4.
for $t = 1$ to $\widetilde{\mathsf{T}}$ do
    Pull the next arm $X_t$ from the sequence $S$.
    Observe reward $r_t$ and update $\mathbf{V} \leftarrow \mathbf{V} + X_t X_t^T$
end for
Set estimate $\widehat{\theta} := \mathbf{V}^{-1}\left(\sum_{t=1}^{\widetilde{\mathsf{T}}} r_t X_t\right)$
Find $\gamma = \max_{z\in\mathcal{X}}\langle z, \widehat{\theta}\rangle$
Update $\widetilde{\mathcal{X}} \leftarrow \{x \in \mathcal{X} : \langle x, \widehat{\theta}\rangle \geq \gamma - 16\sqrt{\frac{3\ \gamma\ d^{\frac{5}{2}}\ \nu\ \log(\mathsf{T})}{\widetilde{\mathsf{T}}}}\}$
$\mathsf{T}' \leftarrow \frac{2}{3}\ \widetilde{\mathsf{T}}$
Part II
while end of time horizon $\mathsf{T}$ is reached do
    Initialize $V = [0]_{d,d}$ to be an all zeros $d \times d$ matrix and $s = [0]_d$ to be an all-zeros vector.
    // Beginning of new phase.
    Find the probability distribution $\lambda \in \Delta(\widetilde{\mathcal{X}})$ by maximizing the following objective

        $$\log \mathsf{Det}(\mathbf{V}(\lambda)) \text{ subject to } \lambda \in \Delta(\widetilde{\mathcal{X}}) \text{ and } \mathsf{Supp}(\lambda) \leq d(d+1)/2. \tag{3.13}$$

    for each arm $a$ in $\mathsf{Supp}(\lambda)$ do
        Pull arm $a$ for the next $\lceil \lambda_a\ \mathsf{T}' \rceil$ rounds.
        Observe rewards and Update $\mathbf{V} \leftarrow \mathbf{V} + \lceil \lambda_a \mathsf{T}' \rceil \cdot aa^T$
        Observe $\lceil \lambda_a\ \mathsf{T}' \rceil$ corresponding rewards $z_1, z_2, \ldots$ and update $s \leftarrow s + (\sum_j z_j)a$.
    end for
    Estimate $\widehat{\theta} = \mathbf{V}^{-1}\left(\sum_{t\in\mathcal{E}} r_t X_t\right)$
    Find $\gamma = \max_{z\in\mathcal{X}}\langle z, \widehat{\theta}\rangle$
    $\widetilde{\mathcal{X}} \leftarrow \{x \in \mathcal{X} : \langle x, \widehat{\theta}\rangle \geq \gamma - 16\sqrt{\frac{\gamma\ d^{\frac{5}{2}}\ \log(\mathsf{T})}{\mathsf{T}'}}\}$
    $\mathsf{T}' \leftarrow 2 \times \mathsf{T}'$    // End of phase.
end while
```

The regret guarantee in Theorem 3.1 depends logarithmically on $|\mathcal{X}|$. Such a dependence makes the guarantee vacuous when the set of arms $\mathcal{X}$ is infinitely large (or even $|\mathcal{X}| = \Omega(2^{\sqrt{\mathsf{T}d^{-1}}})$). To resolve this limitation, we extend LinNash with a modified confidence width that depends only on the largest estimated reward $\gamma := \max_{x\in\mathcal{X}}\langle x, \widehat{\theta}\rangle$. Specifically, we consider the confidence width $16\sqrt{\frac{\gamma\ d^{\frac{5}{2}}\ \nu\ \log(\mathsf{T})}{\mathsf{T}'}}$, for all the arms, and select the set of surviving arms in each phase (of Part II of the algorithm for infinite arms) as follows:

$$\widetilde{\mathcal{X}} = \left\{ x \in \mathcal{X} : \langle x, \widehat{\theta} \rangle \geq \gamma - 16\sqrt{\frac{\gamma\ d^{\frac{5}{2}}\ \nu\ \log(\mathsf{T})}{\mathsf{T}'}} \right\} \tag{3.14}$$

See Algorithm 6 for details. The theorem below is the main result of this section.

Theorem 3.2. *For any given stochastic linear bandits problem with set of arms $\mathcal{X} \subset \mathbb{R}^d$, time horizon $\mathsf{T} \in \mathbb{Z}_+$, and ν-sub-Poisson rewards, Algorithm 5 achieves Nash regret*

$$\mathrm{NR}_{\mathsf{T}} = O\left(\beta \frac{d^{\frac{5}{4}}\sqrt{\nu}}{\sqrt{\mathsf{T}}} \log(\mathsf{T}) \right),$$

Here, $\beta = \max\{1,\ \langle x^*, \theta^* \rangle \log d\}$, with $x^* \in \mathcal{X}$ denoting the optimal arm and θ^* the (unknown) parameter vector.

Proof of Theorem 3.2 and a detailed regret analysis of Algorithm 6 can be found in Appendix 3.9.

3.5 Experiments

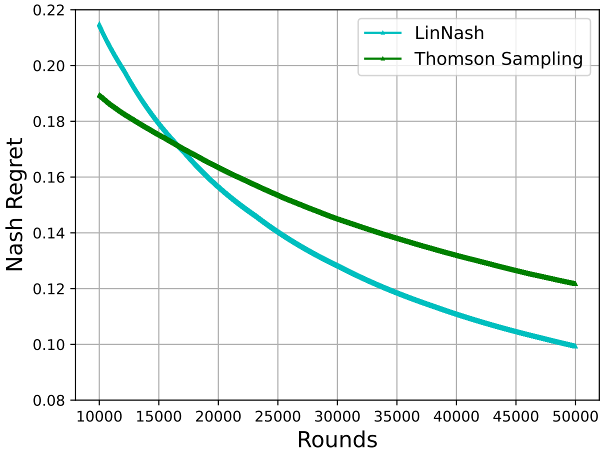
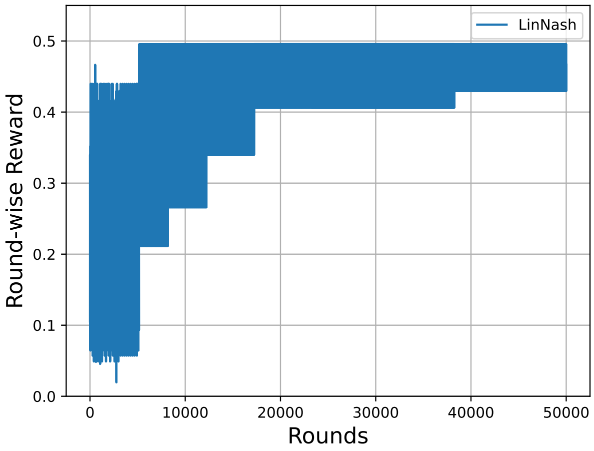


Figure 3.1: figure
Nash Regret comparison of LinNash and Thompson Sampling



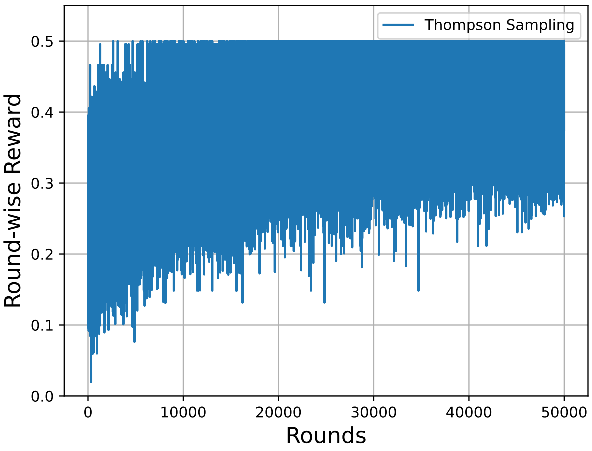


Figure 3.2: figure
Round-wise reward for LinNash

Figure 3.3: figure
Round-wise reward for Thompson Sampling

We conduct experiments to compare the performance of our algorithm LinNash with Thompson Sampling on synthetic data. For a comparison, we select Thompson Sampling (Algorithm 1 in [AG13]), instead of UCB/OFUL, since randomization is essential to achieve meaningful Nash Regret guarantees.

We fine-tune the parameters of both algorithms and evaluate their performance in the following experimental setup: We fix the ambient dimension $d = 80$, the number of arms $|\mathcal{X}| = 10000$, and the number of rounds $\mathsf{T} = 50000$. Both the unknown parameter vector, θ^*, and the arm embeddings are sampled from a multivariate Gaussian distribution. Subsequently, the arm embeddings are shifted and normalized to ensure that all mean rewards are non-negative, with the maximum reward mean being set to 0.5. Upon pulling an arm, we observe a Bernoulli random variable with a probability corresponding to its mean reward.

In this experimental setting, we observe a significant performance advantage of LinNash over Thompson Sampling. We plot our results in Figure 5.1, which shows that the Nash regret of LinNash decreases notably faster than that of Thompson Sampling.

Another notable advantage of LinNash evident from the experiments is due to successive elimination. The variance in the quality of arms pulled decreases as the number of rounds progresses – see Figures 3.2 and 3.3. This is due to the bulk elimination of suboptimal arms at regular intervals. In contrast, Thompson Sampling incurs a large variance in quality of arms being pulled even after several rounds, since no arms are being eliminated at any point.

3.6 Proof of Lemmas 3.1 and 3.2

Lemma 3.1. *Any non-negative random variable $X \in [0, \mathsf{B}]$ is B-sub-Poisson, i.e., if mean $\mathbb{E}[X] = \mu$, then for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[e^{\lambda X}] \leq \exp\left(B^{-1}\mu\left(e^{B\lambda} - 1\right)\right)$.*

Proof. For random variable X we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda X\right)\right] &= \mathbb{E}\left[\exp\left(\lambda B\frac{X}{B} + (1 - \frac{X}{B})0\right)\right] \\
&\leq \mathbb{E}\left[\frac{X}{B}e^{(\lambda B)} + \left(1 - \frac{X}{B}\right)e^{0}\right] && \text{(due to convexity of } e^x\text{)} \\
&= 1 + \frac{\mathbb{E}\left[X\right]}{\mathsf{B}}\left(e^{\lambda \mathsf{B}} - 1\right) \\
&\leq 1 + \frac{\mu}{\mathsf{B}}\left(e^{\lambda \mathsf{B}} - 1\right) \\
&\leq \exp\left(\frac{\mu}{\mathsf{B}}\left(e^{\lambda \mathsf{B}} - 1\right)\right).
\end{aligned}$$

□

Lemma 3.2. *Let X be a non-negative sub-Gaussian random variable X with mean $\mu = \mathbb{E}[X]$ and sub-Gaussian norm σ. Then, X is also $\left(\frac{\sigma^2}{\mu}\right)$-sub-Poisson.*

Proof. Since X is a σ-sub-Gaussian random variable, for any non-negative scalar $s \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[e^{sX}] &\leq \exp\left(s\mu + \frac{(s\sigma)^2}{2}\right) \\
&= \exp\left(\frac{\mu^2}{\sigma^2}\left(\frac{s\sigma^2}{\mu} + \frac{1}{2}\left(\frac{s\sigma^2}{\mu}\right)^2\right)\right) \qquad (3.15)
\end{aligned}$$

The fact that X is a positive random variable implies that the mean $\mu > 0$. Also, the considered scalar $s \geq 0$ and, hence, the term $\frac{s\sigma^2}{\mu} > 0$. Also, recall that $e^x \geq 1 + x + \frac{x^2}{2}$, for any non-negative x. Using these observations and equation (3.15), we obtain

$$\mathbb{E}[e^{sX}] \leq \exp\left(\frac{\mu^2}{\sigma^2}\left(e^{\frac{s\sigma^2}{\mu}} - 1\right)\right) \qquad (3.16)$$

For random variable X, inequality (3.16) ensures that the required mgf bound (equation (3.2)) holds for all non-negative s and with sub-Poisson parameter equal to $\frac{\sigma^2}{\mu}$.

We next complete the proof by showing that the mgf bound holds for negative s as well. Towards this, write $B := \frac{\sigma^2}{\mu}$ and define random variable $Y := \mathbf{1}_{\{X \leq B\}}\ X + \mathbf{1}_{\{X > B\}}\ B$. Note that Y is a positive, bounded random variable. Furthermore, for any negative s, we have $\exp(sY) \geq \exp(sX)$. Therefore, for a negative s, it holds that $\mathbb{E}[\exp(sX)] \leq \mathbb{E}[\exp(sY)]$. Since positive random variable $Y \in [0, B]$, the mgf bound obtained in Lemma 3.1 gives us

$$\mathbb{E}[e^{sX}] \leq \mathbb{E}\left[e^{sY}\right] \leq \exp\left(\frac{\mu}{\mathsf{B}}\left(e^{s\mathsf{B}} - 1\right)\right).$$

Since $B := \frac{\sigma^2}{\mu}$, the mgf bound (equation (3.2)) on X holds for negative s as well. This, overall, shows that X is a $\left(\frac{\sigma^2}{\mu}\right)$-sub-Poission random variable. The lemma stands proved. □

3.7 Proof of Concentration Bounds

Lemma 3.6. *Let $x_1, x_2, \ldots, x_s \in \mathbb{R}^d$ be a fixed set of vectors and let $r_1, r_2, \ldots, r_s$ be independent ν-sub-Poisson random variables satisfying $\mathbb{E}r_s = \langle x_s, \theta^* \rangle$ for some unknown θ^*. Further, let matrix $\mathbf{V} = \sum_{j=1}^{s} x_j x_j^T$ and $\widehat{\theta} = \mathbf{V}^{-1}\left(\sum_j r_j x_j\right)$ be the least squares estimator of θ^*. Consider any $z \in \mathbb{R}^d$ with the property that $z^T \mathbf{V}^{-1} x_j \leq \gamma$ for all $j \in [s]$. Then, for any $\delta \in [0,1]$ we have*

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{\delta^2 \langle z, \theta^*\rangle}{3\nu\gamma}\right) \quad \textit{and} \tag{3.8}$$

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq (1-\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{\delta^2 \langle z, \theta^*\rangle}{2\nu\gamma}\right). \tag{3.9}$$

Proof. We use the Chernoff method to get an upper bound on the desired probabilities, as shown below

$$\begin{aligned}
\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} &= \mathbb{P}\left(\exp(c\ \langle z, \widehat{\theta}\rangle) \geq \exp(c(1+\delta)\langle z, \theta^*\rangle)\right) && \text{(for some constant } c\text{)} \\
&\leq \frac{\mathbb{E}[\exp\left(c\ z^T \mathbf{V}^{-1}\left(\sum_t r_t x_t\right)\right)]}{\exp(c\ (1+\delta)\langle z, \theta^*\rangle)} \\
&= \frac{\prod_{t=1}^{s} \mathbb{E}[\exp\left(c\ r_t \mathbf{V}^{-1} x_t\right)]}{\exp(c\ (1+\delta)\langle z, \theta^*\rangle)} && (r_t\text{'s are independent)} \\
&\leq \frac{\prod_{t=1}^{s} \exp\left(\frac{\mathbb{E}[r_t]}{\nu}\left(e^{c\nu z^T \mathbf{V}^{-1} x_t} - 1\right)\right)}{\exp(c\ (1+\delta)\langle z, \theta^*\rangle)} && (r_t \text{ is sub Poisson)} \\
&= \exp\left(-c\langle z, \theta^*\rangle(1+\delta) + \sum_{t=1}^{s} \frac{\langle x, \theta^*\rangle}{\nu}\left(e^{c\ \nu z^T \mathbf{V}^{-1} x_t} - 1\right)\right).
\end{aligned}$$

Substituting $c = \frac{\log(1+\delta)}{\nu\gamma}$, we get

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{\langle z, \theta^*\rangle}{\nu\gamma}(1+\delta)\log(1+\delta) + \sum_{t=1}^{s}\frac{\langle x_t, \theta^*\rangle}{\nu}\left((1+\delta)^{\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t} - 1\right)\right). \tag{3.17}$$

Since $\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t \leq 1$ we have $(1+\delta)^{\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t} \leq 1 + \delta \cdot \frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t$. Substituting in (3.17) we get

$$\begin{aligned}
&\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} \\
&\leq \exp\left(-\frac{1}{\nu\gamma}\langle z, \theta^*\rangle(1+\delta)\log(1+\delta) + \sum_{t=1}^{s}\langle x_t, \theta^*\rangle \cdot \frac{\delta}{\nu\gamma}z^T\mathbf{V}^{-1}x_t\right) \\
&= \exp\left(-\frac{1}{\nu\gamma}\langle z, \theta^*\rangle(1+\delta)\log(1+\delta) + \frac{\delta}{\nu\gamma}\sum_{t=1}^{s}\theta^{*T}x_t x_t^T\mathbf{V}^{-1}z\right) && \text{(rearranging terms)} \\
&= \exp\left(-\frac{1}{\nu\gamma}\langle z, \theta^*\rangle(1+\delta)\log(1+\delta) + \frac{\delta}{\nu\gamma}\langle z, \theta^*\rangle\right). && \left(\textstyle\sum_{t=1}^{s} x_t x_t^T = \mathbf{V}\right)
\end{aligned}$$

Using the logarithmic inequality $\log(1+\delta) \geq \frac{2\delta}{2+\delta}$, we further simplify as

$$\begin{aligned}
\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \geq (1+\delta)\langle z, \theta^*\rangle\right\} &\leq \exp\left(-\frac{\langle z, \theta^*\rangle}{\nu\gamma}\left((1+\delta)\log(1+\delta) - \delta\right)\right) \\
&\leq \exp\left(\frac{-\delta^2\langle z, \theta^*\rangle}{(2+\delta)\,\nu\gamma}\right) \\
&\leq \exp\left(\frac{-\delta^2\langle z, \theta^*\rangle}{3\nu\gamma}\right). && \text{(since } \delta \in [0,1])
\end{aligned}$$

Following similar steps and substituting $c = \frac{\log(1-\delta)}{\nu\gamma}$, we obtain a bound on the lower tail (inequality 3.9):

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq (1-\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(-\frac{1}{\nu\gamma}\langle z, \theta^*\rangle(1-\delta)\log(1-\delta) - \frac{\delta}{\nu\gamma}\langle z, \theta^*\rangle\right).$$

Now, using the logarithmic inequality $(1-\delta)\log(1-\delta) \geq -\delta + \frac{\delta^2}{2}$, we get

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq (1-\delta)\langle z, \theta^*\rangle\right\} \leq \exp\left(\frac{-\delta^2\langle z, \theta^*\rangle}{2\nu\gamma}\right)$$

□

Combining (3.9) and (3.8) we get the following Corollary.

Corollary 3.1. *Using the notations as in Lemma 3.6, we have*

$$\mathbb{P}\left\{|\langle z,\widehat{\theta}\rangle - \langle z,\theta^*\rangle| \geq \delta\langle z,\theta^*\rangle\right\} \leq 2\,\exp\left(-\frac{\delta^2\langle z,\theta^*\rangle}{3\gamma}\right). \tag{3.18}$$

The next two lemmas are variants of Lemma 3.6 where we bound the error in terms of an upper bound on $\langle z,\theta^*\rangle$.

Lemma 3.9. *Let $x_1, x_2, \ldots, x_s \in \mathbb{R}^d$ be a fixed set of vectors and let $r_1, r_2, \ldots, r_s$ be independent $\nu-$sub Poisson random variables satisfying $\mathbb{E}r_s = \langle x_s,\theta^*\rangle$ for some unknown θ^*. In that case, let matrix $\mathbf{V} = \sum_{j=1}^s x_j x_j^T$ and $\widehat{\theta} = \mathbf{V}^{-1}\left(\sum_j r_j x_j\right)$ be the least squares estimator of θ^*. Consider any $z \in \mathbb{R}^d$ that satisfies $z^T\mathbf{V}^{-1}x_j \leq \gamma$ for all $j \in [s]$ and $\langle z,\theta^*\rangle \leq \alpha$. Then for any $\delta \in [0,1]$ we have*

$$\mathbb{P}\left\{\langle z,\widehat{\theta}\rangle \geq (1+\delta)\alpha\right\} \leq e^{-\frac{\delta^2\alpha}{3\gamma\nu}}. \tag{3.19}$$

Proof. Following the same approach as in the proof of Lemma 3.6, we have

$$\begin{aligned}\mathbb{P}\left\{\langle z,\widehat{\theta}\rangle \geq (1+\delta)\alpha\right\} &\leq \frac{\mathbb{E}[\exp(c\;z^T\mathbf{V}^{-1}\left(\sum_t r_t x_t\right))]}{\exp(c\;(1+\delta)\alpha)}\\ &\leq \exp\left(-c\alpha(1+\delta) + \sum_{t=1}^{s}\frac{\langle x_t,\theta^*\rangle}{\nu}\left(e^{c\nu z^T\mathbf{V}^{-1}x_t} - 1\right)\right)\\ &\qquad\qquad\qquad\qquad (r_t \text{ are sub-poisson and independent})\end{aligned}$$

Now, substituting $c = \frac{1}{\nu\gamma}\log(1+\delta))$ and using $(1+\delta)^{\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t} \leq 1 + \delta\cdot\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t$ we have

$$\begin{aligned}\mathbb{P}\left\{\langle z,\widehat{\theta}\rangle \geq (1+\delta)\alpha\right\} &\leq \exp\left(-\frac{1}{\gamma\nu}\alpha(1+\delta)\log(1+\delta) + \sum_{t=1}^{s}\frac{\langle x_t,\theta^*\rangle}{\nu}\theta^*\left((1+\delta)^{\frac{1}{\gamma}z^T\mathbf{V}^{-1}x_t} - 1\right)\right)\\ &\leq \exp\left(-\frac{1}{\nu\gamma}\alpha(1+\delta)\log(1+\delta) + \frac{\delta}{\nu\gamma}\sum_{t=1}^{s}\theta^{*T}x_t x_t^T\mathbf{V}^{-1}Z\right)\\ &= \exp\left(-\frac{1}{\nu\gamma}\alpha(1+\delta)\log(1+\delta) + \frac{\delta}{\nu\gamma}\langle z,\theta^*\rangle\right)\\ &\leq \exp\left(-\frac{1}{\nu\gamma}\alpha(1+\delta)\log(1+\delta) + \frac{\delta}{\nu\gamma}\alpha\right) && (\alpha \geq \langle z,\theta^*\rangle)\\ &\leq \exp\left(\frac{-\delta^2\alpha}{(2+\delta)\,\nu\gamma}\right) && (\text{using } \log(1+\delta) \geq \tfrac{2\delta}{2+\delta})\end{aligned}$$

Since $\delta \in [0,1]$, we have the desired result. □

Lemma 3.10. *Using the same notations as in Lemma 3.9, for any $\delta \in [0,1]$, the following holds*

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq \langle z, \theta^*\rangle - \delta\alpha\right\} \leq \exp\left(-\frac{\delta^2\alpha}{2\gamma\nu}\right) \tag{3.20}$$

Proof. Using steps similar to the previous lemmas, we obtain

$$\begin{aligned}\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq \langle z, \theta^*\rangle - \delta\alpha\right\} &\leq \frac{\mathbb{E}[\exp(c\; z^T\mathbf{V}^{-1}\left(\sum_t r_t x_t\right))]}{\exp(c\;(\langle z, \theta^*\rangle - \delta\alpha))} \\ &\leq \exp\left(c\alpha\delta + c\langle z, \theta^*\rangle + \sum_{t=1}^{s}\frac{\langle x_t, \theta^*\rangle}{\nu}\left(e^{c\nu z^T\mathbf{V}^{-1}x_t} - 1\right)\right) \\ &\qquad (r_t \text{ are sub-poisson and independent})\end{aligned}$$

Substituting $c = \frac{\log(1-\delta)}{\nu\gamma}$ and simplifing we get

$$\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq \langle z, \theta^*\rangle - \delta\alpha\right\} \leq \exp\left(-\frac{\langle z, \theta^*\rangle}{\nu\gamma}\left(\log\left(1-\delta\right)+\delta\right) + \frac{\alpha}{\nu\gamma}\delta\log\left(1-\delta\right)\right)$$

Note that since $\log(1-\delta)+\delta$ is negative, we can upper bound the above expression by replacing $\langle z, \theta^*\rangle$ with α.

