ALGORITHMS AND THEORY OF BANDITS IN PRACTICAL SETUPS

by

Yao Zhao

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF COMPUTER SCIENCE

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2025

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Yao Zhao, titled: Algorithms and theory of bandits in Practical setups

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Kwang-Sung Jun (Oct 1, 2025 06:10:14 GMT+9)
Dr. Kwang-Sung Jun — Date: 10/01/2025

Dr. Jason Pacheco — Date: 09/29/2025

Chicheng Zhang (Sep 30, 2025 10:28:33 PDT)
Dr. Chicheng Zhang — Date: 09/30/2025

Dr. Ming Li — Date: 09/30/2025

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Kwang-Sung Jun (Oct 1, 2025 06:10:14 GMT+9)
Dr. Kwang-Sung Jun — Date: 10/01/2025
Dissertation Committee Chair
Computer Science Department

ACKNOWLEDGMENTS

Writing this section is one of the most joyful moments as I wrap up my PhD journey. I would like to express my deepest gratitude to my advisor, Kwang-Sung Jun. It is because of him that I am able to complete a PhD journey I am satisfied with.

I still remember five years ago when I first contacted Kwang during the PhD application process. He responded to me very quickly, and his support never stopped since then. He was always transparent about the process, and I accepted the offer without hesitation. Looking back, I feel very happy that I made the right choice, even over offers from higher-ranked universities.

When COVID hit unexpectedly, I was not able to arrive in Tucson on time and missed my first semester. But Kwang was very supportive. He started remote meetings with me to get me ramped up on research. I still remember how he would connect his iPad, share the screen, and write down math step by step to explain things to me, or point to a paper and walk me through a concept. At that time, we didn't know if I would ever make it to Tucson or when the pandemic would end. I was truly moved by the patience and kindness of someone who, back then, I had never even met in person.

Looking back at the past five years with all the ups and downs, I appreciate Kwang's careful guidance in my research. He had an incredible balance knowing when to step in and help me directly and when to let me figure things out by myself. That was more than research training, it was the best way to grow. I also appreciate his patience. Even when progress was slow, he never rushed me, and somehow the results would come together. Over time, I felt the trust he had in me, which meant a lot.

Kwang is not only a great advisor but also a great mentor. He always encouraged me to pursue internships, and I am grateful for the opportunities I had. I learned not just research, but also important skills such as communication. I still remember the time we

defined meeting protocols. Only after using them did I realize how valuable that was. Thanks to his guidance, I have developed habits that help me enjoy communicating and connecting with people.

Hats off to my advisor, Kwang.

I am also grateful to my external collaborators, who gave me many opportunities to learn, grow, and collaborate. Csaba Szepesvári, with his wisdom and stature in the field, still took time to work patiently with me early on. Connor Stephens was always full of energy and ready for discussions. Tanner Fiez was incredibly detail-oriented, I still remember him pointing out a small constant I had overlooked in a long derivation. Lalit Jain took every opportunity to teach me something, from hands-on coding to small concepts in textbooks.

I also thank my first roommate, Tianchi Zhao. He picked me up from the airport the night I arrived and introduced me to Tucson. He was not only a great roommate but also a great friend, and I am grateful for the time we lived together. I thank Zhengzhong Liang for being a close friend since my very first days in Tucson. I lost count of how many dinners we shared before he left for his next step. He is always my go-to person whenever I have questions.

Paulo Soares and I shared a windowless office for a long time when I first arrived in Tucson, and later we became colleagues again during my internship. I really appreciate his help in my career.

I also want to acknowledge my life in Tucson, the place I called home for the past five years. At first, I complained it was the smallest city I had ever lived in, but now I feel sad to leave. I am grateful for the life I built here, not just research, but also everyday life and friendships, even outside of UA. My neighbor's family were especially kind, inviting me for meals every week and making me feel warmly at home. It has been a fixed schedule that every Saturday afternoon, we spent hours talking about life and everything. Those moments were precious to me.

Lastly, I thank my parents, who have always fully trusted me and supported every decision I made.

LAND ACKNOWLEDGMENT

We respectfully acknowledge the University of Arizona is on the land and territories of Indigenous peoples. Today, Arizona is home to 22 federally recognized tribes, with Tucson being home to the O'odham and the Yaqui. Committed to diversity and inclusion, the University strives to build sustainable relationships with sovereign Native Nations and Indigenous communities through education offerings, partnerships, and community service.

DEDICATION

Dedicated to my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Multi-armed bandit is a central paradigm for interactive machine learning, where an agent must learn efficiently from adaptive feedback. From a data collection point of view, bandit algorithms can be applied as a tool for efficient sequential data collection and serve as a powerful method to solve many real-world problems. This perspective leads to the pure exploration problem, where the goal is to focus entirely on data exploration during the learning process and use the collected data for a final task, such as decision-making or model training. A key aspect of this framework is that individual data collection steps do not have an immediate impact on a final objective. This distinguishes the pure exploration framework, which has recently shown its strength in many practical problems, from the more traditional regret minimization framework with a cumulative objective. In this dissertation, three problems are studied around this theme, ranging from theoretical foundations to practical applications.

The first problem addresses a gap in the theoretical understanding of simple regret, a fundamental yet under-explored performance measure in pure exploration; since its proposal, a rigorous, instance-dependent characterization has been lacking. The second problem arises in experimental settings, such as clinical trials, where bandits are a powerful tool for accelerating discovery. However, factors like user non-compliance can confound the experimental results. We address the question of how to design efficient pure exploration algorithms to support critical decision-making in the presence of such confounding factors. The third problem concerns the alignment of Large Language Models (LLMs), which often requires a high-quality preference dataset that is expensive to collect. We study how to design efficient data exploration strategies to improve the sample efficiency of preference learning.

These three problems share the common theme of efficient data collection for a final task, where the immediate impact of each data collection step is disregarded in favor of the terminal objective. This dissertation tackles these three problems by developing a suite of novel

algorithms and analyses to address each of them. A common goal across all contributions is the provision of an instance-dependent guarantee for each problem. This is more practical and fine-grained than the worst-case bounds commonly seen in the literature, providing a more complete, problem instance level characterization that reflects the intrinsic difficulty of each specific task.

Chapter 1

Introduction

1.1 Background and Motivation

Interactive machine learning, in stark contrast to traditional supervised learning, is a paradigm where a learning agent interacts with an environment via an algorithm that repeatedly and adaptively takes actions and collects feedback. Such an environment can be a human-in-the-loop system (e.g., a recommender system) or a physical system (e.g., robotics). Often, the interaction is aimed at collecting valuable, frequently expert, human feedback through a sequential decision-making process. The feedback can take various forms, such as explicit labels, error corrections, result validations, or rankings. The downstream tasks range from collecting insightful data to inform a critical decision to curating a high-quality training dataset for other machine learning models.

A representative model for interactive machine learning is the multi-armed bandit (MAB) (Lattimore and Szepesvári, 2020), which has a long history in the statistics and machine learning literature, with its original formulation dating back to Thompson (1933). The core concept of the MAB model is bandit feedback, which means the learning agent only observes the feedback signal for the action it takes, not for the actions it could have taken. This feedback mechanism is common in real-world applications. For example, in a recommender system, the system only observes user feedback for the recommended items (Li et al., 2010); in a clinical trial, a doctor only observes the outcome for the prescribed treatment (Villar et al., 2015). While the bandit feedback mechanism makes learning more challenging, it also makes the model more realistic and practical. This immediately raises the question of how to design an algorithm that can effectively learn from bandit feedback or, from a different perspective, how to design an efficient data collection process under this mechanism.

Extensive research has been conducted on the multi-armed bandit model, leading to a rich set of algorithms that address various challenges from both theoretical and practical perspectives. For a long while, bandits were studied under the framework of regret minimization (Auer et al., 2002), which aims to minimize the cumulative regret over a time horizon. This framework captures the fundamental trade-off between exploration and exploitation: the agent must explore different actions to gather information about the environment while simultaneously exploiting its current knowledge to maximize immediate rewards. The regret minimization framework has been widely studied in various settings, including stochastic bandits (Auer et al., 2002), contextual bandits (Foster and Rakhlin, 2020), linear bandits (Abbasi-Yadkori et al., 2011), and combinatorial bandits (Combes et al., 2015). More recently, the concept of pure exploration was proposed (Even-Dar et al., 2006; Bubeck et al., 2009), motivated by the fact that in many applications, the agent is more interested in identifying the best action than in maximizing cumulative reward. For example, in an A/B testing scenario, the goal is to identify the best-performing variant, not to maximize the conversion rate during the test itself. Similarly, in hyperparameter tuning, the goal is to find the best configuration for future use (Li et al., 2018b). The pure exploration framework has been studied in settings such as best-arm identification (Even-Dar et al., 2006), top-k arm identification (Chen et al., 2017a), and in more complex structures like linear bandits (Soare et al., 2014).

Building upon this line of research, this dissertation aims to address a set of questions ranging from theoretical foundations to practical applications, including:

Theoretical foundation:

In the pure exploration setting, existing work has primarily focused on developing algorithms that efficiently identify the best arm (or top-k arms) with high probability while minimizing the required number of samples. Much of this research centers on bounding the probability of error, that is, the probability of failing to identify the correct best arm.

However, a fundamental understanding of the limits of pure exploration remains incomplete. As noted as an open problem in Lattimore and Szepesvári (2020), the full output distribution of a pure exploration algorithm is still unknown. The probability of error provides only a partial characterization of this distribution, and its prevalence in the literature

is largely due to its analytical tractability. A more fundamental performance measure is simple regret, first introduced by Bubeck et al. (2009). Despite its importance, obtaining instance-dependent bounds for simple regret remains a significant open problem in the field.

Practical application:

Two emerging topics in real-world applications are strongly tied to the bandit problem, yet they present unique challenges that standard, off-the-shelf algorithms have not yet successfully addressed.

One such area is the use of bandits as an experimentation tool in the presence of non-compliance, a topic of growing importance in industry. While practitioners have increasingly used bandit algorithms to adaptively allocate treatments to identify the best one, a common issue in real-world settings is that the assigned treatment is not always the one applied. This is known as non-compliance. For example, in a clinical trial, a volunteer assigned a specific drug may choose not to take it or to take a different one. Similarly, in an online platform like Netflix, a user might be offered a coupon for a specific subscription tier to gauge engagement, but they may choose to subscribe to a different plan instead. Such scenarios complicate the experimental goal of identifying the best treatment. A key research question is therefore how to design efficient bandit pure exploration solutions for experimentation settings with non-compliance.

Another challenge emerging in recent years is the issue of data collection efficiency for Large Language Model (LLM) alignment. This seemingly less relevant area is, in fact, strongly tied to the bandit problem. A critical step in LLM alignment is collecting human feedback on different LLM-generated answers, often in the form of pairwise comparisons. These comparisons are used to train a reward model for downstream reinforcement learning step. However, collecting human feedback is expensive and time-consuming. Given the need for high-quality data, it is crucial to design an efficient collection process. This task can be naturally formulated as a bandit problem where each arm is a pair of answers to be compared. The challenge here is how to make a minimal number of queries to humans while still collecting a high-quality dataset to train an effective reward model quickly. Due to the inherent bandit feedback in this problem, where we only observe the preference for the pair

that is queried, this topic has attracted significant attention, but an established, optimal solution has not yet been proposed.

1.2 Bandit Model Structure

We briefly review the multi-armed bandit (MAB) model and its notation in this section for completeness.

In a standard MAB model, a learning agent interacts with a set of N arms. Each arm $i \in \{1, \ldots, N\}$ is associated with an unknown reward distribution μ_i. We consider a slotted time system where, at each time step $t = 1, 2, \ldots$, the agent selects an arm $I_t \in \{1, \ldots, N\}$ and receives a reward r_t drawn from the distribution of arm I_t. We assume that reward distributions are independent across arms and that rewards are independent and identically distributed over time for any given arm. Common assumptions on the reward distributions include bounded support, sub-Gaussian tails, or Bernoulli distributions. In this dissertation, we assume all feedback is immediate, i.e., the reward r_t is observed directly after pulling arm I_t. The setting of delayed feedback is not considered.

In addition to the unstructured model described above, many variants of the bandit model incorporate additional structure to capture more complex scenarios, and these models often play a more important role in real-world applications. A prominent example is the linear bandit model. In a linear bandit, each arm i is associated with a feature vector $x_i \in \mathbb{R}^d$, and its expected reward is a linear function of this vector, i.e., $\mu_i = x_i^\top \theta^*$, for some unknown parameter $\theta^* \in \mathbb{R}^d$. In this setting, the arms are no longer independent, and pulling one arm provides information about the rewards of other arms through the shared parameter θ^*. This structure allows an algorithm to generalize information learned from one arm to others, which can significantly improve learning efficiency, especially when the number of arms is large. Standard algorithms designed for the unstructured MAB model would be inefficient in this setting, as they would fail to leverage this shared structure.

This dissertation will consider problems in both the standard MAB setting and in structured settings, including linear and generalized linear models.

1.3 Regret Minimization vs Pure Exploration

While the difference between regret minimization and pure exploration has been briefly mentioned, we provide a more detailed comparison here. This is particularly important as this dissertation focuses on the pure exploration framework, which may be less familiar to some readers. There are two primary frameworks for studying MAB problems: regret minimization and pure exploration. In the regret minimization framework, the goal is to minimize the cumulative regret, defined as the gap between the expected reward of an oracle that always plays the best arm and the expected reward of the algorithm over a time horizon T. In the pure exploration framework, the objective can vary. Typical performance measures include the probability of error ($\mathbb{P}\big(\mu_{J_T} \neq \mu_1\big)$), where J_T is the arm recommended by the algorithm after T rounds; the simple regret ($\mathbb{E}\big[\mu_1 - \mu_{J_T}\big]$); or, in the top-$K$ arm identification problem, the probability of failing to identify the correct set ($\mathbb{P}\big(J_T \notin \{1, \ldots, K\}\big)$).

It was not until the work of Bubeck et al. (2009) that it was widely realized that an algorithm effective for regret minimization can be inefficient for pure exploration, which led to a greater focus on pure exploration-specific algorithms. The pure exploration framework itself contains two different settings: fixed confidence and fixed budget. Within the scope of best-arm identification, for example, the settings are defined as follows. In the fixed confidence setting, the algorithm is given a target confidence level δ and aims to identify the best arm with probability at least $1-\delta$ while minimizing the number of samples used. In this case, the algorithm requires a stopping rule to determine when to stop collecting samples. In the fixed budget setting, the algorithm is given a fixed number of samples T and aims to maximize the probability of correctly identifying the best arm within this budget. This setting does not require a stopping rule since the budget is predetermined. Both of these settings have been widely studied and are used in different applications.

1.4 Experimental Design

We also present an overview of experimental design, which is an important tool used in this dissertation. Experimental design is a branch of statistics that focuses on planning experiments to ensure that the collected data can provide meaningful and valid conclusions.

In the context of bandit problems, experimental design techniques have been widely used for bandits with linear or generalized linear structures. The key idea is to design a sampling strategy in terms of a sampling distribution on the arms to maximize the information gained about the unknown parameters of the model, ahead of time, which can lead to more efficient learning. A well-defined experimental design objective yields an allocation distribution over the arms that can be used to guide the sampling process. Common design objectives include G-optimal, D-optimal, and E-optimal designs, each with its own advantages and purposes.

For example, G-optimal design aims to minimize the maximum prediction variance across the design space (Kiefer and Wolfowitz, 1959):

$$\arg \min_{\lambda \in \Delta(\mathcal{X})} \max_{x \in \mathcal{X}} \|x\|_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}},$$

where $\Delta(\mathcal{X})$ is the probability simplex over the design space $\mathcal{X}$ (the set of all arm feature vectors). D-optimal design aims to minimize the volume of the confidence ellipsoid for the parameter estimates, which is equivalent to maximizing the determinant of the information matrix (Kiefer and Wolfowitz, 1959):

$$\arg \max_{\lambda \in \Delta(\mathcal{X})} \det\left(\sum_{x \in \mathcal{X}} \lambda_x x x^\top\right).$$

E-optimal design aims to maximize the minimum eigenvalue of the information matrix (Ehrenfeld, 1955):

$$\arg \max_{\lambda \in \Delta(\mathcal{X})} \lambda_{\min}\left(\sum_{x \in \mathcal{X}} \lambda_x x x^\top\right).$$

These are all powerful, general-purpose design objectives that can be used in various settings. Depending on the specific requirements, a different design objective can be chosen. For instance, if the goal is to minimize the worst-case prediction error, G-optimal design is a good choice. E-optimal design is useful for collecting data such that the covariance matrix is well-conditioned, ensuring the stability of the parameter estimates. We note that despite the power of these general-purpose design objectives, they are not specifically designed for

a particular task. Consequently, it is often necessary to design a task-specific objective to achieve superior performance.

1.5 Contributions and Organization of the Dissertation

This dissertation contributes to both the theoretical foundations and practical applications of bandit and interactive learning problems. The main contributions are summarized as follows:

- In Chapter 2, we provide a comprehensive analysis of simple regret in the fixed-budget pure exploration setting. Our results answer an open problem mentioned in Lattimore and Szepesvári (2020) by providing the first instance-dependent bounds for simple regret, which help characterize the output distribution of pure exploration algorithms. In addition, we propose a novel algorithm that exhibits an interesting acceleration phenomenon for pure exploration.

- In Chapter 3, we tackle the challenge of best-arm identification with non-compliance. We first formulate a general problem setting that captures key aspects of non-compliance in real-world scenarios. We then propose a novel algorithm that combines ideas from bandit learning and instrumental variable regression to solve this problem. A key contribution is a non-asymptotic confidence bound for the proposed estimator, which is motivated by the well-studied two-stage least squares estimator in econometrics. We further provide a sample complexity analysis for the algorithm and demonstrate its effectiveness through experiments.

- In Chapter 4, we address the problem of efficient data collection for pairwise preference datasets in LLM alignment. Existing solutions often rely on off-the-shelf experimental design objectives that are not specifically designed for this task and overlook its unique structure, leading to suboptimal performance. We start from a simple intuition that leads to a novel experimental design objective that is specifically tailored to the preference data collection problem. This results in an efficient algorithm that, to the best of our knowledge, achieves the first instance-dependent guarantee for this problem.

In parallel, we propose an easy-to-implement greedy algorithm and validate its strong empirical performance on real-world datasets.

Chapter 2

Revisiting simple regret: Fast rates for returning a good arm

Simple regret is a natural and parameter-free performance criterion for pure exploration in multi-armed bandits. However, possibly due to the challenge of characterizing it, in the theory community, it has been largely overlooked in favor of studying the probability of selecting the best, or an ε-good arm.

In this chapter, we make significant progress on minimizing simple regret in both data-rich ($T \geq n$) and data-poor regime ($T \leq n$) where n is the number of arms, and T is the number of samples. At its heart is our improved instance-dependent analysis of the well-known Sequential Halving (SH) algorithm, where we bound the probability of returning an arm whose mean reward is not within ε from the best (i.e., not ε-good) for *any* choice of $\varepsilon > 0$, despite ε not being input to SH. Our bound not only leads to an optimal worst-case simple regret bound of $\sqrt{n/T}$ up to logarithmic factors but also essentially matches the instance-dependent lower bound for returning an ε-good arm reported by Katz-Samuels and Jamieson (2020). For the more challenging data-poor regime, we propose Bracketing SH (BSH) that enjoys the same improvement even without sampling each arm at least once.

Our empirical study shows that BSH outperforms existing methods on real-world tasks.

2.1 Introduction

We consider the pure exploration problem in multi-armed bandits. In this problem, given n arms, the learner sequentially chooses an arm $I_t \in [n] := \{1, \ldots, n\}$ at time t to observe a reward $r_t = \mu_{I_t} + \eta_t$ where μ_i is the mean reward of arm i and η_t is stochastic σ^2-sub-Gaussian noise. Without loss of generality, we assume that $1 \geq \mu_1 \geq \cdots \geq \mu_n \geq 0$. After T time steps, the learner is required to output an arm J_T that is estimated to have the largest

mean reward. The budget T may or may not be given to the algorithm as input. Such a problem was formalized by Even-Dar et al. (2006) and Bubeck et al. (2009), but similar problems were considered much earlier (Bechhofer, 1958; Chernoff, 1959). Recently, pure exploration has found many applications such as efficient crowdsourcing for cartoon caption contests (Tanczos et al., 2017), hyperparameter optimization (Li et al., 2018a, 2020), and improving the time complexity of clustering algorithms (Baharav and Tse, 2019a). Hereafter, we take $\sigma^2 = 1$ for ease of exposition, but all our results can easily be extended to σ^2-sub-Gaussian noise.

Among many performance criteria, *simple regret* proposed by Bubeck et al. (2009) is a natural measure of the quality of the estimated best arm J_T:

$$\mathsf{SReg}_T := \mathbb{E}\left[\mu_1 - \mu_{J_T}\right] . \tag{2.1}$$

Simple regret is an *unverifiable* performance measure, meaning that the algorithm need not verify its performance to a prescribed level. This is in stark contrast to the fixed confidence setting (Even-Dar et al., 2006) such as (ε,δ)-PAC that requires the algorithm to *verify* the quality of the estimated best arm to the prescribed target accuracy ε and confidence level δ. While certain applications do need such verification, there are many applications where the sampling budget is too limited to show meaningful guarantees, such as cartoon caption contests (Jain et al., 2020) and biological experiments (Jun et al., 2016). In these cases, algorithms with verifiable guarantees may not be meaningful. Furthermore, existing algorithms with verifiable guarantees are vacuous in the data-poor regime of $T \le n$ as it requires each arm to be pulled at least once (Katz-Samuels and Jamieson, 2020, Section C.1). The fixed budget setting considers an unverifiable measure of $\mathbb{P}\big(\mu_{J_T} \neq \mu_1\big)$ (Audibert et al., 2010), but this quantity inevitably depends on the smallest gap $\mu_1 - \mu_2$ (Carpentier and Locatelli, 2016), which is usually not meaningful until the sample size is very large.

Despite being attractive, the analysis of simple regret has been elusive. Existing studies focus on either achieving minimax simple regret bound (Lattimore et al., 2016a) or characterization of how simple regret is different from the cumulative regret (Bubeck et al., 2009).

Table 2.1: Sample complexity comparison with ME (Even-Dar et al., 2006), $\mathcal{P}_3$ (Chaudhuri and Kalyanakrishnan, 2019), and BUCB (Katz-Samuels and Jamieson, 2020). The lower bound for Polynomial(α) instances (Theorem 3, Lemma 3.1) is a contribution of this chapter which improves on Katz-Samuels and Jamieson (2020, Theorem 1) to include the correct dependence on δ. The lower bound for EqualGap(m) was first shown by Chaudhuri and Kalyanakrishnan (2017), which works for the fixed budget setting as well. We omit constant and logarithmic factors in n and m. Methods with asterisk consider the fixed confidence setting, and those with dagger have an in expectation bound rather than the high probability one. Anytime algorithms do not require the knowledge of the budget T as input, and parameter-free ones do not require any of $\{\alpha, m, \varepsilon, \delta\}$ as input.

	Polynomial(α) with $\alpha > \frac{1}{2}$	EqualGap(m) with $m \le n/4$	Data-poor	Anytime	Parameter-free
ME*	$\frac{n}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	✗	✗	✗
$\mathcal{P}_3^*$	$\frac{1}{\varepsilon^{2+(1/\alpha)}}\log\left(\frac{1}{\delta}\right) \vee \frac{1}{\varepsilon^2}\log^2\left(\frac{1}{\delta}\right)$	$\frac{1}{\varepsilon^2}\left(\frac{n}{m}\log\left(\frac{1}{\delta}\right) + \log^2\left(\frac{1}{\delta}\right)\right)$	✓	✗	✗
BUCB†	$\frac{n}{\varepsilon^{2-(1/\alpha)}}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	✓	✓	✗
DSH (ours)	$\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right) \vee \frac{n}{\varepsilon^{1/\alpha}}$	$\frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right) \vee \frac{n}{\varepsilon^2}$	✗	✓	✓
BSH (ours)	$\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	✓	✓	✓
Lower bound	$\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right)$			

It is not known whether simple algorithms like Sequential Halving (SH) (Karnin et al., 2013) enjoy the optimal simple regret or not.

In this chapter, we revisit simple regret and make two main contributions. First, we provide a novel and tight analysis of Sequential Halving (SH), one of the state-of-the-art algorithms for pure exploration. We analyze a strong performance criterion that we call *ε-error probability*:

$$\mathbb{P}\left(\mu_1 - \mu_{J_T} > \varepsilon\right) \tag{2.2}$$

where ε is not given to the algorithm as input. This is precisely characterizing the distribution of μ_{J_T}, which was mentioned as an interesting direction of research in Lattimore and Szepesvári (2020, Section 33.4 Note 9). If an algorithm $\mathcal{A}$ does not take ε as input yet has a guarantee for any ε, we say $\mathcal{A}$ enjoys a *uniform ε-error probability bound*. The ε-error

probability is stronger than simple regret since

$$\mathsf{SReg}_T = \mathbb{E}[\mu_1 - \mu_{J_T}] = \int_0^\infty \mathbb{P}(\mu_1 - \mu_{J_T} > \varepsilon)\,\mathrm{d}\varepsilon\ . \tag{2.3}$$

Thus, a uniform ε-error probability bound implies a simple regret bound.

We call an arm i to be ε-good if $\mu_i \geq \mu_1 - \varepsilon$. Let $\Delta_i := \mu_1 - \mu_i$ be the gap of arm i and $g(\varepsilon) := \left|\left\{i \in [n] \mid \mu_i \geq \mu_1 - \varepsilon\right\}\right|$ be the number of ε-good arms. We prove that SH (Karnin et al., 2013) has the following bound for any $\varepsilon > 0$ (Theorem 1):

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) \leq \exp\left(-\tilde{\Theta}\left(\frac{T}{H_2(\varepsilon)}\right)\right), \tag{2.4}$$

where $\tilde{\Theta}(\cdot)$ hides logarithmic factors and

$$H_2(\varepsilon) := \frac{1}{g(\varepsilon/2)} \max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\ .$$

is the sample complexity function. By setting $\varepsilon = 0$, the complexity becomes $H_2(0) = \max_{i \geq 2} i\Delta_i^{-2} =: H_2$, so our bound is never worse than the original bound of SH presented in Karnin et al. (2013): $\mathbb{P}(\mu_{J_T} \neq \mu_1) \leq \exp(-\tilde{\Theta}(\frac{T}{H_2}))$.

Furthermore, our bound essentially matches the lower bound presented in Katz-Samuels and Jamieson (2020). We also derive a minimax-style bound on the ε-error probability, which can be easily integrated out over ε to obtain $\mathsf{SReg}_T = \tilde{O}(\sqrt{n/T})$, and further show that a variant of SH results in removing logarithmic factors from the bound and achieve $\mathsf{SReg}_T = O(\sqrt{n/T})$. Finally, we derive an anytime version of SH that we call Doubling SH (DSH), which enjoys the same guarantees without requiring T as input. We present details of these claims in Section 2.3.

As the second contribution, we propose an improved pure exploration algorithm for the data-poor regime. This is motivated by the fact that the guarantee (2.4) becomes vacuous in the data-poor regime ($T \leq \max_{i \geq g(\varepsilon)+1} i\Delta_i^{-2}$ to be precise) as it does not even allow us to pull every arm once. Inspired by Katz-Samuels and Jamieson (2020), we propose Bracketing SH (BSH) that progressively subsamples a larger set of arms (i.e., brackets) for which we invoke

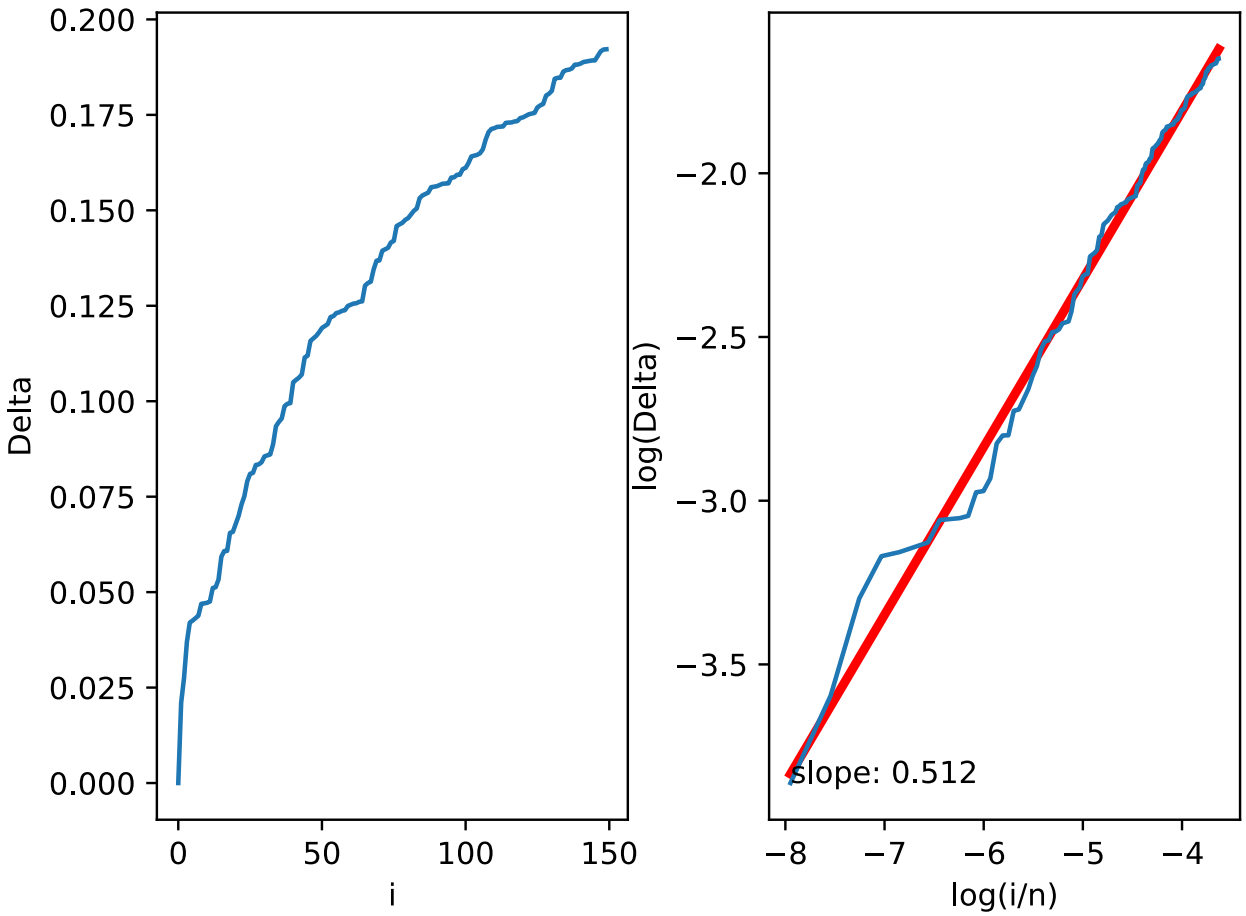


Figure 2.1: The polynomial fitting for the best 150 arms of the New Yorker Cartoon Caption contest 788.

SH to find a good arm. Our analysis shows that BSH enjoys a bound that is essentially the same as (2.4) but works even for $T \leq \max_{i \geq g(\varepsilon)+1} i\Delta_i^{-2}$. Compared to BUCB (Katz-Samuels and Jamieson, 2020), the state-of-the-art algorithm for the data-poor regime, BSH achieves three improvements: (i) BSH is parameter-free whereas BUCB requires the target error rate δ as input, (ii) BSH has a high probability sample complexity bound as opposed to the expected sample complexity bound of BUCB, (iii) BSH shows a much stronger bound than BUCB. We provide details on BSH in Section 2.4.

We compare our bounds with existing bounds in various setups in Table 2.1. For ease of comparison, we consider special instances called Polynomial(α) and EqualGap(m). Polynomial(α) is an instance with gap structure $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ for $2 \leq i \leq n$, where $\alpha > 0.5$ (Jamieson et al., 2013), and EqualGap(m) is an instance that has $m-1$ arms with the gap $\varepsilon/4$ and $n-m$

arms with the gap $\frac{5}{4}\varepsilon$. Both α and m have the role of adjusting the number of good arms. Our bound for BSH matches the lower bounds for the first time to our knowledge, thereby establishing the optimality for these two instances.

Note that many real-world instances follow Polynomial(α); e.g., see Figure 2.1. One interesting aspect of Polynomial(α) is that, although α affects the fraction of good arms, the optimal rate $1/\varepsilon^2$ is independent of α. This is because the hardness of returning an ε-good arm depends not just on how many good arms we have but also on how far they are from bad arms as we explain in Section 2.3.

Our result also resolves the open problem of having a factor of $\ln^2(1/\delta)$ rather than $\ln(1/\delta)$ when obtaining accelerated sample complexities under the presence of multiple good arms. Such an issue appeared repeatedly in the literature; e.g., Chaudhuri and Kalyanakrishnan (2019), Katz-Samuels and Jamieson (2020, Section E), and Aziz et al. (2018, Section 5). They all rely on the subsampling device or its variants. Our improved bound can be attributed to the fact that an accelerated sample complexity can be obtained even *without the subsampling device* – we are the first to achieve such a result, to our knowledge.

We evaluate our algorithm on the cartoon caption contest dataset in Section 2.5 and conclude the chapter with future work in Section 2.6.

2.2 Preliminaries

We focus on the fixed budget and anytime setting in pure exploration. In the *fixed budget* setting (Audibert and Bubeck, 2010), the learner is given a budget $T \geq 1$ as input and is required to return an estimated best arm $J_T \in [n]$ at the end of T-th time step. In the *anytime* setting, there is no prescribed T, and the learner is required to output $J_t \in [n]$ for every time step $t \geq 1$. In these settings, unlike the fixed confidence setting, the learner is not required to verify the quality of the returned arm J_T. We collectively call these settings the *unverifiable* setting and refer to the fixed confidence setting as the *verifiable* setting. Note that for any verifiable setting, one can consider an unverifiable version. For example, the verifiable (m, ε)-good arm identification problem (Chaudhuri and Kalyanakrishnan, 2017) requires that the learner, given (m, ε, δ), identify an arm i that is (m, ε)-good (i.e., $\mu_i \geq \mu_m - \varepsilon$)

w.p. at least $1-\delta$. One can consider the unverifiable (m,ε)-good arm identification problem where the goal is to minimize $\mathbb{P}(\mu_{J_T} \geq \mu_m - \varepsilon)$ after T time steps.

Sample Complexity. When comparing bounds for simple regret or ε-error probability, it is often easier to consider the sample complexity, which measures the number of samples required to achieve the target performance level. For example, the sample complexity for the ε-error probability is the smallest time step τ such that $\forall t \geq \tau, \mathbb{P}(\mu_1 - \mu_t > \varepsilon) \leq \delta$.

Notations. Throughout the chapter, "const" is a universal and positive constant, which may have different values for different expressions. Both $\tilde{\Theta}(\cdot)$ and $\tilde{\mathcal{O}}(\cdot)$ omits logarithmic factors on m, n, but not $1/\delta$.

2.3 Improved Analyses of Sequential Halving

Sequential Halving (SH) (Karnin et al., 2013) is a simple and parameter-free algorithm that has had an outsize impact on pure exploration since its introduction. Many recent practical algorithms for pure exploration (Aziz et al., 2022; Li et al., 2018a, 2020; De Heide et al., 2021) make use of this algorithm, either with slight modifications or as a subroutine. These works have demonstrated that the algorithm's practical and theoretical significance is greater than its original presentation. Even beyond the basic bandit setting, SH also has been used for planning in reinforcement learning (Danihelka et al., 2022), where SH is observed to be easier to tune compared to some UCB style algorithms as it does not have problem-dependent parameters.

We present SH in Algorithm 1. SH consists of $\lceil \log_2(n) \rceil$ stages. In each stage ℓ, the algorithm performs a uniform sampling over the surviving arms S_ℓ followed by eliminating the bottom half of S_ℓ w.r.t. empirical means and sets $S_{\ell+1}$ as the resulting arm set.

SH is known to achieve a nearly optimal instance-dependent bound on the $(\varepsilon = 0)$-error probability, which we detail in Table 2.2. However, it is unknown if SH is a good algorithm for returning an ε-good arm. For $\varepsilon > 0$, though in the fixed confidence setting, Median Elimination (ME) achieves the sample complexity of $\frac{n}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$ (Even-Dar et al., 2006). Does SH achieve a similar sample complexity for ε-error probability? Despite the similarity between SH and ME, the answer is not clear because the sampling scheme of SH is different

from ME. In light of the ubiquity of this algorithm and recent empirical success, a stronger characterization of the performance of SH is of great interest to the community.