$$\begin{aligned}\mathbb{P}\left\{\langle z, \widehat{\theta}\rangle \leq \langle z, \theta^*\rangle - \delta\alpha\right\} &\leq \exp\left(-\frac{\alpha}{\nu\gamma}\left(\log\left(1-\delta\right)+\delta-\delta\log\left(1-\delta\right)\right)\right) \\ &\leq \exp\left(-\frac{\delta^2\alpha}{2\nu\gamma}\right). \qquad (\text{since } (1-\delta)\log(1-\delta) \geq -\delta + \tfrac{\delta^2}{2})\end{aligned}$$

Hence, the lemma stands proved. □

3.8 Regret Analysis of Algorithm 5: Proofs of Lemmas 3.7 and 3.8

We will first define events E_1 and E_2 for each phase of the algorithm and show that they hold with high probability. We will use the events in the regret analysis.

- Event E_1: At the end of Part I, let $\widehat{\theta}$ be the unbiased estimator of θ^* and $\widetilde{\mathsf{T}}$ be as defined in Algorithm 5. All arms $x \in \mathcal{X}$ with $\langle x, \theta^*\rangle < 10\sqrt{\frac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ satisfy

$$\langle x, \widehat{\theta}\rangle \leq 20\sqrt{\frac{d\nu\log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}}} \tag{3.21}$$

In addition, all arms $x \in \mathcal{X}$ with $\langle x, \theta^* \rangle \geq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ satisfy

$$|\langle x, \theta^* \rangle - \langle x, \widehat{\theta} \rangle| \leq 3\sqrt{\frac{d\nu \langle x, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\widetilde{\mathsf{T}}}} \quad \text{and} \tag{3.22}$$

$$\frac{1}{2}\langle x, \theta^* \rangle \leq \langle x, \widehat{\theta} \rangle \leq \frac{4}{3}\langle x, \theta^* \rangle. \tag{3.23}$$

- Event E_2: Let $\widetilde{\mathcal{X}}$ denote the surviving set of arms at the start of a phase in Part II, and T' be as defined in Algorithm 5. For all phases and for all $x \in \widetilde{\mathcal{X}}$ such that $\langle x, \theta^* \rangle \geq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$, the estimator $\widehat{\theta}$ (calculated at the end of a phase) satisfies

$$|\langle x, \theta^* \rangle - \langle x, \widehat{\theta} \rangle| \leq 3\sqrt{\frac{d\nu \langle x, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} \quad \text{and} \tag{3.24}$$

$$\frac{1}{2}\langle x, \theta^* \rangle \leq \langle x, \widehat{\theta} \rangle \leq \frac{4}{3}\langle x, \theta^* \rangle. \tag{3.25}$$

3.8.1 Supporting Lemmas

Lemma 3.11 (Chernoff Bound). *Let $Z_1, \ldots, Z_n$ be independent Bernoulli random variables. Consider the sum $S = \sum_{r=1}^{n} Z_r$ and let $\mu = \mathbb{E}[S]$ be its expected value. Then, for any $\varepsilon \in [0, 1]$, we have*

$$\mathbb{P}\{S \leq (1-\varepsilon)\mu\} \leq \exp\left(-\frac{\mu\varepsilon^2}{2}\right).$$

Lemma 3.12. *During Part* I*, arms from D-optimal design are added to S at least $\widetilde{\mathsf{T}}/3$ times with probability greater than $1 - \frac{1}{\mathsf{T}}$.*

Proof. We use Lemma 3.11 with Z_i as indicator random variables that take value one when an arm from $\mathcal{A}$ (the support of λ in the optimal design) is chosen. By setting $\varepsilon = \frac{1}{3}$ and $\mu = \frac{\widetilde{\mathsf{T}}}{2}$, we obtain the required probability bound. □

Lemma 3.13. *Using the notations in Algorithm 4, if the event in Lemma 3.12 holds, then for each $x \in \mathcal{X}$ and each round t in Part* I *of the algorithm, we have*

$$x^T \mathbf{V}^{-1} X_t \leq \frac{3d}{\widetilde{\mathsf{T}}},$$

where X_t is the arm pulled in round t.

Proof. Let $\mathbf{U}(\lambda)$ and λ denote the optimal design matrix (as defined in (3.4)) and the solution to the D-optimal design problem in Algorithm 4, respectively. That is, λ is the solution of the optimization problem stated in equation (3.5) and $\mathbf{U}(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^T$. Lemma 3.3 implies that $||x||_{\mathbf{U}(\lambda)^{-1}} \leq \sqrt{d}$ for all $x \in \mathcal{X}$.

Next, note that the construction of the sequence $\mathcal{S}$ in Part I (Subroutine `GenerateArmSequence`) and the event specified in Lemma 3.12 give us $\mathbf{V} \succ \frac{\widetilde{\mathsf{T}}}{3}\mathbf{U}(\lambda)$. Hence,

$$
\begin{aligned}
x^T \mathbf{V}^{-1} X_t &\leq \|x\|_{\mathbf{V}^{-1}} \left\|\mathbf{V}^{-1} X_t\right\|_{\mathbf{V}} && \text{(via Hölder's inequality)} \\
&= \|x\|_{\mathbf{V}^{-1}} \|X_t\|_{\mathbf{V}^{-1}} \\
&\leq \|x\|_{\left(\frac{\widetilde{\mathsf{T}}}{3}\mathbf{U}(\lambda)\right)^{-1}} \|X_t\|_{\left(\frac{\widetilde{\mathsf{T}}}{3}\mathbf{U}(\lambda)\right)^{-1}} && \left(\text{since } \mathbf{V} \succ \tfrac{\widetilde{\mathsf{T}}}{3}\mathbf{U}(\lambda)\right) \\
&= \sqrt{\frac{3}{\widetilde{\mathsf{T}}}} \|x\|_{\mathbf{U}(\lambda)^{-1}} \sqrt{\frac{3}{\widetilde{\mathsf{T}}}} \|X_t\|_{\mathbf{U}(\lambda)^{-1}} \\
&\leq \sqrt{\frac{3d}{\widetilde{\mathsf{T}}}} \sqrt{\frac{3d}{\widetilde{\mathsf{T}}}} && \text{(by Lemma 3.3)} \\
&= \frac{3d}{\widetilde{\mathsf{T}}}.
\end{aligned}
$$

□

The next lemma lower bounds the probability of event E_1 (see equations (3.21), (3.22), and (3.23)).

Lemma 3.14. *Event E_1 holds with probability at least $1 - \frac{6}{T}$.*

Proof. First, consider all arms $x \in \mathcal{X}$ for which $\langle x, \theta^* \rangle < 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$. Here, we invoke Lemma 3.9, with $\gamma = \frac{3d}{\widetilde{\mathsf{T}}}$ (as derived in Lemma 3.13), $\alpha = 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$, and $\delta = 1$, to obtain

$$
\begin{aligned}
\mathbb{P}\left\{\langle x, \widehat{\theta}\rangle \leq 20\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}\right\} &\leq \exp\left(-\frac{\delta^2 \alpha}{3\gamma\nu}\right) \\
&\leq \exp\left(-\frac{10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}\,3\sqrt{Td\nu\log(\mathsf{T}|\mathcal{X}|)}}{3\nu d}\right) \\
&\leq \frac{1}{\mathsf{T}|\mathcal{X}|} \qquad (3.26)
\end{aligned}
$$

Next, we consider arms $x \in \mathcal{X}$ such that $\langle x, \theta^* \rangle \geq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ and for such arms establish equations (3.22) and (3.23). Towards this, we invoke Lemma 3.6, with parameters $\gamma = \frac{3d}{\widetilde{\mathsf{T}}}$ and

$\delta = 3\sqrt{\frac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\langle x,\theta^*\rangle\widetilde{\mathsf{T}}}}$. It is relevant to note that here $\delta \in [0,1]$ – this containment follows from the condition $\langle x,\theta^*\rangle \geq 10\sqrt{\frac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ and $\widetilde{\mathsf{T}} = 3\sqrt{\mathsf{T}d\nu\log(\mathsf{T}|\mathcal{X}|)}$. Therefore,

$$
\begin{aligned}
\mathbb{P}\left\{|\langle x,\theta^*\rangle - \langle x,\widehat{\theta}\rangle| \geq 3\sqrt{\frac{d\nu\langle x,\theta^*\rangle\log(\mathsf{T}|\mathcal{X}|)}{\widetilde{\mathsf{T}}}}\right\} &= \mathbb{P}\left\{|\langle x,\theta^*\rangle - \langle x,\widehat{\theta}\rangle| \geq \delta\langle x,\theta^*\rangle\right\} \\
&\qquad \left(\text{since } \delta = 3\sqrt{\tfrac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\langle x,\theta^*\rangle\widetilde{\mathsf{T}}}}\right) \\
&\leq 2\exp\left(-\frac{\frac{9d\nu\log(\mathsf{T}|\mathcal{X}|)}{\langle x,\theta^*\rangle\widetilde{\mathsf{T}}}\langle x,\theta^*\rangle}{3\nu\frac{3d}{\widetilde{\mathsf{T}}}}\right) \quad \text{(Lemma 3.6)} \\
&= \frac{2}{\mathsf{T}|\mathcal{X}|} \qquad (3.27)
\end{aligned}
$$

For establishing equation (3.23), we invoke Lemma 3.6 again, now with $\gamma = \frac{3d}{\widetilde{\mathsf{T}}}$ and $\delta = \frac{1}{3}$:

$$
\begin{aligned}
\mathbb{P}\left\{\langle x,\widehat{\theta}\rangle \geq \frac{4}{3}\langle x,\theta^*\rangle\right\} &\leq \exp\left(-\frac{3\sqrt{\mathsf{T}\nu d\log(\mathsf{T}|\mathcal{X}|)}\,\langle x,\theta^*\rangle}{27\nu d}\right) \\
&\leq \exp\left(-\frac{3\sqrt{\mathsf{T}\nu d\log(\mathsf{T}|\mathcal{X}|)}\;10\sqrt{\frac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}}{27\nu d}\right) \\
&\leq \frac{1}{\mathsf{T}|\mathcal{X}|} \qquad (3.28)
\end{aligned}
$$

Similarly, with $\delta = \frac{1}{2}$, Lemma 3.6 gives us

$$
\mathbb{P}\left\{\langle x,\widehat{\theta}\rangle \leq \frac{1}{2}\langle x,\theta^*\rangle\right\} \leq \frac{1}{\mathsf{T}|\mathcal{X}|} \qquad (3.29)
$$

Finally, we combine (3.26), (3.27), (3.28) and (3.29), and apply a union bound over all arms in $\mathcal{X}$. Then, conditioning on the event in Lemma 3.12 leads to the stated probability bound. The lemma stands proved. □

The next lemma shows that event E_2 (see equations (3.24) and (3.25)) holds with high probability

Lemma 3.15. *Event E_2 holds with probability at least $1 - \frac{3\log\mathsf{T}}{T}$.*

Proof. Consider any phase in Part II and let $\mathbf{U}(\lambda)$ be the optimal design matrix obtained after solving the D-optimal design problem at the start of the phase. By Lemma 3.3, for all $x, z \in \widetilde{\mathcal{X}}$

we have

$$
\begin{aligned}
z^T \mathbf{V}^{-1} x &\leq \|z\|_{\mathbf{V}^{-1}} \left\| \mathbf{V}^{-1} x \right\|_{\mathbf{V}} && \text{(via Hölder's inequality)} \\
&\leq \|z\|_{\mathbf{V}^{-1}} \|x\|_{\mathbf{V}^{-1}} \\
&\leq \sqrt{\frac{d}{\mathsf{T}'}} \sqrt{\frac{d}{\mathsf{T}'}} = \frac{d}{\mathsf{T}'}
\end{aligned}
$$

First, we address equation (3.24). In particular, we instantiate Lemma 3.6 with $\delta = 3\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\langle x, \theta^* \rangle \mathsf{T}'}}$ and $\gamma = \frac{d}{\mathsf{T}'}$. Note that given the lower bound on $\langle x, \theta^* \rangle$ and the inequality $\mathsf{T}' \geq 2\sqrt{\mathsf{T} d\nu \log(\mathsf{T}|\mathcal{X}|)}$ ensure that δ lies in $[0, 1]$. Hence, substituting these values of δ and γ in Lemma 3.6, we obtain

$$
\begin{aligned}
\mathbb{P}\left\{ |\langle x, \theta^* \rangle - \langle x, \widehat{\theta} \rangle| \geq 3\sqrt{\frac{d\nu \langle x, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} \right\} &\leq 2\exp\left(-\frac{\frac{9 d\nu \log(\mathsf{T}|\mathcal{X}|)}{\langle x, \theta^* \rangle \mathsf{T}'} \cdot \langle x, \theta^* \rangle}{3 \frac{d\nu}{\mathsf{T}'}} \right) \\
&\leq \frac{2}{(\mathsf{T}|\mathcal{X}|)^3}
\end{aligned}
$$

Next, following a similar approach as in the proof of Lemma 3.14, we use Lemma 3.6 with $\delta = \frac{1}{3}$ and $\delta = \frac{1}{2}$ to establish the upper and lower bounds of equation (3.25), respectively. Applying a union bound across arms in $\widetilde{\mathcal{X}}$ and over all—at most $\log \mathsf{T}$—phases, we obtain the desired probability bound of $1 - \frac{3 \log \mathsf{T}}{\mathsf{T}}$. □

Corollary 3.2.

$$
\mathbb{P}\{E_1 \cap E_2\} \geq 1 - \frac{4 \log \mathsf{T}}{\mathsf{T}}.
$$

Proof. From Lemma 3.14 we have $\mathbb{P}\{E_1\} \geq 1 - \frac{6}{\mathsf{T}}$. Furthermore, from Lemma 3.15 we have $\mathbb{P}\{E_2\} \geq 1 - \frac{3 \log \mathsf{T}}{\mathsf{T}}$. Applying a union bound on the complements of these two events establishes the corollary. □

Lemma 3.16. *Consider any bandit instance with* $\langle x^*, \theta^* \rangle \geq 192\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$. *If event* E_1 *holds, then any arm with mean* $\langle x, \theta^* \rangle \leq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ *is eliminated after Part* I *of Algorithm 5.*

Proof. We will show that in the given bandit instance and under the event E_1, for each arm $x \in \mathcal{X}$ with mean $\langle x, \theta^* \rangle \leq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$ the upper Nash confidence bound (see equation (3.7)) is less than the lower confidence bound of the optimal arm x^*. Hence, all such arms x are eliminated from consideration in Line 8 of Algorithm 5. This will establish the lemma.

The upper Nash confidence bound of arm x at the end of Part I is defined as

$$\begin{aligned}
\text{UNCB}\left(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) &= \langle x, \widehat{\theta}\rangle + 6\sqrt{\frac{3\langle x, \widehat{\theta}\rangle\ d\ \nu\ \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} \\
&\leq 20\sqrt{\frac{d\ \nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} + 6\sqrt{\frac{3\langle x, \widehat{\theta}\rangle\ d\ \nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} && \text{(via event } E_1\text{)} \\
&\leq 20\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} + 6\sqrt{\frac{3 \cdot 20\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} d\nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{3\sqrt{\mathsf{T}\nu d \log(\mathsf{T}|\mathcal{X}|)}}} && \text{(substituting } \widetilde{\mathsf{T}}\text{)} \\
&\leq 47\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} && (3.30)
\end{aligned}$$

In the given bandit instance and under event E_1, for the optimal arm x^*, we have

$$\begin{aligned}
\langle x^*, \widehat{\theta}\rangle &\leq \langle x^*, \theta^*\rangle + 3\sqrt{\frac{d\nu\langle x^*, \theta^*\rangle \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} \\
&= \langle x^*, \theta^*\rangle \left(1 + 3\sqrt{\frac{d\nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{\langle x^*, \theta^*\rangle 3\sqrt{\mathsf{T}d\ \nu \log(\mathsf{T}|\mathcal{X}|)}}}\right) && \text{(substituting } \widetilde{\mathsf{T}}\text{)} \\
&\leq \langle x^*, \theta^*\rangle \left(1 + 3\sqrt{\frac{d\nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{192\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}3\sqrt{\mathsf{T}\nu d \log(\mathsf{T}|\mathcal{X}|)}}}\right) \\
& && \left(\text{using } \langle x^*, \theta^*\rangle \geq 192\sqrt{\tfrac{d\nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}}}\right) \\
&= \frac{17}{16}\langle x^*, \theta^*\rangle. && (3.31)
\end{aligned}$$

Therefore, the lower Nash confidence bound of x^* satisfies

$$\begin{aligned}
\text{LNCB}\left(x^*, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) &= \langle x^*, \widehat{\theta}\rangle - 6\sqrt{\frac{3\langle x^*, \widehat{\theta}\rangle\ d\ \nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} \\
&\geq \langle x^*, \theta^*\rangle - 3\sqrt{\frac{d\ \nu\ \langle x^*, \theta^*\rangle \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} - 6\sqrt{\frac{3\langle x^*, \widehat{\theta}\rangle\ d\ \nu\ \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} \\
& && \text{(via (3.22) in event } E_1\text{)} \\
&\geq \langle x^*, \theta^*\rangle - \left(3 + 6\sqrt{\frac{51}{16}}\right)\sqrt{\frac{d\ \nu\ \langle x^*, \theta^*\rangle \log\left(\mathsf{T}|\mathcal{X}|\right)}{\widetilde{\mathsf{T}}}} \\
& && \left(\text{since } \langle x^*, \widehat{\theta}\rangle \leq \tfrac{17}{16}\langle x^*, \theta^*\rangle \text{ via (3.31)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \langle x^*, \theta^* \rangle \left(1 - 14\sqrt{\frac{d\nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\langle x^*, \theta^* \rangle \widetilde{\mathsf{T}}}}\right) \\
&\geq \langle x^*, \theta^* \rangle \left(1 - 14\sqrt{\frac{d\nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{192\sqrt{\frac{d\nu \ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}3\sqrt{\mathsf{T}d\nu\log(\mathsf{T}|\mathcal{X}|)}}}\right) \\
&\geq \frac{5}{12}\langle x^*, \theta^* \rangle \\
&\geq 80\sqrt{\frac{d\nu\log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}} \qquad (3.32)
\end{aligned}
$$

Equations (3.32) and (3.30) imply

$$
\mathrm{UNCB}\left(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) < \mathrm{LNCB}\left(x^*, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) \qquad (3.33)
$$

As mentioned previously, Line 8 in Algorithm 5 eliminates all arms x that satisfy inequality (3.33). Hence, the lemma stands proved □

3.8.2 Proofs of Lemmas 3.7 and 3.8

Lemma 3.7. *Consider any bandit instance in which for the optimal arm $x^* \in \mathcal{X}$ we have $\langle x^*, \theta^* \rangle \geq 192\sqrt{\frac{d\ \nu}{T}}\log(\mathsf{T}|\mathcal{X}|)$. Then, with probability at least $\left(1 - \frac{4\log\mathsf{T}}{\mathsf{T}}\right)$, the optimal arm x^* always exists in the surviving set $\widetilde{\mathcal{X}}$ in Part* I *and in every phase in Part* II *of Algorithm 5.*

Proof. We will show that, under events E_1 and E_2, throughout the execution of Algorithm 5 the UNCB of the optimal arm x^* is never less than the LNCB of any arm x. Hence, then the optimal arm x^* never satisfies the elimination criterion in Algorithm 5 and, hence, x^* always exists in the surviving set of arms.

First, we consider arms x with the property that $\langle x, \theta^* \rangle < 10\sqrt{\frac{d\nu\log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}}}$. For any such arm x, at the end of Part I of the algorithm we have

$$
\mathrm{LNCB}\left(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) \leq \mathrm{UNCB}\left(x, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) \underset{\text{via (3.33)}}{<} \mathrm{LNCB}\left(x^*, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right) \leq \mathrm{UNCB}\left(x^*, \widehat{\theta}, \widetilde{\mathsf{T}}/3\right).
$$

Hence, at the end of Part I, arm x^* is not eliminated via the LNCB of any x which satisfies $\langle x, \theta^* \rangle < 10\sqrt{\frac{d\nu\log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}}}$. Further, note that, under event E_1, such arms are eliminated at the end of Part I (Lemma 3.16). Hence, the LNCB of such arms are not even considered in the phases of Part II.

To complete the proof, we next show that the UNCB of the optimal arm x^* is at least the LNCB of all arms x which bear $\langle x, \theta^* \rangle \geq 10\sqrt{\frac{d\nu\log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}}}$. Below, we will consider the Nash confidence bounds for a general T'. Replacing T' by $\widetilde{\mathsf{T}}$ gives us the desired confidence-bounds comparison for the end of Part I – this repetition is omitted.

Under events E_1 and E_2, for any arm x with $\langle x, \theta^* \rangle \geq 10\sqrt{\frac{d\nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}}}$, it holds that

$$\begin{aligned}
\text{LNCB}(x, \widehat{\theta}, \mathsf{T}') &= \langle x, \widehat{\theta} \rangle - 6\sqrt{\frac{\langle x, \widehat{\theta} \rangle \ d \ \nu \ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} \\
&\leq \langle x, \theta^* \rangle + 3\sqrt{\frac{d \ \nu \ \langle x, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} - 6\sqrt{\frac{\langle x, \widehat{\theta} \rangle \ d\nu \ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} && \text{(via (3.24))} \\
&\leq \langle x, \theta^* \rangle - \left(\frac{6}{\sqrt{2}} - 3 \right) \sqrt{\frac{d\nu \langle x, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} && (\langle x, \widehat{\theta} \rangle \geq \tfrac{1}{2} \langle x, \theta^* \rangle \text{ via (3.25)}) \\
&\leq \langle x, \theta^* \rangle. && (3.34)
\end{aligned}$$

Complementarily, for optimal arm x^* we have

$$\begin{aligned}
\text{UNCB}(x^*, \widehat{\theta}, \mathsf{T}') &= \langle x^*, \widehat{\theta} \rangle + 6\sqrt{\frac{\langle x^*, \widehat{\theta} \rangle \ d \ \nu \ \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} \\
&\geq \langle x^*, \theta^* \rangle - 3\sqrt{\frac{d\nu \langle x^*, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} + 6\sqrt{\frac{\langle x^*, \widehat{\theta} \rangle \ d \ \nu \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} \\
&\geq \langle x^*, \theta^* \rangle + \left(\frac{6}{\sqrt{2}} - 3 \right) \sqrt{\frac{d\nu \ \langle x^*, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{\mathsf{T}'}} && (\text{since } \langle x^*, \widehat{\theta} \rangle \geq \tfrac{\langle x^*, \theta^* \rangle}{2}) \\
&\geq \langle x^*, \theta^* \rangle && (3.35)
\end{aligned}$$

Since $\langle x^*, \theta^* \rangle \geq \langle x, \theta^* \rangle$ for all arms x, inequalities (3.34) and (3.35) lead to the confidence-bounds comparison:

$$\text{UNCB}(x^*, \widehat{\theta}, \mathsf{T}') \geq \text{LNCB}(x, \widehat{\theta}, \mathsf{T}').$$

Hence, if events E_1 and E_2 hold, then the optimal arm x^* is never eliminated from Algorithm 5. Further, Corollary 3.2 ensures that the events E_1 and E_2 hold with probability at least $1 - \frac{4 \log \mathsf{T}}{\mathsf{T}}$. Hence, the lemma stands proved. □

Lemma 3.8. *Consider any phase ℓ in Part* II *of Algorithm 5 and let $\widetilde{\mathcal{X}}$ be the surviving set of arms at the beginning of that phase. Then, with $\widetilde{\mathsf{T}} = \sqrt{d\nu \mathsf{T} \log(\mathsf{T}\,|\mathcal{X}|)}$, we have*

$$\Pr\left\{ \langle x, \theta^* \rangle \geq \langle x^*, \theta^* \rangle - 25\sqrt{\frac{3d\nu \langle x^*, \theta^* \rangle \log(\mathsf{T}|\mathcal{X}|)}{2^\ell \cdot \widetilde{\mathsf{T}}}} \ \textit{for all } x \in \widetilde{\mathcal{X}} \right\} \geq 1 - \frac{4 \log \mathsf{T}}{\mathsf{T}} \qquad (3.10)$$

Here, ν is the sub-Poisson parameter of the stochastic rewards.

Proof. For the analysis, assume that events E_1 and E_2 hold. Lemma 3.7 ensures that the optimal arm is contained in the surviving set of arms $\widetilde{\mathcal{X}}$. Furthermore, if an arm $x \in \widetilde{\mathcal{X}}$ at the beginning of the ℓ^{th} phase, then it must be the case that arm x was not eliminated in the previous phase (which executed for $\mathsf{T}'/2$ rounds); in particular, we have $\text{UNCB}(x, \widehat{\theta}, \mathsf{T}'/2) \geq \text{LNCB}(x^*, \widehat{\theta}, \mathsf{T}'/2)$. This inequality reduces to

$$\langle x, \widehat{\theta}\rangle + 6\sqrt{\frac{\langle x, \widehat{\theta}\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\frac{\mathsf{T}'}{2}}} \geq \langle x^*, \widehat{\theta}\rangle - 6\sqrt{\frac{\langle x^*, \widehat{\theta}\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\frac{\mathsf{T}'}{2}}}.$$

Rearranging the terms, we obtain

$$\begin{aligned}
\langle x, \widehat{\theta}\rangle &\geq \langle x^*, \widehat{\theta}\rangle - 6\sqrt{\frac{\langle x^*, \widehat{\theta}\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\frac{\mathsf{T}'}{2}}} - 6\sqrt{\frac{\langle x, \widehat{\theta}\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\frac{\mathsf{T}'}{2}}} \\
&\geq \langle x^*, \widehat{\theta}\rangle - 6\sqrt{\frac{4\langle x^*, \theta^*\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{3\mathsf{T}'/2}} - 6\sqrt{\frac{4\langle x, \theta^*\rangle \ d \ \nu \log\left(\mathsf{T}|\mathcal{X}|\right)}{3\mathsf{T}'/2}} \\
&\qquad\qquad\qquad\qquad (\langle x, \widehat{\theta}\rangle \leq \tfrac{4}{3}\langle x, \theta^*\rangle \text{ via } (3.25)) \\
&\geq \langle x^*, \widehat{\theta}\rangle - 20\sqrt{\frac{\langle x^*, \theta^*\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}'}}.
\end{aligned}$$

Further, invoking equation (3.24) for x^* leads to

$$\begin{aligned}
\langle x, \theta^*\rangle &\geq \langle x^*, \theta^*\rangle - 20\sqrt{\frac{\langle x^*, \theta^*\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}'}} - 3\sqrt{\frac{\langle x^*, \theta^*\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\frac{\mathsf{T}'}{2}}} \\
&\geq \langle x^*, \theta^*\rangle - 25\sqrt{\frac{\langle x^*, \theta^*\rangle \ d \ \nu \ \log\left(\mathsf{T}|\mathcal{X}|\right)}{\mathsf{T}'}}.
\end{aligned}$$

Substituting $\mathsf{T}' = 2^\ell\widetilde{\mathsf{T}}/3$, the above inequality reduces to the desired bound in (3.10). From Corollary 3.2, we have that the events E_1 and E_2 hold with probability at least $1 - \frac{4\log\mathsf{T}}{\mathsf{T}}$. Hence, the lemma stands proved. □

3.9 Regret Analysis of Algorithm 6

Instead of ensuring probability bounds on individual arms, we construct a confidence ellipsoid around θ^*. In the context of Algorithm 6, we define the following events for the regret analysis:

G_1 In Part I, arms from the D-optimal design are chosen at least $\widetilde{\mathsf{T}}/3$ times. If $\langle x^*, \theta^*\rangle \geq$

$196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}}\log\mathsf{T}$, then $\widehat{\theta}$ calculated at the end of Part I satisfies

$$\left\|\widehat{\theta}-\theta^*\right\|_{\mathbf{V}} \leq 7\sqrt{\langle x^*,\theta^*\rangle d^{\frac{3}{2}}\nu\log\mathsf{T}}.$$

G_2 In Part II, for every phase, if $\langle x^*,\theta^*\rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}}\log\mathsf{T}$, the estimators $\widehat{\theta}$ satisfy:

$$\left\|\widehat{\theta}-\theta^*\right\|_{\mathbf{V}} \leq 7\sqrt{\langle x^*,\theta^*\rangle d^{\frac{3}{2}}\nu\log\mathsf{T}}.$$

Without loss of generality, we assume throughout that $\langle x^*,\theta^*\rangle \geq 196\frac{d^{1.25}\sqrt{\nu}}{\sqrt{\mathsf{T}}}\log\mathsf{T}$. Otherwise, the regret bound in Theorem 3.2 trivially holds. Let $\mathcal{B}$ denote the unit ball in $\mathbb{R}^d$. We have

$$\begin{aligned}\left\|\widehat{\theta}-\theta^*\right\|_{\mathbf{V}} &= \left\|\mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\right\|_2 \\ &= \max_{y\in\mathcal{B}} \langle y, \mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\rangle.\end{aligned}$$

We construct an ε-net for the unit ball, denoted as $\mathcal{C}_\varepsilon$. For any $y \in \mathcal{B}$, we define $y_\varepsilon := \operatorname{argmin}_{b\in\mathcal{C}_\varepsilon} \|b-y\|_2$. We can now write

$$\begin{aligned}\left\|\widehat{\theta}-\theta^*\right\|_{\mathbf{V}} &= \max_{y\in\mathcal{B}} \langle y-y_\varepsilon, \mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\rangle + \langle y_\varepsilon, \mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\rangle \\ &\leq \max_{y\in\mathcal{B}} \|y-y_\varepsilon\|_2 \left\|\mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\right\|_2 + |\langle y_\varepsilon, \mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\rangle| \\ &\leq \varepsilon\left\|(\widehat{\theta}-\theta^*)\right\|_{\mathbf{V}} + |\langle y_\varepsilon, \mathbf{V}^{\frac{1}{2}}(\widehat{\theta}-\theta^*)\rangle|.\end{aligned}$$

Rearranging, we obtain

$$\left\|\widehat{\theta}-\theta^*\right\|_{\mathbf{V}} \leq \frac{1}{1-\varepsilon}|\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \widehat{\theta}-\theta^*\rangle|. \tag{3.36}$$

In the following lemmas, we show that $|\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \widehat{\theta}-\theta^*\rangle|$ is small for all values of y_ε.

Lemma 3.17. *Let $x_1, x_2, \ldots, x_n$ be a sequence of fixed arm pulls (from a set $\mathcal{X}$) such that each arm x in the support λ from D-optimal design (for $\mathcal{X}$) is pulled at least $\lceil\lambda_x\tau\rceil$ times. Consider the matrix $\mathbf{V} = \sum_{j=1}^n x_j x_j^\mathsf{T}$ and let z be a vector such that $\|z\|_2 \leq 1$ and $\langle z\mathbf{V}^{\frac{1}{2}}, \theta^*\rangle \geq$*

$6\nu\sqrt{\frac{d}{\tau}}\ \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$. *Then, with probability greater than* $1-\frac{2}{\mathsf{T}|\mathcal{C}_\varepsilon|}$, *we have,*

$$|\langle z\mathbf{V}^{\frac{1}{2}},\theta^*-\widehat{\theta}\rangle| \le \left(3\nu\sqrt{\frac{nd}{\tau}}\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\langle x^*,\theta^*\rangle\right)^{\frac{1}{2}}$$

Proof. We begin by utilizing Lemma 3.6. First, we determine the γ parameter in the lemma as follows, for any $t\in[n]$ we have

$$\begin{aligned}\left(z\mathbf{V}^{\frac{1}{2}}\right)^T\mathbf{V}^{-1}x_t &\le \left\|z\mathbf{V}^{\frac{1}{2}}\right\|_{\mathbf{V}^{-1}}\left\|\mathbf{V}^{-1}x_t\right\|_{\mathbf{V}}\\ &\le \|z\|_2\,\|x_t\|_{\mathbf{V}^{-1}}\\ &\le \|x_t\|_{\mathbf{V}^{-1}}\,. && \text{(since } \|z\|_2\le 1\text{)}\end{aligned}$$

Let A_λ be the optimal design matrix. Since $\mathbf{V}\succ\tau A_\lambda$, we have

$$\begin{aligned}\|x_t\|_{\mathbf{V}^{-1}} &\le \|x_t\|_{\frac{1}{\tau}A_\lambda^{-1}}\\ &\le \sqrt{\frac{d}{\tau}}. && \text{(by Lemma 3.3)}\end{aligned}$$

Now, we use Corollary 3.1 with $\gamma=\sqrt{\frac{d}{\tau}}$ and $\delta=\left(3\sqrt{\frac{d}{\tau}}\frac{\nu\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)}{\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle}\right)^{\frac{1}{2}}$. Note that $\delta\in[0,1]$ since $\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle\ge 6\sqrt{\frac{d}{\tau}}\nu\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$. We obtain the following probability bound

$$\begin{aligned}\mathbb{P}\left\{|\langle z\mathbf{V}^{\frac{1}{2}},\theta^*-\widehat{\theta}\rangle|\ge\left(3\nu\sqrt{\frac{d}{\tau}}\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle\right)^{\frac{1}{2}}\right\} &\le 2\ \exp\left(-\frac{3\sqrt{\frac{d}{\tau}}\frac{\nu\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)}{\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle}\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle}{3\nu\sqrt{\frac{d}{\tau}}}\right)\\ &\le\frac{2}{\mathsf{T}|\mathcal{C}_\varepsilon|}. && (3.37)\end{aligned}$$

Finally, we establish an upper bound on the term $\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle$ as follows

$$\begin{aligned}\langle z\mathbf{V}^{\frac{1}{2}},\theta^*\rangle &\le \|z\|_2\left\|\mathbf{V}^{\frac{1}{2}}\theta^*\right\|_2\\ &\le\sqrt{\theta^{*T}\mathbf{V}\theta^*} && \text{(since } \|z\|_2\le 1\text{)}\\ &=\sqrt{\left(\sum_{i\in[n]}\theta^{*T}x_ix_i^T\theta^*\right)}\end{aligned}$$

$$= \sqrt{n}\langle x^*, \theta^* \rangle. \qquad (\langle x_i, \theta^* \rangle \leq \langle x^*, \theta^* \rangle)$$

Substituting in (3.37) we get the lemma statement. This completes the proof of the lemma. □

Lemma 3.18. *Consider the same notation as in Lemma 3.17. If* $\langle z\mathbf{V}^{\frac{1}{2}}, \theta^* \rangle \in \left[0, 6\nu\sqrt{\frac{d}{\tau}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\right]$, *then with probability greater than* $1 - \frac{2}{\mathsf{T}|\mathcal{X}|}$ *we have*

$$|\langle z\mathbf{V}^{\frac{1}{2}}, \theta^* - \widehat{\theta} \rangle| \leq 12\nu\sqrt{\frac{d}{\tau}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right).$$

Proof. Utilizing Lemma 3.9, with $\delta = 1$, $\alpha = 6\nu\sqrt{\frac{d}{\tau}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$, and $\gamma = \sqrt{\frac{d}{\tau}}$, we have $\langle z\mathbf{V}^{\frac{1}{2}}, \widehat{\theta} \rangle \leq 12\nu\sqrt{\frac{d}{\tau}} \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$. Since $\langle z\mathbf{V}^{\frac{1}{2}}, \theta^* \rangle \geq 0$, it follows, with probability greater than $1 - \frac{1}{\mathsf{T}|\mathcal{X}|}$, that

$$\langle z\mathbf{V}^{\frac{1}{2}}, \widehat{\theta} - \theta^* \rangle \leq 12\nu\sqrt{\frac{d}{\tau}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right).$$

Next, applying Lemma 3.10 with $\delta = 1$ and $\alpha = 6\nu\sqrt{\frac{d}{\tau}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$, we have, with probability greater than $1 - \frac{1}{\mathsf{T}|\mathcal{X}|}$,

$$\langle z\mathbf{V}^{\frac{1}{2}}, \theta^* - \widehat{\theta} \rangle \leq 6\nu\sqrt{\frac{d}{\tau}} \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right) \leq 12\nu\sqrt{\frac{d}{\tau}} \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right).$$

Hence, the lemma stands proved. □

Lemma 3.19. *If* $\langle x^*, \theta^* \rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}} \log \mathsf{T}$, *then*

$$\mathbb{P}\left\{G_1\right\} \geq 1 - \frac{3}{\mathsf{T}} \tag{3.38}$$

Proof. First, we note (from Lemma 3.12) that arms from the solution of the D-optimal design problem are selected (with probability greater than $1 - \frac{1}{\mathsf{T}}$) at least $\widetilde{\mathsf{T}}/3$ times. Hence, we can use Lemmas 3.17 and 3.18 with $\tau = \widetilde{\mathsf{T}}/3$.

Let us consider the case where $\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \theta^* \rangle \geq 6\sqrt{\frac{3d}{\widetilde{\mathsf{T}}}}\, \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)$. We have that the following holds with probability greater than $1 - \frac{1}{\mathsf{T}|\mathcal{C}_\varepsilon|}$:

$$\begin{aligned} \left\|\widehat{\theta} - \theta^*\right\|_{\mathbf{V}} &\leq \frac{1}{1-\varepsilon}\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \widehat{\theta} - \theta^* \rangle && \text{(from (3.36))} \\ &\leq \frac{1}{1-\varepsilon}\left(3\nu\sqrt{\frac{\widetilde{\mathsf{T}}d}{\frac{\widetilde{\mathsf{T}}}{3}}} \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\langle x^*, \theta^* \rangle\right)^{\frac{1}{2}} && \text{(using Lemma 3.17)} \end{aligned}$$

$$\leq \frac{1}{1-\varepsilon}\left(3\sqrt{3d}\;\nu \log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\langle x^*, \theta^*\rangle\right)^{\frac{1}{2}}.$$

Next, we note that $|\mathcal{C}_\varepsilon| \leq \left(\frac{3}{\varepsilon}\right)^d$ [LS20], and by choosing $\varepsilon = 1/2$ we get

$$\left\|\widehat{\theta} - \theta^*\right\|_{\mathbf{V}} \leq 7\left(\nu d^{\frac{3}{2}} \log\left(\mathsf{T}\right)\langle x^*, \theta^*\rangle\right)^{\frac{1}{2}}$$

Taking a union bound over all elements in $\mathcal{C}_\varepsilon$ gives a probability bound of $1 - \frac{1}{\mathsf{T}}$.

Now, for the case where $\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \theta^*\rangle \in \left[0, 6\sqrt{\frac{3d}{\widetilde{\mathsf{T}}}}\;\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right)\right]$, substituting $\tau = \widetilde{\mathsf{T}}/3$ in Lemma 3.18 we have, with probability greater than $1 - \frac{1}{\mathsf{T}|\mathcal{C}_\varepsilon|}$,

$$\begin{aligned}
\left\|\widehat{\theta} - \theta^*\right\|_{\mathbf{V}} &\leq \frac{1}{1-\varepsilon}\langle y_\varepsilon \mathbf{V}^{\frac{1}{2}}, \widehat{\theta} - \theta^*\rangle \\
&\leq \frac{12\nu}{1-\varepsilon}\sqrt{\frac{d}{\tau}}\;\log\left(\mathsf{T}|\mathcal{C}_\varepsilon|\right) && \text{(using Lemma 3.18)} \\
&\leq 24\nu\sqrt{\frac{3d^3}{\widetilde{\mathsf{T}}}}\;\log\left(\mathsf{T}\right) && \text{(substituting } \varepsilon = 0.5) \\
&\leq 7\left(d^{\frac{3}{2}}\nu \log\left(\mathsf{T}\right)\langle x^*, \theta^*\rangle\right)^{\frac{1}{2}}
\end{aligned}$$

The last inequality is due to the fact that $\langle x^*, \theta^*\rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}}\log \mathsf{T}$ and $\widetilde{\mathsf{T}} = 3\sqrt{\mathsf{T}\nu d^{2.5}\log \mathsf{T}}$. We again take a union bound over all elements in $\mathcal{C}_\varepsilon$ to get a probability bound of $1 - \frac{1}{\mathsf{T}}$.

Finally, a union bound over the two cases and the event in Lemma 3.12 proves the lemma. □

Lemma 3.20. *If* $\langle x^*, \theta^*\rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}}\log \mathsf{T}$, *then*

$$\mathbb{P}\left\{G_2\right\} \geq 1 - \frac{\log \mathsf{T}}{\mathsf{T}}. \tag{3.39}$$

Proof. To prove Lemma 3.20, we follow the same steps as in the proof of Lemma 3.19. Utilizing Lemma 3.17 and Lemma 3.18 with $\tau = \mathsf{T}'$, we establish that for any fixed phase, the following inequality holds with probability greater than $1 - \frac{1}{\mathsf{T}}$:

$$\left\|\widehat{\theta} - \theta^*\right\|_{\mathbf{V}} \leq 7\left(d^{\frac{3}{2}}\nu \log \mathsf{T}\langle x^*, \theta^*\rangle\right)^{\frac{1}{2}}.$$

Taking a union bound over all – at most $\log \mathsf{T}$ – phases in Part II of Algorithm 6 gives us the desired lower bound on $\mathbb{P}\left\{G_2\right\}$. □

Corollary 3.3. *If G_1 holds, then for all $x \in \mathcal{X}$, $\widehat{\theta}$ calculated at the end of Part* I *satisfies*

$$|\langle x, \widehat{\theta}\rangle - \langle x, \theta^*\rangle| \leq 7\sqrt{\frac{3\langle x^*, \theta^*\rangle d^{2.5}\nu \log \mathsf{T}}{\widetilde{\mathsf{T}}}}$$

Consider any phase ℓ in Part II*. If G_2 holds, then for every arm in the surviving arm set $\widetilde{\mathcal{X}}$, $\widehat{\theta}$ calculated at the end of the phase satisfies*

$$|\langle x, \widehat{\theta}\rangle - \langle x, \theta^*\rangle| \leq 7\sqrt{\frac{3\langle x^*, \theta^*\rangle d^{2.5}\nu \log \mathsf{T}}{2^\ell \ \widetilde{\mathsf{T}}}}.$$

Proof. First we use Hölder's inequality

$$|\langle x, \theta^* - \widehat{\theta}\rangle| \leq \|x\|_{\mathbf{V}^{-1}} \left\|\theta^* - \widehat{\theta}\right\|_{\mathbf{V}}. \tag{3.40}$$

Since G_1 holds, arms from the optimal design matrix are selected at least $\widetilde{\mathsf{T}}/3$ times; we have by Lemma 3.3

$$\|x\|_{\mathbf{V}^{-1}} \leq \sqrt{\frac{3d}{\widetilde{\mathsf{T}}}}.$$

Similarly, for every phase in Part II with $\mathsf{T}' = 2^\ell \widetilde{\mathsf{T}}/3$ we have

$$\|x\|_{\mathbf{V}^{-1}} \leq \sqrt{\frac{d}{\mathsf{T}'}}.$$

Finally, using bounds on $\left\|\theta^* - \widehat{\theta}\right\|_{\mathbf{V}}$ from events G_1 and G_2, and substituting in (3.40), we get the desired bound. □

Corollary 3.4. *If $\langle x^*, \theta^*\rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}}\log \mathsf{T}$*

$$\frac{7}{10}\langle x^*, \theta^*\rangle \leq \max_{x\in\mathcal{X}}\langle x, \widehat{\theta}\rangle \leq \frac{13}{10}\ \langle x^*, \theta^*\rangle$$

Proof. Since $\mathsf{T}' \geq 2\widetilde{\mathsf{T}}/3$, via Corollary 3.3 any $\widehat{\theta}$ calculated in Part I or during any phase of Part II satisfies

$$|\langle x, \widehat{\theta}\rangle - \langle x, \theta^*\rangle| \leq 7\sqrt{\frac{3\langle x^*, \theta^*\rangle d^{2.5}\nu \log \mathsf{T}}{\widetilde{\mathsf{T}}}}$$

We have

$$\begin{aligned}
\max_{x\in\mathcal{X}}\langle x,\widehat{\theta}\rangle &\geq \langle x^*,\widehat{\theta}\rangle \\
&\geq \langle x^*,\theta^*\rangle - 7\sqrt{\frac{\langle x^*,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\widetilde{\mathsf{T}}}} \\
&\geq \langle x^*,\theta^*\rangle\left(1-7\sqrt{\frac{d^{2.5}\nu\log\mathsf{T}}{\langle x^*,\theta^*\rangle\widetilde{\mathsf{T}}}}\right) \\
&\geq \frac{7}{10}\langle x^*,\theta^*\rangle \qquad \left(\text{since } \langle x^*,\theta^*\rangle \geq 196\sqrt{\tfrac{d^{2.5}\nu}{\mathsf{T}}}\log\mathsf{T} \text{ and } \widetilde{\mathsf{T}} = 3\sqrt{\mathsf{T}d^{2.5}\nu\log(\mathsf{T})}\right)
\end{aligned}$$

Now, for any $x\in\mathcal{X}$,

$$\begin{aligned}
\langle x,\widehat{\theta}\rangle &\leq \langle x,\theta^*\rangle + 7\sqrt{\frac{\langle x^*,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\tau}} \\
&\leq \langle x^*,\theta^*\rangle\left(1+7\sqrt{\frac{d^{2.5}\nu\log\mathsf{T}}{\langle x^*,\theta^*\rangle\tau}}\right) \\
&\leq \frac{13}{10}\langle x^*,\theta^*\rangle
\end{aligned}$$

Hence, the lemma stands proved. □

Lemma 3.21. *If events G_1 and G_2 hold then the optimal arm x^* always exists in the surviving set $\widetilde{X}$ in every phase in Part* II *of Algorithm 6*

Proof. Let $\tau = \widetilde{\mathsf{T}}/3$ for Part I and $\tau = \mathsf{T}'$ for every phase of Part II. From Corollary 3.3 we have

$$\begin{aligned}
\langle x^*,\widehat{\theta}\rangle &\geq \langle x^*,\theta^*\rangle - 7\sqrt{\frac{\langle x^*,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\tau}} \\
&\geq \langle x,\theta^*\rangle - 7\sqrt{\frac{\langle x^*,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\tau}} && (\text{since } \langle x^*,\theta^*\rangle \geq \langle x,\theta^*\rangle) \\
&\geq \langle x,\widehat{\theta}\rangle - 14\sqrt{\frac{\langle x^*,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\tau}} && (\text{using Corollary 3.3}) \\
&\geq \langle x,\widehat{\theta}\rangle - 16\sqrt{\frac{\max_{x\in\widetilde{\mathcal{X}}}\langle x,\theta^*\rangle d^{2.5}\nu\log\mathsf{T}}{\tau}}. && (\text{using Corollary 3.4})
\end{aligned}$$

Hence, the best arm will never satisfy the elimination criteria in Algorithm 6. □

Lemma 3.22. *Given that events G_1 and G_2 hold, consider any phase index ℓ in Part* II *of Alg. 6. For the surviving set of arms $\widetilde{\mathcal{X}}$ at the beginning of that phase, and for $\widetilde{\mathsf{T}} = \sqrt{d^{2.5}\nu\mathsf{T}\log(\mathsf{T})}$,* *the following inequality holds for all $x \in \widetilde{\mathcal{X}}$*

$$\langle x, \theta^* \rangle \geq \langle x^*, \theta^* \rangle - 26\sqrt{\frac{3d^{2.5}\nu\langle x^*, \theta^* \rangle}{2^\ell \cdot \widetilde{\mathsf{T}}}}. \tag{3.41}$$

Proof. Lemma 3.21 ensures that the optimal arm is contained in the surviving set of arms $\widetilde{\mathcal{X}}$. Furthermore, if an arm $x \in \widetilde{\mathcal{X}}$ is pulled in the ℓ^{th} phase, then it must be the case that arm x was not eliminated in the previous phase (with a phase length parameter $\frac{\mathsf{T}'}{2}$); in particular the arms x does not satisfy the inequality on Line 21 of Algorithm 6. This inequality reduces to

$$\begin{aligned}\langle x, \widehat{\theta} \rangle &\geq \langle x^*, \widehat{\theta} \rangle - 16\sqrt{\frac{\max_{x \in \widetilde{\mathcal{X}}} \langle x, \widehat{\theta} \rangle \ d^{2.5} \ \nu \ \log{(\mathsf{T})}}{\frac{\mathsf{T}'}{2}}} \\ &\geq \langle x^*, \widehat{\theta} \rangle - 26\sqrt{\frac{\langle x^*, \theta^* \rangle \ d^{2.5} \ \nu \ \log{(\mathsf{T})}}{\mathsf{T}'}}\end{aligned}$$

(via Corollary 3.4)

Substituting $\mathsf{T}' = 2^l\widetilde{\mathsf{T}}/3$ in the above inequality proves the Lemma. □

Theorem 3.2. *For any given stochastic linear bandits problem with set of arms $\mathcal{X} \subset \mathbb{R}^d$, time horizon $\mathsf{T} \in \mathbb{Z}_+$, and ν-sub-Poisson rewards, Algorithm 5 achieves Nash regret*

$$\mathrm{NR}_{\mathsf{T}} = O\left(\beta \frac{d^{\frac{5}{4}}\sqrt{\nu}}{\sqrt{\mathsf{T}}} \log(\mathsf{T})\right),$$

Here, $\beta = \max\{1, \ \langle x^*, \theta^* \rangle \log d\}$, with $x^* \in \mathcal{X}$ denoting the optimal arm and θ^* the (unknown) parameter vector.