In this section, via a powerful analysis of the ε-error probability incurred by SH, we show that this algorithm achieves an optimal bound for the unverifiable (m, ε)-good arm identification problem that further implies a minimax optimal simple regret bound (Lattimore and Szepesvári, 2020, Exercise 33.1). We also provide a near-optimal instance-dependent uniform ε-error probability bound for the first time.

Our main theorem shows the following instance-dependent uniform ε-error probability bound for SH.

Theorem 1. *For any $\varepsilon \in (0, 1)$, the ε-error probability of SH satisfies*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) \leq \exp\left(-\tilde{\Theta}\left(\frac{T}{H_2(\varepsilon)}\right)\right),$$

for $T \geq \tilde{\Theta}\left(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)$.

The proof is in Section 2.9.5. Note the requirement of $T \geq \tilde{\Theta}\left(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)$ is not a weakness of our result. One can easily see that if $\Delta_i = \varepsilon, \forall i \geq 2$, the standard result of SH (Karnin et al., 2013) becomes vacuous for $T = o(\frac{n}{\varepsilon^2})$. We stress that, unlike all existing work that we are aware of, Theorem 1 bounds the ε-error probability for an algorithm that does not require ε as input. That is to say, ε is a parameter for analysis only. This upper bound of ε-error probability implies a sample complexity of $\tilde{O}(H_2(\varepsilon) \ln(1/\delta))$.

Polynomial gap instances.

The class of polynomial gap instances was introduced by Jamieson et al. (2013), who referred to them as 'α-parameterized' instances. The authors studied these instances in the best-arm identification setting, showing that for $\alpha > 1/2$, the sample complexity $H = \sum_{i=2}^{n} \Delta_i^{-2}$ scales like $\Omega(n^{2\alpha})$, with the increase in sample complexity corresponding to the growing number of close-to-optimal arms as α increases. In one of the highlight results of this work, we show that when $\varepsilon > 0$ is not pathologically small, SH returns ε-good arms on these instances with an optimal sample complexity that is independent of the number of arms n.

Corollary 1.1. *For Polynomial(α) with $\alpha > \frac{1}{2}$, we have*

$$H_2(\varepsilon) = \frac{1}{g\left(\frac{\varepsilon}{2}\right)} \max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} < \frac{4}{\varepsilon^2}.$$

The proof is in Section 2.9.6. While this result initially seems surprising, the lack of dependence on n stems from the ratio between the number of 'well-separated' ε-good arms and the sample complexity of filtering out ε-bad arms (i.e., non-ε-good arms) on these instances. As n increases, due to the polynomial gap structure, the number of well-separated ε-good arms (e.g., the collection of $\varepsilon/2$-good arms) increases linearly with n, which balances the linear growth in n of the sum over the ε-bad arm gaps $\sum_{i=g(\varepsilon)+1}^{n} \Delta_i^{-2}$. This second term is effectively the sample complexity of discarding all ε-bad arms. Since we only require one of the good arms to be returned, we get to divide $\sum_{i=g(\varepsilon)+1}^{n} \Delta_i^{-2}$ by $g(\varepsilon/2)$.

Lower bounds. We compare our upper bound in Theorem 1 with lower bounds. For brevity, we present an informal version here and defer the full version to Section 2.9.15 along with a slightly stronger version.

Theorem 2. *(Informal version of Katz-Samuels and Jamieson (2020, Theorem 1)) For any algorithm, there exists an instance where the expected unverifiable sample complexity for returning an ε-good arm with probability greater than 15/16 satisfies*

$$\Omega\left(H^{low}(\varepsilon) := \frac{1}{g(\varepsilon)} \sum_{i=g(\varepsilon)+1}^{n} \left(\frac{1}{\Delta_i^2}\right) - \frac{1}{\Delta_{g(\varepsilon)+1}^2}\right).$$

Corollary 2.1 (Comparison with previous lower bound)**.** *In the case that $g(\varepsilon)$ and $g\left(\frac{\varepsilon}{2}\right)$ have the same order ($\frac{g\left(\frac{\varepsilon}{2}\right)}{g(\varepsilon)}$ is irrelevant to ε), the sample complexity measure $H_2(\varepsilon)$ satisfies*

$$H^{low}(\varepsilon) \lesssim H_2(\varepsilon) \lesssim H^{low}(\varepsilon) + \frac{2}{\Delta_{g(\varepsilon)+1}^2}$$

where $\lesssim$ hides constants and logarithms of n and $1/\delta$.

The sample complexity implied by Theorem 2 for returning an ε-good arm with probability $1-\delta$, with $\delta \in (0, 1/16)$, is $\Omega\left(H^{\text{low}}(\varepsilon)\right)$, which does not have the factor $\ln(1/\delta)$. We contribute a new lower bound that suggests a sample complexity of $\Omega\left(\tilde{H}(\varepsilon)\log(1/\delta)\right)$, with a complexity

term $\tilde{H}(\varepsilon)$ that we describe next, and with which we show that Theorem 1 is essentially tight on polynomial gap instances with $\alpha > 1/2$.

For an n-armed bandit instance ν, let

$$\tilde{H}(\nu, \varepsilon) = \sum_{i=2}^{n/g(\varepsilon)} \Delta_{i \cdot g(\varepsilon)}^{-2},$$

where for simplicity we assume that n is an integer multiple of $g(\varepsilon)$.

Theorem 3. *Fix $n > 2$, $\varepsilon > 0$ and an n-armed unit-variance Gaussian instance ν. With $a := \tilde{H}(\nu, \varepsilon)$ we define $\mathcal{B}_n(a, \varepsilon)$, the collection of n-armed bandit instances with unit-variance gaussian arms for which $\tilde{H}(\nu', \varepsilon) \le a$ for each $\nu' \in \mathcal{B}_n(a, \varepsilon)$. Then for any algorithm π and sample budget $T \in \mathbb{N}$,*

$$\sup_{\nu' \in \mathcal{B}_n(a,\varepsilon)} \mathbb{P}_{\nu', \pi}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) \ge \max\left\{\frac{1}{2}, \frac{1}{36}\exp\left(-24T/a\right)\right\}.$$

The complexity term $\tilde{H}(\nu, \varepsilon)$ plays the same role in our result as that of $H(\nu) = \sum_{i=2}^{n} \Delta_i^{-2}$ in Carpentier and Locatelli (2016, Theorem 1). The proof of Theorem 3 can be found in Section 2.9.8.

Lemma 3.1. *Consider a Polynomial(α) instances with fixed $\alpha > 1/2$. Then for any $\varepsilon \le 1/2$, $\tilde{H} \ge \frac{c(\alpha)}{\varepsilon^2}$ where $c(\alpha)$ is a problem-dependent constant which does not depend on ε.*

The proof is in Section 2.9.9. Combining Corollary 1.1 and Lemma 3.1, we see that Theorem 1 is optimal up to logarithmic factors in the sample complexity for polynomial gap instances with $\alpha > 1/2$.

2.3.1 Unverifiable (m, ε)-Good Arm Identification

We show the error probability of SH for unverifiably identifying an (m, ε)-good arm (defined in Section 2.2) and discuss its implications.

Algorithm 1: Sequential Halving (SH)

Input: budget: T, arms: $[n]$
Initialize: $S_1 = [n]$
for $\ell = 1, \dots, \lceil \log_2 n \rceil$ **do**
  Sample each arm $i \in S_\ell$ for T_ℓ times where

$$T_\ell = \left\lfloor \frac{T}{|S_\ell| \lceil \log_2 n \rceil} \right\rfloor$$

  Let $S_{\ell+1}$ be the set of $\lceil S_\ell / 2 \rceil$ arms in S_ℓ with the largest empirical rewards.
Set J_T as the only arm in $S_{\lceil \log_2 n \rceil}$.
Output: J_T

Theorem 4. *For any $m \le n$ and any $\varepsilon \in (0,1)$, the error probability of SH for identifying an (m,ε)-good arm satisfies*

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \le \log_2 n \cdot \exp\left(-\text{const} \cdot m \cdot \left(\frac{\varepsilon^2 T}{4n \log_2^2(2m) \log_2 n} - \ln(4e)\right)\right).$$

Thus, there exist an absolute constant $c_1 > 0$ s.t. if $T \ge c_1 \frac{(\ln(4e) + \ln\ln n) n \log_2^2(2m) \log_2 n}{\varepsilon^2} = \tilde{\Theta}\big(\frac{n}{\varepsilon^2}\big)$, we have

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \le \exp\left(-\tilde{\Theta}\left(m \frac{\varepsilon^2 T}{n}\right)\right).$$

The proof is in Section 2.9.3.

Remark 1. *Note that our bound implies a verifiable sample complexity bound for the (m,ε)-good arm identification setup as well because, given (m,ε,δ), one can work out the sufficient samples size T' to control the RHS of Theorem 4 to be at most δ and then run SH with T'.*

Implications for $m = 1$. The worst-case upper bound of Theorem 4 for $m = 1$ corresponds to the sample complexity of $\tilde{\mathcal{O}}\big(\frac{n}{\varepsilon^2} \log \frac{1}{\delta}\big)$. This is the same as the classic lower bound result of the (ε,δ)-PAC problem (Mannor and Tsitsiklis, 2004) up to logarithmic factors. While numerous algorithms achieve a matching upper bound, including Even-Dar et al. (2006) and Hassidim et al. (2020), our result in Theorem 1 indicates a tighter bound when there are many good arms.

Algorithm 2: Doubling Sequential Halving (DSH)

Input arms: $[n]$
Initialize $T_1 = \lceil n \log_2 n \rceil$, $\hat{j} = j$ for some arbitrarily chosen $j \in [n]$, $\hat{\mu}^* = -\infty$. Define the time blocks $\mathcal{T}_1 = \{1, \dots, T_1\}$ and $\mathcal{T}_k = \{T_1(2^{k-1} - 1) + 1, \dots, T_1(2^k - 1)\}, \forall k \geq 2$, which satisfies $|\mathcal{T}_k| = T_1 2^{k-1}$. $k = 1$. An instance of SH denoted by $\mathcal{S}_1$ with budget $|\mathcal{T}_1|$.
for $t = 1, 2, \dots$ **do**
  Pull arm I_t according to the recommendation from $\mathcal{S}_k$
  Receive reward R_t and send it to $\mathcal{S}_k$
  if $t = \max \mathcal{T}_k$ **then**
    Set $(\hat{j}, \hat{\mu}^*)$ as the output arm and its empirical mean from the last stage of $\mathcal{S}_k$
    $k \leftarrow k + 1$
    Initialize a new instance of SH denoted by $\mathcal{S}_k$ with budget $|\mathcal{T}_k|$
  Output: $J_t = \hat{j}$ and $M_t = \hat{\mu}^*$.

We now apply (2.3) to convert a uniform ε-error probability bound into a simple regret bound.

Corollary 4.1. *SH satisfies* $\mathsf{SReg}_T \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{n}{T}}\right)$.

The proof is in Section 2.9.2. Note that our bound is minimax optimal up to logarithmic factors (Lattimore and Szepesvári, 2020, Exercise 33.1). To our knowledge, the only algorithm that we are aware of that achieves the minimax simple regret is MOSS (Audibert et al., 2009; Lattimore et al., 2016a). However, as MOSS is designed to minimize cumulative regret, it cannot enjoy low instance-dependent bounds for pure exploration tasks (Bubeck et al., 2011). While uniform sampling can also achieve a near-optimal simple regret, one can show that uniform sampling can be arbitrarily worse than SH w.r.t. the $(\varepsilon = 0)$-error probability as its sample assignments over the arms are inherently non-adaptive.

Remark 2. *In Section 2.9.12, we show that one can adjust T_ℓ in Algorithm 1 to achieve the minimax optimality up to a constant factor (i.e., no logarithmic factors), but it is not clear if this variant achieves a similar instance-dependent sample complexity as Algorithm 1.*

Implications for $m > 1$. We now show that SH enjoys the optimal worst-case guarantee for the unverifiable (m, ε)-good arm identification problem up to logarithmic factors. Our result is especially attractive since the target number of good arms m is not an input to SH.

Our upper bound of Theorem 4 corresponds to a sample complexity of $\tilde{\mathcal{O}}\left(\frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$, which matches the lower bound in Chaudhuri and Kalyanakrishnan (2017), thereby closing the gap between the upper and lower bounds from previous work; the algorithms therein have suboptimal upper bounds that scale with $\log^2\left(\frac{1}{\delta}\right)$. Furthermore, these algorithms all require (m, ε, δ) as input, which is natural for obtaining verifiable sample complexities but becomes a limitation in other settings. In contrast, Theorem 4 allows (m, ε) to be chosen freely to measure the algorithm's performance since the parameters for measuring the performance are no longer necessary for executing the algorithm. Therefore, one can rewrite Theorem 4 as follows:

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) \leq \min_{\{(m',\varepsilon'):\Delta_{m'}+\varepsilon'\leq\varepsilon\}} \exp\left(-\tilde{\Theta}\left(m'\frac{\varepsilon'^2 T}{n}\right)\right).$$

2.3.2 Anytime version of SH

To support anytime simple regret minimization, we combine SH with the doubling trick (Jun and Nowak, 2016), which we call Doubling SH (DSH). Specifically, DSH repeatedly runs SH. The k-th SH receives a doubled budget compared with the $(k-1)$-th SH. Before it finishes k-th SH, it always returns the output of $(k-1)$-th SH. The pseudocode can be found in Algorithm 2. The following theorem shows that DSH enjoys the same guarantee as SH up to a constant factor.

Theorem 5. *Let $\varepsilon \in (0,1)$. A single run of DSH satisfies the following ε-error probability*

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{T}{H_2(\varepsilon)}\right)\right),$$

simultaneously for all $T \geq \tilde{\Theta}\left(\max_{i\geq g(\varepsilon)+1}\frac{i}{\Delta_i^2}\right)$

This is a direct consequence of the fact that at time t, the latest finished SH was run with a budget of at least $t/4$. The full proof is in Section 2.9.14.

2.4 Simple Regret in the Data-Poor Regime

Despite the optimality of SH shown in the last section, it requires at least $\tilde{\Theta}\Big(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\Big)$ samples, which in the worst case could be $\tilde{\Theta}\big(n/\varepsilon^2\big)$. This requirement is unacceptable when the number of arms n is large (or even infinite) while the fraction of the ε-good arms is kept constant. This setup is commonly referred to as the data-poor regime, where one would like to return a good arm even before the algorithm samples every arm once.

To cope with the data-poor regime, we take inspiration from BUCB (Katz-Samuels and Jamieson, 2020) and propose Bracketing SH (BSH) that enjoys a similar guarantee as DSH but enjoys non-vacuous sample complexity in the data-poor regime. To understand the bracketing trick, suppose that we uniformly sample a subset of size n/m $(0 < m \leq n)$ from the entire arm set $[n]$. With constant probability, this subset includes at least one of the top m arms. By applying any pure exploration algorithm to this subset, we expect to find the best arm within the subset with the sample complexity that scales with n/m rather than n. Such a subset is called a *bracket*. Note that there is a natural trade-off here with the bracket size. If the size is large, we are likely to include many good arms, but the sample complexity of identifying a good arm becomes large because we have a lot of arms to pull at least once. On the other hand, if the size is small, we are not likely to include any good arm, although the sample complexity of identifying an arm that is good relative to the bracket is small. As we show later, one can precisely work out such a tradeoff mathematically and find the best bracket size. The challenge is, however, that the best bracket size typically requires knowledge of the number of good arms.

To avoid requiring such knowledge, BSH adopts the bracketing technique of Katz-Samuels and Jamieson (2020) that progressively creates a larger and larger bracket size as the time step t gets larger while invoking a base algorithm for each bracket by cycling through them (i.e., Round Robin). Specifically, BSH uses DSH as the base algorithm. At each time step, BSH takes in the estimated best arms and their empirical means from all the brackets and outputs the one with the largest empirical mean.

We summarize BSH in Algorithm 3 where the operator $U\big([n], k\big)$ samples k items with replacement from $[n]$ (uniformly at random). If $k \geq n$, $U([n], k)$ returns $[n]$. To define

Algorithm 3: Bracketing SH (BSH)

Input: arms: $[n]$

Initialize: $t = 1$, $B = 0$, $b_1 = 1$. Define $(I(\mathcal{D}), J(\mathcal{D}), M(\mathcal{D}))$ to be the arm to be pulled next, the current estimated best arm, and its empirical mean from an algorithm $\mathcal{D}$, respectively.

for $t = 1, 2, 3 \dots$ **do**

  if $t \geq B2^B$ **then**

    $B \leftarrow B + 1$

    Sample a bracket $A_B = U\big([n], n \vee 2^B\big)$

    Initialize a new instance of DSH, denoted by $\mathcal{D}_B$, with the bracket A_B.

  Pull arm $I_t = I(\mathcal{D}_{b_t})$

  Receive a reward R_t from the environment and send R_t to $\mathcal{D}_{b_t}$

  Output $J_t = J(\mathcal{D}_{\hat{b}})$ where $\hat{b} = \arg\max_{b \in [B]} M(\mathcal{D}_b)$

  $b_{t+1} = \begin{cases} b_t + 1 & \text{if } b_t < B \\ 1 & \text{otherwise} \end{cases}$

terminology, we say a new bracket is *opened* when a new bracket is sampled. We set the bracket-opening schedule such that the B-th bracket is opened at time step $(B-1) \cdot 2^{B-1}$. The arm pulls between the time step $(B-1) \cdot 2^{B-1}$ and $B \cdot 2^B$ are equally allocated to the opened brackets (total B of them). We say an arm *represents* bracket A_B at time step t if it is the output returned by the DSH on bracket A_B at time step t.

Let us first show the properties of the bracketing technique. The bracketing design ensures the diversity of the size of opened brackets. Specifically, the smallest bracket has 2 arms, and the largest bracket has $\tilde{\Theta}(t)$ arms. Moreover, it also ensures that all the opened brackets have received an (order-wise) equal amount of sampling budget at any time.

We now present the ε-error probability bound of BSH.

Theorem 6. *Let $\varepsilon \in (0,1)$. A single run of BSH satisfies the following ε-error probability*

$$\mathbb{P}(\mu_1 - \mu_{J_t} > \varepsilon) \leq \exp\left(-\tilde{\Theta}\left(\frac{t}{\max\left\{H_2\left(\frac{\varepsilon}{2}\right), \frac{1}{\varepsilon^2}\right\}}\right)\right),$$

simultaneously for all $t \geq \tilde{\Theta}\Big(H_2\left(\frac{\varepsilon}{2}\right)\Big)$.

The proof is in Section 2.9.16. The key difference between Theorem 5 and Theorem 6 is the latter does not require the number of samples to be at least $T \geq \tilde{\Theta}\left(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)$.

To show the effectiveness of Theorem 6, we now discuss the implication of our result and compare it with BUCB by Katz-Samuels and Jamieson (2020) using their performance measure called (ε, δ)-unverifiable sample complexity.

Definition 1 ((ε, δ)-unverifiable sample complexity)**.** *(Katz-Samuels and Jamieson, 2020). For an algorithm π and an instance ρ. Let $\tau_{\varepsilon,\delta}$ be a stopping time such that*

$$\mathbb{P}\left(\forall t \geq \tau_{\varepsilon,\delta} : \mu_{J_t} > \mu_1 - \varepsilon\right) \geq 1 - \delta.$$

Then $\tau_{\varepsilon,\delta}$ is called (ε, δ)-unverifiable sample complexity of the algorithm with respect to ρ.

The (ε, δ)-unverifiable sample complexity indicates how many samples are sufficient for an algorithm to begin to output an ε-good arm with high probability. Compared with the verifiable sample complexity, it does not require the algorithm to verify the output is ε-good. More discussion can be found in Katz-Samuels and Jamieson (2020). While the verbatim requirement of (ε, δ)-unverifiable sample complexity is stronger than the uniform ε-error probability bound, we show that the latter implies the former under a mild assumption in Section 2.9.20.

To make explicit comparisons, we turn to specific problem instances and show the (ε, δ)-unverifiable sample complexity bounds for BUCB (Katz-Samuels and Jamieson, 2020, Theorem 7) and BSH.

Corollary 6.1. *Consider the EqualGap(m) instance. BUCB achieves an expected (ε, δ)-unverifiable sample complexity as*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m} \log\left(\frac{1}{\delta}\right)\right).$$

BSH achieves an (ε, δ)-unverifiable sample complexity as, with probability $1 - \delta$,

$$\tau_{\varepsilon,\delta} \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m} \log\left(\frac{1}{\delta}\right)\right).$$

The proof is in Section 2.9.17. In this instance, BUCB and BSH have nearly the same performance guarantees except that BUCB's (ε, δ)-unverifiable sample complexity is stated as an expected sample complexity. Interestingly, Katz-Samuels and Jamieson (2020, Section E) remarks that their attempt to derive a high-probability bound resulted in the factor of $\ln^2(1/\delta)$, which we believe is due to the suboptimal bound (i.e., no accelerated rates) of their choice of the base algorithm.

Scaling with n/m instead of n reveals the merit of the (ε, δ)-unverifiable sample complexity that was not achieved by the algorithms designed to achieve verifiable sample complexity such as Median Elimination (Even-Dar et al., 2006).

Corollary 6.2. *Consider the Polynomial(α) instance; i.e., $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ with $\alpha > 0.5$. For any $\varepsilon \in (0,1)$, let $\hat{\tau}_{\varepsilon,\delta}$ be the upper bound of the expected (ε,δ)-unverifiable sample complexity reported in Katz-Samuels and Jamieson (2020, Theorem 7) for BUCB. Then, BUCB satisfies*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \le \hat{\tau}_{\varepsilon,\delta} = \tilde{\Theta}\left(\varepsilon^{-\frac{2\alpha-1}{\alpha}} n \log\left(\frac{1}{\delta}\right)\right).$$

On the other hand, BSH satisfies, with probability $1-\delta$,

$$\tau_{\varepsilon,\delta} \le \tilde{\mathcal{O}}\left(\varepsilon^{-2} \log\left(\frac{1}{\delta}\right)\right).$$

The proof is in Section 2.9.18. The (ε, δ)-unverifiable sample complexity of BSH for the Polynomial(α) instance does not scale with the number of arms n polynomially, unlike BUCB.

For large-scale instances, BSH provides a much stronger guarantee than BUCB. Even when n is not too large, as long as $\Delta_2 \le \varepsilon$ (which implies $\varepsilon^{\frac{1}{\alpha}} \le n$), BSH still has a better sample complexity bound than BUCB.

2.5 Experiments

We test Bracketing SH on a real-world dataset called The New Yorker Cartoon Caption Contest (Jain et al., 2020). For each cartoon, the editors of The New Yorker collect the eval-

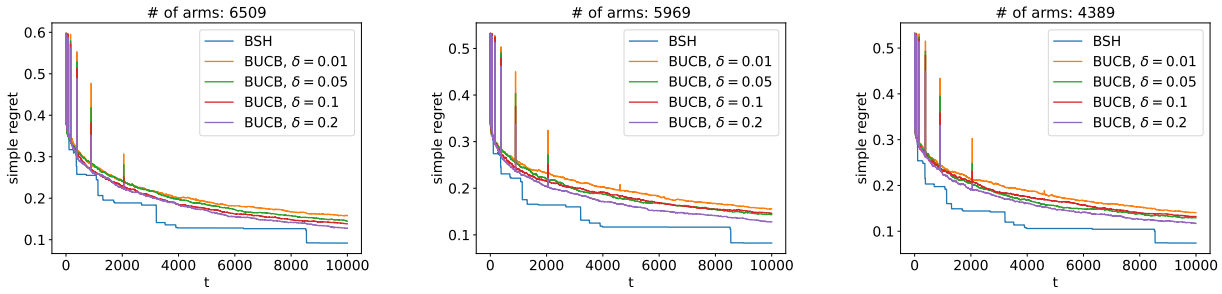


Figure 2.2: The simple regret comparison between Bracketing UCB and Bracketing SH for the New Yorker Cartoon Caption contest 780, 781, 782.

uation score of n captions from the participants. Specifically, upon arrival of a participant, the algorithm sequentially shows a number of captions and receives the evaluation score of "unfunny", "somewhat funny", or "funny". Following Katz-Samuels and Jamieson (2020), we use the proportion of times the score was "somewhat funny" or "funny" as the ground truth mean reward for each caption. We then set the reward distribution as the Bernoulli. We choose contests 780, 781, and 782, which contain 6509 arms, 5969 arms, and 4389 arms, respectively.

As we are especially interested in the data-poor regime, we report the performance of the algorithms up to time step 10,000, which is only about twice larger than the number of arms in our dataset.

We ran BSH and BUCB with various δ choices to see the best version of BUCB, which was repeated 500 times with a different random seed. We used a desktop with AMD Ryzen 5 PRO 4650GE CPU and 16GB RAM to conduct the experiment, which took two hours to produce each plot.

We summarize the results in Figure 2.2, where BSH is a clear winner over BUCB. Specifically, in contest 781, at time step $4,000$, Bracketing UCB achieves simple regret about 0.18 for $\delta = 0.2$, while BSH achieves simple regret about 0.11, amounting to 39% improvement.

We explain more on the implementation in Section 2.8.

2.6 Conclusion

Obtaining a uniform ε-error probability bound is precisely a way to characterize the distribution function of Δ_{J_T} induced by a particular algorithm, based on which we believe that a

uniform ε-error probability bound is a fundamental quantity for any measures for returning an ε-good arm.

As such, we believe our results open up numerous exciting open problems. First, SH does not achieve an optimal worst-case sample complexity of $\Theta(\frac{n}{\varepsilon^2}\ln(1/\delta))$ (due to logarithmic factors). While adjusting its sample allocation scheme does achieve the optimal rate as we report in the appendix, it does not seem to achieve the usual instance-dependent sample complexity of $\tilde{\Theta}(\sum_i(\varepsilon \vee \Delta_i)^{-2})$ achieved by many existing algorithms. We wonder if a modified SH or other algorithms can simultaneously be optimal for both sample complexity measures. Second, another practical pure exploration algorithm is track-and-stop by Garivier and Kaufmann (2016a) and their variants. They are also parameter-free if we only take their sampling strategy and not the stopping strategy. It would be interesting to investigate if track-and-stop achieves similar near-optimal guarantees as DSH.

2.7 Related work

Pure exploration. Pure exploration has a broad position in several closely related research directions, including multi-armed bandits (Audibert and Bubeck, 2010; Karnin et al., 2013; Carpentier and Locatelli, 2016; Aziz et al., 2018; Karnin et al., 2013; Jamieson et al., 2014; Garivier and Kaufmann, 2016b; Chen et al., 2017b; Simchowitz et al., 2017; Hassidim et al., 2020) and reinforcement learning (Even-Dar et al., 2006; Azar et al., 2017; Kaufmann et al., 2021). Even more, the Monte-Carlo tree search with tree depth 1 is also a pure exploration problem (Kocsis and Szepesvári, 2006). Even if we only consider multi-armed bandits, there are various types of problems that have been formulated. Our investigation focuses on the standard K-armed bandits model. But we note here pure exploration has also been studied in structured bandits like kernel bandits (Camilleri et al., 2021) and linear bandits (Soare et al., 2014; Zhu et al., 2022).

Started from Even-Dar et al. (2006); Mannor and Tsitsiklis (2004), pure exploration in K-armed bandits is studied with the celebrated (ε, δ)-PAC framework. Mannor and Tsitsiklis (2004) show a lower bound, which matches the upper bound of Median Elimination in Even-Dar et al. (2006). The recent work by Hassidim et al. (2020) further shows that it is possible to achieve $\frac{n}{2\varepsilon^2}\log\frac{1}{\delta}$ asymptotically, where the constant factor of $1/2$ is the optimal one. Aziz

et al. (2018) study (ε, δ)-PAC for the infinitely-armed bandits model, where they model the infinitely-armed bandits setting as the arm reservoir. Among them, Even-Dar et al. (2006); Hassidim et al. (2020) focus on the worst-case guarantee. Aziz et al. (2018) provide more instance-dependent bounds.

Best arm identification. The best arm identification is the factually dominating setting for studying the pure exploration problem. Specifically, there are two problem setups of the best arm identification, fixed budget (Audibert and Bubeck, 2010; Karnin et al., 2013; Carpentier and Locatelli, 2016) and fixed confidence (Karnin et al., 2013; Jamieson et al., 2014; Garivier and Kaufmann, 2016b; Chen et al., 2017b; Simchowitz et al., 2017; Degenne and Koolen, 2019). In the fixed budget setting, the algorithm takes a budget as the input and is required to return an output before exhausting all the sample budget. The paper by Audibert and Bubeck (2010) is the first one to open up this direction and proposes Successive Rejects algorithm, which achieves the optimality up to logarithmic factors. The follow-up work of Karnin et al. (2013) proposes SH algorithm, which is also an elimination-based algorithm as Successive Reject, but it eliminates empirically bad arms more aggressively. It was not until Carpentier and Locatelli (2016) SH was proven to be optimal up to doubly-logarithmic factors.

Many meaningful results have been developed for the fixed confidence setting. To our knowledge, the paper by Garivier and Kaufmann (2016b) is the first one that claims asymptotic optimality. They propose a non-asymptotic lower bound for the sample complexity and also an algorithm called Track-and-Stop that matches the lower bound asymptotically as δ goes to 0. This result is further extended to the setting where the best arm is not unique (Degenne and Koolen, 2019). Very recently, Balagopalan* et al. (2025) show that there are much more can be improved for the fixed confidence setting. Specifically, they discover that some popular fixed confidence algorithms, such as Successive Elimination of Even-Dar et al. (2006) can run into the situation where the algorithm never stops. A new performance requirement called Exponential-tailed stopping time is proposed to avoid such a situation, and they show that even some simple algorithms such as the doubling Sequential Halving of Zhao et al. (2023) can achieve an Exponential-tailed stopping time. They further propose

a meta algorithm that can transform any fixed confidence algorithm to a stronger one that achieves the Exponential-tailed stopping time.

Simple regret. In our opinion, simple regret is the least understood measure for pure exploration. Though many works claim simple regret as their target, their performance bounds are either on best arm misidentification or (ε, δ)-PAC identification. Of course, by Audibert and Bubeck (2010), one can always use the probability of best arm misidentification to upper bound simple regret. However, in the data-poor regime, i.e., $t < n$, the probability of best arm misidentification is vacuous since we cannot even guarantee each arm has been sampled at least once. For (ε, δ)-PAC identification, the algorithm usually requires ε as the input of the algorithm. However, simple regret itself is a parameter-free performance measure. Perhaps only Carpentier and Valko (2015) truly deals with simple regret directly. The algorithm of Carpentier and Valko (2015) works for the simple regret minimization problem since it does not take such a predefined threshold ε as the input. There is another different problem setup in which the algorithm is required to output one of the top-m arms with ε slackness (Chaudhuri and Kalyanakrishnan, 2017, 2019). Though such a problem setup also characterizes how close the output is to the best one, it still needs m and ε as the input of the algorithm. Recent work by De Heide et al. (2021) proposes an algorithm that guarantees that it returns one of the m best arms with the same mean reward without knowledge of m as input to the algorithm. The algorithm of De Heide et al. (2021) works for both finite and infinite arms settings. However, in light of its limited problem setting that requires the top m arms to have the same mean reward, it is unclear to us if the result can be extended to guarantee simple regret or ε-error probability bound as we do. The model of this chapter can be regarded as a generalization of De Heide et al. (2021) because all the results of De Heide et al. (2021) can be matched with our results up to logarithmic factors if we consider their setting, but our model does not require the identity among the optimal arms. Furthermore, we provide analysis on ε-error probability, a generalization of misidentification probability, and analysis on simple regret.

(ε, δ)-Verifiable and unverifiable sample complexity. The fixed confidence best arm identification shares some similarities with the (ε, δ)-PAC identification. For both settings, the algorithm is required to verify the output can meet the corresponding criteria, simple

regret of 0 or simple regret of ε, with error probability δ. Verifiability requires the algorithm to pull each arm at least once. Therefore the (ε, δ)-PAC is inherently impossible for the data-poor regime, e.g., $t < n$. Katz-Samuels and Jamieson (2020) argue (ε, δ)-PAC is not aligned with practical usage. They propose the (ε, δ)-unverifiable sample complexity, which does not require the algorithm to verify the output is ε-good. Instead, the (ε, δ)-unverifiable sample complexity represents the number of samplings that the algorithm minimally needs such that it has the ability to output an ε-good arm. Note the sample complexity of a fixed budget algorithm (or, to be more accurate, the doubling trick version of the fixed budget algorithm) also captures the nature of the (ε, δ)-unverifiable sample complexity. In the reinforcement learning setting, Dann et al. (2017) also discuss the limitation of (ε, δ)-PAC and propose a notion called Uniform-PAC.

Top-k arm identification. The top-k arm identification problem requires the algorithm to return k arms with the highest mean reward instead of only the best one (Kalyanakrishnan and Stone, 2010; Kalyanakrishnan et al., 2012; Chen et al., 2017a; Simchowitz et al., 2017). The seminal work is Kalyanakrishnan and Stone (2010). Their problem formulation is a direct extension of the (ε, δ)-PAC identification problem. The goal is to return k arms that have mean rewards no less than $\mu_k - \varepsilon$. The worst-case lower bound is shown in Kalyanakrishnan et al. (2012). The first instance-dependent lower bound for the k-identification problem is given by Simchowitz et al. (2017) and Chen et al. (2017a) almost at the same time, together with the near-optimal algorithms. Zhou et al. (2014) proposes the aggregate regret to measure the performance of the algorithms for top-k arm identification, which is defined as the gap between the average reward of the top-k arms and that of the output. The best arm identification can be regarded as a special case of the top-k arm identification problem. The similarity is that we are both interested in good arms, but we are measuring the likelihood of identifying one arm out of the top-k rather than the set of top-k arms.

2.8 Discussion on the Practical Algorithm Implementation

The design of DSH (Algorithm 2) and BSH (Algorithm 3) involves two key ideas, bracketing and doubling trick. In practice, both two techniques could be implemented in a more efficient way. In addition, the base algorithm SH (Algorithm 1) has a commonly used im-

plementation for reusing samples (Baharav and Tse, 2019b; Jun and Nowak, 2016). We summarize these strategies as follows for the practitioner's consideration. However, they do not make any order-wise difference (not more than logarithmic factors) in terms of the theoretical guarantee.

- SH: The whole procedure of SH is divided into $\log_2 n$ stages. All the stages are independent of each other since the samples of prior stages are abandoned when a new stage starts, as described in Algorithm 1. In practical usage, one can keep all the samples since the first stage for the surviving arms.

- DSH: The common doubling trick does not require keeping the samples after finishing one invocation of the base algorithm. Not requiring to keep the samples is convenient to implement as we only need to repeatedly initialize a new instance of the base algorithm with the doubled budget parameter. In fact, for multi-armed bandit problems, keeping all the samples of prior invocations is beneficial. For the implementation, one can create a class for the base algorithm SH, and we initialize a new instance of the class with initial empirical rewards equal to the empirical rewards saved in the prior instance.

- BSH: The bracketing technique does not promise to avoid overlap with the already-opened brackets. A more practical implementation of BSH is to share the samples of the same arm across different brackets. Such that the empirical reward is more accurate. However, reusing sampling across different brackets is meaningless if we consider the infinitely-armed bandit models. Because, for the infinitely-armed bandit models, we will not draw exactly the same arm more than once. Thus it is barely possible to have overlap among the opened brackets.

Note our implementation reported in section 2.5 only uses the first reusing strategy for the base algorithm SH.

2.9 Proofs

2.9.1 Implication of Theorem 4 for $m = 1$

The following theorem corresponds to the implication of $m = 1$ of Theorem 4. However, we prove it independently here.