Proof. Without loss of generality, we assume that $\langle x^*, \theta^* \rangle \geq 196\sqrt{\frac{d^{2.5}\nu}{\mathsf{T}}} \log \mathsf{T}$. Otherwise, the Nash Regret bound is trivially true. For Part I, the product of expected rewards satisfies

$$\begin{aligned}\prod_{t=1}^{\widetilde{\mathsf{T}}} \mathbb{E}[\langle X_t, \theta^* \rangle \mid G_1 \cap G_2]^{\frac{1}{\mathsf{T}}} &\geq \left(\frac{\langle x^*, \theta^* \rangle}{2(d+1)}\right)^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}} && \text{(From Lemma 3.5)} \\ &= \langle x^*, \theta^* \rangle^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}} \left(1 - \frac{1}{2}\right)^{\frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}}} \\ &\geq \langle x^*, \theta^* \rangle^{\frac{\widetilde{\mathsf{T}}}{\mathsf{T}}} \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}}\right).\end{aligned}$$

For Part II, we use Lemma 3.8. Let $\mathcal{E}_i$ denote the time interval of the i^{th} phase, and let T'_i be the phase length parameter in that phase. Recall that $|\mathcal{E}_i| \leq \mathsf{T}'_i + \frac{d(d+1)}{2}$. Also, the algorithm runs for at most $\log \mathsf{T}$ phases. Hence, we have

$$\begin{aligned}
\prod_{t=\widetilde{\mathsf{T}}+1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^* \rangle \mid G_1 \cap G_2]^{\frac{1}{\mathsf{T}}} &= \prod_{\mathcal{E}_j} \prod_{t \in \mathcal{E}_j} \mathbb{E}[\langle X_t, \theta^* \rangle \mid G_1 \cap G_2]^{\frac{1}{\mathsf{T}}} \\
&\geq \prod_{\mathcal{E}_j} \left(\langle x^*, \theta^* \rangle - 26 \sqrt{\frac{d^{2.5} \ \nu \ \langle x^*, \theta^* \rangle \log (\mathsf{T})}{\mathsf{T}'_j}} \right)^{\frac{|\mathcal{E}_j|}{\mathsf{T}}} \\
&\geq \langle x^*, \theta^* \rangle^{\frac{\mathsf{T} - \widetilde{\mathsf{T}}}{\mathsf{T}}} \prod_{i=1}^{\log \mathsf{T}} \left(1 - 26 \sqrt{\frac{d^{2.5} \ \nu \ \log (\mathsf{T})}{\langle x^*, \theta^* \rangle \mathsf{T}'_j}} \right)^{\frac{|\mathcal{E}_j|}{\mathsf{T}}} \\
&\geq \langle x^*, \theta^* \rangle^{\frac{\mathsf{T} - \widetilde{\mathsf{T}}}{\mathsf{T}}} \prod_{i=1}^{\log \mathsf{T}} \left(1 - 52 \frac{|\mathcal{E}_j|}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \ \log (\mathsf{T})}{\langle x^*, \theta^* \rangle \mathsf{T}'_j}} \right)
\end{aligned}$$

The last inequality is due to the fact that $(1-x)^r \geq (1-2rx)$ where $r \in [0,1]$ and $x \in [0, 1/2]$. Note that the term $\sqrt{\frac{d^{2.5} \nu \log (\mathsf{T})}{\langle x^*, \theta^* \rangle \mathsf{T}'_j}} \leq 1/2$ for $\langle x^*, \theta^* \rangle \geq 196 \sqrt{\frac{d^{2.5} \nu}{\mathsf{T}}} \log \mathsf{T}$, $\mathsf{T}' \geq 2 \sqrt{\mathsf{T} d^{2.5} \nu \log \mathsf{T}}$, and $\mathsf{T} \geq e^6$. We can further simplify the expression as follows

$$\begin{aligned}
\prod_{j=1}^{\log \mathsf{T}} \left(1 - 52 \frac{|\mathcal{E}_j|}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \ \log (\mathsf{T})}{\langle x^*, \theta^* \rangle \mathsf{T}'_j}} \right) &\geq \prod_{j=1}^{\log \mathsf{T}} \left(1 - 52 \frac{\mathsf{T}'_j + \frac{d(d+1)}{2}}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \ \log (\mathsf{T})}{\langle x^*, \theta^* \rangle \mathsf{T}'_j}} \right) \\
&\geq \prod_{j=1}^{\log \mathsf{T}} \left(1 - 78 \frac{\sqrt{\mathsf{T}'_j}}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \log (\mathsf{T})}{\langle x^*, \theta^* \rangle}} \right) \\
&\qquad\qquad\qquad\qquad \text{(assuming } \mathsf{T}'_j \geq d(d+1)\text{)} \\
&\geq 1 - 78 \frac{1}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \log (\mathsf{T})}{\langle x^*, \theta^* \rangle}} \left(\sum_{j=1}^{\log \mathsf{T}} \sqrt{\mathsf{T}'_j} \right) \\
&\geq 1 - 78 \frac{1}{\mathsf{T}} \sqrt{\frac{d^{2.5} \ \nu \ \log (\mathsf{T})}{\langle x^*, \theta^* \rangle}} \left(\sqrt{\mathsf{T} \log \mathsf{T}} \right) \\
&\qquad\qquad\qquad\qquad \text{(using Cauchy Schwarz)} \\
&\geq 1 - 78 \sqrt{\frac{d^{2.5} \nu}{\mathsf{T} \langle x^*, \theta^* \rangle}} \log (\mathsf{T}).
\end{aligned}$$

Combining the lower bound for rewards in Part I and Part II of the algorithm, we obtain

$$
\begin{aligned}
\prod_{t=1}^{\mathsf{T}} \mathbb{E}[\langle X_t, \theta^* \rangle]^{\frac{1}{\mathsf{T}}} &\geq \prod_{t=1}^{\mathsf{T}} \Big(\mathbb{E}[\langle X_t, \theta^* \rangle \mid G_1 \cap G_2] \cdot \mathbb{P}\{G_1 \cap G_2\} \Big)^{\frac{1}{\mathsf{T}}} \\
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} \right) \left(1 - 78 \sqrt{\frac{d^{2.5}\nu}{\mathsf{T}\langle x^*, \theta^* \rangle}} \log{(\mathsf{T})} \right) \mathbb{P}\{G_1 \cap G_2\} \\
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} - 78 \sqrt{\frac{d^{2.5}\nu}{\mathsf{T}\langle x^*, \theta^* \rangle}} \log{(\mathsf{T})} \right) \mathbb{P}\{G_1 \cap G_2\} \\
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))\widetilde{\mathsf{T}}}{\mathsf{T}} - 78 \sqrt{\frac{d^{2.5}\nu}{\mathsf{T}\langle x^*, \theta^* \rangle}} \log{(\mathsf{T})} \right) \left(1 - \frac{2 \log \mathsf{T}}{\mathsf{T}} \right) \\
&\geq \langle x^*, \theta^* \rangle \left(1 - \frac{\log(2(d+1))3\sqrt{\mathsf{T} d\nu \log(\mathsf{T})}}{\mathsf{T}} - 78 \sqrt{\frac{d^{2.5}\nu}{\mathsf{T}\langle x^*, \theta^* \rangle}} \log{(\mathsf{T})} - \frac{2 \log \mathsf{T}}{\mathsf{T}} \right) \\
&\geq \langle x^*, \theta^* \rangle - 78 \sqrt{\frac{\langle x^*, \theta^* \rangle d^{2.5}\nu}{\mathsf{T}}} \log{(\mathsf{T})} - 2 \frac{\langle x^*, \theta^* \rangle \log(2(d+1))3\sqrt{d \log(\mathsf{T})}}{\sqrt{\mathsf{T}}}.
\end{aligned}
$$

Hence, the Nash Regret can be bounded as

$$
\begin{aligned}
\mathrm{NR}_T &= \langle x^*, \theta^* \rangle - \left(\prod_{t=1}^{T} \mathbb{E}[\langle X_t, \theta^* \rangle] \right)^{1/T} \\
&\leq 78 \sqrt{\frac{\langle x^*, \theta^* \rangle d^{2.5}\nu}{\mathsf{T}}} \log{(\mathsf{T})} + 2 \frac{\langle x^*, \theta^* \rangle \log(2(d+1))3\sqrt{d\nu \log(\mathsf{T})}}{\sqrt{\mathsf{T}}}.
\end{aligned}
$$

The theorem stands proved. □

3.10 Conclusion and Future Work

Fairness and welfare considerations have emerged as a central design objective in online decision-making. Motivated broadly by such considerations, the current work addresses the notion of Nash regret in the linear bandits framework. We develop essentially tight Nash regret bounds for linear bandit instances with a finite number of arms.

In addition, we extend this guarantee to settings wherein the number of arms is infinite. Here, our regret bound scales as $d^{5/4}$, where d is the ambient dimension. Note that, for linear bandits with infinite arms, [AYPS11] obtains a bound of $d/\sqrt{\mathsf{T}}$ for average regret. We conjecture that a similar dependence should be possible for Nash regret as well and pose this strengthening as a relevant direction of future work. Another important direction would be to study Nash regret for other bandit frameworks (such as contextual bandits and combinatorial bandits) and

Markov Decision Processes (MDPs).

Chapter 4

Full Feedback with Adversarial Rewards

In this chapter, we delve into the realm of online learning where the rewards are generated adversarially. Unlike bandit feedback, where only the chosen action's reward is revealed, in this framework, the algorithm receives feedback in the form of a reward function associated with each action, even for the unchosen ones. The algorithm must sequentially select actions from a given set and update its decision-making based on the received rewards. This setting poses unique challenges and requires robust strategies to handle the adversarial nature of the reward generation process.

Formally, consider a set of actions $\mathcal{X} \subset \mathbb{R}^d$ represented as d-dimensional vectors and sequence of concave reward functions $f_1, f_2, \ldots, f_T$, where $f_t : \mathcal{X} \to \mathbb{R}$ represents the reward function associated with the t-th round. In each round t, the algorithm selects an action $x_t \in \mathcal{X}$ and receives a reward $r_t = f_t(x_t)$. The performance of the algorithm is measured in terms of the Nash Social Welfare (NSW) across rounds. We define Nash Regret as the difference between the algorithm's performance well with respect to the best fixed action in hindsight where the best fixed action in hindsight is defined as

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \left(\prod_{t=1}^{T} f_t(x) \right). \tag{4.1}$$

Nash regret is defined as

$$\mathrm{NR}_T = \left(\prod_{t=1}^{T} f_t(x^*) \right)^{\frac{1}{T}} - \left(\prod_{t=1}^{T} f_t(x_t) \right)^{\frac{1}{T}} \tag{4.2}$$

Next, we look at a special case of online concave optimization, namely the Experts problem, and provide an algorithm that achieves an essentially tight Nash Regret.

4.1 Prediction with Expert Advice

Consider a set of N experts enumerated as $1, 2, \dots N$, each providing predictions or advice on the outcome of a given task. At each round t, the learner receives the predictions of the N experts and selects one expert's prediction. Notably, the learner choice can be randomized; that is, in every round t the learner outputs a distribution p_t over the set of experts, where p_{ti} denotes the probability with which the i^th expert is chosen. Subsequently, the true outcome of the task is revealed, and the learner receives a reward corresponding to the chosen distribution over the expert.

We use r_t to denote the reward vector in round t where r_{ti} denotes the reward corresponding to the i^th expert. The performance of the algorithm is accessed ex-ante, that is, in each round t the algorithms performance is equal to $\langle p_t, r_t \rangle = \sum_{i=1}^{N} p_{ti} \; r_{ti}$ The goal again is to minimize the Nash Regret, which, in the context of experts problem setting, is defined as

$$\mathrm{NR}_T = \max_{i \in [N]} \left(\prod_{t=1}^{T} r_{ti} \right)^{\frac{1}{T}} - \left(\prod_{t=1}^{T} \langle p_t, r_t \rangle \right)^{\frac{1}{T}}$$

Algorithm 7 Prediction with Experts Advice

Input:$1, 2, \dots N$ **Experts**

```
Initialize $w_i = 1 \quad \forall i \in [N]$.
for  $t \in 1, 2, \dots, T$ do
    Choose an expert $i$ with probability $p_{ti} = \frac{w_i}{\sum_{j=1}^{N} w_j}$.
    for  Expert $i \in [N]$ do
        Observe reward $r_{ti}$.
        Update $w_i \leftarrow w_i \times r_{ti}$
    end for
end for
```

Theorem 4.1. *For bounded rewards, that is, $r_{ti} \in [0, 1]$ for all t and i, Algorithm 7 achieves a Nash regret of*

$$\mathrm{NR}_T \leq \frac{\log N}{T}$$

Proof. Consider round t, the probability of choosing arm i is given by

$$p_{ti} = \frac{\prod_{s=1}^{t-1} r_{si}}{\sum_{i\in[N]}\prod_{s=1}^{t-1} r_{si}}$$

The expected reward is given by

$$\begin{aligned}
\langle p_t, x_t\rangle = \sum_{i\in[N]} p_{ti} r_{ti} &= \sum_{i\in[N]} \frac{\prod_{s=1}^{t-1} r_{si}}{\sum_{i\in[N]}\prod_{s=1}^{t-1} r_{si}}\, r_{ti} \\
&= \frac{\sum_{i\in[N]}\prod_{s=1}^{t} r_{si}}{\sum_{i\in[N]}\prod_{s=1}^{t-1} r_{si}}
\end{aligned}$$

We define $W_t \coloneqq \sum_{i\in[N]}\prod_{s=1}^{t} r_{si}$. Using this definition we have for all $t \geq 2$, we have

$$\langle p_t, x_t\rangle = \frac{W_t}{W_{t-1}}$$

Hence, the Nash Social Welfare of the algorithm is equal to

$$\begin{aligned}
\prod_{t=1}^{T}\langle p_t, x_t\rangle &= \langle p_1, x_1\rangle \prod_{t=2}^{T}\frac{W_t}{W_{t-1}} \\
&= \langle p_1, x_1\rangle \frac{W_T}{W_1} \\
&= \frac{\sum_{i\in[N]} r_{1i}}{k}\frac{W_T}{W_1} \\
&= \frac{W_T}{k}
\end{aligned}$$

Now let the best expert in hindsight be denoted as $i^* = argmax_{i\in[N]}\prod_{t=1}^{T} r_{ti}$. The Nash regret bound can be obtained as

$$\begin{aligned}
\mathrm{NR}_T &= \left(\prod_{t=1}^{T} r_{ti^*}\right)^{\frac{1}{T}} - \left(\prod_{t=1}^{T}\langle p_t, x_t\rangle\right)^{\frac{1}{T}} \\
&\leq \left(\prod_{t=1}^{T} r_{ti^*}\right)^{\frac{1}{T}} - \left(\frac{W_T}{k}\right)^{\frac{1}{T}} \\
&\leq \left(\prod_{t=1}^{T} r_{ti^*}\right)^{\frac{1}{T}} \left(1 - \frac{1}{k^{\frac{1}{T}}}\right) && \text{(since } W_T = \textstyle\sum_{i\in[N]}\prod_{s=1}^{T} r_{si} \geq \prod_{t=1}^{T} r_{ti^*}\text{)}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\prod_{t=1}^{T} r_{ti^*}\right)^{\frac{1}{T}} \left(1 - e^{\frac{\log k}{T}}\right) \\
&\leq \left(\prod_{t=1}^{T} r_{ti^*}\right)^{\frac{1}{T}} \left(\frac{\log k}{T}\right) \\
&\leq \frac{\log k}{T} && (\text{ since } r_{ti} \in [0,1])
\end{aligned}
$$

The theorem stands proved. □

4.1.1 Lower Bound

We will now establish a lower bound for the experts problem that matches the upper bound presented in Theorem 4.1 upto log factors.

Theorem 4.2. *For every Algorithm $\mathcal{A}$, there exists an instance of the experts problem such that the Nash regret,* NR_T *satisfies*

$$\mathrm{NR}_T \geq \frac{1}{2T}$$

Proof. Consider a randomized instance with binary rewards. In each round $t \in [T]$ and for every arm $i \in [N]$, we have

$$r_{ti} = \begin{cases} 0 & \text{with probability } \frac{1}{T}, \\ 1 & \text{with probability } 1 - \frac{1}{T}. \end{cases}$$

With respect to this randomized instance, the expected Nash Regret can be calculated as

$$\mathbb{E}\left[\mathrm{NR}_T\right] = \mathbb{E}\left[\max_{i \in [N]} \left(\prod_{t=1}^{T} r_{ti}\right)^{\frac{1}{T}}\right] - \mathbb{E}\left[\left(\prod_{t=1}^{T} \langle p_t, r_t \rangle\right)^{\frac{1}{T}}\right],$$

where the expectation is taken with respect to the random rewards generated according to (4.1.1).

The expected reward for any choice of p_t in round t is given by

$$
\begin{aligned}
\mathbb{E}\left[\langle p_t, r_t \rangle\right] &= \sum_{i \in [N]} p_{ti} \mathbb{E}\left[r_{ti}\right] \\
&= \left(1 - \frac{1}{T}\right) \sum_{i \in [N]} p_{ti}
\end{aligned}
$$

$$= \left(1 - \frac{1}{T}\right).$$

Using Jensen's inequality and the concavity of the geometric mean, we have

$$\begin{aligned}\mathbb{E}\left[\left(\prod_{t=1}^{T} \langle p_t, r_t \rangle\right)^{\frac{1}{T}}\right] &\leq \left(\prod_{t=1}^{T} \mathbb{E}\left[\langle p_t, r_t \rangle\right]\right)^{\frac{1}{T}} \\ &= 1 - \frac{1}{T}\end{aligned}$$

Next, we consider the expected value of the best arm in hindsight. The probability of any arm receiving a reward of 1 in every round is given by

$$\left(1 - \frac{1}{T}\right)^{T} \geq \frac{1}{e}\left(1 - \frac{1}{T}\right) \geq \frac{1}{3} \geq \frac{1}{3}$$

In other words, for any expert i, with probability greater than $\frac{2}{3}$ the Nash Social Welfare ($\prod_{t=1}^{T} r_{ti}$) is equal to 1. The probability with which at least one expert has nonzero NSW value is greater than

$$1 - \left(\frac{2}{3}\right)^{N} \geq 1 - \frac{1}{T^2} \qquad \text{(assuming } N \geq 5 \log T\text{)}$$

That is, with probability greater than $\left(1 - \frac{1}{T^2}\right)$, we have $\max_{i \in [N]} \left(\prod_{t=1}^{T} r_{ti}\right)^{\frac{1}{T}} = 1$. Therefore,

$$\mathbb{E}\left[\max_{i \in [N]} \left(\prod_{t=1}^{T} r_{ti}\right)^{\frac{1}{T}}\right] \geq 1 - \frac{1}{T^2} - ().$$

This gives us a lower bound on the expected Nash Regret of any algorithm for this instance

$$\mathrm{NR}_T \geq 1 - \frac{1}{T^2} - \left(1 - \frac{1}{T}\right) \geq \frac{1}{2T}$$

The theorem stands proved. □

4.2 Online Concave Optimization

Let us consider the setting of Online Concave Optimization (OCO). We define a convex set $\mathcal{X} \subset \mathbb{R}^d$ and a sequence of reward functions $f_1, f_2, \ldots, f_T$, where $f_t : \mathcal{X} \to \mathbb{R}^+$ represents the reward function associated with the t-th round. Each function f_t is concave and positive. It is important to note that these functions are chosen by an oblivious adversary.

During each round, the OCO algorithm selects a point in the set $\mathcal{X}$, and subsequently, the corresponding reward function is revealed. It is worth mentioning that if all the reward functions correspond to the same function, denoted as $f_t = f$ for all $t \in [T]$, the OCO problem reduces to an offline optimization problem.

Now we present Algorithm 8, which achieves a low Nash Regret.

```
Algorithm 8 Nash OCO
Input: A convex set $\mathcal{X}$
Initialize the dummy function $f_0(x) = 1 \forall x \in \mathcal{X}$
for $t \in 1, 2, \ldots, T$ do
    Calculate $x_t$ using the following formula
        $x_t = \int_{x \in \mathcal{X}} \frac{\prod_{s=0}^{T} f_s(x)}{\int_{x \in \mathcal{X}} \prod_{s=0}^{T} f_s(x)\ dx} x\ dx$
    Play $x_t$ and receive reward $f_t(x_t)$.
end for
```

Before deriving a bound on the Nash regret for Algorithm 8, we will first examine the problem of online convex optimization with exp-concave loss functions and discuss known bound on cumulative regret for such functions.

Definition 4.1. *A convex function $h : \mathbb{R}^n \to \mathbb{R}$ is defined to be exp-concave over a set $\mathcal{X}$ if the function g is concave, where $g : \mathcal{X} \to \mathbb{R}$ is defined as*

$$g(x) = e^{-h(x)}$$

We will use the following lemma from [H+16] to obtain a regret bound.

Lemma 4.1 (Theorem 4.4 [H+16])**.** *The exponentially weighted online optimizer (Algorithm 11 in [H+16]) when used for a sequence of exp-concave loss functions, $\{g_1, g_2, \ldots, g_T\}$, and a given convex set, $\mathcal{X} \subset \mathbb{R}^d$, satisfies*

$$\sum_{t=1}^{T} g_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} g_t(x) \le d \log T + 2. \tag{4.3}$$

We will now obtain an upper bound for the Nash regret of Algorithm 8.

Theorem 4.3. *For any sequence of positive concave reward functions* $\{f_1, f_2, \ldots, f_T\}$ *bounded between* $[0, 1]$*, the Nash regret of Algorithm 8 satisfies*

$$NR_T \leq \frac{2d \log T}{T}.$$

Proof. We begin by noting that the function $g_t(x) := -\log f_t(x)$ is exp-concave over the set $\mathcal{X}$. The calculation of x_t in Algorithm 8 can be equivalently expressed in terms of g_t as

$$x_t = \int_{x\in\mathcal{X}} \frac{e^{-\sum_{s=0}^{T} g_s(x)}}{\int_{x\in\mathcal{X}} e^{-\sum_{s=0}^{T} g_s(x)}\ dx} x\ dx.$$

This is exactly the same as the calculation of x_t in Algorithm 11 in [H+16]. Therefore, we can directly use the cumulative regret bound from Lemma 4.1. Now let us examine the left-hand side (LHS) of (4.3). From the definition of g_t, we have

$$\begin{aligned}\sum_{t=1}^{T} g_t(x_t) - \min_{x\in\mathcal{X}} \sum_{t=1}^{T} g_t(x) &= -\log \prod_{t=1}^{T} f_t(x_t) + \max_{x\in\mathcal{X}} \log \prod_{t=1}^{T} f_t(x) \\ &= \log \frac{\prod_{t=1}^{T} f_t(x^*)}{\prod_{t=1}^{T} f_t(x_t)} \qquad \text{(recall } x^* = \text{argmax}_{x\in\mathcal{X}} \Big(\textstyle\prod_{t=1}^{T} f_t(x)\Big)\text{)}\end{aligned}$$

By (4.3), we have

$$\begin{aligned}\log \frac{\prod_{t=1}^{T} f_t(x^*)}{\prod_{t=1}^{T} f_t(x_t)} &\leq d \log T + 2 \\ &\leq 2d \log T \qquad \text{(assuming } T \geq 9\text{)}\end{aligned}$$

Dividing both sides by T and exponentiating, we obtain

$$\begin{aligned}\frac{\left(\prod_{t=1}^{T} f_t(x^*)\right)^{\frac{1}{T}}}{\left(\prod_{t=1}^{T} f_t(x_t)\right)^{\frac{1}{T}}} &\leq e^{\frac{2d \log T}{T}} \\ &\leq 1 + \frac{2d \log T}{T}\end{aligned}$$

Rearranging the terms, we get

$$\begin{aligned}\left(\prod_{t=1}^{T} f_t(x^*)\right)^{\frac{1}{T}} - \left(\prod_{t=1}^{T} f_t(x_t)\right)^{\frac{1}{T}} &\leq \left(\prod_{t=1}^{T} f_t(x_t)\right)^{\frac{1}{T}} \frac{2d\log T}{T} \\ &\leq \frac{2d\log T}{T}. \qquad \text{(since } f_t\text{-s are bounded between } [0,1]\text{)}\end{aligned}$$

This completes the proof of the theorem. □

Chapter 5

Learning Good Interventions in Causal Bayesian Networks

In this chapter, We study the causal bandit problem that entails identifying a near-optimal intervention from a specified set $\mathcal{A}$ of (possibly non-atomic) interventions over a given causal graph. Here, an optimal intervention in $\mathcal{A}$ is one that maximizes the expected value for a designated reward variable in the graph, and we use the notion of simple regret to quantify near optimality. Considering Bernoulli random variables and for causal graphs on N vertices with constant in-degree, prior work has achieved a worst case guarantee of $\widetilde{O}(N/\sqrt{T})$ for simple regret. We utilize the idea of covering interventions (which are not necessarily contained within $\mathcal{A}$) and establishes a simple regret guarantee of $\widetilde{O}(\sqrt{N/T})$. Notably, and in contrast to prior work, our simple regret bound depends only on explicit parameters of the problem instance. We also go beyond prior work and achieve a simple regret guarantee for causal graphs with unobserved variables. Further, we perform experiments to show improvements over baselines in this setting.

5.1 Our Contributions

We present an algorithm to minimize simple regret in the causal bandit problem. Here, the learner is given a causal graph $\mathcal{G}$ on N Bernoulli random variables and a set $\mathcal{A}$ of (possibly non-atomic) interventions over $\mathcal{G}$. The learner's objective is to identify, within $\mathcal{A}$, an intervention that maximizes the expected value for a designated reward variable in $\mathcal{G}$. Furthermore, we consider a model wherein, while a near-optimal intervention is required from the target set $\mathcal{A}$, the learner is not confined to $\mathcal{A}$ during the exploration phase. In particular, we use the construct of covering interventions (see Definition 5.1) during exploration and show that this flexibility leads to multiple improvements over prior work. Indeed, this model is applicable in many settings wherein the learner is not confined to the target set during exploration. Consider, as stylized examples: (i) the display advertising context, wherein, during testing, one can intervene upon features, which during deployment, cannot be altered, and (ii) robotic control, in which, during simulations, hypothetical configurations can be deployed.

In fact, our result is robust enough to be used in settings where certain variables cannot be intervened upon even during exploration. One can consider such 'off-limits' variables as unobserved and then utilize our extension to graphs with unobserved parts (see Section 5.5). The list below summarizes our main contributions:

- For the causal bandit problem, we improve the worst-case guarantee for simple regret from $\tilde{\mathcal{O}}(\sqrt{N^2/T})$ to $\tilde{\mathcal{O}}(\sqrt{N/T})$.[1] Here, the $\tilde{\mathcal{O}}(\cdot)$ notation subsumes the dependence on the maximum in-degree d in the graph and logarithmic factors; see Theorem 5.1 for an explicit bound. Our algorithm can address arbitrary causal graphs. Though, as in prior works [YHS⁺18, ABDK18], our result is particularly relevant for graphs in which the maximum in-degree d is sufficiently smaller than N.

- We obtain a novel simple regret algorithm for causal graphs with unobserved variables. This extension addresses the most general setting for causal Bayesian networks (see Definition 1.3.1 in [Pea00]) and addresses a key limitation of almost all[2] prior works on causal bandits. We detail the extension in Section 5.5.

- Our experiments show a marked improvement on the baselines from prior work (see Section 5.7), thereby substantiating the theoretical guarantees.

Our worst-case guarantee for simple regret is in terms of only the explicit parameters, such as the number of variables N and the maximum in-degree in $\mathcal{G}$; see Theorem 5.1. By contrast, the simple regret bound provided in [YHS$^+$18] depends on analytically complex quantities. In addition, our guarantee holds for time horizon $T \gtrsim N^3$. This is a marked improvement over [YHS$^+$18], which requires $T \gtrsim N^{16}$. In fact, our algorithm (Algorithm 9) is notably simple – we view this as a positive feature, which aids in implementation and adaptation of the developed method. Here, it is also relevant to note that the key technical contribution of the work is the involved regret analysis (see Section 2.3.1).

[1] As mentioned previously, T denotes the time horizon (i.e., number of exploratory interventions) and N denotes the number of vertices in the causal graph.

[2] The exceptions here are the recent works of Maiti et al. [MNS22] along with Xiong and Chen [XC23]. These works are discussed at the end of the section.

Covering interventions as a complementary tool for exploration. We note that covering interventions do not conform to the existing causal-bandit framework of exploring solely within the specified set of interventions $\mathcal{A}$. However, instead of viewing $\mathcal{A}$ as a confined set of 'arms,' one can work with the enriched perspective that causal bandits are an optimization problem. Indeed, the goal of the optimization problem is to identify the best intervention in $\mathcal{A}$, but—similar to many other optimization methods—exploration can happen outside the feasible region (i.e., outside $\mathcal{A}$). In this spirit, the use of covering interventions can be identified as a complementary exploration model. This model leverages the richer context of the causal bandits setting (e.g., the causal graph itself) and, as mentioned previously, is potentially applicable in various real-world contexts. Overall, covering interventions are theoretically interesting and enable notable improvements, including novel simple regret guarantees with unobserved variables.

5.2 Additional Related Work

Lattimore et al. [LLR16] first addressed the causal bandit, though only for parallel causal graphs and with atomic interventions. Maiti et al. [MNS22] extended this work on atomic interventions to provide simple regret guarantees in the presence of unobserved or hidden variables. An importance sampling based approach was studied in [SSDS17] to identify atomic soft interventions that minimize simple regret. Lu et al. [LMTY20] provide guarantees for cumulative regret for general causal graphs (which include hidden variables). Nair et al. [NPS21] looked at cumulative as well as simple regret in case of the budgeted setting where the observation-intervention trade-off was studied when interventions are costlier than observations. Sen et al. [SSK+17] extend the model causal bandits to include contextual causal bandits and study cumulative regret in this context. Lu et al. [LMT21] study cumulative regret in the case where the full graph structure is not known. The work [LMT22] extends the model for causal bandits to include causal Markov decision processes (C-MDPs) using a modification of the algorithm in [AOM17].

There are two recent works that focus on non-atomic interventions in the causal bandit context. The paper by Varici et al. [VSST22] studies cumulative regret for causal bandits with non-atomic interventions, albeit in the specific context of linear structural equation models. Xiong and Chen [XC23] obtain sample-complexity bounds for identification of near-optimal interventions, with a particular focus on binary generalized linear models (BGLMs). The worst-case sample complexity guarantee obtained in [XC23] is proportional to the size of the intervention set $\mathcal{A}$, i.e., proportional to $|\mathcal{A}|$. By contrast, the simple regret bound obtained in the current work has only a logarithmic dependence on $|\mathcal{A}|$; recall that $|\mathcal{A}|$ can be exponentially large. Xiong and Chen [XC23] also address the case of unobserved (hidden) variables. However, this work assumes identifiability (the fact that all interventional distributions can be estimated through observations alone). We require no such assumption.

Apart from these works on causal bandits, we utilize the idea of covering interventions proposed by Acharya et al. [ABDK18]. They use covering interventions for distribution learning and testing problems over causal graphs. On the other hand, we use covering interventions for simple regret minimization. It is important to note that a direct use of the distribution learning algorithm (Algorithm 3) from [ABDK18] leads to a suboptimal regret bound for the causal bandit problem. Specifically, the learning algorithm of Acharya et al. [ABDK18] requires $\tilde{\mathcal{O}}(N^2\varepsilon^{-4})$ samples to learn interventional distributions up to a total variation distance of ε; see Theorem 3.4 in [ABDK18]. Hence, if used for identifying a near-optimal intervention in $\mathcal{A}$, this method would incur $\tilde{\mathcal{O}}\left(\frac{\sqrt{N}}{T^{1/4}}\right)$ simple regret.

5.3 Notation and Preliminaries

We study the causal bandit problem over causal graphs $\mathcal{G} = (\mathcal{V}, E)$. In the given (directed and acyclic) graph $\mathcal{G}$ the vertices, $\mathcal{V}$, correspond to Bernoulli random variables and E is the set of directed edges that capture causal relations between these variables.

We will use V_i or i, interchangeably, to refer to the ith node of the given causal graph $\mathcal{G}$. Since $\mathcal{G}$ is directed and acyclic, it admits a topological ordering. We will, throughout, assume that the vertices in $\mathcal{V}$ are indexed to respect a topological order, i.e., for each pair of indices $i < j$, vertex V_i appears before V_j in the topological order. Note that for any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$ the indexing of the vertices within $\mathcal{U}$ follows the topological ordering of these vertices. Furthermore, in the set $\mathcal{V}$, the last vertex with respect to the indexing (and, equivalently, the topological ordering) is the designated reward variable. That is, in a causal graph with $N \coloneqq |\mathcal{V}|$ vertices, V_N is the reward variable.

Write $\mathrm{Pa}(i)$ to denote the set of parents of node V_i. Also, we define the set of parents for a subset of vertices $\mathcal{U} \subseteq \mathcal{V}$ as $\mathrm{Pa}(\mathcal{U}) \coloneqq (\cup_{V\in\mathcal{U}} \mathrm{Pa}(V)) \setminus \mathcal{U}$. We use the following notations to indicate subsets of the vertices: write $[i,j] \coloneqq \{V_i, V_{i+1}, V_{i+2} \ldots V_j\}$ and, similarly, $(i,j] = [i+1,j]$, $(i,j) = [i+1,j-1]$ and $[i,j) = [i,j-1]$. Write the ancestor set $\mathrm{Ac}(i) \coloneqq [1,i) \setminus \mathrm{Pa}(i)$, i.e., $\mathrm{Ac}(i)$ denotes the set of vertices that precede V_i in the topological ordering, excluding the parents $\mathrm{Pa}(V_i)$.

An intervention is defined as an $N = |\mathcal{V}|$ dimensional vector $A \in \{0, 1, *\}^N$ that encapsulates the values assigned to each vertex in $\mathcal{G}$; in particular, $A_i = *$ denotes that V_i is not intervened upon, while $A_i = 1$ and $A_i = 0$ denote that, in the intervention, V_i is set to 1 and 0, respectively. In addition, $\mathcal{V}(A) \coloneqq \{V_i \in \mathcal{V} : A_i = *\}$ denotes the set of vertices that are not intervened under A. Also, for any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$, write $\mathcal{V}_{\mathcal{U}}(A) \coloneqq \mathcal{U} \cap \mathcal{V}(A)$.

Binary vectors $\mathbf{z} \in \{0,1\}^N$ will be used to denote an assignment to the vertices (random variables) in $\mathcal{V}$. Here, $\mathbf{z}_i$ denotes the assignment to vertex V_i. For any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$, we will use $\mathbf{z}_U \in \{0,1\}^{|\mathcal{U}|}$ to denote an assignment to the vertices in $\mathcal{U}$. Let $Z(A)$ denote the set of all binary assignments that comply with an intervention A and have the reward $V_N = 1$, i.e., $Z(A) \coloneqq \{\mathbf{z} \in \{0,1\}^N : \mathbf{z}_i = A_i, \text{ for all } i \in \mathcal{V} \setminus (\mathcal{V}(A)), \text{ and } \mathbf{z}_N = 1\}$.

We use the following short-hand notations in our analysis to denote the conditional and interventional probability distributions:

$$\begin{aligned}
\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_U\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | U = \mathbf{z}_U\right]. \\
\mathcal{P}_{\mathbf{z}_U}\left(\mathbf{z}_i\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(U = \mathbf{z}_U\right)\right] \\
&= \mathbb{P}_{\mathrm{do}(U = \mathbf{z}_U)}\left[V_i = \mathbf{z}_i\right]. \\
\mathcal{P}_{\mathbf{z}_U}\left(\mathbf{z}_i \mid \mathbf{z}_W\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(U = \mathbf{z}_U\right), W = \mathbf{z}_W\right]. \\
\mathcal{P}_A\left(\mathbf{z}_i \mid \mathbf{z}_W\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(A\right), W = \mathbf{z}_W\right].
\end{aligned}$$

It is important to note that intervening on all parent nodes of a vertex is the same as conditioning on them

$$\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(i)}}\left(\mathbf{z}_i\right) = \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \tag{5.1}$$

We use $\mu\left(A\right)$ to denote the expected reward under intervention A, i.e., $\mu\left(A\right) = \mathbb{P}\left[V_N = 1 | \mathrm{do}(A)\right]$. Specifically,

$$\mu\left(A\right) = \sum_{\mathbf{z} \in Z(A)} \prod_{i \in \mathcal{V}(A)} \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$$

We use $\widehat{\mu}(A)$ and $\widehat{\mathcal{P}}\left(\right)$ to denote the estimates for the corresponding quantities, and $\Delta\mathcal{P}\left(\right)$ to denote the error in the estimates. In particular, for an empirical estimation in which vertex V_i is sampled T_i times, with parents taking value $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0,1\}^{|\mathrm{Pa}(i)|}$, we have estimate

$$\widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) = \frac{\sum_{s=1}^{T_i} \mathbb{I}[Y_{i,s} = \mathbf{z}_i]}{T_i},$$

where $Y_{i,s}$ is the s-th sample of vertex V_i. In addition, we have

$$\begin{aligned}
\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) &= \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) - \widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \\
\widehat{\mu}(A) &= \sum_{\mathbf{z} \in Z(A)} \prod_{i \in \mathcal{V}(A)} \widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \qquad (5.2)
\end{aligned}$$

Recall that, in the causal bandits problem, the objective is to find—from within a specified collection of interventions $\mathcal{A}$—an intervention with maximum possible expected reward. We will write $A^* \in \mathcal{A}$ to denote the optimal intervention and $\mu(A^*)$ for the optimal reward, i.e., $\mu(A^*) = \max_{A \in \mathcal{A}} \ \mu(A)$. Also, for any algorithm, let $A_T \in \mathcal{A}$ be the (randomized) output computed after T rounds; in each round the algorithm performs an intervention and observes a sample under it.[1] The simple regret of the algorithm is defined as

$$R_T = \mathbb{E}\left[\mu(A^*) - \mu(A_T)\right]. \tag{5.3}$$

5.4 Finding Near-Optimal Intervention via Covering

To find a near-optimal intervention from the given set of interventions $\mathcal{A}$ (specifically, to bound simple regret), instead of directly performing each $A \in \mathcal{A}$, we utilize interventions from a curated set of interventions $\mathcal{I}$, referred to as the covering intervention set (see Definition 5.1). The obtained samples are then used to estimate the interventional distribution for each $A \in \mathcal{A}$ and, hence, find a near-optimal intervention within $\mathcal{A}$. The notion of covering intervention set was formulated in [ABDK18] and is defined next.

Definition 5.1 (Covering Intervention Set). *A collection of interventions $\mathcal{I}$ is said to be a* covering intervention set *iff, for each vertex $i \in \mathcal{V}$ and every assignment $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0,1\}^{|\mathrm{Pa}(i)|}$, there exists an intervention $I \in \mathcal{I}$ with the properties that*

- *Vertex i not intervened in I (i.e., $I_i = *$).*
- *Every vertex in $\mathrm{Pa}(i)$ is intervened (i.e., $I_p \neq *$, for all $p \in \mathrm{Pa}(i)$).*
- *I restricted to $\mathrm{Pa}(i)$ has the assignment $\mathbf{z}_{\mathrm{Pa}(i)}$ (i.e., $I_p = \mathbf{z}_{\mathrm{Pa}(i),p}$ for all $p \in \mathrm{Pa}(i)$).*

It is shown in [ABDK18] that, for any causal graph $\mathcal{G}$ with N vertices and in-degree at most d, one can construct—using a randomized method—a covering intervention set $\mathcal{I}$ of size $O\left(d\ 2^d \log(NT)\right)$.

Specifically, for count $k = 3d\ 2^d(\log N + 2d + \log T)$, one can populate k interventions $I \in \{0,1,*\}^N$ as follows: for each variable $i \in \mathcal{V}$, independently, set

$$I_i = \begin{cases} 0 & \text{with probability } \frac{d}{2(1+d)}, \\ 1 & \text{with probability } \frac{d}{2(1+d)}, \\ * & \text{otherwise.} \end{cases}$$

[1] Note that while the computed intervention must be contained in set $\mathcal{A}$, the interventions performed in the T rounds are not necessarily from $\mathcal{A}$.

All the constructed k interventions constitute the set $\mathcal{I}$. This randomized construction is known to succeed (in providing a covering interventions set) with probability at least $(1 - 1/T)$. Formally,[1]

Lemma 5.1 ([ABDK18]). *For any moderately large $T \in \mathbb{Z}_+$, every causal graph $\mathcal{G}$—with N vertices and in-degree at most d—admits a covering intervention set $\mathcal{I}$ of size $k = 3d\ 2^d(\log N + 2d + \log T)$. Furthermore, such a set $\mathcal{I}$ can be found with probability at least $(1 - 1/T)$.*

We will write ConstructCover($\mathcal{G}$) to denote the randomized construction of $\mathcal{I}$ mentioned above. ConstructCover($\cdot$) will be used as a subroutine in our simple-regret algorithm (Algorithm 9).

Theorem 5.1, stated below, is the main result of this section. The theorem asserts that, for causal graphs with constant in-degree and N vertices, Algorithm 9 achieves a simple regret of $\widetilde{O}\left(\sqrt{N/T}\right)$.

Given a causal graph $\mathcal{G}$ and a collection of interventions $\mathcal{A}$, Algorithm 9 first obtains a covering intervention set $\mathcal{I}$, for the graph $\mathcal{G}$, via the subroutine ConstructCover. Then, the algorithms performs, $T/|\mathcal{I}|$ times, each intervention $I \in \mathcal{I}$. Since $\mathcal{I}$ is a covering intervention set, for each vertex $\widehat{i} \in \mathcal{V}$, there exists an intervention $\widehat{I} \in \mathcal{I}$ under which all the parents $\mathrm{Pa}\left(\widehat{i}\right)$ are intervened upon, but $\widehat{i}$ itself is not. The intervention $\widehat{I}$ has already been performed $T/|\mathcal{I}|$ times by the algorithm. Using these $T/|\mathcal{I}|$ independent samples and for a specific assignment $\mathbf{z}_{\mathrm{Pa}\left(\widehat{i}\right)}$ (induced under $\widehat{I}$), we have the estimate $\widehat{\mathcal{P}}\left(\mathbf{z}_{\widehat{i}} \mid \mathbf{z}_{\mathrm{Pa}\left(\widehat{i}\right)}\right)$. Hence, for every vertex $i \in \mathcal{V}$ and every assignment $\mathbf{z}_{\mathrm{Pa}(i)}$, the algorithm has an estimate $\widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$ in hand. Using these probability estimates, the algorithm computes the reward estimates $\widehat{\mu}(A)$ for each intervention $A \in \mathcal{A}$; see equation (5.2). Finally, enumerating over the given set $\mathcal{A}$, the algorithm returns the intervention with the maximum estimated reward. It is relevant to note that this patently simple algorithm requires a technically involved regret analysis (detailed in Section 2.3.1). Indeed, the analysis is a key contribution of the current work.

Theorem 5.1. *Let $\mathcal{G}$ be any given causal graph with N vertices and in-degree at most d. Also, let $\mathcal{I}$ be a covering intervention set of $\mathcal{G}$. Then, Algorithm 9—when executed for any (moderately large) time horizon T—achieves simple regret*

$$R_T = O\left(\sqrt{\frac{N|\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}}\right).$$

[1] This lemma is a direct implication of Lemma 2 from [ABDK18], instantiated with $\delta = \frac{1}{T}$, $K = 2$.