Theorem 7. *For any $\varepsilon \in (0,1)$, the error probability of SH for identifying an ε-good arm satisfies,*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) \le 3\log_2 n \cdot \exp\left(-\frac{\varepsilon^2}{32}\frac{T}{n\log_2 n}\right).$$

Proof. To avoid redundancy and for the sake of readability, we assume n is of a power of 2. It is easy to verify the result for any n. Let $\varepsilon_1 = \varepsilon/4$, $T' := \frac{T}{\log_2 n}$. And define $\varepsilon_{\ell+1} = \frac{3}{4}\cdot\varepsilon_\ell$. For each stage ℓ, define the event G_ℓ as

$$G_\ell := \left\{\max_{i\in S_{\ell+1}} \mu_i \ge \max_{i\in S_\ell} \mu_i - \varepsilon_\ell\right\}.$$

Thus as long as $\bigcap_{\ell=1}^{\log_2 n} G_\ell$ happens, we have that the arm returned after the final stage is an ε-good arm, because

$$\sum_{\ell=1}^{\log_2 n} \varepsilon_\ell < \sum_{\ell=1}^{\infty}\left(\frac{3}{4}\right)^{\ell-1}\cdot\varepsilon_1 = \frac{\varepsilon}{4}\sum_{\ell=1}^{\infty}\left(\frac{3}{4}\right)^{\ell-1} \le \frac{\varepsilon}{4}\lim_{n\to\infty}\frac{1-(3/4)^n}{1-3/4} = \varepsilon.$$

Thus, by a union bound,

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) &\le \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\log_2 n} G_\ell\right)^c\right)\\ &\le \sum_{\ell=1}^{\log_2 n}\mathbb{P}\big(G_\ell^c\big). \end{aligned} \tag{2.5}$$

Let a_ℓ be the best arm in S_ℓ,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c\big) &= \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\big) \\ &\leq \mathbb{P}\big(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\big). \\ &\leq \exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) + \mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\big).\end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\big) &\leq \mathbb{P}\Big(\big|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\big| \geq |S_\ell|/2\Big) \\ &\overset{(a_1)}{\leq} \frac{\mathbb{E}\Big[\big|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\big|\Big]}{|S_\ell|/2} \\ &\leq \frac{|S_\ell| \exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right)}{|S_\ell|/2} \\ &= 2\exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right).\end{aligned}$$

For (a_1), we use Markov's inequality. Then,

$$\mathbb{P}\big(G_\ell^c\big) \leq 3\exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) = 3\exp\left(-\left(\frac{9}{16}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{2^{-(\ell-1)}n}\right) = 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{n}\right).$$

Taking the above into (2.5), we have

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) &\leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\big(G_\ell^c\big) \\ &\leq \sum_{\ell=1}^{\log_2 n} 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{n}\right) \\ &\leq 3\log_2 n \cdot \exp\left(-\frac{\varepsilon^2}{32}\frac{T}{n\log_2 n}\right).\end{aligned}$$

□

2.9.2 *Proof of Corollary 4.1*

We state the following result that includes two different budget allocation schemes for SH. The Section 2.9.12 should be read first where we define the two budget allocation schemes:

$$\text{Option 1: } T_\ell = \left\lfloor \frac{T}{|S_\ell|\lceil \log_2 n \rceil} \right\rfloor, \quad \text{Option 2: } T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor.$$

Corollary 4.1. *The simple regret of SH satisfies*

$$\textit{Option 1: } \mathsf{SReg}_T \le \mathcal{O}\left(\sqrt{\frac{n \log^3 n}{T}}\right), \quad \textit{Option 2: } \mathsf{SReg}_T \le \mathcal{O}\left(\sqrt{\frac{n}{T}}\right).$$

Proof. The simple regret can be calculated by considering the following integral with respect to ε. We use SH with Option 2. For Option 1, the same analysis holds.

$$\begin{aligned}
\mathsf{SReg}_T &= \int_0^\infty \mathbb{P}\left(\mu_1 - \mu_{J_T} > \varepsilon\right) \mathrm{d}\varepsilon \\
&= \int_0^\infty \left(1 \wedge \exp\left(-\varepsilon^2 \frac{T}{n}\right)\right) \mathrm{d}\varepsilon \\
&\le \int_0^{\sqrt{\frac{n}{T}}} 1 \, \mathrm{d}\varepsilon + \int_0^\infty \exp\left(-\varepsilon^2 \frac{T}{n}\right) \mathrm{d}\varepsilon \\
&\le \sqrt{\frac{n}{T}} + \int_0^\infty \exp\left(-\varepsilon^2 \frac{T}{n}\right) \mathrm{d}\varepsilon.
\end{aligned}$$

We can borrow the result from Gaussian distribution with mean 0 and variance σ^2,

$$\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x = \frac{1}{2}.$$

Taking $\frac{1}{2\sigma^2} = \frac{T}{n}$, we have

$$\int_0^\infty \exp\left(-\varepsilon^2 \frac{T}{n}\right) \mathrm{d}\varepsilon = \frac{\sigma\sqrt{2\pi}}{2} = \frac{\sqrt{\pi}}{2}\sqrt{\frac{n}{T}}.$$

Thus, we have

$$\mathsf{SReg}_T \le \mathcal{O}\left(\sqrt{\frac{n}{T}}\right).$$

□

2.9.3 Proof of Theorem 4

Theorem 4 is an equivalent result to Theorem 8, so the proof of one implies the other. While Theorem 8 takes a form that is easier to instantiate the bound for specific instances, Theorem 4 takes a form that is easier to prove. Thus, we choose to prove Theorem 4 directly, leaving the proof of Theorem 8 as a consequence of Theorem 4.

Let us first provide an intuitive explanation. Imagine one possible 'typical' scenario where, in each stage, the set of surviving arms for the next stage happens to maintain the fraction of good arms as at least m/n. This means that, after $\Theta(\log_2 m)$ stages, we expect to have at least one good arm in the surviving arm set. The number of surviving arms at that time is around n/m. At this point, the rest of the procedure can be analyzed by the standard analysis of SH where the goal is to find the arm that is the best in the current surviving arm set. Thus, it remains to bound how likely it is to have the 'typical' event above. That is, we would like to bound the failure probability of this event at each stage as tightly as possible. Bounding this failure probability is our key technical innovation (Proposition 9), which promotes the fast rate of the failure probability that improves as a function of the number of good arms. Another key technical step of the proof is a careful design of events that would guarantee the chosen arm's suboptimality in the last stage to be at most ε, which requires splitting ε into smaller pieces to be distributed over the stages.

Theorem 4. *For any $m \le n$ and any $\varepsilon \in (0,1)$, the error probability of SH for identifying an (m,ε)-good arm satisfies*

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \le \log_2 n \cdot \exp\left(-\mathrm{const}\cdot m\left(\frac{\varepsilon^2 T}{4n\log_2^2(2m)\log_2 n} - \ln(4e)\right)\right).$$

Thus, there exist a positive constant c_1 *such that for* $T \geq c_1 \frac{\left(\ln(4e)+\ln\ln n\right) n \log_2^2(2m)\log_2 n}{\varepsilon^2} = \tilde{\Theta}\left(\frac{n}{\varepsilon^2}\right)$,

$$\mathbb{P}\left(\mu_{J_T} < \mu_m - \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(m\frac{\varepsilon^2 T}{n}\right)\right),$$

where, $\tilde{\Theta}$ *means ignoring the logarithmic factors of* m, n *and constants.*

Proof. Let us consider the case of $m \leq n/2$ first.

Let $\ell^* = \lceil \log_2 m \rceil$, $\varepsilon' = \frac{\varepsilon}{2\log_2 m}$, $T' := \left\lfloor \frac{T}{\log_2 n} \right\rfloor$. And g_ℓ denotes the number of $(m, \ell\cdot\varepsilon')$-good arms after finishing stage $\ell \in \left[\lceil \log_2 m \rceil\right]$. For stage ℓ, we define the event G_ℓ as,

$$G_\ell := \left\{ g_\ell \geq 2^{-\ell}\cdot m \right\}.$$

Specifically, we have G_{ℓ^*} to be the event that the number of $(m, \varepsilon/2)$-good arms after finishing stage ℓ^* is at least 1. We define G_{ℓ^*+1} as the algorithm succeeds to return an arm in $\mathrm{Top}_m(\varepsilon)$,

$$G_{\ell^*+1} := \left\{ \mu_{J_T} \geq \mu_m - \varepsilon \right\}.$$

Then the event $\bigcap_{\ell=1}^{\ell^*+1} G_\ell$ is a possible path the algorithm returns an arm in $\mathrm{Top}_m(\varepsilon)$ in the end. Thus the probability of missing all of the (m, ε)-good arms can be upper bounded as follows. Define $F_\ell = \left(\bigcap_{i=1}^{\ell-1} G_i\right) \cap G_\ell^c$ for $\ell > 1$, $F_1 = G_1^c$. Since $\bigcap_{\ell=1}^{\ell^*+1} G_\ell \subset G_{\ell^*+1}$ implies $G_{\ell^*+1}^c = \left\{\mu_{J_T} < \mu_m - \varepsilon\right\} \subset \left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c$,

$$\begin{aligned}
\mathbb{P}\left(\mu_{J_T} < \mu_m - \varepsilon\right) &\leq \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c\right) \\
&= \mathbb{P}\left(\bigcup_{\ell=1}^{\ell^*+1} F_\ell\right) \\
&\overset{(a_1)}{\leq} \sum_{\ell=1}^{\ell^*+1} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) \\
&= \sum_{\ell=1}^{\ell^*} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) + \mathbb{P}\left(G_{\ell^*+1}^c \mid \bigcap_{i=1}^{\ell^*} G_i\right). \qquad (2.6)
\end{aligned}$$

For (a_1), we use $\mathbb{P}(A, B) \leq \mathbb{P}(A|B)$.

For the first term of (2.6), we apply the result of Proposition 9 to each stage of SH with the parameters therein as $k = 2^{-\ell}\cdot m, s = 2^{-\ell}\cdot n$, $m' = 2^{-\ell+1}\cdot m$, $\varepsilon = \varepsilon'$, and $T = T'$. Thus, for stage ℓ,

$$
\begin{aligned}
&\mathbb{P}\big(G_\ell^c \mid G_{\ell-1}\big) \\
&\le \binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m}\exp\left(-2^{-\ell}\cdot m\frac{\varepsilon'^2T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) + \binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)}\exp\left(-2^{-\ell}\cdot (n-m)\frac{\varepsilon'^2T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\overset{(a_3)}{\le} h_{\ell,1}\exp\left(-2^{-\ell}\cdot m\frac{\varepsilon'^2T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) + h_{\ell,2}\exp\left(-2^{-\ell}\cdot (n-m)\frac{\varepsilon'^2T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\le h_{\ell,1}\exp\left(-\frac{m}{2}\frac{\varepsilon'^2T'}{2n}\right) + h_{\ell,2}\exp\left(-\frac{n-m}{2}\frac{\varepsilon'^2T'}{2n}\right) \\
&\overset{(a_4)}{\le} h_{\ell,1}\exp\left(-\frac{m}{2}\frac{\varepsilon'^2T'}{2n}\right) + h_{\ell,2}\exp\left(-\frac{n}{4}\frac{\varepsilon'^2T'}{2n}\right) \\
&\le \exp\left(-\frac{m}{2}\frac{\varepsilon^2T}{2n\log_2^2 m\log_2 n} + \ln(2e)2^{-\ell}\cdot m\right) \\
&\quad + \exp\left(-\frac{n}{4}\frac{\varepsilon^2T}{2n\log_2^2 m\log_2 n} + \ln(4e)2^{-\ell}\cdot (n-m)\right) \\
&\le \exp\left(-m\left(\frac{\varepsilon^2T}{4n\log_2^2 m\log_2 n} - 2\ln(4e)2^{-\ell}\right)\right) \\
&\quad + \exp\left(-\frac{n}{2}\left(\frac{\varepsilon^2T}{4n\log_2^2 m\log_2 n} - 2\ln(4e)2^{-\ell}\right)\right) \\
&\le 2\exp\left(-m\left(\frac{\varepsilon^2T}{4n\log_2^2 m\log_2 n} - 2\ln(4e)2^{-\ell}\right)\right)
\end{aligned}
\tag{2.7}
$$

where (a_3) is $\binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m} \le (2e)^{2^{-\ell}\cdot m} =: h_{\ell,1}$ (Sterling's formula $\binom{x}{y} \le \left(\frac{ex}{y}\right)^y$) and $\binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)} \le (2en/(n-m))^{2^{-\ell}\cdot (n-m)} \le (4e)^{2^{-\ell}\cdot n} =: h_{\ell,2}$ and (a_4) is by the assumption $m \le n/2$.

Then we have,

$$
\sum_{\ell=1}^{\ell^*}\mathbb{P}\big(G_\ell^c \mid G_{\ell-1}\big) \le \text{const}\cdot\log_2 m\cdot\exp\left(-m\left(\frac{\varepsilon^2T}{4n\log_2^2 m\log_2 n} - \ln(4e)\right)\right). \tag{2.8}
$$

For the second term of (2.6), $G_{\ell^*+1}^c \mid G_{\ell^*}$ indicates the events where SH fails to return an (m,ε)-good arm given the event that there is at least one $(m,\varepsilon/2)$-good arm after finishing

stage ℓ^*. Note the stages from ℓ^* to the last stage can be regarded as a whole process of SH with a budget $\big(\lceil \log_2 n \rceil - \ell^*\big)T' \geq \text{const}\cdot \log_2\big(\frac{n}{m}\big)\frac{T}{\log_2(n)}$. The initial arms are the surviving arms after finishing stage ℓ^*. Note we have $2^{-\ell^*}\cdot n = n/m$ arms surviving, denoted by S_{ℓ^*}, after finishing stage ℓ^*. Let c be the index of the true best arm in S_{ℓ^*}. Due to the event G_{ℓ^*}, we have $\mu_c \geq \mu_m - \frac{\varepsilon}{2}$. Then, by applying the result of Theorem 7,

$$\begin{aligned}
\mathbb{P}\big(G^c_{\ell^*+1} \mid G_{\ell^*}\big) &= \mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon \mid G_{\ell^*}\big) \\
&\leq \mathbb{P}\bigg(\mu_{J_T} < \mu_c - \frac{\varepsilon}{2} \mid G_{\ell^*}\bigg) \\
&\leq 3\log_2\bigg(\frac{n}{m}\bigg)\cdot \exp\bigg(-\frac{m}{128}\frac{\varepsilon^2\big(\lceil \log_2 n \rceil - \ell^*\big)T'}{n\log_2\big(\frac{n}{m}\big)}\bigg) \\
&\leq \log_2\bigg(\frac{n}{m}\bigg)\cdot \exp\bigg(-\text{const}\cdot m\frac{\varepsilon^2 T}{n\log_2 n}\bigg)\,. \qquad (2.9)
\end{aligned}$$

Bringing (2.8),(2.9) into (2.6), we have,

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \leq \log_2 n\cdot \exp\bigg(-\text{const}\cdot m\bigg(\frac{\varepsilon^2 T}{4n\log_2^2(m)\log_2 n} - \ln(4e)\bigg)\bigg).$$

Note the above analysis is for $m > 1$. To incorporate the result of Theorem 7 for $m = 1$, we rewrite the final formula as

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \leq \log_2 n\cdot \exp\bigg(-\text{const}\cdot m\bigg(\frac{\varepsilon^2 T}{4n\log_2^2(2m)\log_2 n} - \ln(4e)\bigg)\bigg).$$

Thus, there exist a positive constant c_1 such that for $T \geq c_1\frac{(\ln(4e)+\ln\ln n)n\log_2^2(2m)\log_2 n}{\varepsilon^2} = \tilde{\Theta}\big(\frac{n}{\varepsilon^2}\big)$,

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \leq \exp\bigg(-\tilde{\Theta}\bigg(m\frac{\varepsilon^2 T}{n}\bigg)\bigg).$$

For the case of $m > n/2$, we consider

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \varepsilon\big) \leq \mathbb{P}\big(\mu_{J_T} < \mu_{n/2} - \varepsilon\big).$$

Thus, one can repeat the same analysis as above with m replaced by $n/2$. Using $m/2 \le n/2 \le m$, the statement of this theorem holds. □

The result of Theorem 4 directly improves the previous works including $\mathcal{O}\Big(\frac{n}{m\varepsilon^2}\log^2\big(\frac{1}{\delta}\big)\Big)$ of Chaudhuri and Kalyanakrishnan (2017), and $\mathcal{O}\Big(\frac{1}{\varepsilon^2}\Big(\frac{n}{m}\log\big(\frac{1}{\delta}\big)+\log^2\big(\frac{1}{\delta}\big)\Big)\Big)$ of Chaudhuri and Kalyanakrishnan (2019). They share a similar strategy that, assuming the knowledge of m, ε and δ, uniformly samples $\Theta\Big(\frac{n}{m}\log\big(\frac{1}{\delta}\big)\Big)$ arms first and then runs the Median-Elimination algorithm of Even-Dar et al. (2006) or LUCB of Kaufmann and Kalyanakrishnan (2013) for this subset.

2.9.4 Proof of Theorem 8

Theorem 8. *For any $\varepsilon \in (0,1)$, the error probability of SH for identifying an ε-good arm satisfies,*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) \le \min_{\{(m',\varepsilon'):\Delta_{m'}+\varepsilon'\le\varepsilon\}} \log_2 n \cdot \exp\left(-\text{const}\cdot m'\left(\frac{\varepsilon'^2 T}{4n\log_2^2(2m')\log_2 n} - \ln(4e)\right)\right).$$

Proof. Note the result we proved in Theorem 4 holds for any $m \le n$ and any $\varepsilon \in (0,1)$. Thus $\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big)$ can be upper bounded for any $(m',\varepsilon') \in \big\{(m',\varepsilon') : \Delta_{m'} + \varepsilon' \le \varepsilon\big\}$. The result is therefore implied. □

2.9.5 Proof of Theorem 1

Theorem 1. *For any $\varepsilon \in [0,1)$, the ε-error probability of SH satisfies*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \varepsilon\big) \le \exp\left(-\tilde{\Theta}\left(\frac{T}{\max_{i\ge g(\varepsilon)+1}\frac{i}{\Delta_i^2}\cdot\frac{1}{g\left(\frac{\varepsilon}{2}\right)}}\right)\right),$$

for $T \ge \tilde{\Theta}\Big(\max_{i\ge g(\varepsilon)+1}\frac{i}{\Delta_i^2}\Big)$.

Proof. To avoid redundancy and for the sake of readability, we assume n is of a power of 2. It is easy to verify the result for any n. Recall $g\big(\frac{\varepsilon}{2}\big)$ is the number of $\frac{\varepsilon}{2}$-good arms in the instance, and $g_\ell\big(\frac{\varepsilon}{2}\big)$ is the number of $\frac{\varepsilon}{2}$-good arms at the beginning of stage ℓ. Thus

$g_1\left(\frac{\varepsilon}{2}\right) = g\left(\frac{\varepsilon}{2}\right)$. Let $\ell' = \max\left\{1 \le \ell \le \log_2(n) - 2 \mid \Delta_{\frac{n}{2^{\ell+1}}} > \varepsilon\right\}$. There are three cases we need to consider:

- $\ell' = \log_2 n - 2$.
- $1 \le \ell' < \log_2 n - 2$.
- Such an ℓ' does not exist.

The last case that such a ℓ' does not exit means $\varepsilon \ge \Delta_{\frac{n}{4}}$, which means that the instance has $\Theta(n)$ amount of ε-good arms, therefore an easy instance. We address this case at the end of the proof.

Case 1. $\ell' = \log_2 n - 2$

In this case, we have $\Delta_2 > \varepsilon$, then the problem is a fixed budget best arm identification problem rather than an ε-good arm identification problem. Since $g(\varepsilon) = g\left(\frac{\varepsilon}{2}\right) = 1$ for $\Delta_2 > \varepsilon$, the theorem statement reduces to

$$\mathbb{P}\left(\mu_1 - \mu_{J_T} > \varepsilon\right) \le \exp\left(-\tilde{\Theta}\left(\frac{T}{\max_{i\ge 2} \frac{i}{\Delta_i^2}}\right)\right),$$

which is a trivial consequence of Karnin et al. (2013, Theorem 4.1) under the theorem's condition on T.

Case 2. $1 \le \ell' < \log_2 n - 2$

For $1 \le \ell' < \log_2 n - 2$, recall that $|A_\ell| = \frac{n}{2^{\ell-1}}$. One can show that

$$|A_{\ell'+1}|/4 \le g(\varepsilon) \le |A_{\ell'}|/4 \quad \text{and} \quad \Delta_{|A_{\ell'+1}|/4} \le \Delta_{g(\varepsilon)} \le \varepsilon < \Delta_{|A_{\ell'}|/4}\ . \tag{2.10}$$

For $\ell \in [\ell' + 1]$, define the good event for stage ℓ as

$$G_\ell := \left\{ g_\ell\left(\frac{\varepsilon}{2}\right) \ge g\left(\frac{\varepsilon}{2}\right) - \left\lceil (\ell - 1)\frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil + 1 \right\}.$$

Note G_1 holds with probability 1. Define $m' := \left\lceil \frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil = \tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right)\right)$. Intuitively, we allow at most m' of $\frac{\varepsilon}{2}$-good arms to be eliminated per stage in the initial phases, i.e., $\ell \in [\ell' + 1]$ of

running a Sequential Halving. The ε-error probability can be expressed as

$$\mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon\Big) = \mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon, \cap_{\ell=2}^{\ell'+1} G_\ell\Big) + \mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon, \Big(\cap_{\ell=2}^{\ell'+1} G_\ell\Big)^c\Big).$$

Since

$$\mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon, \cap_{\ell=2}^{\ell'+1} G_\ell\Big) \le \mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon \mid \cap_{\ell=2}^{\ell'+1} G_\ell\Big),$$

and

$$\mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon, \Big(\cap_{\ell=2}^{\ell'+1} G_\ell\Big)^c\Big) \le \mathbb{P}\Big(\Big(\cap_{\ell=2}^{\ell'+1} G_\ell\Big)^c\Big) \le \sum_{\ell=1}^{\ell'} \mathbb{P}\big(G_{\ell+1}^c \mid \cap_{i=1}^{\ell} G_i\big),$$

we have

$$\mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon\Big) \le \mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon \mid \cap_{\ell=2}^{\ell'+1} G_\ell\Big) + \sum_{\ell=1}^{\ell'} \mathbb{P}\big(G_{\ell+1}^c \mid \cap_{i=1}^{\ell} G_i\big). \tag{2.11}$$

We bound the two terms in (2.11) respectively.

We start with the second term. Let $\ell \in [\ell']$. To bound $\mathbb{P}\big(G_{\ell+1}^c \mid \cap_{i=1}^{\ell} G_i\big)$ with Lemma 9.1, we need to relate the quantities in stage ℓ with those in Lemma 9.1.

We first claim that, given $\cap_{i=1}^{\ell} G_i$, the event $G_{\ell+1}^c$ implies that there are at least m' of $\frac{\varepsilon}{2}$-good arms that are eliminated during the stage ℓ, i.e., $g_\ell(\frac{\varepsilon}{2}) - g_{\ell+1}(\frac{\varepsilon}{2}) \ge m'$.

To see why, we have by $\cap_{i=1}^{\ell} G_i$,

$$g_\ell\left(\frac{\varepsilon}{2}\right) \ge g\left(\frac{\varepsilon}{2}\right) - \left\lceil (\ell-1)\frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil + 1,$$

and by $G_{\ell+1}^c$,

$$g_{\ell+1}\left(\frac{\varepsilon}{2}\right) \le g\left(\frac{\varepsilon}{2}\right) - \left\lceil \ell\frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil.$$

Therefore, using $\lceil x+y\rceil - \lceil x\rceil + 1 \geq \lceil y\rceil$, we have

$$g_\ell(\frac{\varepsilon}{2}) - g_{\ell+1}(\frac{\varepsilon}{2}) = \left\lceil \ell \frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil + 1 - \left\lceil (\ell-1)\frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil \geq \left\lceil \frac{g\left(\frac{\varepsilon}{2}\right)}{2\log_2 n} \right\rceil = m' \ ,$$

which proves the claim.

Define A_ℓ as the set of surviving arms at the beginning of stage ℓ, $A_{\ell,i}$ as i-th best arm in A_ℓ (ties are broken arbitrarily), and $A_\ell[L+1:R] \coloneqq \left\{A_{\ell,i}, i \in [R]\setminus[L]\right\}$. Let $d_\ell(\varepsilon)$ be the smallest gap of the mean rewards between any two arms chosen from $A_\ell\left[\frac{n}{2^{\ell+1}}+1 : \frac{n}{2^{\ell-1}}\right]$ and $A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]$ respectively, i.e.,

$$d_\ell(\varepsilon) = \min\left\{\mu_i - \mu_j \mid j \in A_\ell\left[\frac{n}{2^{\ell+1}}+1 : \frac{n}{2^{\ell-1}}\right], i \in A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]\right\}.$$

The definition of $d_\ell(\varepsilon)$ is same as d_q of Lemma 9.1. By Lemma 8.1, we have $d_\ell(\varepsilon) \geq \frac{1}{2}\Delta_{\frac{n}{2^{\ell+1}}}$. Define $S_\ell(\varepsilon) \coloneqq \frac{\Delta^2_{\frac{n}{2^{\ell+1}}}}{n\cdot 2^{-\ell+1}}$.

We can now apply Lemma 9.1 with $T = T', s = \frac{n}{2^\ell}, k = m'$, and $q = \frac{n}{2^{\ell+1}}$ as follows:

$$\begin{aligned}
&\mathbb{P}\left(G^c_{\ell+1} \mid \cap_{i=1}^{\ell} G_i\right) \\
&\leq \binom{g_\ell\left(\frac{\varepsilon}{2}\right)}{m'} \exp\left(-m'\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}}\right) + \binom{\frac{3}{4}n\cdot 2^{-\ell+1}}{\frac{1}{4}n\cdot 2^{-\ell+1}+m'} \exp\left(-\left(\frac{1}{4}n\cdot 2^{-\ell+1}+m'\right)\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}}\right) \\
&\overset{a_1}{\leq} \binom{g_\ell\left(\frac{\varepsilon}{2}\right)}{m'} \exp\left(-m'\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}}\right) + \binom{n\cdot 2^{-\ell+1}}{n\cdot 2^{-\ell}} \exp\left(-\frac{n}{2^{\ell+1}}\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}}\right) \\
&\overset{a_2}{\leq} \exp\left(-m'\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}} + m'\ln\left(\frac{eg_\ell\left(\frac{\varepsilon}{2}\right)}{m'}\right)\right) + \exp\left(-\frac{n}{2^{\ell+1}}\frac{d^2_\ell(\varepsilon)T'}{8n\cdot 2^{-\ell+1}} + \frac{n}{2^\ell}\ln(2e)\right) \\
&\leq \exp\left(-m'\frac{\Delta^2_{\frac{n}{2^{\ell+1}}}T'}{32n\cdot 2^{-\ell+1}} + m'\ln\left(\frac{eg_\ell\left(\frac{\varepsilon}{2}\right)}{m'}\right)\right) + \exp\left(-\frac{n}{2^{\ell+1}}\frac{\Delta^2_{\frac{n}{2^{\ell+1}}}T'}{32n\cdot 2^{-\ell+1}} + \frac{n}{2^\ell}\ln(2e)\right) \quad (d_\ell(\varepsilon) \geq \tfrac{1}{2}\Delta_{\frac{n}{2^{\ell+1}}}) \\
&\overset{a_3}{\leq} \exp\left(-\tilde{\Theta}\left(m'S_\ell(\varepsilon)T\right)\right) + \exp\left(-\tilde{\Theta}\left(\frac{n}{2^{\ell+1}}S_\ell(\varepsilon)T\right)\right), \qquad \left(T \geq \tilde{\Theta}\left(\tfrac{1}{S_\ell(\varepsilon)}\right)\right)
\end{aligned} \tag{2.12}$$

where,

- b_1 is by $\binom{x}{y} < \binom{x'}{y}$ for $x' > x$ and $\max_y \binom{x}{y} = \binom{x}{\lfloor x/2 \rfloor} = \binom{x}{\lceil x/2 \rceil}$,
- b_2 is by upper bounding the binomial coefficients by Sterling's formula $\binom{x}{y} \le \left(\frac{ex}{y}\right)^y$ and moving the coefficients into the exponent.
- b_3 is by $T \ge \tilde{\Theta}\left(\frac{1}{S_\ell(\varepsilon)}\right)$ which is implied by $T \ge \tilde{\Theta}\left(\max_{i \ge g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)$ in the theorem statement.

By the definition of ℓ' and the fact of $\ell \in [\ell']$, we have

$$\Delta_{\frac{n}{2^{\ell+1}}} > \varepsilon.$$

By the definition of $g\left(\frac{\varepsilon}{2}\right)$, we have

$$\Delta_{g\left(\frac{\varepsilon}{2}\right)} \le \frac{\varepsilon}{2}.$$

Thus

$$\Delta_{\frac{n}{2^{\ell+1}}} - \Delta_{g\left(\frac{\varepsilon}{2}\right)} > \frac{\varepsilon}{2},$$

which means $\frac{n}{2^{\ell+1}} > g\left(\frac{\varepsilon}{2}\right) \ge m'$. We continue to upper bound (2.12). For $T \ge \tilde{\Theta}\left(\frac{1}{S_\ell(\varepsilon)}\right)$, we have

$$\begin{aligned}
\mathbb{P}\left(G_{\ell+1}^c \mid \cap_{i=1}^{\ell} G_i\right) &\le \exp\left(-\tilde{\Theta}\left(m' S_\ell(\varepsilon) T\right)\right) + \exp\left(-\tilde{\Theta}\left(\frac{n}{2^{\ell+1}} S_\ell(\varepsilon) T\right)\right) \\
&\le \exp\left(-\tilde{\Theta}\left(m' S_\ell(\varepsilon) T\right)\right) + \exp\left(-\tilde{\Theta}\left(m' S_\ell(\varepsilon) T\right)\right) \\
&\le \exp\left(-\tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right) S_\ell(\varepsilon) T\right)\right). \qquad \left(m' = \tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right)\right)\right)
\end{aligned} \tag{2.13}$$

Next, we bound the first term in (2.11). The stages $\ell \ge \ell' + 1$ can be viewed as a new Sequential Halving game with $A_{\ell'+1}$ as the initial arm set, $g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)$ of which are $\frac{\varepsilon}{2}$-good.

When $\cap_{i=1}^{\ell'+1} G_i$ happens, we have, as $\ell' < \log_2 n$,

$$g_{\ell'+1}\Big(\frac{\varepsilon}{2}\Big) \geq g\Big(\frac{\varepsilon}{2}\Big) - \left\lceil \ell' \frac{g\big(\frac{\varepsilon}{2}\big)}{2\log_2 n} \right\rceil + 1 \geq g\Big(\frac{\varepsilon}{2}\Big) - \left\lceil \frac{g\big(\frac{\varepsilon}{2}\big)}{2} \right\rceil + 1 \geq \frac{g\big(\frac{\varepsilon}{2}\big)}{2} = \tilde{\Theta}\Big(g\Big(\frac{\varepsilon}{2}\Big)\Big).$$

By Lemma 8.2, a $\Big(g_{\ell'+1}\big(\frac{\varepsilon}{2}\big), \frac{\Delta_{g(\varepsilon)+1}}{2}\Big)$-good arm in $A_{\ell'+1}$ returned after finishing the final stage of Sequential Halving is an ε-good arm. We define $\mu_m(A)$ as the m-th largest value in $\{\mu_j : j \in A\}$. We apply Theorem 4 and set m, n, ε of Theorem 4 as $m = g_{\ell'+1}\big(\frac{\varepsilon}{2}\big), n = |A_{\ell'+1}|, \varepsilon = \frac{\Delta_{g(\varepsilon)+1}}{2}$. To apply Theorem 4, we need the condition of $T \geq \tilde{\Theta}\big(\frac{n}{\varepsilon^2}\big)$ in Theorem 4 to be true, which means we need $T \geq \tilde{\Theta}\Big(\frac{|A_{\ell'+1}|}{\Delta^2_{g(\varepsilon)+1}}\Big)$. Note by (2.10), we have $\frac{n}{2^{\ell'+2}} \leq g(\varepsilon)$, thus $|A_{\ell'+1}| = \frac{n}{2^{\ell'}} \leq 4g(\varepsilon) \leq 4\big(g(\varepsilon)+1\big)$. Using this bound and the condition on T from the theorem statement, we have $T \geq \tilde{\Theta}\Big(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\Big) \geq \tilde{\Theta}\Big(\frac{g(\varepsilon)+1}{\Delta^2_{g(\varepsilon)+1}}\Big) \geq \tilde{\Theta}\Big(\frac{|A_{\ell'+1}|}{\Delta^2_{g(\varepsilon)+1}}\Big)$ So, the condition of Theorem 4 is satisfied. Then, we have

$$\begin{aligned}
&\mathbb{P}\Big(\mu_1 - \mu_{J_T} > \varepsilon \mid \cap_{i=1}^{\ell'+1} G_i\Big) \\
\leq\ &\mathbb{P}\left(\mu_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) - \mu_{J_T} > \frac{1}{2}\Delta_{g(\varepsilon)+1} \mid \cap_{i=1}^{\ell'+1} G_i\right) && \text{(Lemma 8.2)} \\
\leq\ &\exp\left(-\tilde{\Theta}\left(g_{\ell'+1}\Big(\frac{\varepsilon}{2}\Big) \cdot \frac{\Delta^2_{g(\varepsilon)+1} T'}{|A_{\ell'+1}|}\right)\right) && \Big(T \geq \tilde{\Theta}\Big(\frac{g(\varepsilon)+1}{\Delta^2_{g(\varepsilon)+1}}\Big)\Big) \\
\leq\ &\exp\left(-\tilde{\Theta}\left(g\Big(\frac{\varepsilon}{2}\Big) \cdot \frac{\Delta^2_{g(\varepsilon)+1} T'}{4\big(g(\varepsilon)+1\big)}\right)\right). && \Big(g_{\ell'+1}\big(\frac{\varepsilon}{2}\big) = \tilde{\Theta}\Big(g\big(\frac{\varepsilon}{2}\big)\Big)\Big)
\end{aligned} \tag{2.14}$$

The role of considering ℓ' is that as long as the Sequential Halving goes well ($\cap_{i=1}^{\ell'+1} G_i$ happens) the ratio of $\frac{\varepsilon}{2}$-good arms to the all surviving arms is increasing to $g\big(\frac{\varepsilon}{2}\big)/g(\varepsilon)$ after finishing stage ℓ'. Specifically, for the Polynomial(α) instance of our interest where $\Delta_i = \big(\frac{i}{n}\big)^\alpha$ with $\alpha > 0.5$, $g\big(\frac{\varepsilon}{2}\big)/g(\varepsilon)$ can be lower bounded by a constant. The last step is to combine

(2.13) and (2.14). For $T \geq \tilde{\Theta}\left(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)$, we have

$$\begin{aligned}
\mathbb{P}\left(\mu_1 - \mu_{J_T} > \varepsilon\right) &\leq \mathbb{P}\left(\mu_1 - \mu_{J_T} > \varepsilon \mid \cap_{\ell=2}^{\ell'+1} G_\ell\right) + \sum_{\ell=1}^{\ell'} \mathbb{P}\left(G_{\ell+1}^c \mid \cap_{i=1}^{\ell} G_i\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right) \cdot \frac{\Delta_{g(\varepsilon)+1}^2 T}{4\left(g(\varepsilon)+1\right)}\right)\right) + \log(n) \exp\left(-\tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right) \min_{\ell \leq \ell'} S_\ell(\varepsilon) T\right)\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(\min\left\{g\left(\frac{\varepsilon}{2}\right) \cdot \frac{\Delta_{g(\varepsilon)+1}^2}{4\left(g(\varepsilon)+1\right)}, g\left(\frac{\varepsilon}{2}\right) \min_{\ell \leq \ell'} S_\ell(\varepsilon)\right\} T\right)\right) \\
&\qquad\qquad \left(T \geq \tilde{\Theta}\left(\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}\right)\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right) \cdot \min\left\{\frac{\Delta_{g(\varepsilon)+1}^2}{4\left(g(\varepsilon)+1\right)}, \min_{\ell \leq \ell'} \frac{\Delta_{\frac{n}{2^{\ell+1}}}^2}{\frac{4n}{2^{\ell+1}}}\right\} T\right)\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(g\left(\frac{\varepsilon}{2}\right) \cdot \min_{i \geq g(\varepsilon)+1} \frac{\Delta_i^2}{i} T\right)\right) \\
&= \exp\left(-\tilde{\Theta}\left(\frac{T}{\max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} \cdot \frac{1}{g\left(\frac{\varepsilon}{2}\right)}}\right)\right).
\end{aligned}$$

Case 3. ℓ' does not exist

This case condition implies that $\Delta_{\frac{n}{4}} \leq \varepsilon$, which means the number of ε-good arms is $\Theta(n)$, so we can use (2.14) solely to claim our bound. □

Lemma 8.1. *In the case of* $1 \leq \ell' < \log_2 n - 2$, *for* $\ell \in [\ell']$, *define* $d_\ell(\varepsilon)$ *as the minimum gap of the mean rewards between any two arms chosen from* $A_\ell\left[\frac{n}{2^{\ell+1}} + 1 : \frac{n}{2^{\ell-1}}\right]$ *and* $A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]$ *respectively,*

$$d_\ell(\varepsilon) = \min\left\{\mu_i - \mu_j \mid \forall j \in A_\ell\left[\frac{n}{2^{\ell+1}} + 1 : \frac{n}{2^{\ell-1}}\right], \forall i \in A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]\right\}.$$