Algorithm 9 Covering Interventions Algorithm

Input: Causal graph $\mathcal{G}$, target intervention set $\mathcal{A}$, and time horizon $T \in \mathbb{Z}_+$.

```
Set $\mathcal{I} \leftarrow$ ConstructCover($\mathcal{G}$).
For each $I \in \mathcal{I}$, intervene with do($I$) and collect $\frac{T}{|\mathcal{I}|}$ samples.
for each intervention $A \in \mathcal{A}$ do
    Compute $\widehat{\mu}(A)$ using equation (5.2).
end for
return argmax$_{A\in\mathcal{A}}$ $\widehat{\mu}(A)$.
```

Hence, using Lemma 5.1, we obtain the following bound on the simple regret of Algorithm 9

$$R_T = O\left(\sqrt{\frac{N\ d2^d\ \log|\mathcal{A}|}{T}}\log T\right).$$

For graphs with additional structure (e.g. bounded out degree or trees), one can obtain covering intervention sets with size smaller than the one provided in Lemma 5.1 (see Lemma 2 in [ABDK18]). Since the regret guarantee of Algorithm 9 depends on the size of the covering intervention set, the simple regret bound improves for such specific graphs.

5.4.1 Regret Analysis

We first provide a standard concentration bound which will be used in the analysis.

Lemma 5.2 (Hoeffding's Inequality). *Let $Z_1, \dots, Z_n$ be independent bounded random variables with $Z_i \in [a_i, b_i]$, for all $i \in [n]$. Then, for all $\varepsilon \geq 0$:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| \geq \varepsilon\right\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right).$$

To begin the regret analysis, we note that, for each intervention $A \in \mathcal{A}$, the estimate $\widehat{\mu}(A)$ can be expressed as

$$\widehat{\mu}(A) = \sum_{\mathbf{z}\in Z(A)} \prod_{i\in\mathcal{V}(A)} \left(\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) + \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right).$$

Expanding the product, we obtain

$$\widehat{\mu}(A) = \mu(A) + \sum_{\mathbf{z}\in Z(A)} \left(\sum_{i\in\mathcal{V}(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}(A), j\neq i} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \ + \ \mathcal{L}_{\mathbf{z}}\right).$$

Here, $\mathcal{L}_{\mathbf{z}}$ represents all the product entries in the expansion that include more than one error term of the form $\Delta\mathcal{P}(\cdot \mid \cdot)$. Specifically,

$$\mathcal{L}_{\mathbf{z}} = \sum_{k=2}^{|\mathcal{V}(A)|} \sum_{\substack{U \subseteq \mathcal{V}(A) \\ |U|=k}} \left[\left(\prod_{i \in U} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \right) \times \right. \\ \left. \left(\prod_{j \in \mathcal{V}(A) \setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \right) \right] \tag{5.4}$$

We further write $\mathcal{H}_{\mathbf{z}}$ to represent the sum of the entries with a single error term:

$$\mathcal{H}_{\mathbf{z}} := \sum_{i \in \mathcal{V}(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{\substack{j \in \mathcal{V}(A) \\ j \neq i}} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \tag{5.5}$$

Hence,

$$\widehat{\mu}(A) - \mu(A) = \sum_{\mathbf{z} \in Z(A)} \left(\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}}\right).$$

We will establish upper bounds on the sums of $\mathcal{L}_{\mathbf{z}}$s and $\mathcal{H}_{\mathbf{z}}$s and in Lemma 5.4 and Lemma 5.5, respectively. These lemmas show that the sum of the $\mathcal{H}$ terms dominates the sum of the $\mathcal{L}$ terms. Furthermore, these bounds imply that the estimated reward $\widehat{\mu}(A)$ is sufficiently close to the true expected reward $\mu(A)$ for each $A \in \mathcal{A}$.

Lemma 5.3. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 9, write $\mathcal{E}$ to denote the event that $|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)| \leq \sqrt{\frac{|\mathcal{I}|(d+\log(NT))}{T}}$, for all vertices $i \in \mathcal{V}$ and all assignments $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0,1\}^{|\mathrm{Pa}(i)|}$. Then, $\mathbb{P}\{\mathcal{E}\} \geq \left(1 - \frac{2}{T}\right)$.*

Proof. Since $\mathcal{I}$ is a covering intervention set, for each conditional distribution $\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$, we have at least $\frac{T}{|\mathcal{I}|}$ independent samples. Now, we invoke Lemma 5.2, with $\varepsilon = \sqrt{\frac{|\mathcal{I}|\log(2^d NT)}{T}}$, and apply the union bound over all $i \in [N]$ and all assignments to $\mathrm{Pa}(i)$. This gives us the desired probability bound. □

Lemma 5.4. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 9, the following event holds with probability at least $\left(1 - \frac{2}{T}\right)$:*

$$\sum_{\mathbf{z} \in Z(A)} |\mathcal{L}_{\mathbf{z}}| \leq 4(N\eta)^2 \qquad \textit{for all } A \in \mathcal{A}.$$

Here, parameter $\eta = \sqrt{\frac{|\mathcal{I}|(d+\log(NT))}{T}}$ *and* T *is moderately large.*

Proof. We will use the fact that each error term in $\mathcal{L}_{\mathbf{z}}$ satisfies the bound stated in Lemma 5.3. Moreover, we utilize the graph structure to marginalize variables that do not appear in an expansion of $\mathcal{L}_{\mathbf{z}}$.

$$\begin{aligned}\sum_{\mathbf{z}\in Z(A)} |\mathcal{L}_{\mathbf{z}}| &\le \sum_{\mathbf{z}\in Z(A)} \sum_{k=2}^{|\mathcal{V}(A)|} \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \left(\prod_{i\in U} \left|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right|\right) \left(\prod_{j\in\mathcal{V}(A)\setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \\ &= \sum_{k=2}^{|\mathcal{V}(A)|} \sum_{\mathbf{z}\in Z(A)} \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \left(\prod_{i\in U} \left|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right|\right) \left(\prod_{j\in\mathcal{V}(A)\setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right).\end{aligned}$$

First, we upper bound each term considered in the outer-most sum. Towards this, let $U = \{V_{x_1}, V_{x_2}, \ldots, V_{x_k}\}$ to be a subset of vertices that appears in the inner sum. Here, $x_1 < x_2 < \ldots < x_k$ and, as mentioned previously, the indexing of the vertices respects a topological ordering over the causal graph. In the derivation below, we will split the sum into k parts, $\sum_{\mathbf{z}_{[1:x_1]}} \sum_{\mathbf{z}_{(x_1:x_2]}} \ldots \sum_{\mathbf{z}_{(x_k:N]}}$, and individually bound the marginalized probability distribution.

$$\begin{aligned}&\sum_{\mathbf{z}\in Z(A)} \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \left(\prod_{i\in U} \left|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right|\right) \left(\prod_{j\in\mathcal{V}(A)\setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \\ &\le \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \sum_{\mathbf{z}\in Z(A)} \eta^k \left(\prod_{j\in\mathcal{V}(A)\setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \qquad \text{(via Lemma 5.3, } \left|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right| \le \eta\text{)} \\ &= \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \eta^k \sum_{\mathbf{z}_{[1:\mathbf{x}_1]}\in Z_{[1:\mathbf{x}_1]}(A)} \left(\prod_{j_1\in\mathcal{V}_{[1:x_1)}(A)} \mathcal{P}\left(\mathbf{z}_{j_1} \mid \mathbf{z}_{\mathrm{Pa}(j_1)}\right)\right) \sum_{\mathbf{z}_{(x_1:\mathbf{x}_2]}\in Z_{(x_1:x_2]}(A)} \left(\prod_{j_2\in\mathcal{V}_{(x_1:x_2)}(A)} \mathcal{P}\left(\mathbf{z}_{j_2} \mid \mathbf{z}_{\mathrm{Pa}(j_2)}\right)\right) \ldots \\ &\sum_{\mathbf{z}\in Z_{(x_i:x_{i+1}]}(A)} \left(\prod_{j_i\in\mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}\left(\mathbf{z}_{j_i} \mid \mathbf{z}_{\mathrm{Pa}(j_i)}\right)\right) \ldots \sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \left(\prod_{j_k\in\mathcal{V}_{(x_k:N]}(A)} \mathcal{P}\left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}(j_k)}\right)\right)\end{aligned} \tag{5.6}$$

The last term in the above expression can be bounded as follows

$$\begin{aligned}\sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \left(\prod_{j\in\mathcal{V}_{(x_k:N]}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) &= \sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{(x_k:N]}(A) = \mathbf{z}_{(x_k:N]}|\mathrm{Pa}(\mathcal{V}_{(x_k:N]}(A))\right] \\ &= \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1|\mathrm{Pa}\left(\mathcal{V}_{(x_k:N]}(A)\right)\right] \leq 1.\end{aligned}$$

For all other terms, we have the following inequality

$$\begin{aligned}&\sum_{\mathbf{z}\in Z_{(x_i:x_{i+1}]}(A)} \left(\prod_{j\in\mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \\ &\qquad= \sum_{\mathbf{z}_{x_{i+1}}\in\{0,1\}} \sum_{\mathbf{z}_{(x_i:x_{i+1})}\in Z_{(x_i:x_{i+1})}(A)} \left(\prod_{j\in\mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \\ &\qquad= \sum_{\mathbf{z}_{x_{i+1}}\in\{0,1\}} \sum_{\mathbf{z}_{(x_i:x_{i+1})}\in Z_{(x_i:x_{i+1})}(A)} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{(x_i:x_{i+1})}(A) = \mathbf{z}_{(x_i:x_{i+1})}|\mathrm{Pa}\left(\mathcal{V}_{(x_i:x_{i+1})}(A)\right)\right] \\ &\qquad\leq \sum_{\mathbf{z}_{x_{i+1}}\in\{0,1\}} 1 \\ &\qquad= 2.\end{aligned}$$

Substituting in (5.6), we get

$$\sum_{\mathbf{z}\in Z(A)} \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} \left(\prod_{i\in U}\left|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right|\right)\left(\prod_{j\in\mathcal{V}(A)\setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right) \leq \sum_{\substack{U\subseteq\mathcal{V}(A)\\|U|=k}} (2\eta)^k = \binom{N}{k}(2\eta)^k.$$

Therefore, the sum $\sum_{\mathbf{z}\in Z(A)} |\mathcal{L}_{\mathbf{z}}|$ satisfies

$$\begin{aligned}\sum_{\mathbf{z}\in Z(A)} |\mathcal{L}_{\mathbf{z}}| &\leq \sum_{k=2}^{N}\binom{N}{k}(2\eta)^k \\ &= \sum_{k=0}^{N}\binom{N}{k}(2\eta)^k - 2N\eta - 1 \\ &= (1+2\eta)^N - 2N\eta - 1 \\ &\leq e^{2N\eta} - 2N\eta - 1 \\ &\leq 1 + 2N\eta + (2N\eta)^2 - 2N\eta - 1 && \left(\text{with } \eta \leq \tfrac{1}{2N}\right)\end{aligned}$$

$$\le 4N^2\eta^2.$$

The lemma stands proved. □

Lemma 5.5. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 9, the following event holds with probability at least $\left(1-\frac{2}{T}\right)$:*

$$\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \le \sqrt{\frac{N|\mathcal{I}|\log\left(|\mathcal{A}|T\right)}{T}} \qquad \textit{for all } A\in\mathcal{A}.$$

Proof. The definition of $\mathcal{H}_{\mathbf{z}}$ gives us

$$\begin{aligned}
&\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \\
&= \left|\sum_{\mathbf{z}\in Z(A)} \sum_{i\in\mathcal{V}(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}(A), j\neq i} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right| \\
&= \left|\sum_{i\in\mathcal{V}(A)} \sum_{\mathbf{z}\in Z(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}(A), j\neq i} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\right| \\
&= \left|\sum_{i\in\mathcal{V}(A)} \sum_{\substack{\mathbf{z}_{[1:i]}\in\\ Z_{[1:i]}(A)}} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \sum_{\substack{\mathbf{z}_{(i:N]}\in\\ Z_{(i:N]}(A)}} \prod_{k\in\mathcal{V}_{(i:N]}(A)} \mathcal{P}\left(\mathbf{z}_k \mid \mathbf{z}_{\mathrm{Pa}(k)}\right)\right| \\
&= \left|\sum_{i\in\mathcal{V}(A)} \sum_{\substack{\mathbf{z}_{[1:i]}\in\\ Z_{[1:i]}(A)}} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \sum_{\substack{\mathbf{z}_{(i:N]}\in\\ Z_{(i:N]}(A)}} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{(i:N]}(A) = \mathbf{z}_{(i:N]} \mid \mathrm{Pa}\left(\mathcal{V}_{(i:N]}(A)\right)\right]\right| \\
&= \left|\sum_{i\in\mathcal{V}(A)} \sum_{\substack{\mathbf{z}_{[1:i]}\in\\ Z_{[1:i]}(A)}} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1 \mid \mathrm{Pa}\left(\mathcal{V}_{(i:N]}(A)\right)\right]\right| \\
&= \left|\sum_{i\in\mathcal{V}(A)} \sum_{\mathbf{z}_i\in\{0,1\}} \sum_{\substack{\mathbf{z}_{[1:i)}\in\\ Z_{[1:i)}(A)}} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1 \mid \mathrm{Pa}\left(\mathcal{V}_{(i:N]}(A)\right)\right]\right|
\end{aligned}$$

$$= \left| \sum_{i \in \mathcal{V}(A)} \sum_{\mathbf{z}_i \in \{0,1\}} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(i)} \in \\ Z_{\mathrm{Pa}(i)}(A)}} \Delta \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \sum_{\substack{\mathbf{z}_{\mathrm{Ac}(i)} \in \\ Z_{\mathrm{Ac}(i)}(A)}} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1 \mid \mathrm{Pa}\left(\mathcal{V}_{(i:N]}(A)\right)\right] \right| .$$

Recall that $\mathrm{Ac}(i) = [1, i) \setminus \mathrm{Pa}(i)$ and write

$$c_i(z_i, \mathbf{z}_{\mathrm{Pa}(i)}) := \sum_{\substack{\mathbf{z}_{\mathrm{Ac}(i)} \in \\ Z_{\mathrm{Ac}(i)}(A)}} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1 \mid \mathrm{Pa}\left(\mathcal{V}_{(i:N]}(A)\right)\right] \tag{5.7}$$

Also, as a shorthand for $z_i = 1$ and $z_i = 0$ we will write 1_i and 0_i, respectively. With these notations, we have

$$\begin{aligned} \left| \sum_{\mathbf{z} \in Z(A)} \mathcal{H}_{\mathbf{z}} \right| &= \left| \sum_{i \in \mathcal{V}(A)} \sum_{\mathbf{z}_i \in \{0,1\}} \sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \Delta \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) c_i(\mathbf{z}_i, \mathbf{z}_{\mathrm{Pa}(i)}) \right| \\ &= \left| \sum_{i \in \mathcal{V}(A)} \sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \Delta \mathcal{P}\left(1_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \left(c_i(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c_i(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right) \right| \\ &\qquad \left(\text{since } \Delta \mathcal{P}\left(1_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) = -\Delta \mathcal{P}\left(0_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right) \end{aligned}$$

Since $\mathcal{I}$ is a covering intervention set, for each pair $(i, \mathbf{z}_{\mathrm{Pa}(i)})$, there exists an intervention $I \in \mathcal{I}$ such that intervening $\mathrm{do}(I)$ provides a sample from the conditional probability distribution $\mathbb{P}[V_i = 1 \mid \mathrm{Pa}(V_i) = \mathbf{z}_i]$. Hence, Line 2 of the algorithm provides at least $\frac{T}{|\mathcal{I}|}$ independent samples from the conditional distribution $\mathbb{P}[V_i = 1 \mid \mathrm{Pa}(V_i) = \mathbf{z}_i]$. We write the s^{th} sample for this conditional distribution by $Y_s(i, \mathbf{z}_{\mathrm{Pa}(i)})$. Now, we have

$$\left| \sum_{\mathbf{z} \in Z(A)} \mathcal{H}_{\mathbf{z}} \right| = \left| \sum_{i \in \mathcal{V}(A)} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(i)} \in \\ Z_{\mathrm{Pa}(i)}(A)}} \frac{|\mathcal{I}|}{T} \left(\sum_{s=1}^{T/|\mathcal{I}|} Y_s(i, \mathbf{z}_{\mathrm{Pa}(i)}) - \mathcal{P}\left(1_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \right) \left(c_i(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c_i(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right) \right| .$$

We will apply Hoeffding's inequality (Lemma 5.2) to bound the above expression. Note that in this expression, besides $Y_s(i, \mathbf{z}_{\mathrm{Pa}(i)})$-s, all the other terms are deterministic. In particular, we show in Claim 5.1 (stated and proved below) that $\sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} (c(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c(0_i, \mathbf{z}_{\mathrm{Pa}(i)}))^2 \leq 1$, for all i. Hence, for any $A \in \mathcal{A}$, Lemma 5.2 gives us

$$\mathbb{P}\left(\left| \sum_{\mathbf{z} \in Z(A)} \mathcal{H}_{\mathbf{z}} \right| \geq \varepsilon \right) \leq 2 \exp\left(\frac{-T\varepsilon^2}{|\mathcal{I}| \sum_{i \in \mathcal{V}(A)} \sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} (c_i(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c_i(0_i, \mathbf{z}_{\mathrm{Pa}(i)}))^2} \right)$$

$$
\begin{aligned}
&\leq 2\,\exp\left(\frac{-T\varepsilon^2}{|\mathcal{I}|\,|\mathcal{V}(A)|}\right) && \text{(via Claim 5.1)}\\
&\leq 2\,\exp\left(\frac{-T\varepsilon^2}{|\mathcal{I}|\,N}\right).
\end{aligned}
$$

Setting $\varepsilon = \sqrt{\frac{N\ |\mathcal{I}| \log(|\mathcal{A}|T)}{T}}$ and taking union bound over all $A \in \mathcal{A}$, gives us the required probability bound. This completes the proof of the lemma. □

We next establish the claim used in the proof of Lemma 5.5.

Claim 5.1.

$$\sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \left(c(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right)^2 \leq 1.$$

Proof. The definition of $c(\mathbf{z}_i, \mathbf{z}_{\mathrm{Pa}(i)})$ (see equation (5.7)) gives us

$$
\begin{aligned}
&\left|c(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right|\\
&= \Bigg| \sum_{\mathbf{z} \in Z_{\mathrm{Ac}(i)}(A)} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}\left(\mathcal{V}_{[1:i]}(A)\right) = (\mathbf{z}_{[1:i)} \cup 1_i)\right] -\\
&\qquad \sum_{\mathbf{z} \in Z_{\mathrm{Ac}(i)}(A)} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}\left(\mathcal{V}_{[1:i]}(A)\right) = (\mathbf{z}_{[1:i)} \cup 0_i)\right] \Bigg|\\
&= \Bigg| \sum_{\mathbf{z} \in Z_{\mathrm{Ac}(i)}(A)} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \Big(\mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}(\mathcal{V}_{[1:i]}(A)) = (\mathbf{z}_{[1:i)} \cup 1_i)\ \right] -\\
&\qquad \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}(\mathcal{V}_{[1:i]}(A)) = (\mathbf{z}_{[1:i)} \cup 0_i)\right] \Big) \Bigg|\\
&\leq \sum_{\mathbf{z} \in Z_{\mathrm{Ac}(i)}(A)} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \Big| \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}(\mathcal{V}_{[1:i]}(A)) = (\mathbf{z}_{[1:i)} \cup 1_i)\right] -\\
&\qquad \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{[i+1:N]}(A) = \mathbf{z}_{[i+1:N]} \mid \mathrm{Pa}(\mathcal{V}_{[1:i]}(A)) = (\mathbf{z}_{[1:i)} \cup 0_i)\right] \Big|\\
&\leq \sum_{\mathbf{z} \in Z_{\mathrm{Ac}(i)}(A)} \prod_{j \in \mathcal{V}_{[1:i)}(A)} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right)\\
&= \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{\mathrm{Pa}(i)}(A) = \mathbf{z}_{\mathrm{Pa}(i)}\right].
\end{aligned}
$$

Hence, under intervention $A \in \mathcal{A}$, we have

$$\begin{aligned}\sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \left(c(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right)^2 &\leq \sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \left|c(1_i, \mathbf{z}_{\mathrm{Pa}(i)}) - c(0_i, \mathbf{z}_{\mathrm{Pa}(i)})\right| \\ &\leq \sum_{\mathbf{z}_{\mathrm{Pa}(i)} \in Z_{\mathrm{Pa}(i)}(A)} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{\mathrm{Pa}(i)}(A) = \mathbf{z}_{\mathrm{Pa}(i)}\right] \\ &\leq 1.\end{aligned}$$

This completes the proof of the claim. □

Recall that the random variables $\mathcal{L}_{\mathbf{z}}$ and $\mathcal{H}_{\mathbf{z}}$ depend on the error terms $\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$. Moreover, in Lemma 5.4 and 5.5, the considered sums can range over exponentially many such variables. The technically involved contribution of these lemmas is that we obtain small error bounds even in such settings of exponentially large sums.

5.4.2 Proof of Theorem 5.1

Lemma 5.1 implies that, with probability at least $\left(1 - \frac{1}{T}\right)$, the set $\mathcal{I}$ obtained in Line 1 of Algorithm 9 is indeed a covering intervention set. We combine this guarantee with Lemmas 5.4 and 5.5. In particular, with probability at least $\left(1 - \frac{5}{T}\right)$, we have, for all $A \in \mathcal{A}$:

$$\begin{aligned}&|\mu(A) - \widehat{\mu}(A)| \\ &\quad = \left|\sum_{\mathbf{z} \in Z(A)} (\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}})\right| \\ &\quad \leq \sqrt{\frac{N|\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}} + \frac{4N^2|\mathcal{I}|(d + \log\left(NT\right))}{T} \\ &\quad \leq 2\sqrt{\frac{N|\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}} && \text{(for } T \gtrsim N^3\text{)}\end{aligned}$$

Let $A_T \in \mathcal{A}$ be the intervention returned by Algorithm 9 (after T rounds of interventions), i.e., $A_T = \mathrm{argmax}_{A \in \mathcal{A}} \ \widehat{\mu}(A)$. In addition, $A^* = \mathrm{argmax}_{A \in \mathcal{A}} \ \mu(A)$ be the optimal intervention. Hence, with probability at least $\left(1 - \frac{5}{T}\right)$, we have

$$\mu(A^*) - \mu(A_T) \leq 4\sqrt{\frac{N|\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}} \tag{5.8}$$

This guarantee gives us the desired upper bound on the simple regret, R_T, of Algorithm 9:

$$\begin{aligned} R_T &= \mathbb{E}\left[\mu(A^*) - \mu(A_T)\right] \\ &\leq \left(4\sqrt{\frac{N|\mathcal{I}|\log\left(|\mathcal{A}|T\right)}{T}}\right)\left(1-\frac{5}{T}\right)+\frac{5}{T} \\ &\leq \ 5\sqrt{\frac{N|\mathcal{I}|\log\left(|\mathcal{A}|T\right)}{T}}. \end{aligned}$$

Since the size of the covering intervention set satisfies $|\mathcal{I}| = 3d \cdot 2^d(\log N + 2d + \log T)$ (see Lemma 5.1), we also have the following explicit form of the simple regret bound

$$R_T = O\left(\sqrt{\frac{N\ d2^d\ \log|\mathcal{A}|}{T}}\log T\right).$$

The theorem stands proved.