Then $d_\ell(\varepsilon)$ *satisfies*

$$d_\ell(\varepsilon) \geq \frac{1}{2} \Delta_{\frac{n}{2^{\ell+1}}}.$$

Proof. By definition, we have $A_\ell \subset A_1$, which implies the i-th best arm in A_ℓ has an mean reward lower than or equal to the i-th best arm in A_1. Then we have

$$\begin{aligned}\max\left\{\mu_i \mid i \in A_\ell\left[\frac{n}{2^{\ell+1}}+1 : \frac{n}{2^{\ell-1}}\right]\right\} &\le \max\left\{\mu_i \mid i \in A_1\left[\frac{n}{2^{\ell+1}}+1 : \frac{n}{2^{\ell-1}}\right]\right\} \\ &\le \mu_{\frac{n}{2^{\ell+1}}} \\ &= \mu_1 - \Delta_{\frac{n}{2^{\ell+1}}}.\end{aligned}$$

Also,

$$\min\left\{\mu_i, i \in A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]\right\} \ge \mu_1 - \frac{\varepsilon}{2}.$$

Therefore for $\forall j \in A_\ell\left[\frac{n}{2^{\ell+1}}+1 : \frac{n}{2^{\ell-1}}\right]$ and $\forall i \in A_\ell\left[1 : g_\ell\left(\frac{\varepsilon}{2}\right)\right]$,

$$\mu_i - \mu_j \ge \mu_1 - \frac{\varepsilon}{2} - \left(\mu_1 - \Delta_{\frac{n}{2^{\ell+1}}}\right) = \Delta_{\frac{n}{2^{\ell+1}}} - \frac{\varepsilon}{2}.$$

Since $\ell \in [\ell']$, by the definition of ℓ' we have

$$\Delta_{\frac{n}{2^{\ell+1}}} > \varepsilon.$$

Thus

$$\mu_i - \mu_j \ge \Delta_{\frac{n}{2^{\ell+1}}} - \frac{\varepsilon}{2} = \frac{\Delta_{\frac{n}{2^{\ell+1}}}}{2} + \frac{\Delta_{\frac{n}{2^{\ell+1}}}}{2} - \frac{\varepsilon}{2} > \frac{\Delta_{\frac{n}{2^{\ell+1}}}}{2}.$$

□

Lemma 8.2. *A $\left(g_{\ell'+1}\left(\frac{\varepsilon}{2}\right), \frac{\Delta_{g(\varepsilon)+1}}{2}\right)$-good arm in $A_{\ell'+1}$ returned after finishing the final stage of Sequential Halving is an ε-good arm.*

Proof. We first claim that

$$\Delta_{J_T} > \varepsilon \iff \Delta_{J_T} > \Delta_{g(\varepsilon)}\ .$$

To see this, for the forward direction, we have $\Delta_{J_T} > \varepsilon \geq \Delta_{g(\varepsilon)}$. For the other direction, if $\Delta_{J_T} > \Delta_{g(\varepsilon)}$ is true, then $J_T > g(\varepsilon)$. This means that $\Delta_{J_T} > \varepsilon$. Then, we can also show that

$$\Delta_{J_T} > \varepsilon \iff \Delta_{J_T} \geq \Delta_{g(\varepsilon)+1},$$

which can be shown easily by showing

$$\Delta_{J_T} \geq \Delta_{g(\varepsilon)+1} \iff \Delta_{J_T} > \Delta_{g(\varepsilon)}.$$

Define $\Delta_m(A)$ to be the m-th smallest value in $\{\Delta_i : i \in A\}$. Then, we have $\Delta_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) \leq \frac{1}{2}\varepsilon$. Using the fact that $\varepsilon < \Delta_{g(\varepsilon)+1}$, we have

$$\begin{aligned}
\Delta_{J_T} \geq \Delta_{g(\varepsilon)+1} &= \Delta_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) + \Delta_{g(\varepsilon)+1} - \Delta_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) \\
&\geq \Delta_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) + \Delta_{g(\varepsilon)+1} - \frac{1}{2}\varepsilon \\
&> \Delta_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) + \frac{1}{2}\Delta_{g(\varepsilon)+1}. && (\varepsilon < \Delta_{g(\varepsilon)+1})
\end{aligned}$$

By defining $\mu_m(A)$ as the m-th largest value in $\{\mu_j : j \in A\}$, the above equivalently means

$$\mu_{J_T} < \mu_{g_{\ell'+1}\left(\frac{\varepsilon}{2}\right)}(A_{\ell'+1}) - \frac{1}{2}\Delta_{g(\varepsilon)+1}.$$

Therefore the output is not a $\left(g_{\ell'+1}\left(\frac{\varepsilon}{2}\right), \frac{\Delta_{g(\varepsilon)+1}}{2}\right)$-good arm in $A_{\ell'+1}$. Our claim holds by contraposition. □

2.9.6 Proof of Corollary 1.1

Corollary 1.1. *For the Polynomial(α) instance where $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ with $\alpha > 0.5$, we have*

$$H_2(\varepsilon) = \max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} \cdot \frac{1}{g\left(\frac{\varepsilon}{2}\right)} < \frac{4}{\varepsilon^2}.$$

Proof.

$$\max_{i\geq g(\varepsilon)+1}\frac{i}{\Delta_i^2}\cdot\frac{1}{g\left(\frac{\varepsilon}{2}\right)}=\max_{i\geq g(\varepsilon)+1}i^{1-2\alpha}n^{2\alpha}\frac{1}{n\left(\frac{\varepsilon}{2}\right)^{1/\alpha}}<\left(n\varepsilon^{1/\alpha}\right)^{1-2\alpha}n^{2\alpha}\frac{1}{n\left(\frac{\varepsilon}{2}\right)^{1/\alpha}}<\frac{4}{\varepsilon^2}.$$

where both the last inequality and the last equality are by $\alpha>0.5$. □

2.9.7 Proof of Corollary 2.1

Corollary 2.1. *Katz-Samuels and Jamieson (2020) shows a lower bound of the sample complexity for identifying an ε-good arm that scales as*

$$\Omega\left(H^{low}(\varepsilon):=\frac{1}{g(\varepsilon)}\sum_{i=g(\varepsilon)+1}^{n}\left(\frac{1}{\Delta_i^2}\right)-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\right).$$

In the case that $g(\varepsilon)$ and $g\left(\frac{\varepsilon}{2}\right)$ have the same order ($\frac{g\left(\frac{\varepsilon}{2}\right)}{g(\varepsilon)}$ is irrelevant to ε), the sample complexity measure $H_2(\varepsilon)$ satisfies

$$H^{low}(\varepsilon)\lesssim H_2(\varepsilon)\lesssim H^{low}(\varepsilon)+\frac{2}{\Delta_{g(\varepsilon)+1}^2},$$

where $\lesssim$ hides logarithms of n and constants.

Proof.

$$\begin{aligned}
H^{\text{low}}(\varepsilon)=&\frac{1}{g(\varepsilon)}\sum_{i=g(\varepsilon)+1}^{n}\frac{1}{\Delta_i^2}-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\\
=&\frac{1}{g(\varepsilon)}\sum_{i=g(\varepsilon)+1}^{n}\frac{1}{i}\frac{i}{\Delta_i^2}-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\\
\leq&\frac{1}{g(\varepsilon)}\max_{i\geq g(\varepsilon)+1}\frac{i}{\Delta_i^2}\cdot\sum_{i=g(\varepsilon)+1}^{n}\frac{1}{i}-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\\
\overset{(a_1)}{\leq}&\frac{1}{g(\varepsilon)}\max_{i\geq g(\varepsilon)+1}\frac{i}{\Delta_i^2}\cdot\log(2n)-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\\
=&\frac{g\left(\frac{\varepsilon}{2}\right)}{g(\varepsilon)}H_2(\varepsilon)\cdot\log(2n)-\frac{1}{\Delta_{g(\varepsilon)+1}^2}\\
\leq&\frac{g\left(\frac{\varepsilon}{2}\right)}{g(\varepsilon)}H_2(\varepsilon)\cdot\log(2n),
\end{aligned}$$

for (a_1), we use $\sum_{i=g(\varepsilon)}^{n} \frac{1}{i} \leq \sum_{i\in[n]} \frac{1}{i} \leq \log(2n)$, which can be found in Audibert and Bubeck (2010). On the other hand,

$$\begin{aligned}
H^{\text{low}}(\varepsilon) =& \frac{1}{g(\varepsilon)} \sum_{i=g(\varepsilon)+1}^{n} \frac{1}{\Delta_i^2} - \frac{1}{\Delta_{g(\varepsilon)+1}^2} \\
\geq& \frac{1}{g(\varepsilon)} \max_{i\geq g(\varepsilon)+1} \frac{i - g(\varepsilon)}{\Delta_i^2} - \frac{1}{\Delta_{g(\varepsilon)+1}^2} \\
=& \frac{1}{g(\varepsilon)} \left(\max_{i\geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} - \frac{g(\varepsilon)}{\Delta_i^2} \right) - \frac{1}{\Delta_{g(\varepsilon)+1}^2} \\
\geq& \frac{1}{g(\varepsilon)} \left(\max_{i\geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} - \frac{g(\varepsilon)}{\Delta_{g(\varepsilon)+1}^2} \right) - \frac{1}{\Delta_{g(\varepsilon)+1}^2} \\
=& \frac{1}{g(\varepsilon)} \max_{i\geq g(\varepsilon)+1} \frac{i}{\Delta_i^2} - \frac{2}{\Delta_{g(\varepsilon)+1}^2} \\
=& \frac{g\left(\frac{\varepsilon}{2}\right)}{g(\varepsilon)} H_2(\varepsilon) - \frac{2}{\Delta_{g(\varepsilon)+1}^2}.
\end{aligned}$$

□

2.9.8 Proof of Theorem 3

For an n-armed bandit instance ν and a fixed value of $\varepsilon > 0$, consider a partition of the arms of ν, $\mathcal{A} = \left\{[g(\varepsilon)], [g(\varepsilon)+1 : 2{\cdot}g(\varepsilon)], \ldots, [n - g(\varepsilon) + 1 : n]\right\}$, where we avoid rounding by making the assumption that n is an integer multiple of $g(\varepsilon)$. We use $\mathcal{A}_i := [(i-1){\cdot}g(\varepsilon) : i{\cdot}g(\varepsilon)]$ to denote the subsets of the partition $\mathcal{A}$. Making use of this partition, we define

$$\tilde{H}(\nu, \varepsilon) := \sum_{i=2}^{n/g(\varepsilon)} \tilde{\Delta}_{\mathcal{A}_i}^{-2},$$

where for an arbitrary set of arms $A \subseteq [n]$ we define $\tilde{\Delta}_A := \max_{j\in A} \Delta_j$, the largest gap among the arms in $\mathcal{A}$.

Proof of Theorem 3. The proof structure follows Carpentier and Locatelli (2016, Theorem 1). Fix $n > 2$, $\varepsilon > 0$ an n-armed unit-variance gaussian instance ν and an arbitrary policy π. Let $\mathbb{P}_{\nu,\pi}$ denote the probability measure over interaction histories induced by the interaction of π with instance ν. Without loss of generality, we take the arm means of ν to be ordered

such that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$. Understanding $g(\varepsilon)$ to refer to the number of ε-good arms on ν, we define instances ν^i for $i \in [2 : n/g(\varepsilon)]$ from ν by setting $\nu^i_j = \nu_j$ for $j \notin \mathcal{A}_i$, and setting $\mu^i_j := \mu_j + 2\Delta_j$ for $j \in \mathcal{A}_i$. Each instance ν^i has $g(\varepsilon)$ ε-good arms with indices in $\mathcal{A}_i$, and $\tilde{H}(\nu^i, \varepsilon) \leq \tilde{H}(\nu, \varepsilon) = a$ for all $i \in [2 : n/g(\varepsilon)]$.

We will make use of a change of measure between ν and these alternate instances, restricted on an event on which the interaction histories on ν and ν^i are hard to distinguish from one another:

$$\mathcal{E}_i = \left\{ \sum_{t=1}^{T} \sum_{j=1}^{n} \mathbb{1}\left\{I_t = j\right\} \left[\Lambda^i_t(j) - (1+\rho)\mathsf{KL}(\nu^i_j, \nu_j)\right] \leq \frac{1}{\rho} \log(1/\delta) \right\},$$

where $\Lambda^i_t(j) = \log\left(\dfrac{\mathrm{d}\nu_j(X_t)}{\mathrm{d}\nu^i_j(X_t)}\right)$ for $i \in [2 : n/g(\varepsilon)]$, and $\rho > 0$ is a parameter to be chosen later.

Lemma 8.3 (Jun and Zhang (2022, Lemma 10))**.** *For all $i \in [2 : n/g(\varepsilon)]$ and arbitrary π, $\rho > 0$, and $\delta \in (0, 1)$,*

$$\mathbb{P}_{\nu,\pi}(\mathcal{E}_i) \geq 1 - \delta.$$

We can now prove the result. Denoting the number of times arm $j \in [n]$ has been pulled in T rounds of interaction by $N_j(T)$, let $t_i = \mathbb{E}_{\nu,\pi}\left[\sum_{j \in \mathcal{A}_i} N_j(T)\right]$ be the expected number of pulls that arm group $\mathcal{A}_i$ receives, for $i \in [2 : n/g(\varepsilon)]$. By Markov's inequality, for all $i \in [2 : n/g(\varepsilon)]$ we have

$$\mathbb{P}_{\nu,\pi}\left(\sum_{j \in \mathcal{A}_i} N_j(T) \geq 6t_i \right) \leq 1/6.$$

Defining the event

$$E_i = \left\{J_t \in \mathcal{A}_1\right\} \cap \left\{\mathcal{E}_i\right\} \cap \left\{ \sum_{j \in \mathcal{A}_i} N_j(T) \leq 6t_i \right\}$$

for $i \in [2 : n/g(\varepsilon)]$ and setting $\delta = 1/6$ in Lemma 8.3, for all $i \in [2 : n/g(\varepsilon)]$ and arbitrary π

$$\mathbb{P}_{\nu,\pi}(E_i) \geq 1 - (1/6 + 1/6 + \xi) = 2/3 - \xi,$$

where $\xi := \mathbb{P}_{\nu,\pi}(J_t \notin \mathcal{A}_1)$.

It follows that for $i \in [2 : n/g(\varepsilon)]$

$$\begin{aligned}\mathbb{P}_{\nu^i,\pi}(J_t \notin \mathcal{A}_i) &\geq \mathbb{P}_{\nu^i,\pi}(J_t \in \mathcal{A}_1) \\ &\geq \mathbb{P}_{\nu^i,\pi}(E_i).\end{aligned}$$

Making use of a standard change of measure technique for bandits (Audibert et al., 2010), we have

$$\begin{aligned}\mathbb{P}_{\nu^i,\pi}(E_i) &= \mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{E_i\}\exp\left(-\sum_{t=1}^{T}\sum_{j\in\mathcal{A}_i}\mathbb{1}\{I_t = a\}\,\Lambda_t^i(j)\right)\right] \\ &\geq \mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{E_i\}\exp\left(-\sum_{t=1}^{T}\sum_{j\in\mathcal{A}_i}\mathbb{1}\{I_t = j\}\,2\mathsf{KL}(\nu_j^i,\nu_j) - \log(6)\right)\right] \\ &\qquad\qquad \text{(on } E_i\text{, taking } \rho = 1 \text{ and } \delta = 1/6) \\ &\geq \mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{E_i\}\exp\left(-\sum_{j\in\mathcal{A}_i}4N_j(T)\cdot\max_{k\in\mathcal{A}_i}\Delta_k^2 - \log(6)\right)\right] \\ &\qquad\qquad \text{(evaluating the KL-divergence)} \\ &= \frac{1}{6}\,\mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{E_i\}\exp\left(-\sum_{a\in\mathcal{A}_i}4N_j(T)\tilde{\Delta}_{\mathcal{A}_i}^2\right)\right] \qquad \text{(definition of } \tilde{\Delta}_{\mathcal{A}_i}) \\ &\geq \frac{1}{6}\,\mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{E_i\}\exp\left(-24t_i\tilde{\Delta}_{\mathcal{A}_i}^2\right)\right] \qquad \text{(on } E_i) \\ &= \frac{1}{6}\exp\left(-24t_i\tilde{\Delta}_{\mathcal{A}_i}^2\right)\mathbb{P}_{\nu,\pi}(E_i).\end{aligned}$$

If $\xi \geq 1/2$, the result follows. Otherwise if $\xi \leq 1/2$, then $\mathbb{P}_\nu(E_i) \geq 1/6$ and it follows that $\mathbb{P}_{\nu^i}(J_t \notin \mathcal{A}_i) \geq \frac{1}{36}\exp\left(-24t_i\tilde{\Delta}_{\mathcal{A}_i}^2\right)$.

Observing that $\tilde{H}(\varepsilon,\nu) = \sum_{i=2}^{n/g(\varepsilon)}\tilde{\Delta}_{\mathcal{A}_i}^{-2}$, and $\sum_{i=1}^{n/m} t_i = T$, there exists some $i \in [2 : n/g(\varepsilon)]$ such that

$$t_i \leq \frac{T}{\tilde{H}(\varepsilon,\nu)\tilde{\Delta}_{\mathcal{A}_i}^2} = \frac{T}{a\tilde{\Delta}_{\mathcal{A}_i}^2},$$

which can easily be shown by way of contradiction. For this i, by our display above, we have

$$\max\left\{\mathbb{P}_{\nu^i,\pi}(J_t \notin \mathcal{A}_i), \mathbb{P}_{\nu,\pi}(J_t \notin \mathcal{A}_1)\right\} \geq \max\left\{1/2, \frac{1}{36}\exp\left(-24T/a\right)\right\}.$$

□

2.9.9 Proof of Lemma 3.1

Proof of Lemma 3.1. To see this, for a given value of n and $\alpha > 1/2$, fix $\varepsilon \le 1/2$. We have

$$\begin{aligned}
\tilde{H}(\varepsilon) &= \sum_{i=2}^{n/g(\varepsilon)} \tilde{\Delta}_{\mathcal{A}_i}^{-2} \\
&= \sum_{i=2}^{n/g(\varepsilon)} \left(\frac{n}{i \cdot g(\varepsilon)}\right)^{2\alpha} && \text{(Polynomial}(\alpha)\text{ instance)} \\
&\ge \left(\frac{n}{g(\varepsilon)}\right)^{2\alpha} \left[\sum_{i=2}^{n/g(\varepsilon)} \int_i^{i+1} x^{-2\alpha} dx\right] && (x^{-2\alpha} \text{ decreases monotonically on } x > 0, \text{ for } \alpha > 0) \\
&= \left(\frac{n}{g(\varepsilon)}\right)^{2\alpha} \int_2^{n/g(\varepsilon)+1} x^{-2\alpha} dx \\
&\ge \left(\frac{n}{g(\varepsilon)}\right)^{2\alpha} \int_2^{\varepsilon^{-1/\alpha}+1} x^{-2\alpha} dx && \left(\left(\tfrac{g(\varepsilon)}{n}\right)^{\alpha} < \varepsilon, \text{ by definition of } g(\varepsilon).\right) \\
&= \frac{1}{\varepsilon^2} \frac{2^{-2\alpha+1} - \left(\varepsilon^{-1/\alpha} + 1\right)^{-2\alpha+1}}{2\alpha - 1} \\
&\ge \frac{1}{\varepsilon^2} \left[\frac{2^{-2\alpha+1} - \left(2^{1/\alpha} + 1\right)^{-2\alpha+1}}{2\alpha - 1}\right]. && \text{(assuming } \varepsilon < 1/2\text{)}
\end{aligned}$$

□

2.9.10 Analysis of the uniform sampling

Uniform sampling means the algorithm that takes a budget T as the input and equally allocates the budget to all the arms. Define $\mathrm{Top}_m(\varepsilon) := \{i : \mu_i \ge \mu_m - \varepsilon\}$ and a shortcut of $\mathrm{Top}_m := \mathrm{Top}_m(0)$.

Proposition 9. *Suppose we run the uniform sampling on a n-armed bandit with a budget T and output the s arms with the largest empirical rewards. Let $C_T(s, m', \varepsilon)$ be the (m', ε)-good arms of the output. Then,*

$$\mathbb{P}\Big(\big|C_T(s, m', \varepsilon)\big| \le k\Big) \le \binom{m'}{m'-k} \exp\left(-(m'-k)\frac{\varepsilon^2 T}{8n}\right) + \binom{n - \big|\mathrm{Top}_{m'}(\varepsilon)\big|}{s-k} \exp\left(-(s-k)\frac{\varepsilon^2 T}{8n}\right).$$

This proposition is where our key contribution is rooted at. We show that our bound is tight for the uniform sampling in Section 2.9.11.

Proof. **Step 1**

We claim that the intersection of the two conditions below is a sufficient condition for $\left|C_T(s, m', \varepsilon)\right| > k$.

- $\left|\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}\right| \geq k + 1.$
- $\left|\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i < \mu_i + \frac{\varepsilon}{2}\right\}\right| \geq \left|\mathrm{Top}^c_{m'}(\varepsilon)\right| - s + k + 1.$

If all arms in $\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}$ are included in the output, the claim naturally holds. Thus we focus on the situation where at least one of the arms in $\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}$ is not included in the output. We prove the claim by contradiction. Suppose $\left|C_T(s, m', \varepsilon)\right| \leq k$. Then the output includes no less than $s - k$ non-(m', ε)-good arms. Since

$$s - k + \left|\mathrm{Top}^c_{m'}(\varepsilon)\right| - s + k + 1 = \left|\mathrm{Top}^c_{m'}(\varepsilon)\right| + 1 > \left|\mathrm{Top}^c_{m'}(\varepsilon)\right|,$$

at least one arm in $\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i < \mu_i + \frac{\varepsilon}{2}\right\}$ is included in the output. To be more specific, there must exist overlap between $\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i < \mu_i + \frac{\varepsilon}{2}\right\}$ and the set of the non-(m', ε)-good arms included in the output; otherwise the union of these two sets is even larger than the set of all the non-(m', ε)-good arms we have in the entire arm set. Additionally, all arms in $\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i < \mu_i + \frac{\varepsilon}{2}\right\}$ have less empirical rewards than any arms in $\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}$. Thus the output includes all arms in $\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}$ as well. This leads to the contradiction with our supposition that at least one of the arms in $\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}$ are not included in the output.

Step 2

We apply union bound and Hoeffding's inequality for the event that at least one of the two

conditions does not hold.

$$\begin{aligned}
\mathbb{P}\Big(\big|C_T(s,m',\varepsilon)\big| \le k\Big) &\le \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\varepsilon}{2}\right\}\right| < k+1\right) \\
&\quad + \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i < \mu_i + \frac{\varepsilon}{2}\right\}\right| < \left|\mathrm{Top}^c_{m'}(\varepsilon)\right| - s + k + 1\right) \\
&= \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i \le \mu_i - \frac{\varepsilon}{2}\right\}\right| \ge m' - k\right) \\
&\quad + \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}^c_{m'}(\varepsilon) : \hat{\mu}_i \ge \mu_i + \frac{\varepsilon}{2}\right\}\right| \ge s - k\right) \\
&\le \mathbb{P}\left(\exists \mathcal{A} \subset \mathrm{Top}_{m'}, \text{s.t. } |\mathcal{A}| = m' - k \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \le \mu_i - \frac{\varepsilon}{2}\right) \\
&\quad + \mathbb{P}\left(\exists \mathcal{A} \subset \mathrm{Top}^c_{m'}(\varepsilon), \text{s.t. } |\mathcal{A}| = s - k \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \ge \mu_i + \frac{\varepsilon}{2}\right) \\
&\le \binom{m'}{m'-k} \exp\left(-(m'-k)\frac{\varepsilon^2 T}{8n}\right) + \binom{n - \left|\mathrm{Top}_{m'}(\varepsilon)\right|}{s-k} \exp\left(-(s-k)\frac{\varepsilon^2 T}{8n}\right).
\end{aligned}$$

□

Lemma 9.1. *Suppose we run the uniform sampling on a n-armed bandit with a budget of T and output the s arms with the largest empirical rewards. Let $C_T(s,\varepsilon)$ be the ε-good arms in the output. Assume $s > g(\varepsilon)$ with $g(\varepsilon)$ being the number of ε-good arms. Let q be a parameter satisfying $g(\varepsilon) < q < s$ and $d_q = \mu_{g(\varepsilon)} - \mu_{q+1}$, then*

$$\mathbb{P}\Big(\big|C_T(s,\varepsilon)\big| \le g(\varepsilon) - k\Big) \le \binom{g(\varepsilon)}{k} \exp\left(-k\frac{d_q^2 T}{8n}\right) + \binom{n-q}{s-q+k} \exp\left(-(s-q+k)\frac{d_q^2 T}{8n}\right).$$

Especially for Sequential Halving where $s = \frac{n}{2}$, choosing $q = \frac{n}{4}$ we have

$$\mathbb{P}\left(\left|C_T\left(\frac{n}{2},\varepsilon\right)\right| \le g(\varepsilon) - k\right) \le \binom{g(\varepsilon)}{k} \exp\left(-k\frac{d_q^2 T}{8n}\right) + \binom{\frac{3n}{4}}{\frac{n}{4}+k} \exp\left(-\left(\frac{n}{4}+k\right)\frac{d_q^2 T}{8n}\right).$$

Proof. The event

$$\Big\{\big|C_T(s,\varepsilon)\big| \le g(\varepsilon) - k\Big\}$$

implies that there are at least k ε-good arms eliminated. Since $g(\varepsilon) < q < s$, the s output arms include at most $q-k$ arms that are from the best q arms $\left\{i \mid \mu_i \geq \mu_q\right\}$. Such that, the s output arms include at least $s-q+k$ arms that are from the worst $n-q$ arms $\left\{i \mid \mu_i < \mu_q\right\}$; otherwise, the number of outputs is even less than s. Denote the set of eliminated arms that are ε-good as A', and the set of output arms that are from the worst $n-q$ arms as B', we have $|A'| \geq k, |B'| \geq s-q+k$. We then have the existence of A, B satisfying

$$\left\{\exists A \subset \left[g(\varepsilon)\right], |A| = k, \exists B \subset [n] \setminus [q], |B| = s-q+k, \text{s.t.}, \ \forall i \in A, j \in B, \hat{\mu}_i \leq \hat{\mu}_j\right\}.$$

The above implication can also be seen as a generalization of the technique of Karnin et al. (2013). We then have

$$\left\{\exists A \subset \left[g(\varepsilon)\right], |A| = k, \exists B \subset [n] \setminus [q], |B| = s-q+k, \text{s.t.}, \ \forall i \in A, j \in B, \hat{\mu}_i \leq \mu_i - \frac{d_q}{2} \text{ or } \hat{\mu}_j \geq \mu_j + \frac{d_q}{2}\right\}.$$

By the distributive law of Boolean algebra, we have at least one of the following two holds,

$$\left\{\exists A \subset \left[g(\varepsilon)\right], |A| = k, \text{s.t.}, \ \forall i \in A, \hat{\mu}_i \leq \mu_i - \frac{d_q}{2}\right\},$$

or

$$\left\{\exists B \subset [n] \setminus [q], |B| = s-q+k, \text{s.t.}, \ j \in B, \hat{\mu}_j \geq \mu_j + \frac{d_q}{2}\right\}.$$

Thereby the probability of $\left\{\left|C_T(s,\varepsilon)\right| \le g(\varepsilon) - k\right\}$ can be upper bounded as

$$
\begin{aligned}
&\mathbb{P}\left(\left|C_T(s,\varepsilon)\right| \le g(\varepsilon) - k\right) \\
\le &\mathbb{P}\left(\exists A \subset \left[g(\varepsilon)\right], |A| = k, \text{s.t.}, \ \forall i \in A, \hat{\mu}_i \le \mu_i - \frac{d_q}{2}\right) \\
&+ \mathbb{P}\left(\exists B \subset [n] \setminus [q], |B| = s - q + k, \text{s.t.}, \ j \in B, \hat{\mu}_j \ge \mu_j + \frac{d_q}{2}\right) \\
\le &\binom{g(\varepsilon)}{k} \exp\left(-k\frac{\left(\frac{d_q}{2}\right)^2 T}{2n}\right) + \binom{n-q}{s-q+k} \exp\left(-(s-q+k)\frac{\left(\frac{d_q}{2}\right)^2 T}{2n}\right) \\
= &\binom{g(\varepsilon)}{k} \exp\left(-k\frac{d_q^2 T}{8n}\right) + \binom{n-q}{s-q+k} \exp\left(-(s-q+k)\frac{d_q^2 T}{8n}\right).
\end{aligned}
$$

□

2.9.11 Lower bound for uniform sampling

Theorem 10. *Consider a family $\mathcal{E}(n,m,\varepsilon)$ of all two-level univariate Gaussian bandit instances with n arms, of which m arms have mean $\varepsilon > 0$ and all other arms have to mean 0. We consider the problem in which samples are drawn from some instance $\theta \in \mathcal{E}(n,m,\varepsilon)$ in an off-line, uniform fashion with B samples from each arm (i.e., with a total sampling budget of T samples, $B = T/n$ ignoring integer effects) and a subset of arms $\hat{S} \subseteq [n]$ of size s is then chosen based only on the observed samples (i.e., without knowledge of θ) in a symmetric fashion. Here 'symmetric' is taken to mean that for any subset $S \subseteq [n]$, $\mathbb{P}_\theta(\hat{S} = S) = \mathbb{P}_{\sigma(\theta)}(\hat{S} = \sigma(S))$ where $\sigma \in \Sigma_n$ is an element of the permutation group on sets of size n, and $\sigma(S)$ is taken to mean the element-wise application of σ to the elements of S.*[1] *That is to say that the choice of $\hat{S}$ is independent of the specific indices of the arms chosen.*

[1]Note that the assumption of a symmetric selection $\hat{S}$ is equivalent to considering a lower bound for general selections $\hat{S}$ in the face of a uniform mixture over all permutations of θ (see Simchowitz et al. (2017) for a more in-depth discussion)

Suppose $m \le s \wedge n/3$ and $s < \frac{6}{11}n$. Let $\tilde{M}_\theta$ be the number of false negatives (i.e. $\|\ge$) Then, there exists $\theta \in \mathcal{E}(n, m, \varepsilon)$ s.t.

$$\mathbb{P}_\theta(\tilde{M}_\theta \ge \frac{1}{4}m) \ge \min\left(\frac{1}{2}, \quad 2\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)^{\frac{3}{16}m}\exp\left(-\left(1+16\cdot\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}\right)mB\varepsilon^2\right)\right)$$

where $\mathbb{P}_\theta$ is the probability measure induced by sampling from instance θ.

Proof. Let $\theta = (\varepsilon, \dots, \varepsilon, 0, \dots, 0) \in \mathbb{R}^n$ with $\varepsilon > 0$ be a vector of mean rewards for each arm where the first m coordinates are nonzero. Let $\hat{S} \in [n]$ be the output of the algorithm (recall $|\hat{S}| = s$). Let $\tilde{M}_\theta$ be the number of false negatives when taking $\hat{S}$ as the prediction for the true support $[s]$ of θ. Let $\tilde{Q}_a \subset 2^{[n]}$ be the collection of all subsets of $[n]$ of size s such that $\tilde{M}_\theta = a$. For convenience, let $\gamma = \frac{1}{8}m$. Let $k = \frac{3}{4}s$.

$$\begin{aligned}\xi := \mathbb{P}_\theta(\tilde{M}_\theta \ge m-k) &= \sum_{a=m-k}^{k}\mathbb{P}_\theta(\tilde{M}_\theta = a)\\ &\ge \sum_{a=3\gamma}^{5\gamma}\mathbb{P}_\theta(\tilde{M}_\theta = a).\end{aligned}$$

Let $\mathsf{first}(\tilde{Q}_a)$ be the first member of $\tilde{Q}_a$ in lexicographic order. For example, $\mathsf{first}(tQ_a) = [a+1 : a+s]$ where $[A : B] := \{A, A+1, \dots, B\}$. Then, by symmetry, one can see that $\mathbb{P}_\theta(\tilde{M}_\theta = a) = |\tilde{Q}_a|\,\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a))$.

We consider the event on which the 'empirical KL-divergence' concentrates, which can be shown to be true with probability at least $1-\delta$ by Jun and Zhang (2020, Lemma 5):

$$\mathsf{conc}(\theta', \theta) := \left\{\sum_{i=1}^{n}\sum_{t=1}^{B}\ln\left(\frac{p_{\theta'}(X_{i,t}|A_t = i)}{p_\theta(X_{i,t}|A_t = i)}\right) - (1+\rho)B\sum_{i=1}^{n}\mathsf{KL}(\theta'_i, \theta_i) \le \frac{1}{\rho}\ln(1/\delta)\right\}.$$

We now employ a change of measure argument to $\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a))$ and switch θ with θ' that would result in $\tilde{M}_{\theta'} = a - 3\gamma =: b$. For $a \in [3\gamma : 5\gamma]$, this can be achieved with the choice of $\theta' = (0, \dots, 0, \varepsilon, \dots, \varepsilon, 0, \dots, 0)$ that is supported on $[3\gamma + 1 : m + 3\gamma]$. Let Θ be the set of

permutations of θ. Then,

$$
\begin{aligned}
&\mathbb{P}_\theta(\tilde{M}_\theta = a) \\
&= |\tilde{Q}_a|\,\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a)) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_{\theta'}(\hat{S} = \mathsf{first}(\tilde{Q}_a), \mathsf{conc}(\theta', \theta)) \exp(-(1+\rho)2mB\cdot(\varepsilon^2/2) - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(change of measure; } \rho > 0) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_{\theta'}(\hat{S} = \mathsf{first}(\tilde{Q}_a), \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta', \sigma(\theta'))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(Σ: symmetric group of } [n]) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_b), \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta, \sigma(\theta))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \quad \text{(symmetry)} \\
&= \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|}\,\mathbb{P}_\theta(\tilde{M}_\theta = b, \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta, \sigma(\theta))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \quad \text{(symmetry)} \\
&\geq \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|}\left(\mathbb{P}_\theta(\tilde{M}_\theta = b) - |\Theta|\delta\right)\exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)\right)\ . \\
&\qquad\qquad \text{($\mathbb{P}(A,B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$; union bound over } \{\sigma(\theta) : \sigma \in \Sigma\})
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{a=3\gamma}^{5\gamma} \mathbb{P}_\theta(\tilde{M}_\theta = a) \\
&\geq \underbrace{\min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{Q}_a|}{|\tilde{Q}_{a-3\gamma}|}}_{=:\,Y}\left(\underbrace{\mathbb{P}_\theta(\tilde{M}_\theta \in [0:2\gamma])}_{\geq 1-\xi} - (2\gamma+1)|\Theta|\delta\right)\exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)\right)\ .
\end{aligned}
$$

Let us choose $\delta = \frac{1}{2}\cdot\frac{1-\xi}{(2\gamma+1)|\Theta|}$. Since the LHS above is at most ξ, we have

$$
\xi \geq Y\frac{1-\xi}{2}\exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln\left(\frac{2(2\gamma+1)|\Theta|}{1-\xi}\right)\right)
$$

One can consider two cases, namely $\xi \geq \frac{1}{2}$ and $\xi < \frac{1}{2}$, to arrive at

$$
\xi \geq \min\left(\frac{1}{2},\quad \exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln\left(4(2\gamma+1)|\Theta|\right) + \ln(4Y)\right)\right)
$$

It remains to find an appropriate value of ρ. One simple choice is

$$\rho^{-1} = \frac{1}{2}\frac{\ln(4Y)}{\ln\left(4(2\gamma+1)|\Theta|\right)} \ ,$$

which satisfies that $\rho^{-1} > 0$ as show later. Thus, we have

$$\xi \geq \min\left(\frac{1}{2}, \ \ 2\sqrt{Y}\exp\left(-\left(1+2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)}\right)mB\varepsilon^2\right)\right)$$

It remains to figure out bounds for Y and $|\Theta|$. For Y, note that $|\tilde{Q}_a| = \binom{m}{m-a}\binom{n-m}{s-(m-a)}$. So, for $a \in [3\gamma : 5\gamma]$ and $b = a - 3\gamma$,

$$\begin{aligned}
\min_a \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|} &= \frac{\binom{m}{m-a}\binom{n-m}{s-m+a}}{\binom{m}{m-b}\binom{n-m}{s-m+b}} \\
&= \frac{(m-b)(m-b-1)\cdots(m-a+1)}{a(a-1)\cdots(b+1)} \cdot \frac{(n-s-b)(n-s-b-1)\cdots(n-s-a+1)}{(s-m+a)(s-m+a-1)\cdots(s-m+b+1)} \\
&\overset{(a)}{\geq} \left(\frac{m-b}{a}\right)^{a-b} \cdot \left(\frac{n-s-b}{s-m+a}\right)^{a-b} \\
&\geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s-2\gamma}{s-3\gamma}\right)^{3\gamma} \geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s}{s}\right)^{3\gamma} \\
\implies Y = \min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{Q}_a|}{|\tilde{Q}_{a-2\gamma}|} &\geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s}{s}\right)^{3\gamma}
\end{aligned}$$

where (a) is due to the fact that $\frac{m-b}{a} > 1$ implies $(m-b-i)/(a-i) \geq (m-b)/a$ for $i \in [0 : a-b-1]$ and, with a similar reasoning, $n \geq s/2 \implies \frac{n-s-a+3\gamma}{s-m+a} > 1 \implies (n-s-b-i)/(s-m+a-i) \geq (n-s-b)/(s-m+a)$. Then, using $|\Theta| = \binom{n}{m} \leq \left(\frac{en}{m}\right)^m$, we have

$$\begin{aligned}
\rho = 2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)} &\leq 2\frac{\ln(m+4) + m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\
&\overset{(a)}{\leq} 4\frac{m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\
&\leq 16\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}
\end{aligned}$$

where (a) is by $m \leq n/3 \implies \ln(m+4) \leq m\ln(en/m)$.

Altogether,

$$\mathbb{P}_\theta(\tilde{M} \geq \frac{1}{4}m) \geq \min\left(\frac{1}{2}, \quad 2\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)^{\frac{3}{16}m} \exp\left(-\left(1+16\cdot\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}\right)mB\varepsilon^2\right)\right)$$

To verify that our choice of ρ is nonnegative, it suffices to show that $\frac{6}{5}\cdot\frac{n-s}{s} > 1$, which is true for $s < \frac{6}{11}n$.

□

2.9.12 A minimax optimal budget allocation scheme for SH

We show that with a different budget allocation scheme $T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor$, SH can achieve a sample complexity of $\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right)$ (without any logarithmic terms of n). We call this budget allocation scheme Option 2 and the original one of Karnin et al. (2013) Option 1. The price of achieving this exact minimax optimality is the potential loss of the near instance optimality (or at least the standard proof technique does not work). We summarize the comparison in Table 2.2. It is not clear to us whether there exists a different value of T_ℓ in SH that can achieve both the minimax and instance-dependent optimality. It is also not clear whether Option 2 can achieve a similar bound as Theorem 1. Note that for the fixed budget setting, there is still a $\log\log(n)$ gap between the instance-dependent upper and lower bounds.

Theorem 11. *For any $\varepsilon \in (0,1)$, with Option 2, there exists an absolute constant c_1, such that the error probability of SH for identifying an ε-good arm satisfies,*

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) \leq \exp\left(-\Theta\left(\frac{\varepsilon^2 T}{n}\right)\right),$$

for $T > c_1\cdot\frac{n}{\varepsilon^2}$.

Proof. Still let $\varepsilon_1 = \varepsilon/4$. And define $\varepsilon_{\ell+1} = \frac{3}{4}\cdot\varepsilon_\ell$. For each stage ℓ, define the event G_ℓ as

$$G_\ell := \left\{\max_{i\in S_{\ell+1}}\mu_i \geq \max_{i\in S_\ell}\mu_i - \varepsilon_\ell\right\}.$$

We have

$$\sum_{\ell=1}^{\log n} \varepsilon_\ell < \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \cdot \varepsilon_1 = \frac{\varepsilon}{4} \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \leq \frac{\varepsilon}{4} \lim_{n\to\infty} \frac{1-(3/4)^n}{1-3/4} = \varepsilon.$$

For stage ℓ, we assign budget $T_\ell \cdot n2^{-(\ell-1)} := \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \cdot n2^{-(\ell-1)} = \frac{T}{81} \cdot \left(\frac{8}{9}\right)^{\ell-1} \cdot \ell$. We can verify this allocation scheme does not exceed the budget as follows,

$$\sum_{\ell=1}^{\log n} T_\ell \cdot n2^{-(\ell-1)} \leq \sum_{\ell=1}^{\infty} \frac{T}{81} \cdot \left(\frac{8}{9}\right)^{\ell-1} \cdot \ell \leq \frac{T}{9} \lim_{\ell\to\infty} \frac{1-\left(8/9\right)^{\ell}}{1-8/9} = T.$$

Let a_ℓ be the best arm in S_ℓ,

$$\begin{aligned}
\mathbb{P}\left(G_\ell^c\right) &= \mathbb{P}\left(G_\ell^c, \hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\right) + \mathbb{P}\left(G_\ell^c, \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\right) + \mathbb{P}\left(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right).
\end{aligned}$$

For the first term

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\right) &\leq \exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T_\ell}{|S_\ell|}\right) \\
&\leq \exp\left(-\frac{\varepsilon_1^2\left(\frac{9}{16}\right)^{\ell-1}}{2}\frac{\frac{T}{81n}\cdot\left(\frac{16}{9}\right)^{\ell-1}\cdot\ell\cdot n2^{-(\ell-1)}}{n2^{-(\ell-1)}}\right) \\
&\leq \exp\left(-\frac{\varepsilon^2 T}{32\cdot 81n}\ell\right).
\end{aligned}$$

For the second term,

$$\begin{aligned}
\mathbb{P}\left(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right) &\leq \mathbb{P}\left(\left|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\right| \geq |S_\ell|/2\right) \\
&\leq \frac{\mathbb{E}\left[\left|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\right|\right]}{|S_\ell|/2} \\
&\leq \frac{|S_\ell|\exp\left(-\frac{\varepsilon^2 T}{32\cdot 81n}\ell\right)}{|S_\ell|/2} \\
&= 2\exp\left(-\frac{\varepsilon^2 T}{32\cdot 81n}\ell\right).
\end{aligned}$$

Table 2.2: The ε-error probability of SH (worst-case) and ($\varepsilon = 0$)-error probability (instance-dependent) for Option 1 and 2 where $H_2 := \max_i \frac{i}{\Delta_i^2}$ is the problem hardness parameter. Here, we omit constant factors. Check marks for the worst-case means that it is optimal, and those for the instance-dependent means that it is optimal up to $\log\log(n)$ factors in the sample complexity (Mannor and Tsitsiklis, 2004; Carpentier and Locatelli, 2016).

	Worst-case	Instance-dependent
Option 1: $T_\ell = \left\lfloor \frac{T}{\lvert S_\ell \rvert \lceil \log_2 n \rceil} \right\rfloor$	✗ $\log_2 n \cdot \exp\left(-\frac{\varepsilon^2 T}{n \log_2 n}\right)$	✓ $\log_2 n \cdot \exp\left(-\frac{T}{\log_2 n H_2}\right)$
Option 2: $T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor$	✓ $\exp\left(-\frac{\varepsilon^2 T}{n}\right)$	✗ $\log_2 n \cdot \exp\left(-\frac{T}{n \log_2 n H_2}\right)$

By contraposition,

$$\begin{aligned}\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) &\le \sum_{\ell=1}^{\log n} \mathbb{P}\left(G_\ell^c\right) \\ &\le \sum_{\ell=1}^{\infty} 3\exp\left(-\frac{\varepsilon^2 T}{32 \cdot 81n}\ell\right).\end{aligned}$$

Let $X := \frac{\varepsilon^2 T}{32 \cdot 81n} > 0$,

$$\begin{aligned}\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) &\le \sum_{\ell=1}^{\infty} 3\exp(-\ell X) \\ &\le 3\exp(-X)\frac{1-\exp(-\ell X)}{1-\exp(-X)} \\ &\le 3\exp(-X)\frac{1}{1-\exp(-X)}.\end{aligned}$$

As $\frac{1}{1-\exp(-X)}$ is a monotonically decreasing function, there exists an absolute constant c_1, such that

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \varepsilon\right) \le \exp\left(-\Theta\left(\frac{\varepsilon^2 T}{n}\right)\right),$$

holds for $T > c_1 \cdot \frac{n}{\varepsilon^2}$. □

2.9.13 The number of opened brackets

Lemma 11.1. *The number of opened brackets till round t, denoted by L_t, satisfies*

$$0.63\cdot\log_2\big(1+\ln(2)t\big) < L_t \le 1+\log_2\big(1+\ln(2)t\big).$$

The size of the largest bracket satisfies

$$2^{L_t} \ge \tilde{\Theta}(t).$$

Proof. Recall that whenever $t \ge B2^B$ is true, we open a new bracket and increase B by 1. $B = 0$ before the game starts. First, note that for $t = 1$ we have $L_t = 1$. Then, let us focus on $t \ge 2$. It is not hard to verify that

$$L_t = \max\big\{B : t \ge (B-1)2^{B-1}\big\}.$$

Fix $t \ge 2$. Then, we can define

$$t' := (L_t - 1)2^{L_t-1}, \tag{2.15}$$

the first time point that the bracket L_t was opened. Clearly $t' \le t$. Solve (2.15) for L_t.

$$\begin{aligned} t' &= (L_t-1)\exp\big(\ln(2)(L_t-1)\big),\\ \ln(2)t' &= \ln(2)(L_t-1)\exp\big(\ln(2)(L_t-1)\big)\\ &=: X\exp(X). \end{aligned}$$

Solving it for X is exactly the Lambert W function. By Lemma 17 of Orabona and Pal (2016), we have

$$0.6321\cdot\ln\big(1+\ln(2)t'\big) \le X \le \ln\big(1+\ln(2)t'\big).$$

This implies that

$$X = \ln(2)(L_t - 1) \le \ln\big(1 + \ln(2)t'\big) \le \ln\big(1 + \ln(2)t\big) \implies L_t \le 1 + \log_2\big(1 + \ln(2)t\big).$$

To obtain a lower bound, define

$$t'' := L_t 2^{L_t},$$

which is strictly larger than t. Using similar techniques as above, we can derive

$$L_t \ge 0.6321 \cdot \log_2\big(1 + \ln(2)t''\big) > 0.63 \cdot \log_2\big(1 + \ln(2)t\big).$$

Together, we have

$$0.63 \cdot \log_2\big(1 + \ln(2)t\big) < L_t \le 1 + \log_2\big(1 + \ln(2)t\big).$$

Note that we have

$$L_t 2^{L_t} \le t \le (L_t + 1)2^{L_t+1}.$$

Thus

$$2^{L_t} \ge \frac{t}{2(L_t + 1)} \ge \frac{t}{2\Big(2 + \log_2\big(1 + \ln(2)t\big)\Big)} \ge \frac{t}{4\log_2 t} = \tilde{\Theta}(t). \qquad (t \ge 4)$$

□

2.9.14 The number of pulls for the representative arms

Lemma 11.2. *At round t, the number of times the representative arm of bracket B has been pulled in the latest finished SH, denoted as $D_{B,t}$, satisfies,*

$$D_{B,t} \ge \text{const} \cdot \frac{t}{\ln(t) \cdot \log_2(n)}.$$

Proof. We call the round interval $(L_t-1)2^{L_t-1} \le t < L_t 2^{L_t}$ as *phase* L_t, whose length is $L_t 2^{L_t} - (L_t-1)2^{L_t-1} = (L_t+1)2^{L_t-1}$. Correspondingly, sampling budget of $\frac{(L_t+1)2^{L_t-1}}{L_t}$ is assigned to each bracket. For each opened bracket $A_B, B \in [L_t-1]$, the sampling budget for it is accumulated from phase B to phase L_t, while phase L_t is not finished yet. The sampling budget bracket A_B receives satisfies

$$\begin{aligned} T_{B,t} &= \sum_{i=B}^{L_t-1} \frac{2^{i-1}(i+1)}{i} + \frac{t-2^{L_t-1}(L_t-1)}{L_t} \\ &> \frac{2^{L_t-2}L_t}{L_t-1} + \frac{t-2^{L_t-1}(L_t-1)}{L_t} \\ &> \frac{t}{4L_t} \\ &> \frac{t}{4+4\log\big(1+\ln(2)t\big)} > \frac{t}{4\ln t}. \end{aligned} \qquad (t>16)$$

Recall we run the SH with the doubling trick on each bracket individually. We claim that the latest finished SH receives a budget of at least $T_{B,t}/4$. To see this, note that we initialize Algorithm 2 with budget starting from $\lceil n\log_2 n\rceil$. Then restart the SH algorithm with budget $2\cdot\lceil n\log_2 n\rceil$, and so on. Before finishing the k-th doubling trick, the best arm from $(k-1)$-th doubling trick is output. Thus the portion of the lasted finished SH ranges in

$$\left[\frac{|\mathcal{T}_{k-1}|}{\sum_{i=1}^{k}|\mathcal{T}_i|}, \frac{|\mathcal{T}_{k-1}|}{\sum_{i=1}^{k-1}|\mathcal{T}_i|}\right),$$

which, by some simple calculations, is

$$\left[\frac{b_{r-1}}{4b_{r-1}-b_0}, \frac{b_{r-1}}{2b_{r-1}-b_0}\right). \qquad (2.16)$$

That is to say, the most recently finished SH, which gives the current output, has a budget of at least $T_{B,t}/4$. We notice that $T_{B,t}$ is essentially irrelevant to B. This means each bracket receives an order-wise equal amount of budget to query the SH algorithm. In the SH algorithm, the output arm is pulled a fixed number of times when the number of arms and the sampling budget is fixed. It is pulled $\Theta\Big(\frac{T}{\log_2 n}\Big)$ times for budget T. We have all we

need to determine the lower bound for $D_{B,t}$.

$$D_{B,t} \geq \text{const} \cdot \frac{t}{\ln(t) \cdot \log_2(n)}.$$

□

2.9.15 Lower bounds and their implications in special instances

While Katz-Samuels and Jamieson (2020) shows a lower bound for the (ε, δ)-unverifiable sample complexity that can match the upper bound in special instances, the final result they presented has a negative term, which results in looseness, and the bound can even go to 0 in certain instances.

Recall the definition of $\tau_{\varepsilon,\delta}$, the (ε, δ)-unverifiable sample complexity (Definition 1).

Theorem 12 (Theorem 1 of Katz-Samuels and Jamieson (2020))**.** *Fix $\varepsilon > 0, \delta \in (0, 1/16)$, and a vector $\mu \in \mathbb{R}^n$. Consider n arms where rewards from the i-th arm are distributed according to $\mathcal{N}(\mu_i, 1)$. For every permutation $\pi \in \mathbb{S}^n$, let $(\mathcal{F}_t^\pi)_{t\in\mathbb{N}}$ be the filtration generated by the algorithm playing on instance $\pi(\mu)$. Then the (ε, δ)-unverifiable sample complexity satisfies,*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \geq \frac{1}{64}\left(-\left(\mu_1 - \mu_{g(\varepsilon)+1}\right)^{-2} + \frac{1}{g(\varepsilon)} \sum_{i=g(\varepsilon)+1}^{n} \left(\mu_1 - \mu_i\right)^{-2}\right),$$

where the expectation is with respect to $\pi \in \mathbb{S}^n$ and $\pi(\mu)$.

In fact, the proof of Katz-Samuels and Jamieson (2020, Theorem 1) shows a stronger bound, which is more useful for deriving lower bounds for specific instances.

Theorem 13 (A stronger version of Theorem 1 of Katz-Samuels and Jamieson (2020))**.** *Fix $\varepsilon > 0, \delta \in (0, 1/16)$, and a vector $\mu \in \mathbb{R}^n$. Consider n arms where rewards from the i-th arm are distributed according to $\mathcal{N}(\mu_i, 1)$. For every permutation $\pi \in \mathbb{S}^n$, let $(\mathcal{F}_t^\pi)_{t\in\mathbb{N}}$ be the filtration generated by the algorithm playing on instance $\pi(\mu)$. Then the (ε, δ)-unverifiable*

sample complexity satisfies,

$$\mathbb{E}\big[\tau_{\varepsilon,\delta}\big] \geq \frac{1}{64}\frac{1}{g(\varepsilon)}\sum_{i=2g(\varepsilon)+1}^{n}(\mu_1-\mu_i)^{-2},$$

where the expectation is with respect to $\pi \in \mathbb{S}^n$ *and* $\pi(\mu)$.

Corollary 13.1. *For the EqualGap*(m) *instance, Proposition 13 shows a lower bound of the* (ε,δ)*-unverifiable sample complexity as*

$$\mathbb{E}[\tau_\pi] \geq \frac{1}{100}\left(\frac{n}{m}\frac{1}{\varepsilon^2}-\frac{2}{\varepsilon^2}\right).$$

Proof.

$$\begin{aligned}\frac{1}{64}\frac{1}{g(\varepsilon)}\sum_{i=2g(\varepsilon)+1}^{n}(\mu_1-\mu_i)^{-2} &= \frac{1}{64}\frac{1}{m}\sum_{i=2m+1}^{n}\frac{16}{25\varepsilon^2}\\ &= \frac{1}{64}\frac{n-2m}{m}\frac{16}{25\varepsilon^2}\\ &= \frac{1}{100}\left(\frac{n}{m}\frac{1}{\varepsilon^2}-\frac{2}{\varepsilon^2}\right).\end{aligned}$$

□

Corollary 13.2. *For the Polynomial*(α) *instance, Proposition 13 shows a lower bound of the* (ε,δ)*-unverifiable sample complexity as*

$$\mathbb{E}[\tau_\pi] \geq \frac{1}{128}\frac{1}{2\alpha-1}\left(2^{1-2\alpha}\varepsilon^{-2}-\varepsilon^{-\frac{1}{\alpha}}\right).$$

Proof. For the Polynomial(α) instance,

$$\begin{aligned}\frac{1}{64}\frac{1}{g(\varepsilon)}\sum_{i=2g(\varepsilon)+1}^{n}(\mu_1-\mu_i)^{-2} &= \frac{1}{64}\frac{1}{g(\varepsilon)}\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{i}{n}\right)^{-2\alpha}\\ &= \frac{1}{64}\frac{n^{2\alpha}}{n\varepsilon^{1/\alpha}}\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{1}{i}\right)^{2\alpha}\\ &= \frac{1}{64}\frac{n^{2\alpha-1}}{\varepsilon^{1/\alpha}}\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{1}{i}\right)^{2\alpha}.\end{aligned}$$

The summation of $\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{1}{i}\right)^{2\alpha}$ can be bounded as

$$\begin{aligned}
&\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{1}{i}\right)^{2\alpha} \overset{(a_1)}{\geq} \frac{1}{2}\int_{2g(\varepsilon)}^{n}\left(\frac{1}{i}\right)^{2\alpha} di \\
&= \frac{1}{2}\Big(f(n) - f\big(2g(\varepsilon)\big)\Big) && \left(f(x) = \tfrac{1}{-2\alpha+1}x^{-2\alpha+1}\right) \\
&= \frac{1}{2}\left(\frac{1}{-2\alpha+1}n^{-2\alpha+1} - \frac{1}{-2\alpha+1}(2g(\varepsilon))^{-2\alpha+1}\right) \\
&= \frac{1}{2}\left(\frac{1}{-2\alpha+1}n^{-2\alpha+1} - \frac{1}{-2\alpha+1}(2n\varepsilon^{1/\alpha})^{-2\alpha+1}\right),
\end{aligned}$$

where (a_1) is by the geometric meaning of integral and

$$\frac{1}{2}\left(\frac{1}{i}\right)^{2\alpha} = \left(\frac{1}{2^{\frac{1}{2\alpha}}i}\right)^{2\alpha} \leq \left(\frac{1}{2i}\right)^{2\alpha} \leq \left(\frac{1}{i+1}\right)^{2\alpha}. \qquad (i \geq 1, \alpha > 0.5)$$

Thus

$$\begin{aligned}
\frac{1}{64}\frac{1}{g(\varepsilon)}\sum_{i=2g(\varepsilon)+1}^{n}(\mu_1-\mu_i)^{-2} &= \frac{1}{64}\frac{n^{2\alpha-1}}{\varepsilon^{1/\alpha}}\sum_{i=2g(\varepsilon)+1}^{n}\left(\frac{1}{i}\right)^{2\alpha} \\
&\geq \frac{1}{128}\frac{n^{2\alpha-1}}{\varepsilon^{1/\alpha}}\left(\frac{1}{-2\alpha+1}n^{-2\alpha+1} - \frac{1}{-2\alpha+1}(2n\varepsilon^{1/\alpha})^{-2\alpha+1}\right) \\
&= \frac{1}{128}\frac{1}{\varepsilon^{1/\alpha}}\left(\frac{1}{-2\alpha+1} - \frac{2^{1-2\alpha}}{-2\alpha+1}\varepsilon^{\frac{-2\alpha+1}{\alpha}}\right) \\
&= \frac{1}{128}\varepsilon^{-\frac{1}{\alpha}}\frac{1}{2\alpha-1}\left(2^{1-2\alpha}\varepsilon^{\frac{-2\alpha+1}{\alpha}} - 1\right) \\
&= \frac{1}{128}\frac{1}{2\alpha-1}\left(2^{1-2\alpha}\varepsilon^{-2} - \varepsilon^{-\frac{1}{\alpha}}\right).
\end{aligned}$$

□

2.9.16 Proof of Theorem 6

Theorem 6. *For any $\varepsilon \in (0,1)$, the ε-error probability of BSH satisfies*

$$\mathbb{P}(\mu_1 - \mu_{J_t} > \varepsilon) \leq \exp\left(-\tilde{\Theta}\left(\frac{t}{\frac{1}{\varepsilon^2} + \max_{i \geq g\left(\frac{\varepsilon}{2}\right)+1}\frac{i}{\Delta_i^2}\cdot\frac{1}{g\left(\frac{\varepsilon}{4}\right)}}\right)\right),$$

for $t \geq \tilde{\Theta}\Big(H_2\big(\frac{\varepsilon}{2}\big)\Big)$.

Proof. Due to the doubling nature of the bracketing scheme, it is not hard to see that each bracket receives an order-wise equal amount of sampling budget $\tilde{\Theta}(t)$ (details in Section 2.9.14). Our analysis is centered around finding the ideal bracket whose representative arm is expected to be of the highest quality, which we call the *best bracket*. Intuitively, on the one hand, if the best bracket is too large, the per-arm sampling budget will be too small to guarantee a meaningful output. On the other hand, if the best bracket is too small, it may not well represent the entire instance. The core idea of the proof is balancing this trade-off.

We define the following notations.

- A^*: the best bracket, which is a multiset satisfying $\forall i \in A^*, i \in [n]$. We have $|A^*| = L$ with L being specified later.
- $r = \frac{L}{n}$: the subsampling ratio of the best bracket A^*.
- $\varepsilon_k := 2^{k-1}.\frac{\varepsilon}{4}$, for $k \in [K]$, where $K = \Big\lceil \log_2\big(\frac{8}{\varepsilon}\big) \Big\rceil$. Also K satisfies $\varepsilon_{K-1} \leq 1, \varepsilon_K \geq 1$.
- $Z_k = \Big\{ i \in [n] \mid \Delta_i \in \big[0, \min\{\varepsilon_k, 1\}\big] \Big\}$.
- $g(\varepsilon_k) = |Z_k|$.
- $Z'_k = \Big\{ i \in A^* \mid \Delta_i \in \big[0, \min\{\varepsilon_k, 1\}\big] \Big\}$.
- $g'(\varepsilon_k) = \big|Z'_k\big|$.
- $\mu_i(A)$: the mean reward of the i-th best arm in A.
- $\Delta'_i = \max\big\{\mu_j(A^*) \mid j \in A^*\big\} - \mu_i(A^*)$: the suboptimality gap of i-th best arm in A^* with respect to the best arm in A^*.
- $\Delta''_i = \mu_1 - \mu_i(A^*)$: the suboptimality gap of i-th best arm in A^* with respect to the best arm in the entire instance.
- $U_1\big(\frac{\varepsilon}{2}\big) = \max_{i \geq g(\varepsilon_2)+1} \frac{i}{\Delta_i^2}$, and $U_2\big(\frac{\varepsilon}{2}\big) = \max_{i \geq g'(\varepsilon_2)+1} \frac{i}{\Delta_i'^2}$. We call this the un-accelerated sample complexity on the entire instance and the best bracket for identifying an $\frac{\varepsilon}{2}$-good arm.

We define the following events:

- E_1: $\forall k \in [K], \left|\frac{g'(\varepsilon_k)}{L} - \frac{g(\varepsilon_k)}{n}\right| < \frac{g(\varepsilon_k)}{2n}$.
- E_2: an $\frac{\varepsilon}{2}$-good arm represents the best bracket.

The event E_1 ensures that the best bracket can well represent the entire arm set in the sense that, for each peeling piece k, the ratio of the arms therein to all the arms in the best bracket $\frac{g'(\varepsilon_k)}{L}$ is close to its original ratio $\frac{g(\varepsilon_k)}{n}$. The ε-error probability can now be bounded as

$$\begin{aligned}
\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) &= \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1, E_2^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1, E_2) \\
&\le \mathbb{P}(E_1^c) + \mathbb{P}(E_1, E_2^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_2) \\
&\le \mathbb{P}(E_1^c) + \mathbb{P}\big(E_2^c \mid E_1\big) + \mathbb{P}\big(\mu_{J_t} < \mu_1 - \varepsilon \mid E_2\big). \qquad (2.17)
\end{aligned}$$

We bound each term separately. For the first term.

$$\begin{aligned}
\mathbb{P}(E_1^c) &= \mathbb{P}\left(\exists k, \left|\frac{g'(\varepsilon_k)}{L} - \frac{g(\varepsilon_k)}{n}\right| > \frac{g(\varepsilon_k)}{2n}\right) \\
&\le \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{g'(\varepsilon_k)}{L} - \frac{g(\varepsilon_k)}{n}\right| > \frac{g(\varepsilon_k)}{2n}\right) \\
&\le \sum_{k=1}^{K} \mathbb{P}\left(\frac{g'(\varepsilon_k)}{L} < \frac{g(\varepsilon_k)}{2n} \text{ or } \frac{g'(\varepsilon_k)}{L} > \frac{3g(\varepsilon_k)}{2n}\right) \\
&\le \sum_{k=1}^{K} \mathbb{P}\left(\frac{g'(\varepsilon_k)}{L} < \frac{g(\varepsilon_k)}{2n}\right) + \sum_{k=1}^{K} \mathbb{P}\left(\frac{g'(\varepsilon_k)}{L} > \frac{3g(\varepsilon_k)}{2n}\right).
\end{aligned}$$

Let $\mathsf{KL}(p, q)$ be the Kullback-Leibler divergence between two Bernoulli distributions with parameters p, q. The first summation can be upper bounded by the additive Chernoff bound for binomial distribution (Arratia and Gordon, 1989),

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{P}\left(\frac{g'(\varepsilon_k)}{L} < \frac{g(\varepsilon_k)}{2n}\right) &\le \sum_{k=1}^{K} \exp\left(-L\mathsf{KL}\left(\frac{g(\varepsilon_k)}{2n}, \frac{g(\varepsilon_k)}{n}\right)\right) \\
&\le \sum_{k=1}^{K} \exp\left(-\frac{1}{8} L \frac{\frac{g^2(\varepsilon_k)}{n^2}}{\frac{g(\varepsilon_k)}{n}}\right) \qquad \left(\mathsf{KL}(p, q) \ge \tfrac{(p-q)^2}{2\max\{p,q\}}\right) \\
&= \sum_{k=1}^{K} \exp\left(-\frac{1}{8} L \frac{g(\varepsilon_k)}{n}\right).
\end{aligned}$$

The second summation can be upper bounded by

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{P}\left(\frac{g'(\varepsilon_k)}{L} > \frac{3g(\varepsilon_k)}{2n}\right) &\le \sum_{k=1}^{K}\exp\left(-L\mathsf{KL}\left(1-\frac{3g(\varepsilon_k)}{2n}, 1-\frac{g(\varepsilon_k)}{n}\right)\right)\\
&\overset{(a_1)}{\le} \sum_{k=1}^{K}\exp\left(-L\mathsf{KL}\left(\frac{3g(\varepsilon_k)}{2n}, \frac{g(\varepsilon_k)}{n}\right)\right)\\
&\le \sum_{k=1}^{K}\exp\left(-\frac{1}{12}L\frac{\frac{g^2(\varepsilon_k)}{n^2}}{\frac{g(\varepsilon_k)}{n}}\right) \qquad \left(\mathsf{KL}(p,q)\ge\tfrac{(p-q)^2}{2\max\{p,q\}}\right)\\
&\le \sum_{k=1}^{K}\exp\left(-\frac{1}{12}L\frac{g(\varepsilon_k)}{n}\right),
\end{aligned}$$

where (a_1) is by the definition of KL divergence for Bernoulli distribution. Since $g(\varepsilon_k)$ is a non-decreasing function of k, we have

$$\begin{aligned}
\mathbb{P}(E_1^c) &\le \sum_{k=1}^{K}\exp\left(-\frac{1}{8}L\frac{g(\varepsilon_k)}{n}\right) + \sum_{k=1}^{K}\exp\left(-\frac{1}{12}L\frac{g(\varepsilon_k)}{n}\right)\\
&\le 2\left\lceil\log_2\left(\frac{8}{\varepsilon}\right)\right\rceil\exp\left(-\frac{1}{12}L\frac{g(\varepsilon_1)}{n}\right).
\end{aligned}$$

We choose $L = \tilde{\Theta}\left(\frac{nt}{g(\varepsilon_1)H_2\left(\frac{\varepsilon}{2}\right)}\right)$. Then

$$\begin{aligned}
\mathbb{P}(E_1^c) &\le 2\left\lceil\log_2\left(\frac{8}{\varepsilon}\right)\right\rceil\exp\left(-\tilde{\Theta}\left(\frac{t}{H_2\left(\frac{\varepsilon}{2}\right)}\right)\right)\\
&\le \exp\left(-\tilde{\Theta}\left(\frac{t}{H_2\left(\frac{\varepsilon}{2}\right)}\right)\right). \qquad \left(t \ge \tilde{\Theta}\left(H_2\left(\tfrac{\varepsilon}{2}\right)\right)\right)
\end{aligned} \tag{2.18}$$

Our choice of L here also ensures that a bracket of size L is opened because the largest bracket is in the size of $\tilde{\Theta}(t)$ by Lemma 11.1. Furthermore,

$$L = \tilde{\Theta}\left(\frac{nt}{g(\varepsilon_1)H_2\left(\frac{\varepsilon}{2}\right)}\right) = \tilde{\Theta}\left(\frac{nt}{\max_{i\ge g(\varepsilon_2)+1}\frac{i}{\Delta_i^2}}\right) \le \tilde{\Theta}\left(\frac{nt}{n/\Delta_n^2}\right) = \tilde{\Theta}\left(\Delta_n^2 t\right) \le \tilde{\Theta}(t).$$

The second term of (2.17) means the probability that SH fails to return an $\frac{\varepsilon}{2}$-good arm from bracket A^* given that bracket A^* can well represent the entire instance. Denote $H_2'\left(\frac{\varepsilon}{2}\right)$ as the sample complexity of returning an $\frac{\varepsilon}{2}$-good from the bracket A^*, i.e.,

$$H_2'\left(\frac{\varepsilon}{2}\right) = \max_{i \geq g'(\varepsilon_2)+1} \frac{i}{\Delta_i'^2} \cdot \frac{1}{g'(\varepsilon_1)}.$$

Let $H_2\left(\frac{\varepsilon}{2}\right)$ be the sample complexity of returning an $\frac{\varepsilon}{2}$-good arm from the entire instance, i.e.,

$$H_2\left(\frac{\varepsilon}{2}\right) = \max_{i \geq g(\varepsilon_2)+1} \frac{i}{\Delta_i^2} \cdot \frac{1}{g(\varepsilon_1)}.$$

We use Theorem 1 to upper bound the second term of (2.17). To do so, we need to verify that the best bracket is allocated to enough budget, i.e., the budget allocated to the best bracket should be no less than $U_2\left(\frac{\varepsilon}{2}\right)$. By event E_1, we have

$$\frac{1}{2} r g(\varepsilon_k) \leq g'(\varepsilon_k) \leq 2 r g(\varepsilon_k). \tag{2.19}$$

Then, by Lemma 13.1 and our choice of L, we have

$$U_2\left(\frac{\varepsilon}{2}\right) = g'(\varepsilon_1) H_2'\left(\frac{\varepsilon}{2}\right) \leq 2 r g(\varepsilon_1) H_2'\left(\frac{\varepsilon}{2}\right) = 2\frac{L}{n} g(\varepsilon_1) H_2'\left(\frac{\varepsilon}{2}\right) = \tilde{\Theta}\left(\frac{H_2'\left(\frac{\varepsilon}{2}\right)}{H_2\left(\frac{\varepsilon}{2}\right)} t\right) = \tilde{\Theta}(t).$$

Meanwhile, each bracket receives $\tilde{\Theta}(t)$ budget by the bracketing design. Thus the best bracket is allocated to enough budget. By applying Theorem 1 to (a_1) and Lemma 13.1 to (a_2) below, we have

$$\mathbb{P}\left(E_2^c \mid E_1\right) \overset{(a_1)}{\leq} \exp\left(-\tilde{\Theta}\left(\frac{t}{H_2'\left(\frac{\varepsilon}{2}\right)}\right)\right) \overset{(a_2)}{\leq} \exp\left(-\tilde{\Theta}\left(\frac{t}{H_2\left(\frac{\varepsilon}{2}\right)}\right)\right). \tag{2.20}$$

The third term of (2.17) means an arm whose mean reward is less than $\mu_1 - \varepsilon$ has a larger empirical reward than an arm whose mean reward is larger than $\mu_1 - \varepsilon/2$. Note the recently opened brackets may not have finished their first SH yet. In this case, the DSH always returns an empirical reward of negative infinity. Thus the arms returned from these brackets

will never be selected by BSH. Denote $\dot{A}(t) := \left\{a \in \bigcup_{k=1, A_k \neq A^*}^{L_t} A_k, \mu_a < \mu_1 - \varepsilon\right\}$. We have

$$\left|\dot{A}(t)\right| \leq \sum_{k=1}^{L_t} 2^k \leq 2 \cdot \left(2^{1+\log_2(1+\ln(2)t)} - 1\right) \leq 8 \cdot \left(1 + \ln(2)t\right).$$

By the number of arm pulls in Lemma 11.2,

$$\begin{aligned}
\mathbb{P}\left(\mu_{J_t} < \mu_1 - \varepsilon \mid E_2\right) \leq & \mathbb{P}\left(\exists a_1 \in A_{r_i^*(t)}, \exists a_2 \in \dot{A}(t), \text{s.t.}, \mu_{a_1} \geq \mu_1 - \varepsilon/2, \hat{\mu}_{a_1} < \hat{\mu}_{a_2}\right) \\
\leq & \left|\dot{A}(t)\right| \left|A_{r_i^*(t)}\right| \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t\right)\right) \\
\leq & \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t\right) + \ln\left(\left|\dot{A}(t)\right| \left|A_{r_i^*(t)}\right|\right)\right) \\
\leq & \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t\right) + \ln\left(8 \cdot \left(1 + \ln(2)t\right)t\right)\right) \\
\leq & \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t\right) + \mathcal{O}(\ln t)\right) \\
\leq & \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t\right)\right). \qquad (2.21)
\end{aligned}$$

Given (2.18), (2.20) and (2.21), we have the result proved. □

Lemma 13.1. *Let $H_2'\left(\frac{\varepsilon}{2}\right)$ be the complexity measure of returning an $\frac{\varepsilon}{2}$-good from the bracket A^*, i.e.,*

$$H_2'\left(\frac{\varepsilon}{2}\right) = \max_{i \geq g'(\varepsilon_2)+1} \frac{i}{\Delta_i'^2} \cdot \frac{1}{g'(\varepsilon_1)}.$$

Let $H_2\left(\frac{\varepsilon}{2}\right)$ be the complexity measure of returning an $\frac{\varepsilon}{2}$-good from the entire instance, i.e.,

$$H_2\left(\frac{\varepsilon}{2}\right) = \max_{i \geq g(\varepsilon_2)+1} \frac{i}{\Delta_i^2} \cdot \frac{1}{g(\varepsilon_1)}.$$

Given the event E_1 happens, we have

$$H_2'\left(\frac{\varepsilon}{2}\right) \leq 64 H_2\left(\frac{\varepsilon}{2}\right).$$

Proof. For the entire instance, by Lemma 13.2,

$$\max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\varepsilon_k^2} \le U_1\left(\frac{\varepsilon}{2}\right) \le \max_{k\in\{3,\cdots,K\}} \frac{4g(\varepsilon_k)}{\varepsilon_k^2}. \tag{2.22}$$