5.5 Algorithm for Graphs with Unobserved Variables

We now extend our algorithm to causal graphs with unobserved variables. In particular, we study Semi Markovian Bayesian Networks (SMBNs) where we have the causal graph defined as $\mathcal{G} = (\mathcal{V}, E, E')$. Here, E is the set of directed edges, and E' is the set of bi-directed edges denoting the presence of an unobserved common parent. Any general causal graph can be projected to an equivalent SMBN [TP02]. Hence, without loss of generality and throughout this section, we assume that the causal graph is an SMBN. It is relevant to note that in an SMBN all the vertices in $\mathcal{V}$ are observable and the unobserved variables are encapsulated by the edges E'.

Assume that the vertices $\mathcal{V}$ are topologically ordered (based on the directed edges E) and the ordering is preserved in any subset $\mathcal{U} \subset \mathcal{V}$. The SMBN graph $\mathcal{G}$ can be decomposed into a disjoint set of vertices known as *confounded components* (c-components), where each c-component is the maximal set of vertices that are connected through a bi-directed edge in E'. Let $\mathcal{C}(A)$ denote all the c-components of $\mathcal{G}$ under intervention A. We use C_i to denote the i^{th} c-component in $\mathcal{C}(A)$. We assume that any C_i maintains the topological order (induced by the directed edges E). Now, the joint distribution of the vertices for an assignment $\mathbf{z} \in Z(A)$, under intervention A, can be written as

$$\mathbb{P}\left[V = \mathbf{z} \mid \mathrm{do}(A)\right] = \prod_{C_i \in \mathcal{C}(A)} \mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}\left(\mathbf{z}_{C_i}\right).$$

Under an empirical estimation, we represent the s^{th} sample from the distribution $\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}(\mathbf{z}_{C_i})$ via the indicator random variable $Y_s(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)})$, which takes the value one when $\mathcal{V}_{C_i} = \mathbf{z}_{C_i}$, else it takes the value zero. Let $n\left(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)$ be the total number of samples in this for the pair $(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)})$. We compute the probability estimates as follows

$$\widehat{\mathcal{P}}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}(\mathbf{z}_{C_i}) = \frac{\sum_{s=1}^{T_i} Y_s\left(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)}{n\left(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)} \tag{5.9}$$

$$\widehat{\mu}(A) = \sum_{\mathbf{z} \in Z(A)} \prod_{C_i \in \mathcal{C}(A)} \widehat{\mathcal{P}}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}(\mathbf{z}_{C_i}) \tag{5.10}$$

Next, we extend the definition of covering intervention set (Definition 5.1) for SMBNs:

Definition 5.2. *A set of intervention $\mathcal{I}$ is a covering intervention set if for all subsets S of every c-component in $\mathcal{G}$, and every assignment $\mathbf{z}_{\mathrm{Pa}(S)} \in \{0,1\}^{|\mathrm{Pa}(S)|}$ there exists and $I \in \mathcal{I}$ with the properties that*

- *No vertex in S is intervened in I.*
- *Every vertex in $\mathrm{Pa}(S)$ is intervened in I.*
- $\mathrm{Pa}(S)$ *is intervened with assignment* $\mathbf{z}_{\mathrm{Pa}(S)}$.

We construct a covering intervention set as before using the randomized method in [ABDK18]. The next lemma states that the randomized method provides a covering intervention set of size $\widetilde{O}(\log N)$ even in the case of SMBNs. This result is a direct implication of Lemma 2 in [ABDK18].

Lemma 5.6 ([ABDK18])**.** *For any moderately large $T \in \mathbb{Z}_+$ and any causal graph $\mathcal{G}$—with in-degree at most d and c-components of size at most ℓ—there exists a covering intervention set $\mathcal{I}$ of size $k = (3d)^{\ell}\, 2^{\ell d}(\log N + 2\ell d + \log T)$. Furthermore, such a set $\mathcal{I}$ can be found with probability at least $\left(1 - \frac{1}{T}\right)$.*

The simple regret algorithm for SMBNs is exactly the same as Algorithm 9, except for the following two changes:

- The ConstructCover subroutine returns a covering intervention set of size $(3d)^{\ell} 2^{\ell d}(\log N + 2\ell d + \log T)$.
- We use equation (5.10) to compute the estimates $\widehat{\mu}(A)$ for each $A \in \mathcal{A}$.

The theorem below is the main result of this section.

Theorem 5.2. *Let $\mathcal{G}$ be any given causal graph over N vertices and with c-components of size at most ℓ. Also, let the in-degree of the vertices in $\mathcal{G}$ be at most d. Then, for any (moderately large) time horizon T and given any covering intervention set $\mathcal{I}$ of $\mathcal{G}$, Algorithm 9 achieves simple regret*

$$\mathrm{R}_T = O\left(\sqrt{\frac{N\ 2^d\ 4^\ell\ |\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}}\right).$$

Hence, using Lemma 5.6, we obtain the following bound on the simple regret

$$\mathrm{R}_T = O\left(\sqrt{\frac{N\ (3d\ 8^d)^\ell\ \log|\mathcal{A}|}{T}} \log T\right).$$

5.5.1 Regret Analysis for SMBNs

We introduce the notion of *pseudo parents* of a vertex in a Semi Markov Bayesian Networks (SMBN) graph $\mathcal{G}$, which we will use throughout the proof. Recall that $\mathcal{V}$ denotes the set of vertices, and they conform to a topological ordering. We assume that each c-component C_i maintains the ordering. For an intervention A, consider any c-component $C \in \mathcal{C}(A)$ with vertices $(U_1, U_2, \ldots, U_m)$, the pseudo parents of a vertex U_j is defined as

$$\mathrm{Pa}'(j) \coloneqq \mathrm{Pa}(\{U_1, U_2, \ldots U_j\}) \cup \{U_1, U_2, \ldots U_{j-1}\} \tag{5.11}$$

For any SMBN graph with in-degree at most d and c-components of size at most ℓ, the size $|\mathrm{Pa}'(j)|$ is at most $d\ell + \ell$. Furthermore, note that the set $\mathrm{Pa}'(j)$ will always precede the vertex V_j in any topological ordering of the graph.

The next lemma shows that the distribution of any c-component conditioned on its parents, $\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C)}}(\mathbf{z}_C)$, can be factorized into the distribution of individual vertices conditioned on its pseudo parents. This allows us to extend the techniques used for the regret analysis of fully observable graphs (Section 2.3.1) to the case of SMBNs. Intuitively, one can view the factorization of an SMBN (under an intervention A) as a factorization over a fully observable graph where each vertex V_j has the set $\mathrm{Pa}'(j)$ as its parents.

Lemma 5.7. *For any intervention A and any c-component $C \in \mathcal{C}(A)$, consisting of vertices $\{U_1, U_2 \ldots U_m\}$, we have*

$$\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C)}}(\mathbf{z}_C) = \prod_{j \in C} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right).$$

Here, $\mathrm{Pa}'(j)$ *denotes the set of pseudo parents as defined in equation (5.11).*

A proof of Lemma 5.7 appears in section 5.6.

Now, recall that the estimate $\widehat{\mu}(A)$ can be written as

$$\begin{aligned}
\widehat{\mu}(A) &= \sum_{\mathbf{z}\in Z(A)} \prod_{C_i\in\mathcal{C}(A)} \widehat{\mathcal{P}}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}(\mathbf{z}_{C_i}) \\
&= \sum_{\mathbf{z}\in Z(A)} \prod_{i\in\mathcal{C}(A)} \left(\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) + \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right) \\
&= \mu(A) + \sum_{\mathbf{z}\in Z(A)} \Bigg(\sum_{C_i\in\mathcal{C}(A)} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) \prod_{C_j\in\mathcal{C}(A), j\neq i} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right) + \\
&\quad \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=2}} \left(\prod_{C_i\in U} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right) \left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) + \\
&\quad \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=3}} \left(\prod_{C_i\in U} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right) \left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) + \cdots \Bigg) \\
&\qquad\qquad \text{(expanding product terms)}
\end{aligned}$$

Here, $\Delta\mathcal{P}\left(\right)$ denotes the error in the estimate of the conditional probabilities. Let $\mathcal{L}_{\mathbf{z}}$ represent all the product entries in the expansion that include more than one error term ($\Delta\mathcal{P}\left(\right)$). Specifically,

$$\begin{aligned}
\mathcal{L}_{\mathbf{z}} &= \sum_{k=2}^{|\mathcal{C}(A)|} \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \left(\prod_{C_i\in U} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right) \left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) \\
&= \sum_{k=2}^{|\mathcal{C}(A)|} \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \left(\prod_{C_i\in U} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right) \left(\prod_{\substack{C\in\mathcal{C}(A)\setminus C_i,\\ j\in C}} \mathcal{P}_A\left(\mathbf{z}_{j} \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) \qquad \text{(via Lemma 5.7)}
\end{aligned}$$

We further represent all the entries with a single $\Delta\mathcal{P}\left(\right)$ term as

$$\begin{aligned}
\mathcal{H}_{\mathbf{z}} &= \sum_{C_i\in\mathcal{C}(A)} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) \prod_{\substack{C_k\in\mathcal{C}(A)\\k\neq i}} \mathcal{P}_A\left(\mathbf{z}_{C_k} \mid \mathbf{z}_{\mathrm{Pa}(C_k)}\right) \\
&= \sum_{C_i\in\mathcal{C}(A)} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) \prod_{j\in\mathcal{V}(A)\setminus C_i} \mathcal{P}_A\left(\mathbf{z}_{j} \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right) \qquad (5.12)
\end{aligned}$$

Here, the last equality follows from Lemma 5.7. Hence, we have

$$\widehat{\mu}(A) - \mu(A) = \sum_{\mathbf{z} \in Z(A)} (\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}}) \tag{5.13}$$

We will establish upper bounds on the sums of $\mathcal{L}_{\mathbf{z}}$s and $\mathcal{H}_{\mathbf{z}}$s in Lemma 5.9 and Lemma 5.10, respectively. These lemmas show that the sum of the $\mathcal{H}$ terms dominates the sum of $\mathcal{L}$ terms. Furthermore, these bounds imply that the estimated reward $\widehat{\mu}(A)$ is sufficiently close to the true expected reward $\mu(A)$ for each intervention $A \in \mathcal{A}$.

Lemma 5.8. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 9, write $\mathcal{E}$ to denote the event that $|\Delta \mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}(\mathbf{z}_{C_i})| \leq \sqrt{\frac{|\mathcal{I}|(\ell d + \ell + \log(NT))}{T}}$ for all c-components $C_i \in \mathcal{C}(A)$ and for all $A \in \mathcal{A}$. Then,* $\Pr\{\mathcal{E}\} \geq \left(1 - \frac{2}{T}\right)$.

Proof. Since $\mathcal{I}$ is a covering intervention set (see Defintion 5.2), for each distribution $\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(i)}}(\mathbf{z}_{C_i})$, we have at least $\frac{T}{|\mathcal{I}|}$ independent samples. Also, note that the total number of distributions to be estimated is at most $2^{(\ell d + \ell)} N$. This follows from the fact that each c-component—under any intervention—is a subset of a c-component in the original graph $\mathcal{G}$, and the number of c-components in $\mathcal{G}$ is at most N. Hence, the number of possible distinct c-components (across all intervention) is at most $N 2^{\ell}$. Furthermore, each c-component can have at most ℓd parents with at most $2^{\ell d}$ distinct binary assignments to the parents.

With this count in hand, we invoke Lemma 5.2, with $\varepsilon = \sqrt{\frac{|\mathcal{I}|(\log(2^{\ell d + \ell} NT))}{T}}$ and apply the union bound over all $(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)})$ pairs. This gives us the desired probability bound and completes the proof of the lemma. □

Lemma 5.9. *For estimates obtained via a covering intervention set $\mathcal{I}$, the following event holds with probability at least $(1 - \frac{2}{T})$:*

$$\sum_{\mathbf{z} \in Z(A)} |\mathcal{L}_{\mathbf{z}}| \leq 4^{\ell} (N\eta)^2 \qquad \textit{for all } A \in \mathcal{A}.$$

Here, parameter $\eta = \sqrt{\frac{|\mathcal{I}|(\ell d + \ell + \log(NT))}{T}}$ and T is moderately large.

Proof. We use the fact that each error term in $\mathcal{L}_{\mathbf{z}}$ satisfies the bound stated in Lemma 5.8. Moreover, we use the graph structure to marginalize variables that do not appear in the error terms. The idea is to split the sum $\sum_{\mathbf{z} \in Z(A)}$ into $\sum_{\mathbf{z}_{[1:x_1]}} \sum_{\mathbf{z}_{(x_1:x_2]}} \dots \sum_{\mathbf{z}_{(x_k:N]}}$, where $\{x_1, x_2 \dots, x_k\}$

denotes all the indices in $\mathcal{C}(A)$ that show up as $\Delta\mathcal{P}\left(\right)$ in the expression for $\mathcal{L}_{\mathbf{z}}$.

$$\begin{aligned}
\sum_{\mathbf{z}\in Z(A)} |\mathcal{L}_{\mathbf{z}}| &\leq \sum_{\mathbf{z}\in Z(A)} \sum_{k=2}^{|\mathcal{C}(A)|} \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \left(\prod_{C_i\in U} \left|\Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right|\right)\left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) \\
&= \sum_{k=2}^{|\mathcal{C}(A)|} \sum_{\mathbf{z}\in Z(A)} \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \left(\prod_{C_i\in U} \left|\Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right|\right)\left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) \\
&\leq \sum_{k=2}^{|\mathcal{C}(A)|} \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \sum_{\mathbf{z}\in Z(A)} \eta^k \left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) \\
&\qquad\qquad\qquad\qquad \left(\text{via Lemma 5.8, } \left|\Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right| \leq \eta\right)
\end{aligned}$$

First, we upper bound each term considered in the outer-most sum. Towards this, let U denote the set of c-components that show up as $\Delta\mathcal{P}\left(\right)$, we define $X \coloneqq \cup_{C_i\in U} C_i = \{x_1, x_2, \cdots, x_m\}$ where x_i denotes the vertex $V_{x_i} \in \mathcal{V}(A)$. Note that since c-components are at most of size ℓ and for $|U| = k$, we have $|X| \leq \ell k$. Now, using Lemma 5.7, we obtain

$$\begin{aligned}
&\sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \sum_{\mathbf{z}\in Z(A)} \eta^k \left(\prod_{C_j\in\mathcal{C}(A)\setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right)\right) \\
&= \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \sum_{\mathbf{z}\in Z(A)} \eta^k \left(\prod_{j\in\mathcal{V}(A)\setminus X} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) \\
&= \sum_{\substack{U\subseteq\mathcal{C}(A)\\|U|=k}} \eta^k \sum_{\mathbf{z}_{[1:\mathbf{x}_1]}\in Z_{[1:\mathbf{x}_1]}(A)} \left(\prod_{j\in\mathcal{V}_{[1:x_1)}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) \sum_{\mathbf{z}_{(x_1:\mathbf{x}_2]}\in Z_{(x_1:x_1]}(A)} \left(\prod_{j\in\mathcal{V}_{(x_1:x_2)}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) \cdots \\
&\sum_{\mathbf{z}\in Z_{(x_i:x_{i+1}]}(A)} \left(\prod_{i\in\mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) \cdots \sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \left(\prod_{j\in\mathcal{V}_{(x_k:N]}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right)
\end{aligned} \tag{5.14}$$

The last term in the above expression can be bounded as follows

$$\sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \left(\prod_{i\in\mathcal{V}_{(x_k:N]}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right) = \sum_{\mathbf{z}_{(x_k:N]}\in Z_{(x_k:N]}(A)} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{(x_k:N]}(A) = \mathbf{z}_{(x_k:N]} | \mathrm{Pa}'(\mathcal{V}_{(x_k:N]}(A))\right.$$

$$= \mathbb{P}_{\mathrm{do}(A)}\left[V_N = 1 | \mathrm{Pa}'\left(\mathcal{V}_{(x_k:N]}(A)\right)\right] \leq 1.$$

For all the other terms, we have the following bound

$$\begin{aligned}
&\sum_{\mathbf{z} \in Z_{(x_i:x_{i+1}]}(A)} \left(\prod_{i \in \mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right) \right) \\
&\quad = \sum_{\mathbf{z}_{x_{i+1}} \in \{0,1\}} \sum_{\mathbf{z}_{(x_i:x_{i+1})} \in Z_{(x_i:x_{i+1})}(A)} \left(\prod_{i \in \mathcal{V}_{(x_i:x_{i+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right) \right) \\
&\quad = \sum_{\mathbf{z}_{x_{i+1}} \in \{0,1\}} \sum_{\mathbf{z}_{(x_i:x_{i+1})} \in Z_{(x_i:x_{i+1})}(A)} \mathbb{P}_{\mathrm{do}(A)}\left[\mathcal{V}_{(x_i:x_{i+1})}(A) = \mathbf{z}_{(x_i:x_{i+1})} | \mathrm{Pa}'\left(\mathcal{V}_{(x_i:x_{i+1})}(A)\right)\right] \\
&\quad \leq \sum_{\mathbf{z}_{x_{i+1}} \in \{0,1\}} 1 \\
&\quad = 2.
\end{aligned}$$

Substituting in (5.14), we get

$$\begin{aligned}
\sum_{\substack{U \subseteq \mathcal{C}(A) \\ |U| = k}} \sum_{\mathbf{z} \in Z(A)} \eta^k \left(\prod_{C_j \in \mathcal{C}(A) \setminus U} \mathcal{P}_A\left(\mathbf{z}_{C_j} \mid \mathbf{z}_{\mathrm{Pa}(C_j)}\right) \right) &\leq \sum_{\substack{U \subseteq \mathcal{C}(A) \\ |U| = k}} \eta^k \, 2^{\ell k} && (\text{since } |X| \leq \ell k) \\
&= \binom{N}{k} \left(2^\ell \eta\right)^k.
\end{aligned}$$

Therefore, the sum $\sum_{\mathbf{z} \in Z(A)} |\mathcal{L}_{\mathbf{z}}|$ satisfies

$$\begin{aligned}
\sum_{\mathbf{z} \in Z(A)} |\mathcal{L}_{\mathbf{z}}| &\leq \sum_{k=2}^{N} \binom{N}{k} \left(2^\ell \eta\right)^k \\
&= \sum_{k=0}^{N} \binom{N}{k} \left(2^\ell \eta\right)^k - 2^\ell N \eta - 1 \\
&= (1 + 2^\ell \eta)^N - 2^\ell N \eta - 1 \\
&\leq e^{2^\ell N \eta} - 2^\ell N \eta - 1 \\
&\leq 1 + 2^\ell N \eta + (2^\ell N \eta)^2 - 2^\ell N \eta - 1 && (\text{with } \eta \leq \tfrac{1}{2^\ell N}) \\
&\leq 4^\ell N^2 \eta^2.
\end{aligned}$$

The lemma stands proved. □

Lemma 5.10. *For estimates obtained via a covering intervention set $\mathcal{I}$, the following event holds with probability at least $1-\frac{2}{T}$:*

$$\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \leq \sqrt{\frac{N\ 4^{\ell}\ 2^{d}\ |\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}} \qquad \textit{for all } A\in\mathcal{A}.$$

Proof. Equation (5.12) gives us

$$\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| = \left|\sum_{C_i\in\mathcal{C}(A)} \sum_{\mathbf{z}\in Z(A)} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) \prod_{j\in\mathcal{V}(A)\setminus C_i} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)\right|.$$

Let $X := \{x_1, x_2 \cdots x_m\}$ be the vertices in a c-component C_i considered in the outer summation. Furthermore, for ease of exposition, write $(x_k : x_{k+1})' := (x_k : x_{k+1}) \setminus \mathrm{Pa}(C_i)$, i.e., the set $(x_k : x_{k+1})'$ excludes the parents of the c-component C_i. We have

$$\begin{aligned}
&\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \\
&= \Bigg|\sum_{C_i\in\mathcal{C}(A)} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(C_i)}\in\\ Z_{\mathrm{Pa}(C_i)}(A)}} \sum_{\substack{\mathbf{z}_{C_i}\in\\ Z_{C_i}(A)}} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) \sum_{\substack{\mathbf{z}_{[1:x_1)'}\in\\ Z_{[1:x_1)'}(A)}} \prod_{j_1\in\mathcal{V}_{[1:x_1)}(A)} \mathcal{P}\left(\mathbf{z}_{j_1} \mid \mathbf{z}_{\mathrm{Pa}'(j_1)}\right) \\
&\quad \sum_{\substack{\mathbf{z}_{(x_1:x_2)'}\in\\ Z_{(x_1:x_2)'}(A)}} \prod_{j_2\in\mathcal{V}_{(x_1:x_2)}(A)} \mathcal{P}_A\left(\mathbf{z}_{j_2} \mid \mathbf{z}_{\mathrm{Pa}'(j_2)}\right) \cdots \sum_{\substack{\mathbf{z}_{(x_k:x_{k+1})'}\in\\ Z_{(x_k:x_{k+1})'}(A)}} \prod_{j_2\in\mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}'(j_k)}\right) \cdots\Bigg| \\
&= \left|\sum_{C_i\in\mathcal{C}(A)} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(C_i)}\in\\ Z_{\mathrm{Pa}(C_i)}(A)}} \sum_{\substack{\mathbf{z}_{C_i}\in\\ Z_{C_i}(A)}} \Delta\mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right) c_i\left(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right|.