For the best bracket, by Lemma 13.3, we have

$$\max_{i\ge g'(\varepsilon_2)+1} \frac{i}{\Delta_i''^2} \le \max_{i\ge g'(\varepsilon_2)+1} \frac{i}{\Delta_i'^2} \le 4 \max_{i\ge g'(\varepsilon_2)+1} \frac{i}{\Delta_i''^2}.$$

Note the term $\max_{i\ge g'(\varepsilon_2)+1} \frac{i}{\Delta_i''^2}$ can be considered the un-accelerated sample complexity on a virtual instance where all the arms are the same as the ones of A^*, except the best arm of A^* is replaced as the best arm in the entire instance (if the best arm of A^* is same as the best arm in the entire instance, then the virtual instance is same as A^*). The above inequalities essentially show that on event E_1, setting the best arm in A^* to have the same mean reward as the best arm of the entire instance does not change its un-accelerated sample complexity (up to a constant of 4). Denote

$$U_2'\left(\frac{\varepsilon}{2}\right) = \max_{i\ge g'(\varepsilon_2)+1} \frac{i}{\Delta_i''^2}.$$

Thus,

$$U_2'\left(\frac{\varepsilon}{2}\right) \le U_2\left(\frac{\varepsilon}{2}\right) \le 4U_2'\left(\frac{\varepsilon}{2}\right). \tag{2.23}$$

By Lemma 13.2,

$$\max_{k\in\{3,\cdots,K\}} \frac{g'(\varepsilon_k)}{\varepsilon_k^2} \le U_2'\left(\frac{\varepsilon}{2}\right) \le \max_{k\in\{3,\cdots,K\}} \frac{4g'(\varepsilon_k)}{\varepsilon_k^2}.$$

By event E_1, we have for $k \in [K]$

$$\frac{1}{2} r g(\varepsilon_k) \le g'(\varepsilon_k) \le 2r g(\varepsilon_k). \tag{2.24}$$

Thus

$$\max_{k\in\{3,\cdots,K\}} \frac{rg(\varepsilon_k)}{2\varepsilon_k^2} \le U_2'\left(\frac{\varepsilon}{2}\right) \le \max_{k\in\{3,\cdots,K\}} \frac{8rg(\varepsilon_k)}{\varepsilon_k^2}. \tag{2.25}$$

By (2.23) and (2.25),

$$\max_{k\in\{3,\cdots,K\}} \frac{rg(\varepsilon_k)}{2\varepsilon_k^2} \le U_2\left(\frac{\varepsilon}{2}\right) \le \max_{k\in\{3,\cdots,K\}} \frac{32rg(\varepsilon_k)}{\varepsilon_k^2}. \tag{2.26}$$

Therefore, by (2.22) and (2.26), we have

$$\frac{1}{32r} \le \frac{U_1\left(\frac{\varepsilon}{2}\right)}{U_2\left(\frac{\varepsilon}{2}\right)} \le \frac{8}{r}. \tag{2.27}$$

By $U_1\left(\frac{\varepsilon}{2}\right) = g(\varepsilon_1)H_2\left(\frac{\varepsilon}{2}\right)$, $U_2\left(\frac{\varepsilon}{2}\right) = g'(\varepsilon_1)H_2'\left(\frac{\varepsilon}{2}\right)$, (2.24) and (2.27), we have

$$H_2'\left(\frac{\varepsilon}{2}\right) = \frac{g(\varepsilon_1)U_2\left(\frac{\varepsilon}{2}\right)}{g'(\varepsilon_1)U_1\left(\frac{\varepsilon}{2}\right)} H_2\left(\frac{\varepsilon}{2}\right) \le \frac{2}{r}\cdot 32r\cdot H_2\left(\frac{\varepsilon}{2}\right) = 64H_2\left(\frac{\varepsilon}{2}\right).$$

□

Lemma 13.2. *Let $U\left(\frac{\varepsilon}{2}\right)$ be the un-accelerated sample complexity on any bandit instance for identifying an $\frac{\varepsilon}{2}$ good arm (note $\varepsilon_2 = \frac{\varepsilon}{2}$), i.e.,*

$$U\left(\frac{\varepsilon}{2}\right) = \max_{i\ge g(\varepsilon_2)+1} \frac{i}{\Delta_i^2}.$$

Then up to a constant of 4, the un-accelerated sample complexity can be approximated by $\max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\varepsilon_k^2}$, i.e.,

$$\max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\varepsilon_k^2} \le U\left(\frac{\varepsilon}{2}\right) \le \max_{k\in\{3,\cdots,K\}} \frac{4g(\varepsilon_k)}{\varepsilon_k^2}.$$

Proof. Let $i_0 \ge g(\varepsilon_2) + 1$ satisfy

$$U\left(\frac{\varepsilon}{2}\right) = \max_{i\ge g(\varepsilon_2)+1} \frac{i}{\Delta_i^2} = \frac{i_0}{\Delta_{i_0}^2}.$$

Denote k_0 as the smallest index of the peeling interval to which i_0 belongs, i.e., $\varepsilon_{k_0-1} < \Delta_{i_0} \leq \varepsilon_{k_0}$. Since $i_0 \geq g(\varepsilon_2) + 1$, we know $k_0 \geq 3$. By $\varepsilon_{k_0-1} = \frac{\varepsilon_{k_0}}{2}$, we have $\frac{\varepsilon_{k_0}}{2} < \Delta_{i_0}$. Thus,

$$U\left(\frac{\varepsilon}{2}\right) = \frac{i_0}{\Delta_{i_0}^2} < \frac{4i_0}{\varepsilon_{k_0}^2} \leq \frac{4g(\varepsilon_{k_0})}{\varepsilon_{k_0}^2} \leq \max_{k\in\{3,\cdots,K\}} \frac{4g(\varepsilon_k)}{\varepsilon_k^2},$$

where the second inequality is by $i_0 \leq g(\varepsilon_{k_0})$. By the definition of $U\left(\frac{\varepsilon}{2}\right)$, we have

$$U\left(\frac{\varepsilon}{2}\right) \geq \max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\Delta_{g(\varepsilon_k)}^2} \geq \max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\varepsilon_k^2}$$

Thus,

$$\max_{k\in\{3,\cdots,K\}} \frac{g(\varepsilon_k)}{\varepsilon_k^2} \leq U\left(\frac{\varepsilon}{2}\right) \leq \max_{k\in\{3,\cdots,K\}} \frac{4g(\varepsilon_k)}{\varepsilon_k^2}.$$

□

Lemma 13.3. *Given that the event E_1 holds, for $i \geq g'(\varepsilon_2) + 1$, Δ_i' the suboptimality gap of i-th best arm in the best bracket with respect to the best arm in the best bracket can be approximated by Δ_i'', the suboptimality gap of i-th best arm in the best bracket with respect to the best arm in the entire instance, i.e.,*

$$\frac{1}{\Delta_i''^2} \leq \frac{1}{\Delta_i'^2} \leq \frac{4}{\Delta_i''^2}.$$

Proof. The first inequality holds trivially, and the equality holds when the best arm out of the entire arm set is included in the best bracket. For the second inequality, as the best arm in the best bracket is $\frac{\varepsilon}{4}$-good, we have

$$\Delta_i' - \frac{\varepsilon}{4} \geq \Delta_i'' - \frac{\varepsilon}{2} \geq \frac{\Delta_i''}{2} - \frac{\varepsilon}{4},$$

which results in the second inequality. □

2.9.17 Proof of Corollary 6.1

Corollary 6.1. *Consider the EqualGap(m) instance. BUCB achieves an expected (ε,δ)-unverifiable sample complexity as*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

BSH achieves an (ε,δ)-unverifiable sample complexity as, with probability $1-\delta$,

$$\tau_{\varepsilon,\delta} \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

Proof. For BUCB, Theorem 7 of Katz-Samuels and Jamieson (2020) gives the following instance dependent upper bound for the (ε,δ)-unverifiable sample complexity,

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \min_{j\in[g(\varepsilon)]}\frac{1}{j}\left(\sum_{i=1}^{g(\varepsilon)}\left(\Delta_{i\vee j,g(\varepsilon)+1}\right)^{-2}\ln\left(\frac{n}{j\delta}\right)+\sum_{i=g(\varepsilon)+1}^{n}\Delta_{j,i}^{-2}\ln\left(\frac{1}{\delta}\right)\right).$$

Plug the instance

$$\mu_1=\frac{5}{4}\varepsilon,\mu_i=\varepsilon,\forall i\in\{2,\ldots,m\},\mu_i=0,\forall i\geq m+1 \text{ for some } m\geq 2,\varepsilon>0\ .$$

We easily have

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

For BSH,

$$\max_{i\geq g(\varepsilon/2)+1}\frac{i}{\Delta_i^2}\cdot\frac{1}{g\left(\frac{\varepsilon}{4}\right)}=\frac{16n}{25m\varepsilon^2}.$$

□

2.9.18 Proof of Corollary 6.2

Corollary 6.2. *Consider the Polynomial(α) instance; i.e., $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ with $\alpha > 0.5$. For any $\varepsilon \in (0,1)$, let $\hat{\tau}_{\varepsilon,\delta}$ be the upper bound of the expected (ε,δ)-unverifiable sample complexity reported in Katz-Samuels and Jamieson (2020, Theorem 7) for BUCB. Then, BUCB satisfies*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \le \hat{\tau}_{\varepsilon,\delta} = \tilde{\Theta}\left(\varepsilon^{-\frac{2\alpha-1}{\alpha}} n \log\left(\frac{1}{\delta}\right)\right).$$

On the other hand, BSH satisfies, with probability $1-\delta$,

$$\tau_{\varepsilon,\delta} \le \tilde{\mathcal{O}}\left(\varepsilon^{-2}\log\left(\frac{1}{\delta}\right)\right).$$

Proof. We first show the result for Bracketing UCB. Theorem 7 of Katz-Samuels and Jamieson (2020) gives the following instance dependent upper bound for the (ε,δ)-unverifiable sample complexity,

$$\hat{\tau}_{\varepsilon,\delta} = \min_{j\in[g(\varepsilon)]} \frac{1}{j}\left(\sum_{i=1}^{g(\varepsilon)} \left(\Delta_{i\vee j, g(\varepsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right) + \sum_{i=g(\varepsilon)+1}^{n} \Delta_{j,i}^{-2} \ln\left(\frac{1}{\delta}\right)\right). \tag{2.28}$$

Plug the instance $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ into the above,

$$
\begin{aligned}
\hat{\tau}_{\varepsilon,\delta} &= \min_{j\in[g(\varepsilon)]} \frac{1}{j}\left(\sum_{i=1}^{g(\varepsilon)} \left(\Delta_{i\vee j, g(\varepsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right) + \sum_{i=g(\varepsilon)+1}^{n} \Delta_{j,i}^{-2} \ln\left(\frac{1}{\delta}\right)\right) \\
&> \min_{j\in[g(\varepsilon)]} \frac{1}{j}\left(\sum_{i=1}^{g(\varepsilon)} \left(\Delta_{i\vee j, g(\varepsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right)\right) \\
&> \min_{j\in[g(\varepsilon)]} \frac{1}{j} \ln\left(\frac{1}{\delta}\right)\left(\sum_{i=1}^{g(\varepsilon)} \left(\Delta_{i\vee j, g(\varepsilon)+1}\right)^{-2}\right) \\
&> \min_{j\in[g(\varepsilon)]} \frac{1}{j} \ln\left(\frac{1}{\delta}\right)\left(\left(\Delta_{g(\varepsilon), g(\varepsilon)+1}\right)^{-2}\right) \\
&> \min_{j\in[g(\varepsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\left(g(\varepsilon)+1\right)^\alpha - \left(g(\varepsilon)\right)^\alpha\right)^2}\right) \\
&\overset{(a_1)}{>} \min_{j\in[g(\varepsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\left(\alpha\left(g(\varepsilon)+1\right)^{\alpha-1}\right)^2}\right) \\
&> \min_{j\in[g(\varepsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\alpha^2\left(2g(\varepsilon)\right)^{2\alpha-2}}\right) \\
&= \frac{2^{-2\alpha+2}\alpha^{-2}n^{2\alpha}}{\left(g(\varepsilon)\right)^{2\alpha-1}} \ln\left(\frac{1}{\delta}\right) \\
&\overset{(a_2)}{=} \frac{2^{-2\alpha+2}\alpha^{-2}n}{\varepsilon^{\frac{2\alpha-1}{\alpha}}} \ln\left(\frac{1}{\delta}\right).
\end{aligned}
$$

Note here we lower bound the upper bound in Theorem 7 of Katz-Samuels and Jamieson (2020). The inequality (a_1) results by

$$
(g+1)^\alpha - g^\alpha = \int_g^{g+1} \alpha x^{\alpha-1}\,\mathrm{d}x \le \alpha(g+1)^{\alpha-1}.
$$

The equation (a_2) is because of $g(\varepsilon) = n\varepsilon^{1/\alpha}$ by the definition of the instance.

We also upper bound (2.28). Note that for $i > j$, we have

$$
i^\alpha - j^\alpha = \int_j^i \alpha x^{\alpha-1}\,\mathrm{d}x \ge \alpha(i-j)\cdot j^{\alpha-1}\ .
$$

With this, we have

$$\hat{\tau}_{\varepsilon,\delta} \le n^{2\alpha}\left(\min_{j\le g(\varepsilon)} \frac{1}{(g(\varepsilon)+1-j)^2 j^{2\alpha-2}} + \frac{1}{j}\sum_{i=j+1}^{g(\varepsilon)} \frac{1}{(g(\varepsilon)+1-i)^2 i^{2\alpha-2}} + \frac{1}{j}\sum_{i=g(\varepsilon)+1}^{n} \frac{1}{(i-j)^2 j^{2\alpha-2}}\right)$$

We can bound the third sum by

$$\frac{1}{j}\sum_{i=g(\varepsilon)+1}^{n} \frac{1}{(i-j)^2 j^{2\alpha-2}} \lesssim \frac{1}{j^{2\alpha-1}}\cdot\frac{1}{g(\varepsilon)+1-j}$$

Take $j = g(\varepsilon) - \sqrt{g(\varepsilon)}$. This means that $j = \Theta(g(\varepsilon))$ and $g(\varepsilon) - j = \sqrt{g(\varepsilon)}$. With this choice,

$$\begin{aligned}
\hat{\tau}_{\varepsilon,\delta} &\lesssim n^{2\alpha}\left(\frac{1}{g(\varepsilon)\cdot g(\varepsilon)^{2\alpha-2}} + \frac{1}{g(\varepsilon)}\cdot\left(\sum_{i=g(\varepsilon)-\sqrt{g(\varepsilon)}+1}^{g(\varepsilon)} \frac{1}{(g(\varepsilon)+1-i)^2}\cdot\frac{1}{g(\varepsilon)^{2\alpha-2}}\right) + \frac{1}{g(\varepsilon)^{2\alpha-1}}\cdot\frac{1}{1+\sqrt{g(\varepsilon)}}\right) \\
&\le n^{2\alpha}\left(\frac{1}{g(\varepsilon)^{2\alpha-1}} + \frac{1}{g(\varepsilon)^{2\alpha-1}}\cdot\frac{\pi^2}{6} + \frac{1}{g(\varepsilon)^{2\alpha-1}}\cdot\frac{1}{1+\sqrt{g(\varepsilon)}}\right) \\
&\lesssim n^{2\alpha}\cdot\frac{1}{g(\varepsilon)^{2\alpha-1}} = n^{2\alpha}\frac{1}{(n\varepsilon^{1/\alpha})^{2\alpha-1}} = \frac{n}{\varepsilon^{\frac{2\alpha-1}{\alpha}}}.
\end{aligned}$$

For Bracketing SH,

$$\max_{i\ge g(\varepsilon/2)+1} \frac{i}{\Delta_i^2}\cdot\frac{1}{g\left(\frac{\varepsilon}{4}\right)} = \max_{i\ge g(\varepsilon/2)+1} i^{1-2\alpha} n^{2\alpha}\frac{1}{n\left(\frac{\varepsilon}{4}\right)^{1/\alpha}} < \left(n\left(\frac{\varepsilon}{2}\right)^{1/\alpha}\right)^{1-2\alpha} n^{2\alpha}\frac{1}{n\left(\frac{\varepsilon}{4}\right)^{1/\alpha}} < \frac{16}{\varepsilon^2}.$$

where both the last inequality and the last equality are by $\alpha > 0.5$.

□

2.9.19 An alternative upper bound for BSH

The following bound of BSH is similar to the bound of Theorem 6. Though the bound of Theorem 14 is minimax in nature, we show, in Corollary 14.1 and 14.2, that it still achieves the same upper bound as BUCB for the EqualGap(m) instance and a better upper bound for Polynomial(α) than BUCB when n is large enough because it does not scale with n polynomially. Since this bound involves an optimization problem, we speculate it could be tighter than the bound of Theorem 6 for certain instances.

Theorem 14. *For any $\varepsilon \in (0,1)$, the error probability of Bracketing SH satisfies*

$$\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) \le \exp\left(-\tilde{\Theta}\left(\min\left\{\max_{i\in[g(\varepsilon/2)]} \frac{i\Delta^2_{i,g(\varepsilon/2)+1}}{n}, \varepsilon^2\right\} t'\right)\right),$$

where $t' = \frac{t}{4\ln t}$. Accordingly, $\tilde{\mathcal{O}}\left(\max\left\{\min_{i\in[g(\varepsilon/2)]} \frac{n}{i\Delta^2_{i,g(\varepsilon/2)+1}} \log\left(\frac{1}{\delta}\right), \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right\}\right)$ samples are sufficient for Bracketing SH to output an ε-good arm with probability $1-\delta$.

Proof. The proof idea is similar to the proof of Theorem 6. The core idea is to find the best bracket whose size is well-balanced. At round t, define the best bracket as bracket $r_i^*(t) = \left\lceil \log_2\left(c\Delta^2_{i,g(\varepsilon/2)+1} t'\right)\right\rceil$, where $i \in [g(\varepsilon/2)]$ is a free parameter and c is a logarithmic term for shorthand $c = \left(16\ln(4e)\log_2^2\left(\frac{i\Delta^2_{i,g(\varepsilon/2)+1}t'}{n}\right)\log_2\left(2\Delta^2_{i,g(\varepsilon/2)+1}t'\right)\right)^{-1}$. Thus the size of bracket $r_i^*(t)$ satisfies,

$$c\Delta^2_{i,g(\varepsilon/2)+1}t' \le \left|A_{r_i^*(t)}\right| \le 2^{\log_2\left(c\Delta^2_{i,g(\varepsilon/2)+1}t'\right)+1} = 2c\Delta^2_{i,g(\varepsilon/2)+1}t'.$$

Define the following events:

- E_1: the number of arms with mean reward at least μ_i included in bracket $r_i^*(t)$ is at least j_t, where $j_t = \left\lfloor \frac{ci\Delta^2_{i,g(\varepsilon/2)+1}t'}{2n}\right\rfloor$.
- E_2: $\mu_{a_{r_i^*(t)}} \ge \mu_1 - \varepsilon/2$, where $a_{r_i^*(t)}$ is the arm representing bracket $r_i^*(t)$ at round t.

Then the error probability of Bracketing SH can be expressed as follows,

$$\begin{aligned}\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) &= \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1, E_2^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_1, E_2)\\ &\le \mathbb{P}(E_1^c) + \mathbb{P}(E_1, E_2^c) + \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_2). \qquad (2.29)\end{aligned}$$

We bound the three terms respectively. For the first term, we use the same technique as in the proof sketch.

$$\begin{aligned}
\mathbb{P}(E_1^c) &\le \exp\left(-\left|A_{r_i^*(t)}\right| \cdot \mathsf{KL}\left(\frac{j_t}{\left|A_{r_i^*(t)}\right|}, \frac{i}{n}\right)\right) \\
&\overset{(a_1)}{\le} \exp\left(-\frac{\left|A_{r_i^*(t)}\right| n}{2i}\left(\frac{j_t}{\left|A_{r_i^*(t)}\right|} - \frac{i}{n}\right)^2\right) \\
&\le \exp\left(-\frac{\Delta_{i,g(\varepsilon/2)+1}^2 t' n}{2i}\left(\frac{\frac{i\Delta_{i,g(\varepsilon/2)+1}^2 t'}{2n}}{\Delta_{i,g(\varepsilon/2)+1}^2 t'} - \frac{i}{n}\right)^2\right) \\
&\le \exp\left(-\frac{ci\Delta_{i,g(\varepsilon/2)+1}^2 t'}{8n}\right) \\
&\le \exp\left(-\tilde{\Theta}\left(\frac{i\Delta_{i,g(\varepsilon/2)+1}^2 t'}{n}\right)\right).
\end{aligned} \tag{2.30}$$

The inequality (a_1) is due to $\mathsf{KL}(p,q) \ge \frac{(p-q)^2}{2\max\{p,q\}}$. For the second term of (2.29), we bound the event that SH returns a non-$(j_t, \Delta_{i,g(\varepsilon/2)+1})$-good arm, where j_t is with respect to the best bracket. We use $\mu_{(j_t)}$ for the mean reward of the j_t-th best arm in bracket $r_i^*(t)$. Thus $\mu_{(j_t)} - \Delta_{i,g(\varepsilon/2)+1} \ge \mu_1 - \varepsilon/2$. By Theorem 4,

$$\begin{aligned}
\mathbb{P}(E_1, E_2^c) &\le \mathbb{P}\left(\mu_{a_{r_i^*(t)}} < \mu_1 - \varepsilon/2 \mid E_1\right) \\
&\le \mathbb{P}\left(\mu_{a_{r_i^*(t)}} < \mu_{(j_t)} - \Delta_{i,g(\varepsilon/2)+1} \mid E_1\right) \\
&\overset{(a_1)}{\le} \log_2\left|A_{r_i^*(t)}\right| \cdot \exp\left(-\text{const}\cdot j_t\left(\frac{\Delta_{i,g(\varepsilon/2)+1}^2 t'}{4\left|A_{r_i^*(t)}\right| \log_2^2(2j_t)\log_2\left|A_{r_i^*(t)}\right|} - \ln(4e)\right)\right) \\
&\le \log_2\left|A_{r_i^*(t)}\right| \cdot \exp\left(-\text{const}\cdot\frac{ci\Delta_{i,g(\varepsilon/2)+1}^2 t'}{n}\right) \\
&\le \exp\left(-\tilde{\Theta}\left(\frac{i\Delta_{i,g(\varepsilon/2)+1}^2 t'}{n}\right)\right).
\end{aligned} \tag{2.31}$$

For the inequality (a_1), we bound the term in parenthesis as a constant,

$$\begin{aligned}
&\frac{\Delta^2_{i,g(\varepsilon/2)+1}t'}{4\left|A_{r_i^*(t)}\right|\log_2^2(2j_t)\log_2\left|A_{r_i^*(t)}\right|} - \ln(4e)\\
>&\frac{\Delta^2_{i,g(\varepsilon/2)+1}t'}{8c\Delta^2_{i,g(\varepsilon/2)+1}t'\log_2^2(2c\frac{i\Delta^2_{i,g(\varepsilon/2)+1}t'}{2n})\log_2\left(2c\Delta^2_{i,g(\varepsilon/2)+1}t'\right)} - \ln(4e)\\
>&\frac{1}{8c\log_2^2(\frac{i\Delta^2_{i,g(\varepsilon/2)+1}t'}{n})\log_2\left(2\Delta^2_{i,g(\varepsilon/2)+1}t'\right)} - \ln(4e)\\
>&2\ln(4e) - \ln(4e)\\
=&\ln(4e).
\end{aligned}$$

The third term of (2.29) means an arm whose mean reward is less than $\mu_1 - \varepsilon$ has a larger empirical reward than an arm whose mean reward is larger than $\mu_1 - \varepsilon/2$. Note the recently opened brackets may not have finished their first SH yet. In this case, the DSH always returns an empirical reward of negative infinity. Thus the arms returned from these brackets will never be selected by BSH. Denote $\dot{A}(t) := \left\{a \in \bigcup_{k=1,k\neq r_i^*(t)}^{L_t} A_k, \mu_a < \mu_1 - \varepsilon\right\}$. We have

$$\left|\dot{A}(t)\right| \le \sum_{k=1}^{L_t} 2^k \le 2\cdot\left(2^{1+\log_2(1+\ln(2)t)} - 1\right) \le 8\cdot\left(1 + \ln(2)t\right).$$

By the number of arm pulls in Lemma 11.2,

$$\begin{aligned}
\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon, E_2) \le& \mathbb{P}\left(\exists a_1 \in A_{r_i^*(t)}, \exists a_2 \in \dot{A}(t), \text{s.t.}, \mu_{a_1} \ge \mu_1 - \varepsilon/2, \hat{\mu}_{a_1} < \hat{\mu}_{a_2}\right)\\
\le& \left|\dot{A}(t)\right|\left|A_{r_i^*(t)}\right|\exp\left(-\left(\frac{\varepsilon}{2}\right)^2\cdot\text{const}\cdot\frac{t'}{\log_2 n}\right)\\
\le& \exp\left(-\text{const}\cdot\varepsilon^2\frac{t'}{\log_2 n} + \ln\left(\left|\dot{A}(t)\right|\left|A_{r_i^*(t)}\right|\right)\right)\\
\le& \exp\left(-\text{const}\cdot\varepsilon^2\frac{t'}{\log_2 n} + \ln\left(8\cdot\left(1+\ln(2)t\right)2\Delta^2_{i,g(\varepsilon/2)+1}t'\right)\right)\\
\le& \exp\left(-\text{const}\cdot\varepsilon^2\frac{t'}{\log_2 n} + \mathcal{O}(\ln t)\right)\\
\le& \exp\left(-\tilde{\Theta}\left(\varepsilon^2 t'\right)\right).
\end{aligned} \tag{2.32}$$

Combine (2.30)(2.31)(2.32) and the fact $i \in [g(\varepsilon/2)]$ is a free parameter, then we have

$$\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) \leq \exp\left(-\tilde{\Theta}\left(\min\left\{\max_{i\in[g(\varepsilon/2)]} \frac{i\Delta^2_{i,g(\varepsilon/2)+1}}{n}, \varepsilon^2\right\} t'\right)\right).$$

□

Corollary 14.1. *Consider the EqualGap(m) instance. BUCB achieves an expected (ε, δ)-unverifiable sample complexity as*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

For BSH, the bound of Theorem 14 shows that BSH achieves an (ε, δ)-unverifiable sample complexity as

$$\tau_{\varepsilon,\delta} \leq \tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^2 m}\log\left(\frac{1}{\delta}\right)\right)$$

with probability $1-\delta$.

Proof. For BUCB, the proof is presented in the proof of Corollary 6.1. For BSH, the result is straightforward by plugging the instance,

$$\min_{i\in[g(\varepsilon/2)]} \frac{n}{i\Delta^2_{i,g(\varepsilon/2)+1}}\log\left(\frac{1}{\delta}\right) \leq \min\left\{\frac{n}{(\varepsilon/2)^2}\log\left(\frac{1}{\delta}\right), \frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right\} = \frac{n}{m\varepsilon^2}\log\left(\frac{1}{\delta}\right).$$

□

The following corollary shows that even if we use the upper bound of Theorem 14 that is minimax in nature, we can still achieve an upper bound for the Polynomial(α) instance that does not scale with the instance size n, and for moderately large ε or large enough n, the upper bound of Theorem 14 is always better than the upper bound of BUCB.

Corollary 14.2. *Consider the Polynomial(α) instance $\Delta_i = \left(\frac{i}{n}\right)^\alpha$. BUCB achieves an expected (ε, δ)-unverifiable sample complexity as*

$$\mathbb{E}\left[\tau_{\varepsilon,\delta}\right] \leq \hat{\tau}_{\varepsilon,\delta} = \tilde{\Theta}\left(\varepsilon^{-\frac{2\alpha-1}{\alpha}} n \log\left(\frac{1}{\delta}\right)\right).$$

For BSH, the bound of Theorem 14 shows that BSH achieves an (ε, δ)-unverifiable sample complexity as

$$\tau_{\varepsilon,\delta} \leq \tilde{\mathcal{O}}\left(\varepsilon^{-\frac{2\alpha+1}{\alpha}} \log\left(\frac{1}{\delta}\right)\right).$$

with probability $1-\delta$.

Proof. For BUCB, the proof is presented in the proof of Corollary 6.2. For BSH, first, notice that

$$\begin{aligned}\Delta^2_{i,j+1} > \Delta^2_{i,j} =& \left(\left(\frac{j}{n}\right)^\alpha - \left(\frac{i}{n}\right)^\alpha\right)^2 \\ =& \left(\frac{j}{n}\right)^{2\alpha} + \left(\frac{i}{n}\right)^{2\alpha} - 2\left(\frac{ij}{n^2}\right)^\alpha.\end{aligned}$$

The sample complexity satisfies,

$$\begin{aligned}\min_{i\in[g(\varepsilon/2)]} \frac{n}{i\Delta^2_{i,g(\varepsilon/2)+1}} \log\left(\frac{1}{\delta}\right) \leq& \min_{i\in[g(\varepsilon/2)]} \frac{n}{i\left(\left(\frac{g(\varepsilon/2)}{n}\right)^{2\alpha} + \left(\frac{i}{n}\right)^{2\alpha} - 2\left(\frac{ig(\varepsilon/2)}{n^2}\right)^\alpha\right)} \log\left(\frac{1}{\delta}\right) \\ =& \min_{i\in[g(\varepsilon/2)]} \frac{n\cdot n^{2\alpha}}{i\left(\left(g(\varepsilon/2)\right)^{2\alpha} + i^{2\alpha} - 2\left(ig(\varepsilon/2)\right)^\alpha\right)} \log\left(\frac{1}{\delta}\right) \\ \overset{(a_1)}{\leq}& \frac{2\cdot n^{2\alpha+1}}{(g(\varepsilon/2))^{2\alpha+1}(1+2^{-2\alpha}-2^{-\alpha+1})} \log\left(\frac{1}{\delta}\right) \\ =& \frac{2\cdot n^{2\alpha+1}}{(n(\varepsilon/2)^{1/\alpha})^{2\alpha+1}(1+2^{-2\alpha}-2^{-\alpha+1})} \log\left(\frac{1}{\delta}\right) \\ =& 2^{\frac{2\alpha+1}{\alpha}+1}\left(1+2^{-2\alpha}-2^{-\alpha+1}\right)^{-1} \varepsilon^{-\frac{2\alpha+1}{\alpha}} \log\left(\frac{1}{\delta}\right).\end{aligned}$$

The inequality (a_1) is by taking $i = g(\varepsilon/2)/2$. □

2.9.20 (ε, δ)-unverifiable sample complexity of algorithms with ε-error probability bound

Definition 1 ((ε, δ)-unverifiable sample complexity (Katz-Samuels and Jamieson, 2020))**.** *For an algorithm π and an instance ρ. Let $\tau_{\varepsilon,\delta}$ be a stopping time such that*

$$\mathbb{P}\Big(\forall t \geq \tau_{\varepsilon,\delta} : \mu_{J_t} > \mu_1 - \varepsilon\Big) \geq 1 - \delta.$$

Then, $\tau_{\varepsilon,\delta}$ is called (ε, δ)-unverifiable sample complexity of the algorithm with respect to ρ.

The (ε, δ)-unverifiable sample complexity requires all the outputs at and after $t = \tau_{\varepsilon,\delta}$ are ε-good. This is slightly different from the sample complexity of BSH. However, we show they are order-wise equivalent. To see this, let us consider an anytime algorithm that achieves an exponentially decreasing error probability,

$$\mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) \leq \exp(-c{\cdot}t).$$

Let $t_{\varepsilon,\delta}$ be the first time step that satisfies $\mathbb{P}\Big(\mu_{J_{t_{\varepsilon,\delta}}} < \mu_1 - \varepsilon\Big) \leq \delta$. Note this only guarantees the output at round $t_{\varepsilon,\delta}$ is ε-good with probability $1 - \delta$, instead of all the outputs at and after $t_{\varepsilon,\delta}$. We prove our claim by showing the probability that there is a non-ε-good output at or after $t_{\varepsilon,\delta}$ has an exponentially decreasing rate with the same parameter.

$$\begin{aligned}
\mathbb{P}\Big(\exists t \geq t_{\varepsilon,\delta} : \mu_{J_t} < \mu_1 - \varepsilon\Big) &\leq \sum_{t=t_{\varepsilon,\delta}}^{\infty} \mathbb{P}(\mu_{J_t} < \mu_1 - \varepsilon) \\
&\leq \sum_{t=t_{\varepsilon,\delta}}^{\infty} \exp(-c{\cdot}t) \\
&= \lim_{t\to\infty} \exp\Big(-c{\cdot}t_{\varepsilon,\delta}\Big)\frac{1 - \exp(-c{\cdot}t)}{1 - \exp(-c)} \\
&< \exp\Big(-c{\cdot}t_{\varepsilon,\delta}\Big)\frac{1}{1 - \exp(-c)} \\
&= \exp\Big(-\Theta(c{\cdot}t_{\varepsilon,\delta})\Big).
\end{aligned}$$

Chapter 3

Adaptive Experimentation When You Can't Experiment

This paper introduces the *confounded pure exploration transductive linear bandit* (`CPET-LB`) problem. As a motivating example, often online services cannot directly assign users to specific control or treatment experiences either for business or practical reasons. In these settings, naively comparing treatment and control groups that may result from self-selection can lead to biased estimates of underlying treatment effects. Instead, online services can employ a properly randomized encouragement that incentivizes users toward a specific treatment. Our methodology provides online services with an adaptive experimental design approach for learning the best-performing treatment for such *encouragement designs*. We consider a more general underlying model captured by a linear structural equation and formulate pure exploration linear bandits in this setting. Though pure exploration has been extensively studied in standard adaptive experimental design settings, we believe this is the first work considering a setting where noise is confounded. Elimination-style algorithms using experimental design methods in combination with a novel finite-time confidence interval on an instrumental variable style estimator are presented with sample complexity upper bounds nearly matching a minimax lower bound. Finally, experiments are conducted that demonstrate the efficacy of our approach.

3.1 Introduction

In this study, we present a methodology for adaptive experimentation in scenarios characterized by potential confounding. Online services routinely conduct thousands of A/B tests annually (Kohavi et al., 2020). In most online A/B/N experimentation, meticulous user-level randomization is essential to ensure unbiased estimates of *treatment effects at the*

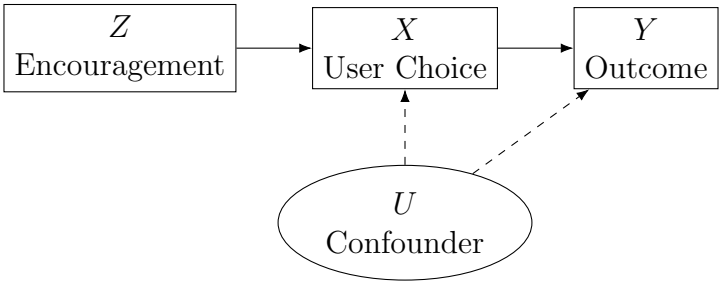


Figure 3.1: Causal graph of the model.

population level, commonly known as average treatment effects (ATE). In this setting, firms are often interested in understanding the treatment with the highest average outcome if presented to all members of the population. However, in many settings firms may not be able to randomize, for example if a feature must be rolled out to all users for various business reasons (Mummalaneni et al., 2022). In such instances, users may choose to engage with a feature or not based on potentially unobservable preferences. Thus the resulting measured outcome may be correlated with the decision to engage in a specific feature. I.e. the underlying characteristics of the user *confound* the relationship between the decision to use the feature being evaluated and the effect of the feature. Thus, naively comparing the average outcome for users who engage with a feature with those who do not suffers from a (selection) bias. This setting is captured in Figure 3.1.

A potential solution is for services to employ *encouragement designs* where users are presented with incentives that encourage users to engage with a specific feature (Bakshy et al., 2014; Bradlow, 1998; Elbers, 2023). As a concrete example, many online services have introduced membership levels with different offerings and prices available to all users. Given a set of membership level options, the service is interested in knowing the counterfactual of which level has the optimal outcome (e.g., total revenue) if every user chooses to join that membership level. In this setting, encouragements could be coupons or trials for corresponding membership levels. In these settings, the firm can use intent-to-treat (ITT) estimates for the treatment effect which naively compare the average outcomes between the groups given different encouragements. In practice, given an encouragement a user may not engage with the corresponding feature choosing either the control or a different feature. Hence, the resulting ITT estimate may be a diluted estimate of the ATE (Angrist et al., 1996). However,

all is not lost: if the encouragement presented to a user is properly randomized, and the service guarantees that the encouragement only affects the outcome through the choice of user treatment, then the encouragement acts as an *instrumental variable*. Standard analysis from the econometrics and compliance literature show that two-stage least squares (`2SLS`) estimators can then be used to provide consistent estimates of treatment effects.

At the same time, firms are also increasingly utilizing adaptive experimentation techniques, often known as *pure exploration multi-armed bandits* (`MAB`) algorithms (Lattimore and Szepesvári, 2020; Fiez et al., 2019), to accelerate traditional A/B/N testing. Pure exploration `MAB` techniques promise to deliver accurate inferences in a fraction of the time and cost as traditional methods. Similar to A/B testing, bandit methods assume users are properly randomized and can fail to learn the optimal treatment if naively used and may be sample inefficient if they fail to take the confounded structure into account.

Contributions. In this work, we provide a methodology for experimenters seeking to use adaptive experimentation in settings with confounding where encouragements are available. We formulate this work in the more general and novel setting of *confounded pure exploration transductive linear bandits* (`CPET-LB`) (Section 3.1.1). We present algorithms using experimental design for the `CPET-LB` problem and analyze the resulting sample complexity. As we demonstrate, even in the simple multi-armed bandit setting described above, computing an effective sampling pattern requires using the machinery of linear bandits. Without knowledge of the underlying structural model, existing linear bandit approaches could lead to sub-optimal sampling. The main technical challenge we face is simultaneously improving our estimate of the structural model while designing with inaccurate estimates (Section 3.3). This approach crucially relies on our development of novel finite-time confidence bounds for two-stage least squares (`2SLS`) style estimators that may be of independent interest (Section 3.2.2). Moreover, we provide worst-case sample complexity lower bounds that are nearly matched by our sample complexity upper bounds (Section 3.7.3). Though the goal of this work is primarily theoretical, we empirically show the efficacy of our method over existing solutions (Section 3.4).

3.1.1 General Problem Formulation

A confounded pure exploration transductive linear bandits (`CPET-LB`) instance consists of finite collections of measurement vectors $\mathcal{Z} \subset \mathbb{R}^d$ and evaluation vectors $\mathcal{W} \subset \mathbb{R}^d$. At each time $t \in \mathbb{N}$, the learner selects $z_t \in \mathcal{Z}$ and observes a pair of noisy responses $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_t \in \mathbb{R}$ generated via the structural equation model

$$x_t = \Gamma^\top z_t + \eta_t, \quad y_t = x_t^\top \theta + \varepsilon_t, \tag{3.1}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ and $\theta \in \mathbb{R}^d$ are model parameters.[1] We define the history $\mathcal{H}_{t-1} = \{(z_s, x_s, y_s)\}_{s<t}$ and $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot|\mathcal{H}_{t-1}]$ denoting the conditional expectation under the filtration generated by $\mathcal{H}_{t-1}$. The noise $\{\eta_t\}_{t=1}^\infty$ and $\{\varepsilon_t\}_{t=1}^\infty$ satisfy the following set of assumptions unless otherwise noted.

Assumption 1. *We assume $\varepsilon_t \mid \mathcal{H}_{t-1}$ is 1-sub-Gaussian (and thus $\mathbb{E}[\varepsilon_t \mid \mathcal{H}_{t-1}] = 0$). Furthermore, $\eta_t \mid \mathcal{H}_{t-1}$ is σ_η^2-sub-Gaussian vectors (and thus $\mathbb{E}[\eta_t \mid \mathcal{H}_{t-1}] = 0$), i.e.,*

$$\forall \beta \in \mathbb{R}, \max_{a:\|a\|_2\le 1} \mathbb{E}[\exp(\beta\langle a, \eta_t\rangle)] \le \exp\left(\frac{\beta^2\sigma_\eta^2}{2}\right).$$

Goal. The objective is to identify $w^* := \arg\max_{w\in\mathcal{W}} w^\top\theta$ with probability at least $1-\delta$ for $\delta \in (0,1)$ while taking a minimum number of measurements.

In the setting where $\Gamma = I, \eta = 0$ and $\mathbb{E}_{t-1}[\varepsilon_t|x_t] = 0$, our setting reduces to the standard *pure exploration transductive linear bandit* problem (Soare et al., 2014; Fiez et al., 2019). In general, the joint noise process $\left[\eta_t, \varepsilon_t\right]$ may be dependent across the entries. In particular, we are allowing for the data generating process to be *endogenous*, meaning that $\mathbb{E}_{t-1}[\varepsilon_t|x_t] \neq 0$. That is, ε_t can affect not just y_t, but also x_t given a choice of z_t. The presence of endogeneity is a key challenge in the `CPET-LB` problem.

Assumption 2 (Exclusion Restriction)**.** *We assume that $\mathbb{E}_{t-1}[z_t\varepsilon_t] = 0$, or alternatively that z_t is uncorrelated with ε_t.*

[1] We assume throughout that $\Gamma \in \mathbb{R}^{d\times d}$ is an invertible matrix.

The variable z_t is commonly referred to as an *instrumental variable* (Angrist et al., 1996). We consider algorithms for the `CPET-LB` problem that stop at a $\mathcal{H}_t$-measurable time $\tau \in \mathbb{N}$, and produce a recommendation $\widehat{w} \in \mathcal{W}$. The goal is δ-PAC algorithms with efficient sample complexity guarantees.

Definition 2 (δ-PAC)**.** *We say an algorithm is δ-PAC for a `CPET-LB` problem with $\mathcal{W}, \mathcal{Z} \subset \mathbb{R}^d$ if for all $\theta \in \mathbb{R}^d$ and $\Gamma \in \mathbb{R}^{d \times d}$, it holds that $\mathbb{P}_{\theta,\Gamma}(\widehat{w} \neq w^*) \leq \delta$ for $\delta \in (0,1)$.*

3.1.2 Encouragement Designs

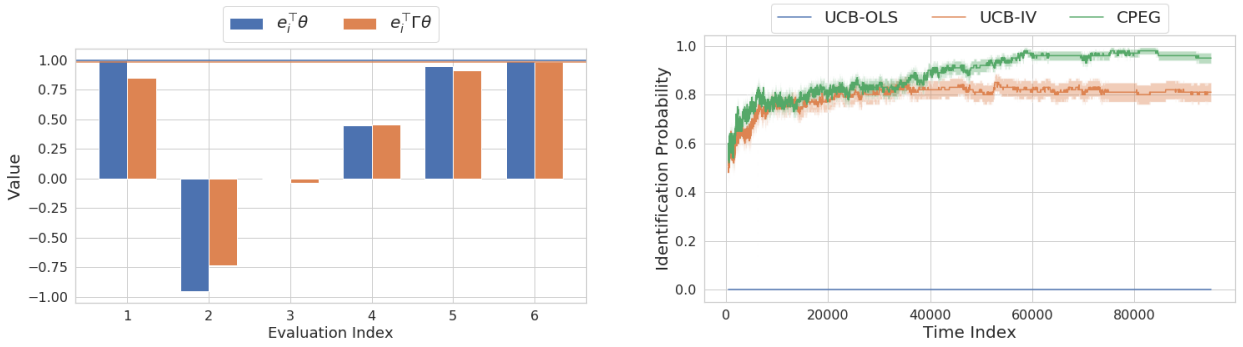
The `CPET-LB` feedback model generalizes the classical *compliance* setting.

Compliance as a Special Case. In *compliance* problems, a decision-maker has access to a set of treatment that can be offered to users, while the users themselves have the option to accept the treatment they are presented or instead opt-in to a different treatment. The goal is to identify the treatment with the optimal average outcome if all users were to accept it. Specifically, given a finite set $\mathcal{A} = \{1, 2, \ldots, d\}$, a decision-maker presents user $t \in \mathbb{N}$ with an encouragement for a treatment $i \in \mathcal{A}$, the user then selects treatment $j \in \mathcal{A}$ where potentially $j \neq i$, and an outcome y_t results. To capture compliance with the `CPET-LB` framework, we set $\mathcal{Z} = \mathcal{X} = \mathcal{W} = \{e_1, \cdots, e_d\}$ and the parameter Γ captures the probability of accepting a treatment given an encouragement for a potentially different treatment. Specifically, $\Gamma(i,j) = \mathbb{P}\big(x_t = e_i \mid z_t = e_j\big)$, and a straightforward computation shows that $x_t = \Gamma^\top z_t + \eta_t$ where $\mathbb{E}[\eta_t | z_t] = 0$ with

$$\eta_t = x_t - \Big[\mathbb{P}\big(e_1 \mid z_t\big), \cdots, \mathbb{P}\big(e_d \mid z_t\big)\Big]^\top. \tag{3.2}$$

Moreover, $y_t = x_t^\top \theta + \varepsilon_t$ gives the resulting reward, which is clearly correlated with the user choice so that $\mathrm{cov}(\eta_t, \varepsilon_t) \neq 0$. Finally, $e_i^\top \theta = \theta_i$ gives the expected value of treatment i over the population and our goal is to identify $w^* = \arg\max_{e_i \in \mathcal{W}} e_i^\top \theta$. [2] Note that when $\Gamma = I$, we

[2] Our setting differs slightly from the traditional compliance setting based on a potential outcomes framework (Angrist et al., 1996). Our setting is equivalent to one where we assume that there is a constant treatment effect. See Chapter 4 of Angrist et al. (1996) for further discussion of the differences. Thus in our setting, learning θ_i and the local average treatment effect (LATE) on compilers are the same.



(a) $w^\top\theta$ and $w^\top\Gamma\theta$ for $w \in \{e_1, \ldots, e_d\}$

(b) $\mathbb{P}(\widehat{w}_t = w^*)$ for the instance

Figure 3.2: (a) A bar chart showing $\mathbb{E}[y|x = w] = w^\top\theta$ and $\mathbb{E}[y|z = w] = z^\top\Gamma\theta$ for all $w \in \mathcal{W}$. This chart shows that the optimal evaluation vector is $w^* = e_1 = \arg\max_{w\in\mathcal{W}} \mathbb{E}[y|x = w]$, while $e_6 = \arg\max_{w\in\mathcal{W}} \mathbb{E}[y|z = w]$ and consequently estimation based on this quantity is problematic. (b) The probability of identifying $w^* = e_1$ for a collection of algorithms on the `CPET-LB` instance from Section 3.1.2.

automatically have that $\eta_t = 0$ and there is no confounding. This reduces to the standard MAB setting.

Motivating Compliance Example.

As a motivating compliance example representing the membership level discussion from the introduction, consider a location model that assumes each user $t \in \mathbb{N}$ arriving online has an underlying unobserved one-dimensional preference $u_t \sim \mathcal{N}(0, \sigma_u^2)$. If an algorithm presents the user with encouragement $z_{I_t} = e_{I_t}$ for $I_t \in \mathcal{A}$, then the user selects into the membership level given by $J_t = \min_{j\in\mathcal{A}} |I_t + u_t - j|$ so that $x_t = e_{J_t}$. This process captures a user being more likely to opt-in to membership levels that are closer to the encouragement that they were presented. The outcome is then given by $y_t = x_t^\top\theta + u_t$.

We conduct an experiment with this problem instance (see Fig. 3.2 and Section 3.7.1 for more details). Specifically, $d = 6$, $\theta = \begin{bmatrix} 1 & -0.95 & 0 & 0.45 & 0.95 & 0.99 \end{bmatrix}$, and $\sigma_u^2 = 0.35$. Observe that $w^* = e_1 = \arg\max_{w\in\mathcal{W}} w^\top\theta$. An upper confidence bound (UCB) selection strategy is simulated that maintains estimates of the mean reward of each encouragement $i \in \mathcal{A}$, namely $\widehat{\mu}_{i,t} = \sum_{s=1}^t \mathbf{1}\{z_t = e_i\} y_t$, and then pulls the one with the highest UCB.

The UCB selection strategy is combined with a pair of recommendation strategies. The UCB-OLS algorithm estimates the mean reward of each treatment using an OLS estimator, namely $\widehat{\theta}_{\text{LS}}^{i,t} = \sum_{s=1}^t \mathbf{1}\{x_t = e_i\} y_t / \sum_{s=1}^t \mathbf{1}\{x_t = e_i\}$, and recommends $\arg\max_{a\in\mathcal{A}} \widehat{\theta}_{\text{LS}}^{i,t}$. Moreover,

the UCB-IV algorithm uses an instrumental variable-estimator (see the next section) that incorporates knowledge of Γ similar to 2SLS to deconfound estimates of the mean rewards of each treatment and recommends the treatment with the maximum estimate. The results over 100 simulations are shown in Fig. 3.2. UCB-OLS completely fails to identify $\theta_1 = \arg\max_{i \in d} \theta_i$ due to a biased estimate, whereas UCB-IV does better. However, UCB-IV methods seem to have a constant probability of error. To see why, note that the expected reward from pulling $z = e_i$ is $e_i^\top \Gamma \theta$. These values are plotted in orange in Figure 3.2a. In particular, with some constant probability, UCB zeroes in on arm 6 becauses of the mean estimates on the z's, and as a result fails to give enough samples to learn that arm 1 is indeed the best. In contrast, our proposed method CPEG, Algorithm 4 manages to find the best arm with significantly higher probability.

Notation. Let $\Delta(\mathcal{Z}) = \{\lambda \in \mathbb{R}^{|\mathcal{Z}|} : \lambda \geq 0, \sum_{z \in \mathcal{Z}} \lambda_z = 1\}$ denote the set of probability distributions over the set $\mathcal{Z}$. Given a distribution $\lambda \in \Delta(\mathcal{Z})$ and matrix $\Gamma \in \mathbb{R}^{d \times d}$, define the operator $A(\lambda, \Gamma) := \sum_{z \in \mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma$. Given $Z \in \mathbb{R}^{T \times d}$ and $\Gamma \in \mathbb{R}^{d \times d}$, define the operator $\bar{A}(Z, \Gamma) := \sum_{t=1}^{T} \Gamma^\top z_t z_t^\top \Gamma = \Gamma^\top Z^\top Z \Gamma$ where $z_t \in \mathbb{R}^d$ denotes row t of Z. Given a vector $x \in \mathbb{R}^d$ and a symmetric positive-definite matrix $A \in \mathbb{R}^{d \times d}$ we let $\|x\|_A^2 = x^\top A x$. We adopt the standard notation that $(a \vee b) \equiv \max\{a, b\}$ and $(a \wedge b) \equiv \min\{a, b\}$ for $a, b \in \mathbb{R}$. $\sigma_{\min}(A), \sigma_{\max}(A)$ denote the minimum and maximum singular value of a matrix A. We denote by $\text{polylog}(x_1, \ldots, x_n)$ any polylogarithmic factors of $x_1, \ldots, x_n$.

3.2 Estimators and Inference

We now present estimators for the unknown parameter θ and prove the associated statistical properties. The estimators discussed in this section are critical to our algorithmic solution outlined in Section 3.3.

3.2.1 Estimators

Before describing our solution concept, we quickly review potential options for estimating θ based on a dataset $Z_T = [z_1, \cdots, z_T]^\top \in \mathbb{R}^{T \times d}$, $X_T = [x_1, \cdots, x_T]^\top \in \mathbb{R}^{T \times d}$, $Y_T = [y_1, \cdots, y_T]^\top \in \mathbb{R}^T$, assumed to be generated according to the model in Eq. (3.1). Recall that the *ordinary-least-*

squares (OLS) estimator for θ is given by

$$\widehat{\theta}_{\texttt{LS}} := \arg\min_{\widehat{\theta}\in\mathbb{R}^d} \sum_{t=1}^{T}(y_t - x_t^\top\widehat{\theta})^2 = (X_T^\top X_T)^{-1}X_T^\top Y_T = \theta + (X_T^\top X_T)^{-1}X_T^\top\varepsilon_T. \tag{3.3}$$

Observe that $\widehat{\theta}_{\texttt{LS}}$ is potentially a biased and inconsistent estimator for θ in the presence of endogenous noise since $\mathbb{E}[\varepsilon_t|x_t] \neq 0$. To remediate this problem, we define a general class of estimators that includes several standard estimators. Given an invertible matrix $\Psi \in \mathbb{R}^{d\times d}$, let $\bar{X}_T := Z_T\Psi$, and consider corresponding estimators termed Ψ-IV estimators of the form

$$\widehat{\theta}_{\Psi} := (\bar{X}_T^\top\bar{X}_T)^{-1}\bar{X}_T^\top Y_T = (\Psi^\top Z_T^\top Z_T\Psi)^{-1}\Psi^\top Z_T^\top Y_T = (Z_T^\top Z_T\Psi)^{-1}Z_T^\top Y_T. \tag{3.4}$$

When $\Psi = I$ we recover the OLS estimator. In the rest of the paper, we will focus on two different potential options for Ψ.

Case 1: Oracle. $\Psi = \Gamma$. To begin, observe that the structural equation model from Eq. (3.1) can be combined by substituting the second equation into the first to obtain the reduced form

$$y_t = z_t^\top\Gamma\theta + \eta_t^\top\theta + \varepsilon_t. \tag{3.5}$$

Since z_t is independent of the i.i.d. process $\eta_t^\top\theta + \varepsilon_t$, the least squares estimator which regresses y_t onto $z_t^\top\Gamma$ is unbiased for estimation of θ and given by

$$\widehat{\theta}_{\texttt{oracle}} = (\bar{X}_T^\top\bar{X}_T)^{-1}\bar{X}_T^\top Y_T = (Z_T^\top Z_T\Gamma)^{-1}Z_T^\top Y_T. \tag{3.6}$$

This estimator will be used to design our general solution concept presented in Section 3.3. Of course in practice we cannot expect to know Γ, but we may be able to estimate it.

Case 2: P-2SLS. $\Psi = \widehat{\Gamma}$. We consider a setting where $\widehat{\Gamma}$ is an (unbiased) estimator of Γ, learned using least-squares from an *independent* dataset $Z_{T_1} = [z'_1, \cdots, z'_{T_1}], X_{T_1} = [z'_1, \cdots, z'_{T_1}]$ collected non-adaptively.[3] That is, $\widehat{\Gamma} = (Z_{T_1}^\top Z_{T_1})^{-1}Z_{T_1}^\top X_{T_1}$ and:

$$\widehat{\theta}_{\texttt{P-2SLS}} = (\widehat{\Gamma}^\top Z_T^\top Z_T\widehat{\Gamma})^{-1}\widehat{\Gamma}^\top Z_T^\top Y_T.$$

[3]Formally, we say that a set of data (Z_T, X_T, Y_T) generated via the model in Eq. (3.1) is collected non-adaptively from an experimental design if z_t is $\mathcal{H}_0$ measurable for all $1 \leq t \leq T$.

We refer to the resulting estimator as a pseudo two stage least squares (P-2SLS) estimator. The main advantage of the P-2SLS estimator over standard 2SLS (given in Section 3.7.2) is easier inference since now $\{\varepsilon_t\}_{t\leq T}$ of our dataset is independent of the measurements of the first dataset Z_{T_1}, X_{T_1}. In the econometrics literature, such an estimator is referred to as a *two-sample 2SLS estimator* (Inoue and Solon, 2010).

3.2.2 Confidence Intervals

In the section that follows, we develop a general algorithmic approach that relies on *experimental design* aimed at reducing the uncertainty in our estimates of the optimal treatment. To this end, we first develop finite-time confidence intervals for estimators presented in the previous section given data generated according to the model in Eq. (3.1) and collected from non-adaptive designs.

We begin by characterizing the properties of the noise structure in the combined model of Eq. (3.5) with the following set of results.

Lemma 14.1. *Under Assumption 1, the noise process $\nu := \eta^\top\theta + \varepsilon$ is σ_ν^2-sub-Gaussian where $\sigma_\nu^2 = 2(\sigma_\eta^2\|\theta\|_2^2 + 1)$, specifically when the instance is compliance, $\sigma_\nu^2 = 2(4\|\theta\|_2^2 + 1)$.*

Oracle Confidence Interval. As in the last section, we assume that we have access to a dataset (Z_T, X_T, Y_T) generated according to Eq. 3.1 and collected non-adaptively. Given Lemma 14.1, it can be shown that $w^\top\widehat{\theta}_{\texttt{oracle}}$ is a sub-Gaussian random variable satisfying the following.

Lemma 14.2. *With probability at least $1-\delta$ for $\delta \in (0,1)$ and $w \in \mathbb{R}^d$,*

$$|w^\top(\widehat{\theta}_{\texttt{oracle}} - \theta)| \leq \sqrt{2\sigma_\nu^2\|w\|^2_{A(Z_T,\Gamma)^{-1}}\log\big(2/\delta\big)},$$

where σ_ν^2 is the sub-Gaussian parameter of the noise $\nu := \eta^\top\theta+\varepsilon$ characterized in Lemma 14.1.

The proof of this result is in Section 3.7.7.

P-2SLS Confidence Interval. We now present a novel finite-time confidence interval for the P-2SLS estimator. As discussed in the previous section with respect to this estimator, we assume access a set of data (Z_{T_1}, X_{T_1}) generated according to Eq. (3.1) and collected

non-adaptively for the purpose of estimating Γ. Moreover, assume access to a separate set of data $(Z_{T_2}, X_{T_2}, Y_{T_2})$ generated according to Eq. (3.1) and collected non-adaptively for the purpose of estimating θ.

Theorem 15. *Suppose that* $\widehat{\Gamma} = (Z_{T_1}^\top Z_{T_1})^{-1} Z_{T_1}^\top X_{T_1}$ *and* $\widehat{\theta}_{\texttt{P-2SLS}} = (\widehat{\Gamma}^\top Z_{T_2} Z_{T_2} \widehat{\Gamma})^{-1} \widehat{\Gamma} Z_{T_2}^\top Y_{T_2}$. *Then, for any* $w \in \mathcal{W}$, *with probability at least* $1-\delta$ *for* $\delta \in (0,1)$,

$$|w^\top(\widehat{\theta}_{\texttt{P-2SLS}} - \theta)| \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4}{\delta}\right)} + \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}} \|\theta\|_2 \sqrt{\sigma_\eta^2 \overline{\log}\big(Z_{T_1}, \delta/4\big)},$$

where

$$\overline{\log}(Z_T, \delta) := 8d \ln\left(1 + \frac{2TL_z^2}{d(2 \wedge \sigma_{\min}(Z_{T_1}^\top Z_{T_1}))}\right) + 16 \ln\left(\frac{2 \cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(Z_{T_1}^\top Z_{T_1})}\right)\right).$$

The proof is presented in Section 3.7.8. Observe that the first term in the `P-2SLS` estimator confidence interval given by $\sqrt{2\sigma_\nu^2 \|w\|^2_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}} \log(4/\delta)}$ matches the `Oracle` estimator confidence interval in Lemma 14.2 when $\widehat{\Gamma} = \Gamma$. The second term scaling like $\mathcal{O}(\|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}} \|\theta\|_2 \sigma_\eta \sqrt{d + \log(1/\delta)})$, is an upper bound on the approximation error $w^\top(\widehat{\Gamma}^{-1}\Gamma - I)\theta$ for any $w \in \mathbb{R}^d$, assuming that $\widehat{\Gamma}$ is learned from an `OLS` estimator (see Theorem 19 for details).

We will see that the form of this confidence interval is particularly convenient for our algorithmic approach given in Section 3.3. In particular, the form of the variance $\|w\|^2_{\bar{A}(Z_{T_2},\Gamma)^{-1}}$ on the first term only depends on a design over instruments. Thus, we can choose an experimental design over Z's which reduces this variance optimally.

Remark 3. *In practice we expect the first stage of samples,* (Z_{T_1}, X_{T_1}) *to be collected from either a burn-in period or from existing historical data. We remark that assuming two stages of samples is common in the orthogonal and double machine learning for estimating nuisance parameters in the data generating process (e.g.* Γ*) (Chernozhukov et al., 2018). Our result matches the existing literature on the asymptotic variance of two sample* `2SLS` *estimators (e.g., Theorem 1 of Inoue and Solon (2010)).*

Remark 4. *The asymptotic variance of standard* **2SLS** *is known to involve a factor* σ_ε^2*, instead of* σ_ν^2 *as we have (Greene, 2003). Recent work by Della Vecchia and Basu (2023) shows a variance involving* $d\sigma_\varepsilon^2$*. However, it's unclear how to use the form of their confidence interval directly for experimental design. In addition, their work is not sufficiently general to handle the general forms of noise that we consider in Lemma 14.1.*

3.3 Adaptive Experimental Design Algorithms

We now present adaptive experimental design algorithms for the `CPET-LB` problem. Our main insight utilizes Eq. 3.1 by plugging the model for x into the top equation resulting in the relationship

$$y = z^\top \Gamma \theta + \theta^\top \eta + \varepsilon.$$

When Γ is known, by Eq. 3.5, we see that `CPET-LB` reduces to a standard pure exploration transductive linear bandit problem where the measurement set is given by $\{\Gamma^\top z\}_{z \in \mathcal{Z}} \subset \mathbb{R}^d$, the evaluation set is $\mathcal{W} \subset \mathbb{R}^d$, and the feedback model is given by $y = v^\top \theta + \nu$ where the noise $\nu = \theta^\top \eta + \varepsilon$ is sub-Gaussian and as before the goal is to identify $\arg\max_{w \in \mathcal{W}} w^\top \theta$. An existing approach to this problem is given by the `RAGE` algorithm (Fiez et al., 2019), which we use as the basis of our approach. Addressing the case of unknown Γ is our major algorithmic contribution, where we develop solutions to improve our estimate of Γ and learn w^* simultaneously. As a warm-up to this approach, we first consider the setting when Γ is known.

3.3.1 Warm-Up: Known Structural Model

Algorithm 4 assumes a parameter L_ν, which acts as an upper bound on the sub-Gaussian constant of the noise $\nu = \theta^\top \eta + \varepsilon$. In each round k, an active set of potentially optimal vectors $\widehat{\mathcal{W}}_k \subset \mathcal{W}$ is maintained. CPEG aims to sample in such a way that reduces the uncertainty of the estimates on the *gaps* $(w - w')^\top \theta$ for each pair $w, w' \in \widehat{\mathcal{W}}_k$ maximally each round. In any given round the algorithm takes N_k samples Z_{N_k}, the confidence interval of Lemma 14.2 shows that the error in estimating $(w - w')^\top \theta$ scales with $\|w - w'\|^2_{(\Gamma^\top Z_{N_k}^\top Z_{N_k} \Gamma)^{-1}}$. This motivates utilizing an *experimental design* approach where we choose a distribution

$\lambda_k \in \Delta(\mathcal{Z})$ to minimize $\max_{w,w' \in \widehat{\mathcal{W}}_k} \|w - w'\|_{(\sum_{z \in \mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)^{-1}}$. The number of resulting samples taken from this design N_k is chosen to guarantee that the confidence interval of Lemma 14.2 is less than 2^{-k}. Then, the elimination step in Line 8 guarantees that all $w \in \mathcal{W}$ such that $(w^* - w)^\top \theta > 2 \cdot 2^{-k}$ are then eliminated from the active set by round $k+1$ of the procedure. To actually choose our samples, as is common in this literature Fiez et al. (2019), we use an efficient rounding procedure, ROUND that requires a minimum number of samples $r(\omega)$.

Sample Complexity Guarantee. The sample complexity of Algorithm 4 depends on the following problem-dependent quantity $\rho^*(\gamma)$ that captures the underlying hardness of a problem instance in terms of $(\mathcal{W}, \mathcal{Z}, \Gamma, \theta)$, when $\gamma = 0$, we abbreviate $\rho^*(0) = \rho^*$,

$$\rho^*(\gamma) = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w \in \mathcal{W} \setminus \{w^*\}} \frac{\|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)^{-1}}}{\langle w^* - w, \theta \rangle^2 \vee \gamma^2}. \tag{3.7}$$

Algorithm 4: CPEG: Confounded pure exploration with Γ

Input: $\mathcal{Z}, \mathcal{W}, \Psi = \Gamma, \delta, L_\nu \geq \sigma_\nu^2, \omega$

Initialize: $k = 1, \mathcal{W}_1 = \mathcal{W}, \zeta_1 = 1$;

Set: $f(w, w', \Gamma, \lambda) := \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)^{-1}}$;

while $|\mathcal{W}_k| > 1$ **do**

- $\lambda_k = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} f(w, w', \Gamma, \lambda)$;
- $\rho(\mathcal{W}_k) = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} f(w, w', \Gamma, \lambda)$;
- $N_k := \lceil 2(1+\omega) 2^{2k} \rho(\mathcal{W}_k) L_\nu \log\big(4k^2 |\mathcal{W}|/\delta\big) \rceil \vee r(\omega)$;
- Pull arms in $Z_{N_k} = \text{ROUND}(\lambda_k, N_k)$ and observe Y_{N_k};
- Compute $\widehat{\theta}^k_\Gamma = \left(Z_{N_k}^\top Z_{N_k} \Gamma\right)^{-1} Z_{N_k}^\top Y_{N_k}$;
- $\mathcal{W}_{k+1} = \mathcal{W}_k \backslash \{w \in \mathcal{W}_k | \exists w' \in \mathcal{W}_k, \langle w' - w, \widehat{\theta}^k_\Gamma \rangle > 2^{-k}\}$;
- $k \leftarrow k + 1$;

Output: The single remaining element in $\mathcal{W}_k$

Theorem 16. *Algorithm 4 is δ-PAC and terminates in at most $c(1+\omega) L_\nu \rho^* \log(1/\delta) + c r(\omega)$ samples, where c hides logarithmic factors of $\Delta := \min_w \langle w^* - w, \theta \rangle$ and $|\mathcal{W}|$, as well as constants.*

The proof of this result is in Section 3.7.9. In the unconfounded case when $\Gamma = I$ and $\eta = 0, \mathbb{E}_{t-1}[\varepsilon_t | x_t] = 0$ this matches the sample complexity of Fiez et al. (2019). In particular,

for the case where $\mathcal{Z} = \mathcal{X} = \mathcal{W}$, the problem further reduces to a standard multi-armed bandit, and if ε is 1-sub-Gaussian noise, Soare et al. (2014) shows that $\rho^* = O(\sum_{i=2}^{d}(\theta_1 - \theta_i)^{-2}))$, which is the optimal sample complexity of best-arm identification for multi-armed bandits. The following lemma shows that the conditioning of Γ can have a strong impact on the resulting sample complexity.

Lemma 16.1. *For the compliance setting, we have* $\min_{\lambda \in \Delta^d} \max_{j,j'} \|e_j - e_{j'}\|^2_{(\sum_{i=1}^{d} \lambda_i \Gamma^\top e_i e_i^\top \Gamma)^{-1}} \leq d \max_{j,j'} \|\Gamma^{-1}(e_j - e_{j'})\|_2^2$. *Furthermore,* $\rho^* \leq \frac{d\sigma^2_{\min}(\Gamma)^{-1}}{\Delta^2_{\min}}$.

To further illustrate the impact of Γ, imagine an extreme setting where $\Gamma = (1-\varepsilon)/d\mathbf{1}\mathbf{1}^\top + \varepsilon I$ and $\varepsilon \approx 0$, i.e. Γ is a perturbation of $1/d\mathbf{1}\mathbf{1}^\top$. It's straightforward to show that the upper bound in the first display of Lemma 16.1 is of the order $O(d\varepsilon^{-2})$ (this is also a lower bound - see Section 3.7.13). In particular, the upper bound on the sample complexity is of the form $d\varepsilon^{-2}/\Delta^2_{\min}$. This is in sharp contrast to the linear bandit case, when $\Gamma = I$ and we are guaranteed a sample complexity of no more than $d/\Delta^2_{\min}$ samples. To gain some intuition, regardless of the choice of λ, $\sum_{i \leq d} \lambda_i \Gamma^\top e_i e_i^\top \Gamma \approx \Gamma$. As a result, $\rho^* \to \infty$ as $\varepsilon \to 0$. Intuitively in the limit, regardless of which instrument $i \leq d$ is being pulled, the resulting distribution on the treatments is uniform (the instruments are *weak*). Thus, it is impossible to deconfound the measurement noise, and recover an estimate of θ. This is a phenomenon which does not arise in the standard multi-armed bandit case with unconfounding.

Remark 5. *We also consider a setting where instead of given* Γ *directly, we are given an estimate* $\widehat{\Gamma}$ *of* Γ *based on offline data. We discuss such an adaptation of Algorithm 4 to this setting in Section 3.7.10 and provide a sample complexity which reflects the error in* Γ *(scaling with* $\rho^*(\gamma)$ *for* $\gamma > 0$*). We remark that this result is subsumed by the approach of Section 3.3.2 and so we omit it in the main text.*

Lower bound. Due to the noise model from confounding and the dependence of the noise $\theta^\top \eta + \varepsilon$, the instance-dependent lower bounds of Fiez et al. (2019) do not immediately apply. We develop a lower bound tailored for the confounding setting **that nearly match the upper bounds** of our algorithms. What's more, our lower bound illustrates the additional difficulty that arises from confounding by an additional factor of d^2 compared to the standard

transductive linear bandit problem in the most general setting where entries of η are sub-Gaussian, but not necessarily independent nor bounded. Due to space limit, we defer it to Section 3.7.3.

3.3.2 Fully Unknown Structural Model

We now consider the setting where Γ is fully unknown. The difficulty of this setting is that the data collection process needs to support both estimation of Γ and θ simultaneously.

Our algorithm, built upon Algorithm 4, is summarized in Algorithm 6. At its core, each phase of the algorithm is divided into two sub-phases, for estimating Γ and θ respectively. Specifically, the second sub-phase is essentially same as Algorithm 4 with $\widehat{\Gamma}_k$ in place of Γ where $\widehat{\Gamma}_k$ is estimated from the first sub-phase. The main novelty of our algorithmic design lies in the first sub-phase, which resolves the challenge of performing the optimal design for estimating Γ. To explain this challenge, the confidence interval for P-2SLS estimators of Theorem 15 indicates that one should pull arms so that we control both $D_2 \coloneqq \max_{w,w'} \|w - w'\|^2_{\bar{A}(Z_{T_2},\widehat{\Gamma}_k)}$ (error from $\hat{\theta}_{\mathsf{P\text{-}2SLS}}$) and $D_1 \coloneqq \max_{w,w'} \|w - w'\|^2_{\bar{A}(Z_{T_1},\widehat{\Gamma}_k)}$ (error from $\widehat{\Gamma}_k$) to be below the target error $O(\zeta_k^2)$ at each phase (ignoring unimportant factors for discussion). Controlling D_2 is trivial, which is done in the second sub-phase as we described above.

However, for D_1, a similar strategy cannot be done because the estimate $\widehat{\Gamma}_k$ is computed directly by sampling arms in Z_{T_1}. That is, the ideal design, based on which we *will* collect data points $z_1, \ldots, z_n$, requires access to the random matrix $\widehat{\Gamma}_k$ that can only be computed *after* sampling $z_1, \ldots, z_n$,. This creates a cycle that seems impossible to resolve. Such an issue, to our knowledge, has not been seen in existing work on pure exploration, and thus resolving it is our key technical contribution.

Our solution is to compute the design based on $\widehat{\Gamma}_k$ from the previous phase. We then perform a doubling trick where we double the sample size (while following the computed design) until D_1 becomes smaller than the target error $O(\zeta_k^2)$. The intuition is that in later phases the estimate $\widehat{\Gamma}_k$ from the previous phase will be accurate enough to ensure that the design is efficient. Note that this novel algorithm induces extra randomness in how many samples we end up collecting in the first sub-phase, which remains random even

Algorithm 6: CPEUG: Confounded pure exploration with **unknown** Γ

Input: $\mathcal{Z}, \mathcal{W}, \delta, L_\nu \geq \sigma_\nu^2, L_\eta \geq \sigma_\eta^2, \omega, \gamma_{\min} \leq \lambda_{\min}(\Gamma), \lambda_E, \kappa_0$

Initialize: $k = 1, \mathcal{W}_1 = \mathcal{W}, \widehat{\Gamma}_0 = \perp, \zeta_1 = 1$;

Define: $f(w, w', \Gamma, \lambda) := \|w - w'\|^2_{(\sum_{z\in\mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}$, $M := \frac{32 L_\eta}{\gamma_{\min}^2 \sigma_{\min}(A(\lambda_E, I))} \vee 1$, $\delta_\ell := \frac{\delta}{4\ell^2}$;

while $|\mathcal{W}_k| > 1$ **do**

$\widehat{\Gamma}_k = \Gamma - \texttt{estimator}\Big(\mathcal{W}_k, \widehat{\Gamma}_{k-1}, \zeta_k, \delta/k^2, \omega, \lambda_E, M, L_\eta\Big)$ `// Step 1: update` $\widehat{\Gamma}$

$\widehat{\theta}^k_{\texttt{P-2SLS}} = \theta - \texttt{estimator}\Big(\mathcal{W}_k, \delta/k^2, \zeta_k, \widehat{\Gamma}_k, \omega, L_\nu\Big)$ `// Step 2: update` $\widehat{\theta}$

$\mathcal{W}_{k+1} = \mathcal{W}_k \backslash \Big\{ w \in \mathcal{W}_k \mid \exists w' \in \mathcal{W}_k, \text{s.t.}, \Big\langle w' - w, \widehat{\theta}^k_{\texttt{P-2SLS}} \Big\rangle > \zeta_k \Big\}$ `// Step 3: elimination`

$k \leftarrow k + 1$, $\zeta_k = 2^{-k}$;

Output: The single remaining element in $\mathcal{W}_k$

after conditioning on the history, unlike the second sub-phase. This makes the analysis challenging, which we describe after the main result.