\end{aligned}$$

Here,

$$\begin{aligned}
&c_i(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) := \\
&\sum_{\substack{\mathbf{z}_{[1:x_1)'}\in\\ Z_{[1:x_1)'}(A)}} \prod_{j_1\in\mathcal{V}_{[1:x_1)}(A)} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j_1)}\right) \sum_{\substack{\mathbf{z}_{(x_1:x_2)'}\in\\ Z_{(x_1:x_2)'}(A)}} \prod_{j_2\in\mathcal{V}_{(x_1:x_2)}(A)} \mathcal{P}_A\left(\mathbf{z}_{j_2} \mid \mathbf{z}_{\mathrm{Pa}'(j_2)}\right) \cdots
\end{aligned}$$

$$\sum_{\substack{\mathbf{z}_{(x_k:x_k+1)'}\in\\ Z_{(x_k:x_{k+1})'}(A)}} \prod_{j_k\in\mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}'(j_k)}\right)\cdots \sum_{\substack{\mathbf{z}_{(x_m:N]'}\in\\ Z_{(x_m:N]'}(A)}} \prod_{j_m\in\mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A\left(\mathbf{z}_{j_m} \mid \mathbf{z}_{\mathrm{Pa}'(j_m)}\right).$$

We show in Claim 5.2 (proved below) that $c_i(z_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) \leq 1$. Therefore,

$$\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \leq \left|\sum_{C_i\in\mathcal{C}(A)} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(C_i)}\in\\ Z_{\mathrm{Pa}(C_i)}(A)}} \sum_{\substack{\mathbf{z}_{C_i}\in\\ Z_{C_i}(A)}} \Delta\mathcal{P}\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right)\right| \tag{5.15}$$

Since $\mathcal{I}$ is a covering intervention set, for each pair $(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)})$, there exits an intervention $I \in \mathcal{I}$ such that intervening $\mathrm{do}(I)$ provides a sample for the distribution $\mathbb{P}[\mathcal{V}_{C_i} \mid \mathrm{do}(\mathrm{Pa}(C_i) = \mathbf{z}_{\mathrm{Pa}(C_i)})]$. Hence, we have at least $\frac{T}{|\mathcal{I}|}$ samples for the distribution $\mathbb{P}[\mathcal{V}_{C_i} \mid \mathrm{do}(\mathrm{Pa}(C_i) = \mathbf{z}_{\mathrm{Pa}(C_i)})]$. We represent the s^{th} sample for the distribution by indicator random variable $Y_s(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)})$ which takes value one when $\mathcal{V}_{C_i} = \mathbf{z}_{C_i}$, else its zero. Hence, inequality (5.15) reduces to

$$\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \leq \left|\sum_{C_i\in\mathcal{V}(A)} \sum_{\substack{\mathbf{z}_{\mathrm{Pa}(C_i)}\in\\ Z_{\mathrm{Pa}(C_i)}(A)}} \frac{|\mathcal{I}|}{T}\sum_{s=1}^{T/|\mathcal{I}|}\left(\sum_{\mathbf{z}_{C_i}\in Z_{C_i}(A)} Y_s(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) - \mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)\right)\right|.$$

In the above expression, the term $\sum_{\mathbf{z}_{C_i}\in Z_{C_i}(A)} Y_s(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) - \mathcal{P}_A\left(\mathbf{z}_{C_i} \mid \mathbf{z}_{\mathrm{Pa}(C_i)}\right)$ is an independent random quantity bounded between $[-2^{|C_i|}, 2^{|C_i|}]$. We now apply Heoffding's inequality (Lemma 5.2)

$$\begin{aligned}\mathbb{P}_{\mathrm{do}(A)}\left[\left|\sum_{\mathbf{z}\in Z(A)} \mathcal{H}_{\mathbf{z}}\right| \geq \varepsilon\right] &\leq 2\mathrm{exp}\left(\frac{-T\varepsilon^2}{2|\mathcal{I}|\sum_{C_i\in\mathcal{C}(A)}\sum_{\mathbf{z}_{\mathrm{Pa}(i)}\in\mathbf{z}_{\mathrm{Pa}(i)}} 2^{2|C_i|}}\right)\\ &\leq 2\mathrm{exp}\left(\frac{-T\varepsilon^2}{2|\mathcal{I}|\sum_{C_i\in\mathcal{C}(A)}\sum_{\mathbf{z}_{\mathrm{Pa}(i)}\in\mathbf{z}_{\mathrm{Pa}(i)}} 2^{2\ell}}\right) \leq 2\mathrm{exp}\left(\frac{-T\varepsilon^2}{2|\mathcal{I}|N2^{\ell d}\cdot 2^{2\ell}}\right).\end{aligned}$$

Setting $\varepsilon = \sqrt{\frac{2N\ |\mathcal{I}|\ 2^{\ell d}\ 4^{\ell}\log(|\mathcal{A}|\cdot T)}{T}}$ and taking union bound over all of $A \in \mathcal{A}$, gives us the required probability bound. This completes the proof of the lemma. □

We next establish the claim used in the proof of Lemma 5.10.

Claim 5.2.

$$c_i(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) \leq 1.$$

Proof. It holds that

$$
\begin{aligned}
&c_i(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}) = \\
&\sum_{\substack{\mathbf{z}_{[1:x_1)'} \in \\ Z_{[1:x_1)'}(A)}} \prod_{j_1 \in \mathcal{V}_{[1:x_1)}(A)} \mathcal{P}_A \left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right) \sum_{\substack{\mathbf{z}_{(x_1:x_2)'} \in \\ Z_{(x_1:x_2)'}(A)}} \prod_{j_2 \in \mathcal{V}_{(x_1:x_2)}(A)} \mathcal{P}_A \left(\mathbf{z}_{j_2} \mid \mathbf{z}_{\mathrm{Pa}'(j_2)}\right) \cdots \\
&\sum_{\substack{\mathbf{z}_{(x_k:x_k+1)'} \in \\ Z_{(x_k:x_{k+1})'}(A)}} \prod_{j_k \in \mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A \left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}'(j_k)}\right) \cdots \sum_{\substack{\mathbf{z}_{(x_m:N]'} \in \\ Z_{(x_m:N]'}(A)}} \prod_{j_k \in \mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A \left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}'(j_k)}\right)
\end{aligned}
$$

We can upper bound each term in the above expression as shown below,

$$
\begin{aligned}
&\sum_{\substack{\mathbf{z}_{(x_k:x_k+1)'} \in \\ Z_{(x_k:x_{k+1})'}(A)}} \prod_{j_k \in \mathcal{V}_{(x_k:x_{k+1})}(A)} \mathcal{P}_A \left(\mathbf{z}_{j_k} \mid \mathbf{z}_{\mathrm{Pa}'(j_k)}\right) \\
&= \sum_{\substack{\mathbf{z}_{(x_k:x_k+1)'} \in \\ Z_{(x_k:x_{k+1})}(A)}} \mathbb{P}_{\mathrm{do}(A)} \left[\mathcal{V}_{(x_k:x_{k+1})}(A) = \mathbf{z}_{(x_k:x_k+1)} | \mathrm{Pa}'(x_k : x_{k+1})\right] \\
&= \mathbb{P}_{\mathrm{do}(A)} \left[\mathcal{V}_{(x_k:x_{k+1}) \cap \mathrm{Pa}(C_i)}(A) = \mathbf{z}_{(x_k:x_k+1) \cap \mathrm{Pa}(C_i)} | \mathrm{Pa}'(x_k : x_{k+1})\right] \\
&\leq 1.
\end{aligned}
$$

Substituting this in the expression for $c_i(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)})$, we get the required bound. □

5.5.2 Proof of Theorem 5.2

Lemma 5.6 implies that, with probability at least $\left(1 - \frac{1}{T}\right)$, the set $\mathcal{I}$ is indeed a covering intervention set for the graph $\mathcal{G}$. We combine this guarantee with Lemmas 5.9 and 5.10. In particular, with probability at least $\left(1 - \frac{5}{T}\right)$, we have, for all $A \in \mathcal{A}$:

$$
\begin{aligned}
|\mu(A) - \widehat{\mu}(A)| &= \left| \sum_{\mathbf{z} \in Z(A)} (\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}}) \right| \\
&\leq \sqrt{\frac{N\ 4^\ell\ 2^d\ |\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}} + \frac{4^\ell N^2 |\mathcal{I}| (\ell d + \ell + \log\left(NT\right))}{T} \\
&\leq 2\sqrt{\frac{N\ 4^\ell\ 2^d\ |\mathcal{I}| \log(|\mathcal{A}|T)}{T}} \qquad \text{(For } T \gtrsim N^3\text{)}
\end{aligned}
$$

Let A_T be the output after T rounds of interventions, i.e., $A_T = \mathrm{argmax}_{A \in \mathcal{A}}\ \widehat{\mu}(A)$. In addition, let $A^* = \mathrm{argmax}_{A \in \mathcal{A}}\ \mu(A)$ be the optimal intervention. Hence, with probability at least $1 - \frac{5}{T}$

we have,

$$\mu(A^*) - \mu(A_T) \leq 4\sqrt{\frac{N\ 4^{\ell}\ 2^{d}\ |\mathfrak{I}| \log(|\mathcal{A}|T)}{T}} \tag{5.16}$$

This gives the desired upper bound on the simple regret, R_T:

$$R_T = \mathbb{E}\left[\mu(A^*) - \mu(A_T)\right] \leq \left(4\sqrt{\frac{N\ 4^{\ell}\ 2^{d}\ |\mathfrak{I}| \log(|\mathcal{A}|T)}{T}}\right)\left(1 - \frac{5}{T}\right) + \frac{5}{T} \leq 5\sqrt{\frac{N\ 4^{\ell}\ 2^{d}\ |\mathfrak{I}| \log(|\mathcal{A}|T)}{T}}.$$

For SMBNs, since the size of the covering intervention set satisfies $|\mathfrak{I}| = (3d)^{\ell} \cdot 2^{\ell d}(\log N + 2\ell d + \log T)$ (see Lemma 5.6), we also have the following explicit form of the simple regret bound

$$\mathrm{R}_T = O\left(\sqrt{\frac{N\ (3d\ 8^{d})^{\ell}\ \log|\mathcal{A}|}{T}} \log T\right).$$

The theorem stands proved.

5.6 Missing Proof from Section 5.5.1

This section provides a proof of Lemma 5.7.

Lemma 5.7. *For any intervention A and any c-component $C \in \mathfrak{C}(A)$, consisting of vertices $\{U_1, U_2 \ldots U_m\}$, we have*

$$\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C)}}\left(\mathbf{z}_C\right) = \prod_{j \in C} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right).$$

Here, $\mathrm{Pa}'(j)$ *denotes the set of pseudo parents as defined in equation (5.11).*

Proof. First, note that intervening on parent vertices of a c-component (under intervention A) is the same as conditioning on them. Specifically,

$$\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C)}}\left(\mathbf{z}_C\right) = \mathcal{P}_A\left(\mathbf{z}_C \mid \mathbf{z}_{\mathrm{Pa}(C)}\right)$$

Further, the chain rule of conditional probability gives us

$$\mathcal{P}_A\left(\mathbf{z}_C \mid \mathbf{z}_{\mathrm{Pa}(C)}\right) = \prod_{j \in C} \mathbb{P}_{\mathrm{do}(A)}\left[U_j = \mathbf{z}_j \mid \mathrm{Pa}(C) = \mathbf{z}_{\mathrm{Pa}(C)}, (U_1 \ldots U_{j-1}) = \mathbf{z}_{(U_1 \ldots U_{j-1})}\right]$$

Next, we use the notion of d-separation (see [Pea09]) to argue that conditioning on just the set $\mathrm{Pa}'(j)$ is sufficient. In particular, note that the set $Y = \mathrm{Pa}(\{U_{j+1} \ldots U_m\})$ is d-separated from vertex U_j by the set $X = \mathrm{Pa}(\{U_1 \ldots U_j\}) \cup (\{U_1 \ldots U_{j-1}\})$. This is due to the fact that all paths from a vertex in Y to U_j are either blocked by a collider vertex in $\{U_{j+1} \ldots U_m\}$ (and the collider vertex is not included X), or the path is blocked by a vertex in X. This implies that conditioned on X, U_j is independent of all vertices in Y [Pea09]. Formally, we write

$$\begin{aligned}
&\mathbb{P}_{\mathrm{do}(A)} \left[U_j = \mathbf{z}_j \mid \mathrm{Pa}(C) = \mathbf{z}_{\mathrm{Pa}(C)}, (U_1 \ldots U_{j-1}) = \mathbf{z}_{(U_1 \ldots U_{j-1})}\right] \\
&= \mathbb{P}_{\mathrm{do}(A)}[U_j = \mathbf{z}_j \mid \mathrm{Pa}(U_1 \ldots U_{j-1}) = \mathbf{z}_{\mathrm{Pa}(U_1 \ldots U_j)}, \mathrm{Pa}(U_{j+1} \ldots U_m) = \mathbf{z}_{\mathrm{Pa}(U_{j+1} \ldots U_m)}, (U_1 \ldots U_{j-1}) = \mathbf{z}_{(U_1 \ldots U_{j-1})} \\
&= \mathbb{P}_{\mathrm{do}(A)} \left[U_j = \mathbf{z}_j \mid \mathrm{Pa}(U_1 \ldots U_{j-1}) = \mathbf{z}_{\mathrm{Pa}(U_1 \ldots U_j)}, (U_1 \ldots U_{j-1}) = \mathbf{z}_{(U_1 \ldots U_{j-1})}\right] \\
&\qquad\qquad \text{(since } \mathrm{Pa}(\{U_1 \ldots U_j\}) \cup \{U_1 \ldots U_{j-1}\} \text{ d-separates } U_j \text{ from } \mathrm{Pa}(\{U_{j+1} \ldots U_m\})) \\
&= \mathbb{P}_{\mathrm{do}(A)} [U_j = \mathbf{z}_j \mid \mathrm{Pa}'(j)] \qquad\qquad \text{(by definition of } \mathrm{Pa}'(j))
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C)}}(\mathbf{z}_C) &= \mathcal{P}_A\left(\mathbf{z}_C \mid \mathbf{z}_{\mathrm{Pa}(C)}\right) \\
&= \prod_{j \in C} \mathbb{P}_{\mathrm{do}(A)}\left[V_j = \mathbf{z}_j \mid \mathrm{Pa}'(j) = \mathbf{z}_{\mathrm{Pa}'(j)}\right] \\
&= \prod_{j \in C} \mathcal{P}_A\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}'(j)}\right)
\end{aligned}$$

This completes the proof of the lemma. □

5.7 Experiments

This section provides empirical evaluations of our algorithm. In the experiments, we compare our algorithm, CoveringInterventions (Algorithm 9) with PropInf, the propagating inference algorithm of Yabe et al. [YHS$^+$18]. As in implementation of [YHS$^+$18] (see Section 5 of the cited paper), we uniformly sample and do not explicitly solve their proposed optimization problem. The source code of our implementations is available at https://github.com/sawarniayush/learning-good-interventions-using-covering

For the experiments, we consider a causal graph $\mathcal{G} = (\mathcal{V}, E)$ (over Bernoulli random variables) with number of nodes (variables) $N = |\mathcal{V}| = 17$ and in-degree $d = 4$. The vertex set $\mathcal{V}$ is partitioned into four subsets with cardinalities $|\mathcal{V}_1| = 7$, $|\mathcal{V}_2| = 5$, $|\mathcal{V}_3| = 4$, and $|\mathcal{V}_4| = 1$, respectively. Here, the singleton $\mathcal{V}_4$ consists of the reward variable, which is connected to all the 4 nodes in $\mathcal{V}_3$. Furthermore, the graph $\mathcal{G}$ is layered in the sense that, for each index $\ell \in \{2, 3, 4\}$ and each node $V_i \in \mathcal{V}_\ell$, the parents $\mathrm{Pa}(i) \subset \mathcal{V}_{\ell-1}$. Also, $\mathcal{V}_1$ is the set of leaf vertices – the vertices in $\mathcal{V}_1$ do not have any incoming edges.

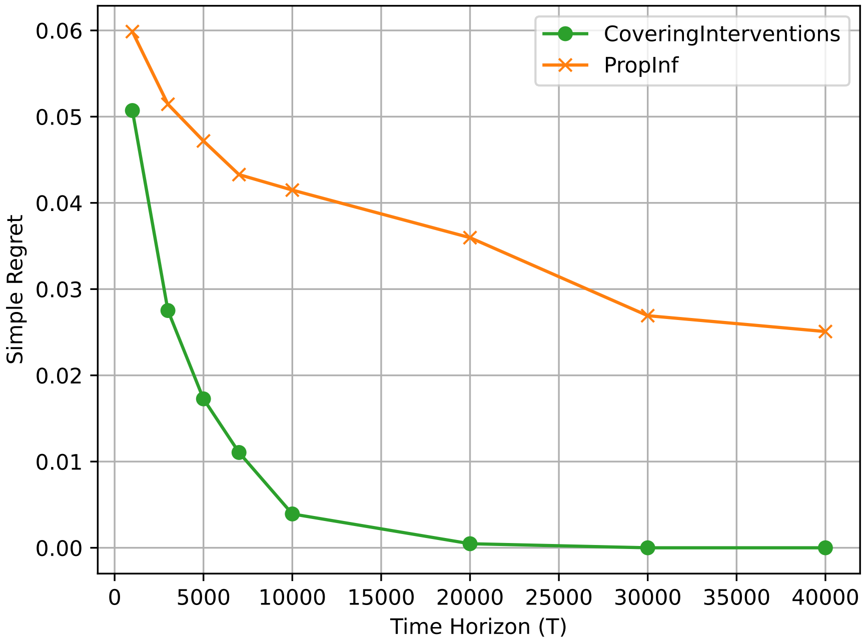


Figure 5.1: figure
Plot of simple regret with rounds of exploration.

For each non-reward variable, V_i, we set the condition probability $\mathbb{P}\{V_i = 1 \mid \text{Pa}(i) = \mathbf{1}\} = 0.8$. That is, when all the parents of V_i are equal to 1,[1] then $V_i = 1$, with probability 0.8. For any other realization of the parents, the conditional probability of $V_i = 1$ is set to be 0.4, i.e., $\mathbb{P}\{V_i = 1 \mid \text{Pa}(i) \neq \mathbf{1}\} = 0.4$. For the reward node V_{17} we have $\mathbb{P}\{V_{17} = 1 \mid \text{Pa}(17) = \mathbf{1}\} = 0.9$ and $\mathbb{P}\{V_{17} = 1 \mid \text{Pa}(17) \neq \mathbf{1}\} = 0.4$.

The set of interventions $\mathcal{A}$ is composed of all possible interventions on the leaf nodes, $\mathcal{A} = \{\text{do}(\mathcal{V}_1 = s) \mid s \in \{0,1\}^7\}$; recall that $|\mathcal{V}_1| = 7$. Note that setting each leaf node to 1 yields the optimal intervention $A^* = \text{do}(\mathcal{V}_1 = \mathbf{1})$.

Simple Regret vs. Time: In our experiments, for the two algorithms, we compare the simple regret with time horizon T. In particular, for each relevant T, we execute the two algorithms 140 times and average the simple regret across these runs. We plot our results in Figure 5.1 and show that CoveringInterventions converges to low regret faster than PropInf.

Runtime: For this experimental setup, CoveringInterventions ran at least 8 times faster

[1] Recall that intervening on all parent nodes of a vertex is the same as conditioning on them.

than PropInf across all the executions.[1] This runtime gap between the two implementations, highlights that CoveringInterventions scales better with the number of variables N.

5.8 Conclusion and Future Work

Using the idea of covering interventions, this chapter obtains improved simple regret guarantees for the causal bandit problem. We also generalize the guarantee to causal graphs with unobserved variables. Notably, and in contrast to prior works, our regret guarantees only depend on the explicit problem parameters. Our experiments empirically highlight that our algorithm provides improvements over baselines. Establishing lower bounds in the general causal bandit setup is an important direction of future work. It is also interesting to develop computationally efficient (simple regret) algorithms for settings in which the target set $\mathcal{A}$ is large and implicitly specified.

[1] The computation of the β parameters is a time consuming step in PropInf.

Bibliography

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. *Advances in Neural Information Processing Systems*, 31, 2018. 96, 98, 100, 101, 102, 111

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002. 9

[AG13] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013. 63

[AJK21] Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021. 48

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017. 97

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011. 85

[BBLB20a] Ilai Bistritz, Tavor Baharav, Amir Leshem, and Nicholas Bambos. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *International Conference on Machine Learning*, pages 930–940. PMLR, 2020. 10

[BBLB20b] Ilai Bistritz, Tavor Baharav, Amir Leshem, and Nicholas Bambos. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *International Conference on Machine Learning*, pages 930–940. PMLR, 2020. 49

[BCS14] Guy Bresler, George H Chen, and Devavrat Shah. A latent source model for online collaborative filtering. *Advances in neural information processing systems*, 27, 2014. 2

[BKMS22] Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. Fairness and welfare quantification for regret in multi-armed bandits. *arXiv preprint arXiv:2205.13930*, 2022. 53, 56

[BPQC$^+$13] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013. 4

[BS10] Dirk Bergemann and Maher Said. Dynamic auctions: A survey. 2010. 2

[CCLY19] Michael B Cohen, Ben Cousins, Yin Tat Lee, and Xin Yang. A near-optimal algorithm for approximating the john ellipsoid. In *Conference on Learning Theory*, pages 849–873. PMLR, 2019. 60

[CKD$^+$15] Pascal Caillet, Sarah Klemm, Michel Ducher, Alexandre Aussem, and Anne-Marie Schott. Hip fracture in the elderly: a re-analysis of the epidos study with causal bayesian networks. *PLoS One*, 10(3):e0120125, 2015. 4

[CKM$^+$19] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3):1–32, 2019. 10, 49

[CKSV19] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 160–169, 2019. 10

[DB15] Arnoud V Den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015. 1

[DP09] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009. 33

[Eck93] Jürgen Eckhoff. Helly, radon, and carathéodory type theorems. In *Handbook of convex geometry*, pages 389–448. Elsevier, 1993. 54

[EG59] Edmund Eisenberg and David Gale. Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics*, 30(1):165–168, 1959. 10

[Git79] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979. 1

[GLS12] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012. 52

[H$^+$16] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. 92, 93

[HMS21a] Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34, 2021. 10

[HMS21b] Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017, 2021. 49

[How97] Ralph Howard. The john ellipsoid theorem. *University of South Carolina*, 1997. 48

[JKMR16a] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016. 10

[JKMR16b] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016. 49

[KEG17] Daniel Koch, Robert S Eisinger, and Alexander Gebharter. A causal bayesian network model of disease progression mechanisms in chronic myeloid leukemia. *Journal of theoretical biology*, 433:94–105, 2017. 4

[KN79] Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435, 1979. 10

[KQ21] William Kuszmaul and Qi Qi. The multiplicative version of azuma's inequality, with an application to contention analysis. *arXiv preprint arXiv:2102.05077*, 2021. 56

[KW60] Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. 51

[LKG16] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016. 2

[LLR16] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016. 4, 97

[LMT21] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34:24817–24828, 2021. 97

[LMT22] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient reinforcement learning with prior causal knowledge. In *Conference on Causal Learning and Reasoning*, pages 526–541. PMLR, 2022. 97

[LMTY20] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020. 97

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. 2, 9, 12, 13, 19, 42, 48, 51, 56, 60, 80

[LWB$^+$18] Sujee Lee, Sijie Wang, Philip A Bain, Christine Baker, Tammy Kundinger, Craig Sommers, and Jingshan Li. Reducing copd readmissions: A causal bayesian network model. *IEEE Robotics and Automation Letters*, 3(4):4046–4053, 2018. 4

[LZ07] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007. 2

[MNS22] 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*, pages 1328–1338. PMLR, 2022. 4, 96, 97

[Mou04] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004. 2, 3, 10, 46

[MY16] Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016. 48

[NJ50] John F Nash Jr. The bargaining problem. *Econometrica: Journal of the econometric society*, pages 155–162, 1950. 10

[NPS21] Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2021. 97

[PACJ07] Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 216–227. SIAM, 2007. 2

[Pea00] Judea Pearl. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000. 96

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009. 4, 120, 121

[PGNN20] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5379–5386, 2020. 10

[PGNN21] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research*, 22(1):7885–7915, 2021. 49

[RKJ08] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791, 2008. 2

[S$^+$19] Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. 2

[SBF17] Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017. 2, 10

[Sev20] Jaime Sevilla. Explaining data using causal bayesian networks. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 34–38, 2020. 4

[SSDS17] Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pages 3057–3066. PMLR, 2017. 97

[SSK$^+$17] Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017. 97

[SV14] Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *ACM SIGecom Exchanges*, 12(2):4–23, 2014. 2

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933. 1

[TM17] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517, 2017. 1

[Tod16] Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016. 60

[TP02] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527, 2002. 110

[Var74] Hal R Varian. Equity, envy, and efficiency. *Journal of Economic Theory*, 9(1):63–91, 1974. 10

[VB22] Yogatheesan Varatharajah and Brent Berry. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022. 1

[VBW15] Sofia Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30:199–215, 05 2015. 1

[VSST22] Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *arXiv preprint arXiv:2208.12764*, 2022. 97

[XC23] Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. In *International Conference on Learning Representations*, 2023. 96, 97

[YHS$^+$18] Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *International Conference on Machine Learning*, pages 5512–5520. PMLR, 2018. 96, 121

[YN12] Takami Yoshida and Kazuhiro Nakadai. Active audio-visual integration for voice activity detection based on a causal bayesian network. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 370–375. IEEE, 2012. 4