Algorithm 5: $\theta - \texttt{estimator}$

Input: $\mathcal{W}, \delta, \zeta, \widehat{\Gamma}, \omega, L_\nu$

Output: $\widehat{\theta}_{\texttt{P-2SLS}}$

$\widehat{\lambda} = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{w,w'\in\mathcal{W}} f(w, w', \widehat{\Gamma}, \lambda)$;

$\rho(\mathcal{W}) = \min_{\lambda\in\Delta(\mathcal{Z})} \max_{w,w'\in\mathcal{W}} f(w, w', \widehat{\Gamma}, \lambda)$;

$N_2 = \left\lceil 2(1+\omega)\zeta^{-2}\rho(\mathcal{W}) L_\nu \log\Big(\frac{4|\mathcal{W}|}{\delta}\Big) \right\rceil \vee r(\omega)$;

Get N_2 samples per design $\widehat{\lambda}$ denoted as $\{Z_2, X_2, Y_2\}$ `// via ROUND`

Update $\widehat{\theta}_{\texttt{P-2SLS}} = (\widehat{\Gamma}^\top Z_2^\top Z_2 \widehat{\Gamma})^{-1} \widehat{\Gamma}^\top Z_2^\top Y_2$;

Our algorithm additionally employs the so-called E-optimal design to ensure that the covariance matrix of the collected data used to estimate Γ is well-conditioned. This conditioning is required to ensure that $\widehat{\Gamma}_k$ concentrates fast enough to Γ as shown in the analysis. The E-optimal design is a well-known design objective in experimental design that aims to maximize the smallest singular value: $\lambda_E^* := \arg\min_{\lambda\in\Delta(\mathcal{Z})} \sigma_{\max}(V^{-1}(\lambda))$, where $V = \sum_{z\in\mathcal{Z}} \lambda_z z z^\top$. We denote $\kappa_0^{-1} := \sigma_{\max}(V^{-1}(\lambda_E^*)) = \sigma_{\min}^{-1}(V(\lambda_E^*))$ as the smallest singular value achieved by the E-optimal design.

We present our analysis result Theorem 17 where we show that, even without knowledge of Γ, the sample complexity scales with the key problem difficulty ρ^* almost matching the sample complexity of Algorithm 4 which relies on knowledge of Γ.

Algorithm 7: Γ – `estimator`

Input: $\mathcal{W}, \widehat{\Gamma}, \zeta, \delta, \omega, \lambda_E, M, L_\eta$

Define: $\texttt{Stop}(\mathcal{W}, Z, \Gamma, \delta) := \max_{w,w' \in \mathcal{W}} \|w - w'\|_{\bar{A}(Z,\Gamma)^{-1}} \|\theta\|_2 \sqrt{L_\eta \overline{\log}(Z, \delta)}$

Initialize: $\ell = 1$, $N_{0,0} = 0$ `// doubling trick initialization`

if $\widehat{\Gamma} = \perp$ **then**

  while $\ell = 1$ `or` $\texttt{Stop}\Big(\mathcal{W}, Z_{0,\ell}, \widehat{\Gamma}', \delta_\ell\Big) > 1$ **do**

    Get $2^{\ell-1}\Big(r(\omega) \vee \frac{2}{\kappa_0}\Big)$ samples denoted as $\big\{Z_{0,\ell}, X_{0,\ell}, Y_{0,\ell}\big\}$ per design λ_E `// via ROUND`

    Update $\widehat{\Gamma}'$ by `OLS` on $\big\{Z_{0,\ell}, X_{0,\ell}\big\}$, $\ell \leftarrow \ell + 1$

else

  $\tilde{\lambda} = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}} f(w, w', \widehat{\Gamma}, \lambda)$

  $N' = \left\lfloor 4gdM \ln\Big(1 + 2M\big(d + L_z^2\big) + 2M2gdM\Big) + 8M \ln\Big(\frac{2 \cdot 6^d}{\delta}\Big) \vee r(\omega) \right\rfloor$

  while $\ell = 1$ `or` $\texttt{Stop}\Big(\mathcal{W}, Z_{0,\ell} \cup Z_{1,\ell}, \widehat{\Gamma}, \delta_\ell\Big) > \zeta$ **do**

    $N_{1,\ell} = 2^\ell N'$ `// doubling trick update`

    Get $N_{1,\ell}$ samples per $\tilde{\lambda}$ denoted as $\big\{Z_{1,\ell}, X_{1,\ell}, Y_{1,\ell}\big\}$ `// via ROUND`

    $N_{0,\ell} = \left\lceil 2gdM \ln\Big(M\big(d + N_{1,\ell} + L_z^2\big)\Big) + 4M \ln\Big(\frac{2 \cdot 6^d}{\delta_\ell}\Big) \vee r(\omega) \vee \frac{2}{\kappa_0} \right\rceil$

    Get $(N_{0,\ell} - N_{0,\ell-1})$ samples per λ_E augmented to $\big\{Z_{0,\ell-1}, X_{0,\ell-1}\big\}$ and get $\big\{Z_{0,\ell}, X_{0,\ell}\big\}$

    Update $\widehat{\Gamma}'$ by `OLS` on $\big\{Z_{0,\ell} \cup Z_{1,\ell}, X_{0,\ell} \cup X_{1,\ell}\big\}$, $\ell \leftarrow \ell + 1$

Output: $\widehat{\Gamma}'$

Theorem 17. *Algorithm 6 is δ-PAC and terminates in at most*

$$(1+\omega)((L_\nu \log(1/\delta) + L_\eta \|\theta\|_2^2 (d + \log(1/\delta)))\rho^* + (d + \log(1/\delta))(L_\eta \|\theta\|_2^2 \rho_0 + M))$$

pulls, ignoring both of the additive and multiplicative logarithms of $\Delta, |\mathcal{W}|, \rho^*, \rho_0, M$, where

$$\rho_0 = \max_{w \in \mathcal{W} \setminus \{w^*\}} \|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \lambda_{E,z} \Gamma^\top z z^\top \Gamma)^{-1}}, \text{ and } M = \frac{32 L_\eta}{\gamma_{\min}^2 \sigma_{\min}\big(A(\lambda_E, I)\big)} \vee 1.$$

Note that ρ_0 does not get hurt by $\langle w^* - w, \theta \rangle$, ($\rho^*$ does). It comes from the fact that in the first phase, we initialize that algorithm with E-optimal design.

The challenge of the analysis can be summarized in two-fold. First, since the concentration result in Theorem 15 is w.r.t. $\widehat{\Gamma}_k$, we need to analyze how the random matrix $\widehat{\Gamma}_k$ concentrates around Γ and how this impacts the sample complexity. For this, we develop

a novel concentration inequality that relates the confidence width involving $\hat{\Gamma}$ from Theorem 15 with the same quantity involving Γ in place of $\hat{\Gamma}$. Second, our algorithm creates a long-range error propagation, which is highly nontrivial to analyze. To see this, the quality of $\hat{\Gamma}_k$ is affected by the design objective function $\max_{w,w'} f(w, w', \hat{\Gamma}_{k-1}, \lambda)$, which depends on the error of the estimate $\hat{\Gamma}_{k-1}$ from the previous phase. This error is, in turn, affected by the error of $\hat{\Gamma}_{k-2}$ by the same mechanism. This is repeated all the way back to the first phase. Thus, any abnormal behavior from the first iteration will have a cumulative impact to even the end. In our analysis, we successfully analyze how the error is propagated from the previous iterations, which forms a complicated recursion. Resolving this recursion is our key novelty in the analysis.

Remark 6. *Our algorithm requires knowledge of a lower bound $\gamma_{\min}$ of $\lambda_{\min}(\Gamma)$. The knowledge of $\gamma_{\min}$ is for simplicity only as one can obtain such a lower bound that is at least half of the true value $\lambda_{\min}(\Gamma)$ via an efficient sampling procedure that we describe in Section 3.7.12.*

3.4 Experiments

We now present experiments a collection of experiments on `CPET-LB` problem instances. The experiments demonstrate that our approach produces efficient designs for inference and estimation.

3.4.1 Comparison Algorithms

The baselines that our approaches are compared with are discussed below. We run experiments both when Γ is known and when Γ is fully unknown.

Known Γ.

To standardize the experiments, the baselines considered run in rounds mirroring the structure of Algorithm 4. Specifically, in round $k \in \mathbb{N}$ a sampling algorithm selects a design $\lambda_k \in \Delta(\mathcal{Z})$, collects N_k samples from the design, and forms a $\Psi - \texttt{IV}$ estimate of θ with $\Psi = \Gamma$ that is combined with a confidence interval (Lemma 14.2 to either eliminate evaluation vectors or validate a stopping condition. The number of samples N_k taken in round $k \in \mathbb{N}$

by any of the algorithms is given by $N_k = \lceil 2(1+\omega)\zeta_k^{-2}\rho(\mathcal{W}_k)L_\nu \log(4k^2|\mathcal{W}|/\delta)\rceil \vee r(\omega)$ where $\rho(\mathcal{W}_k) = \max_{w,w'\in\mathcal{W}_k} f(w,w',\Gamma,\lambda)$ for design $\lambda_k \in \Delta(\mathcal{Z})$ and an active set of evaluation vectors $\mathcal{W}_k$. The round sample count guarantees that given any experimental design, all vectors $w \in \mathcal{W}$ such that $(w^* - w)^\intercal\theta > 2\cdot 2^{-k}$ can be determined to be suboptimal by the end of round k. The sampling methods we consider are now described.

- *Static Oracle.* This design selects $\lambda_k = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{w\in\mathcal{W}\setminus\{w^*\}} f(w^*,w',\Gamma,\lambda)$.
- *Static XY-Optimal.* This design selects $\lambda_k = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{w,w'\in\mathcal{W}} f(w,w',\Gamma,\lambda)$.
- *Static Uniform.* This design selects $\lambda_{k,z} = 1/|\mathcal{Z}| \;\forall\; z \in \mathcal{Z}$.
- *Adaptive Uniform (SE).* This design selects $\lambda_{k,w} = 1/|\mathcal{W}_k| \;\forall\; w \in \mathcal{W}_k$. Note that this algorithm is effectively an adaption of action-elimination (Even-Dar et al., 2006) .

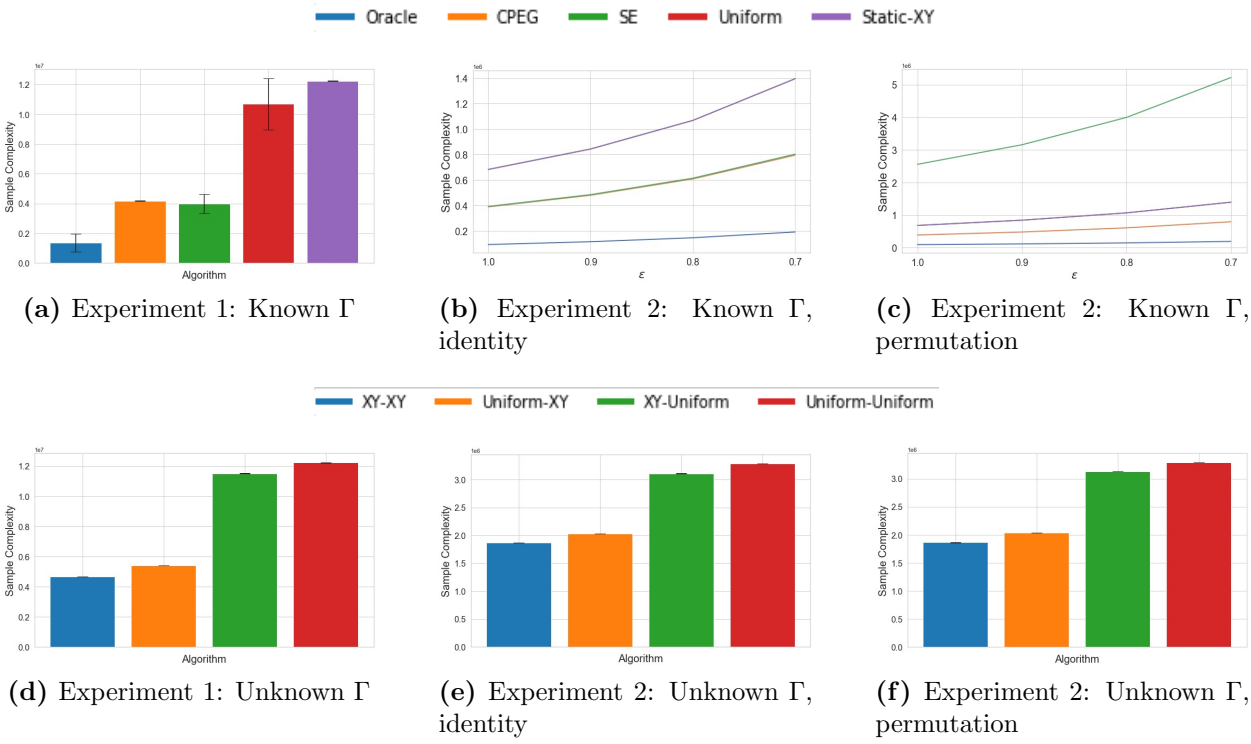
The static designs are independent of the round and simply terminate when all evaluation vectors can be eliminated except for a recommended optimal vector $\widehat{w}$.

Unknown Γ.

For this set of experiments, we compare Algorithm 6 against a collection of variations of the sampling procedures. Specifically, we compare against methods that either replace only the experimental design for estimating Γ, or only the experimental design for estimating θ, or both with uniform sampling. We label the approaches as $N - N$, where N represents the sampling approaches (XY or uniform) for Γ and θ respectively. Moreover, to make our approach more practical, we modify the algorithm so that $\overline{\log}(Z_T,\delta) = 4d + \log(1/\delta)$. The step of incrementally adding more E-optimal design samples is also removed, so we collect E-optimal design samples only once in the beginning of Algorithm 7. We find that even with these modifications to the algorithm, correctness is maintained empirically.

3.4.2 Experiment 1: Jump-Around Instance

We first return back to the location model of Section 3.7.1. Recall that $\mathcal{Z} = \mathcal{W} = \mathcal{X} = \{e_1,\cdots,e_d\}$. For this experiment, we take $d = 6$, let $\theta = \begin{bmatrix}1 & -0.95 & 0.45 & 0.45 & 0.95 & 0.45\end{bmatrix}$ and $\sigma_u^2 = 0.275$. The results of the experiment are shown in Figure 3.3a for the case of known



(a) Experiment 1: Known Γ

(b) Experiment 2: Known Γ, identity

(c) Experiment 2: Known Γ, permutation

(d) Experiment 1: Unknown Γ

(e) Experiment 2: Unknown Γ, identity

(f) Experiment 2: Unknown Γ, permutation

Figure 3.3: Sample complexity for algorithms on `CPET-LB` problems. Our approach is consistently competitive across the experiments.)

Γ. We see that Algorithm 4 performs much better than the baselines and nearly matches the oracle design. Delving into the approach, it is able to quickly eliminate all but w_1 and w_{d-1} and then puts more mass on z_1 and z_{d-1} to reduce the uncertainty on w_1 and w_{d-1}. For the case of unknown Γ, the results are shown in Figure 3.3d, where θ_5 is reduced to 0.9 so that all approaches could finish.

3.4.3 Experiment 2: Interpolation Instance

Let $\mathcal{Z} = \mathcal{W} = \mathcal{X} = \{e_1, \cdots, e_d\}$ define the measurement, evaluation, and observation sets. We first consider that $\Gamma := \frac{(1-\varepsilon)}{d} 1_d 1_d^\intercal + \varepsilon I_d$ for a parameter $\varepsilon \in (0,1)$ where 1_d is a d-dimensional vector of 1's and I_d is the d-dimensional identity matrix. For this experiment, we take $d = 4$ and let $\theta = \begin{bmatrix} 0.5 & 0.583 & 0.67 & 0.75 \end{bmatrix}$. As in all compliance instances, $\eta_t = x_t - \Gamma^\intercal z_t$, and in this simulation $\eta_t = 0.4\eta_t^\intercal v_t$, where $v_t = \bar{v}_t / \|\bar{v}_t\|_2$ and $\bar{v}_t \sim \mathcal{N}(0, I_d)$. The results of the experiment are shown in Figure 3.3b for $\varepsilon \in \{1, 0.9, 0.8, 0.7\}$ with Γ known. Note that Static-XY and Uniform overlap, and SE and CPEG overlap. We see that Algorithm 4 and the adaptive

uniform strategy perform similarly and near optimally. This is to be expected since the most efficient way to gather observations for treatments is to encourage that treatment, given that if the encouragement is not followed each of the alternatives is equally likely and provides no additional information of interest. Moreover, as discussed earlier, the problem gets more challenging as $\Gamma \to 1_d 1_d^\intercal / d$. Note that the identity matrix could be replaced with a permutation matrix, in which case uniform sampling with elimination becomes highly suboptimal. The results for the case of Γ unknown are shown in Figure 3.3e with $\varepsilon = 0.99$. This shows the value that comes from the experimental design for estimating both Γ and θ.

To demonstrate the superiority of our algorithm over SE, we also consider that $\Gamma := \frac{(1-\varepsilon)}{d} 1_d 1_d^\intercal + \varepsilon I_d^p$, where I_d^p is a permutation matrix as follows,

$$I_d^p = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

All other settings remain the same as in the previous interpolation instance. The results for the known Γ case, shown in Figure 3.3c, indicate that SE exhibits significant underperformance due to its sampling rule not accounting for the permutation effect in Γ. In contrast, CPEG consistently achieves near-optimal performance. Note that Static-XY and Uniform still overlap. Figure 3.3f presents the results for the unknown Γ case, where we can notice that, comparing with Figure 3.3e, estimating different permutation matrices (with the identity matrix as a special case) does not affect problem difficulty.

3.5 Conclusion

This work introduces the `CPET-LB` problem in which the learning protocol is characterized by a linear structural equation model governed by parameters Γ and θ. We provide a general solution that simultaneously estimates the structural model while optimally designing to learn the best-arm. The key ideas behind our approach are based on linear experimental design techniques, an instrumental variable estimator whose variance can be controlled by

the design, and novel finite-time confidence intervals on this estimator. This paper presents a number of directions for future work including considering situations where the $d_z \neq d_x$, analysis to improve the dependence on the underlying noise variance, and the pursuit of a tight information-theoretic instance-dependent lower-bound. We hope that this line of work motivates increased discussion of the real impact of confounding on applicability of adaptive experimentation.

3.6 Related Work

Our work is at the intersection of several parallel tracks of literature, pure exploration linear bandits, causal bandits, and econometrics. The most relevant work on pure exploration in linear bandits is the `RAGE` algorithm of Fiez et al. (2019); Soare et al. (2014). `RAGE` is nearly instance optimal for linear bandits in the non-confounded setting. Extensions of `RAGE` to various noise models including logistic and heteroskedastic noise have been considered (Weltz et al., 2023; Jun et al., 2021). Other algorithms for pure exploration linear bandits have been proposed - and we leave it for future work to extend the ideas of this paper to those settings (Li et al., 2023; Degenne et al., 2020).

Confounding in bandits was first considered in the regret minimization setting by Bareinboim et al. (2015). They introduces the Multi-armed bandit with unobserved confounders (MABUC) problem. They empirically demonstrate traditional bandit algorithms can have linear regret in this setting and provide an algorithm that effectively employs observed intuition. The early work of Kallus (2018) also assumes there is an additional unobserved latent class at each time that determines confounding in a compliance setting. They provide novel notions of regret, relative to the instrument with the highest reward ($\arg\max Z^\top \Gamma^\top \theta$ in our notation), the highest treatment ($\arg\max_w w^\top \theta$), regret relative to the best latent class at each time, and regret on the set of "compliers". They discuss the suitability of these various notions of regret, and discuss when sublinear regret is possible. We remark that their approach is similar to ours in the sense that they assume a form of homogeneous effects across the population, and use an estimate of Γ. Recently Kveton et al. (2023) also consider the problem of compliance, however they don't take explicit non-confounding into account and

assume an explicit parametric model that determines the non-compliance. This is analogous to the Heckman selection model considered in econometrics (Greene, 2003).

The recent works of Della Vecchia and Basu (2023); Zhang et al. (2022); Gong and Zhang (2022) considered an online setting where at each time they observe a set $\{(x_t, z_t)\}$ where x_t is the action of interest and z_t is an associated instrument. If action I_t is selected, the reward observed is $y_t = x_{I_t}^\top \theta_t + \varepsilon_t$, where x_t may be endogeneous. Similar to the standard linear bandit setting (Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2020), the goal is to minimize regret relative to the best action at each time. We remark that this setting is very different from ours. Effectively, we are choosing which instrument to select at each time to learn the best-performing treatment - in particular we can't choose a particular intervention. In their setting, they are choosing an intervention at each time and using the instrument purely for de-confounding the result. Experimental design for instruments to have more effective estimation has been considered by Chandak et al. (2023).

In the causal bandit problem, an underlying causal graph between a set of interventions and a reward value is assumed. Actions correspond to intervening (i.e. a "do" operation (Pearl, 2009)) at one or more specific nodes in the causal graph and then observing the corresponding value at the reward node. Causal bandits have been studied extensively in the regret setting (Lattimore et al., 2016b; Lu et al., 2020; Bilodeau et al., 2022) and the pure exploration setting (Sen et al., 2017). Though past works have allowed for unobserved confounders in the graph e.g. Malek et al. (2023), their goal is to learn the best performing intervention, which in our setting would be $\arg\max_{z \in \mathcal{Z}} z^\top \Gamma \theta$ instead of w^*.

Encouragement designs have been considered in many applications in online and offline settings. One of the earliest works on encouragement designs is Bradlow (1998), which considers the problem of using encouragements to determine the impact of coupons at a grocery store. More recent applications include Bakshy et al. (2014); Mummalaneni et al. (2022); Engineering (2023) all in the context of online services and treatments that are required to be served to all users. Most of these works consider a small number of treatments and a heterogeneous treatment effect - hence are interested in LATE estimator. As far as we are aware, we are the only work that considers adaptive encouragement design in the context of the model given in Equation 3.1 and for multiple treatments.

3.7 Proofs

3.7.1 Illustrative Example

We now present an illustrative experiment that highlights the challenges of endogenous noise and the insufficiency of standard experimentation approaches used in the absence of confounding.

Instance Definition. Toward connecting back to membership example in Section 3.1, consider that a service has d membership options given by the set $\mathcal{A} = \{1, \dots, d\}$. Let the set $\mathcal{Z} = \{e_1, \cdots, e_d\}$ represent encouragements (incentives or advertisements) for the corresponding membership options given by $\mathcal{W} = \mathcal{X} = \{e_1, \cdots, e_d\}$. We consider a location model that assumes each user $t \in \mathbb{N}$ arriving online has an underlying unobserved one-dimensional preference $u_t \sim \mathcal{N}(0, \sigma_u^2)$. If an algorithm presents the user with encouragement $z_{I_t} = e_{I_t}$ for $I_t \in \mathcal{A}$, then the user selects into the membership level given by $J_t = \min_{j \in \mathcal{A}} |I_t + u_t - j|$ so that $x_t = e_{J_t}$. For a visual depiction, see Figure 3.4a. This process captures a user being more likely to opt-in to membership levels that are closer to the encouragement that they were presented. The outcome is then given by $y_t = x_t^\top \theta + u_t$. This problem instance is a specific compliance instance. For this experiment, we take $d = 6$, let $\theta = \begin{bmatrix} 1 & -0.95 & 0 & 0.45 & 0.95 & 0.99 \end{bmatrix}$ and $\sigma_u^2 = 0.35$. Observe that the optimal evaluation vector is $w^* = e_1 = \arg\max_{w \in \mathcal{W}} w^\top \theta$.

We simulate a UCB strategy which maintains estimates of the average reward of each of the possible d incentives, namely $\widehat{\mu}_{i,t} = \sum_{s=1}^{t} \mathbf{1}\{z_t = e_i\} y_t$ and then pulls the one with the highest upper confidence bound. This models current practice of using a bandit algorithm to select which incentive to show a user. Our results averaged over 100 simulations are in Figure 3.4d. At each round we estimate the average reward of each level using an OLS estimator, i.e. $\widehat{\theta}_{i,t}^{OLS} = \sum_{s=1}^{t} \mathbf{1}\{x_t = e_i\} y_t / \sum_{s=1}^{t} \mathbf{1}\{x_t = e_i\}$, and check whether it matches the true value (denoted as UCB-OLS). We also consider an instrumental variable-estimator (see the next Section) which incorporates knowledge of Γ similar to 2SLS to deconfound our estimate (UCB-IV). As the plot demonstrates, UCB-OLS completely fails to identify $\theta_1 = \arg\max_{i \in d} \theta_i$ (this line is hard to see it is at 0) due to a biased estimate, whereas UCB-IV does better. However, UCB-IV methods seem to have a constant probability of error. To see why, note that the expected reward from pulling $z = e_i$ is $e_i^\top \Gamma \theta$. These values are

plotted in orange in Figure 3.4c. In particular, with some constant probability, UCB runs on the empirical rewards from pulling z's zeroes in on arm 6, and as a result fails to give enough samples to learn that arm 1 is indeed the best. In contrast, our proposed method CPEG, Algorithm 4 manages to find the best arm with significantly higher probability (the algorithm was run with $\delta = .1$) in the given time horizon.

3.7.2 Standard 2SLS estimator

Consider $\Psi = \widehat{\Gamma}_{2SLS} = (Z_T^\top Z_T)^{-1} Z_T^\top X_T$. In this setting, we recover the standard two-stage-least-squares (2SLS) estimator,

$$\widehat{\theta}_{\texttt{2SLS}} = (X_T^\top Z_T (Z_T^\top Z_T)^{-1} Z_T^\top X_T)^{-1} X_T^\top (Z_T^\top Z_T)^{-1} Z_T^\top Y_T = (Z_T^\top X_T)^{-1} Z_T^\top Y_T.$$

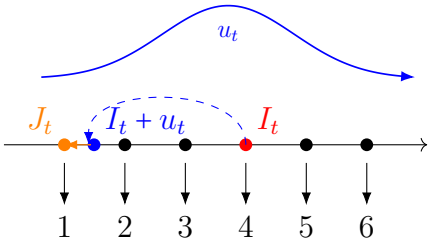
Note that the 2SLS estimator is a biased, but consistent estimator of the parameter θ (Angrist et al., 1996; Greene, 2003).

Note that in particular, the asymptotic variance of 2SLS is known to be $\sigma_\varepsilon^2 \|w\|_{(\widehat{\Gamma}_{\texttt{2SLS}}^\top Z_T^\top Z_T \widehat{\Gamma}_{\texttt{2SLS}})^{-1}}$ (Greene, 2003). Recent work by Della Vecchia and Basu (2023) provides a confidence interval of the form $|w^\top(\widehat{\theta}_{\texttt{2SLS}} - \theta)| \leq O(d\sigma_\varepsilon^2 \|w\|_{(\widehat{\Gamma}_{\texttt{2SLS}}^\top Z_T^\top Z_T \widehat{\Gamma}_{\texttt{2SLS}})^{-1}} \sqrt{\log(T/\delta)})$. However, it's unclear how to use the form of their confidence interval directly for experimental design due to the dependence of $\hat{\Gamma}_{2SLS}$ on the random quantity X. In addition, their work is not sufficiently general to handle the general forms of noise that we consider in Lemma 14.1.

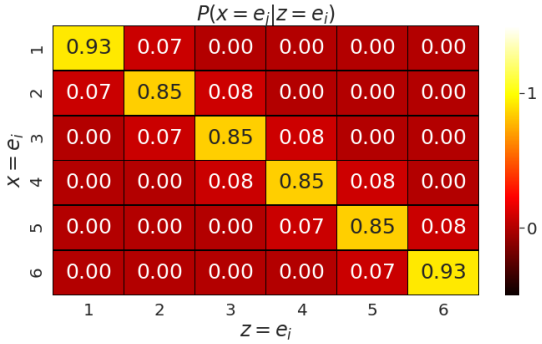
3.7.3 A non-interactive lower bound

Due to the noise model from confounding and the dependence of the noise $\theta^\top \eta + \varepsilon$, the instance-dependent lower bounds of Fiez et al. (2019) do not immediately apply. In this section, we develop a lower bound tailored for the confounding setting.

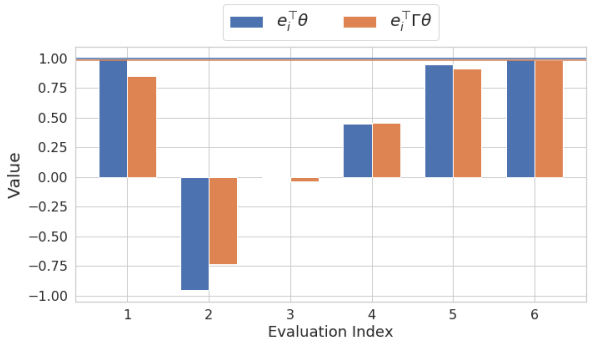
Toward characterizing the optimal sample complexity, we develop a lower bound for a specific non-adaptive algorithm $\mathcal{A}$ that has access to the matrix Γ governing the structural equation model. In particular, suppose that the non-adaptive algorithm $\mathcal{A}$ is allowed to select a sequence of T measurements $\{z_{I_1}, \ldots z_{I_t} \ldots, z_{I_T}\}$ to query prior to collecting any observations, where I_t represents the index of the vector $z \in \mathcal{Z}$ chosen at time $t \in \{1, \ldots, T\}$.



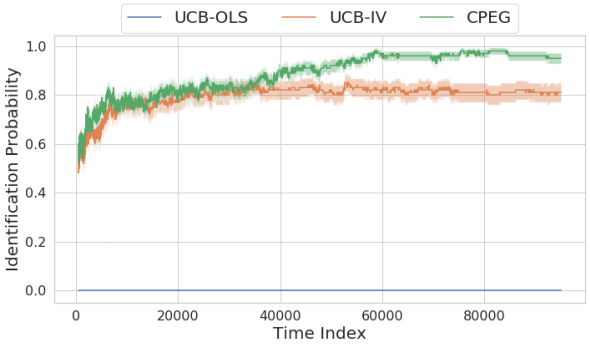
(a) Depiction of the problem instance.



(b) Heat-map of Γ.



(c) $w^\top \theta$ and $w^\top \Gamma \theta$ for $w \in \{e_1, \ldots, e_d\}$



(d) $\mathbb{P}(\widehat{w}_t = w^*)$ for the instance

Figure 3.4: (a) A visual depiction of the problem instance from Section 3.7.1. The user is presented with encouragement $I_t \in \mathcal{A}$ and the user choice is given by J_t where $J_t = \min_{j \in \mathcal{A}} |I_t + u_t - j|$ and $u_t \sim \mathcal{N}(0, \sigma_u^2)$. (b) A heat-map showing the structural parameter Γ for the problem instance from Section 3.7.1. (c) A bar chart showing $\mathbb{E}[y|x = w] = w^\top \theta$ and $\mathbb{E}[y|z = w] = z^\top \Gamma \theta$ for all $w \in \mathcal{W}$. This chart shows that the optimal evaluation vector is $w^* = e_1 = \arg\max_{w \in \mathcal{W}} \mathbb{E}[y|x = w]$, while $e_6 = \arg\max_{w \in \mathcal{W}} \mathbb{E}[y|z = w]$ and consequently estimation based on this quantity is problematic. (d) The probability of identifying $w^* = e_1$ for a collection of algorithms on the `CPET-LB` instance described in Section 3.7.1. Standard optimistic sampling approaches in combination with an ordinary least squares estimator leads to faulty inferences. Given an instrumental variable estimator, these experimental designs eventually give high probability identification but do so inefficiently compared to our proposed approach (see Section 3.3).

Then, given the observations $\{y_1, \ldots, \ldots y_t, \ldots, y_T\}$ generated by the environment, a candidate optimal vector $\widehat{w} \in \mathcal{W}$ is returned by the algorithm. We are interested in the necessary number of observations T that must be collected in order to ensure $\mathbb{P}(\widehat{w} \neq w^*) \leq \delta$ for some $\delta \in (0, 1)$. Thus, it is natural that the optimal non-adaptive algorithm $\mathcal{A}$ using the estimator $\widehat{\theta}_{\texttt{oracle}}$ forms a recommendation rule such that $\widehat{w} = \arg\max_{w \in \mathcal{W}} w^\top \widehat{\theta}_{\texttt{oracle}}$. We now state our lower bound result with respect to the non-adaptive oracle algorithm.

Theorem 18 (Non-Adaptive Oracle Lower Bound)**.** *Consider a problem instance characterized by $\mathcal{W} \subset \mathbb{R}^d, \mathcal{Z} \subset \mathbb{R}^d$, $\Gamma \in \mathbb{R}^{d \times d}$, and $\theta \in \mathbb{R}^d$. Assume Γ is known, θ is unknown, and the noise process is jointly Gaussian and defined by $\gamma := \begin{bmatrix} \eta & \varepsilon \end{bmatrix} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{(d+1)\times(d+1)}$ is an arbitrary correlation matrix. For $\delta \in (0, 0.05]$, if the non-adaptive oracle algorithm acquires $T \leq \sigma^2 \rho^* \log(1/\delta)/2$ samples on the problem instance where $\sigma^2 := v^\top \Sigma v$ and $v := \begin{bmatrix} \theta & 1 \end{bmatrix} \in \mathbb{R}^{d+1}$, then $\mathbb{P}(\widehat{w} \neq w^*) \geq \delta$.*

Corollary 18.1. *There exists a problem instance characterized by $\mathcal{W} \subset \mathbb{R}^d, \mathcal{Z} \subset \mathbb{R}^d$, $\Gamma \in \mathbb{R}^{d \times d}$, and $\theta \in \mathbb{R}^d$ with a noise process satisfying Assumption 1 such that if the non-adaptive oracle algorithm acquires $T \leq \max\{d\|\theta\|_2^2, \sqrt{d}\|\theta\|_2\}\rho^* \log(1/\delta)/2$ samples, then $\mathbb{P}(\widehat{w} \neq w^*) \geq \delta$ for $\delta \in (0, 0.05]$.*

The proof of Theorem 18 is in Section 3.7.4. Notably, the result is reminiscent of lower bounds for the standard pure exploration transductive linear bandit problem without confounding (Fiez et al., 2019; Katz-Samuels et al., 2020) when given the measurement set $\{\Gamma^\top z\}_{z \in \mathcal{W}}$, evaluation set $\mathcal{Z}$, and parameter θ.

Notably, the upper bounds for our algorithms nearly match the lower bound of Theorem 18. However, it is interesting to observe that the sample complexity incurs an additional factor of d^2 relative to the standard transductive linear bandit problem in the most general setting where entries of η are sub-Gaussian, but not necessarily independent nor bounded. This illustrates the additional difficulty that arises from confounding. We point out that this is not likely to be a tight lower bound. In particular, it is a lower bound with respect to a non-adaptive algorithm that uses the particular choice of estimator. We leave improved lower bounds to future work.

3.7.4 Proof of Theorem 18

Theorem 18. *Consider a problem instance characterized by* $\mathcal{W} \subset \mathbb{R}^d, \mathcal{Z} \subset \mathbb{R}^d$, $\Gamma \in \mathbb{R}^{d\times d}$, *and* $\theta \in \mathbb{R}^d$. *Assume* Γ *is known,* θ *is unknown, and the noise process is jointly Gaussian and defined by* $\gamma := \begin{bmatrix}\eta & \varepsilon\end{bmatrix} \sim \mathcal{N}(0, \Sigma)$ *where* $\Sigma \in \mathbb{R}^{(d+1)\times(d+1)}$ *is an arbitrary correlation matrix. For* $\delta \in (0, 0.05]$, *if the non-adaptive oracle algorithm acquires* $T \le \sigma^2\rho^* \log(1/\delta)/2$ *samples on the problem instance where* $\sigma^2 := v^\top \Sigma v$ *and* $v := \begin{bmatrix}\theta & 1\end{bmatrix} \in \mathbb{R}^{d+1}$, *then* $\mathbb{P}(\widehat{w} \neq w^*) \ge \delta$.

Proof. We begin by recalling the framework of the non-adaptive oracle algorithm and discussing the properties of its estimator for the noise structure described in the statement of the result.

Non-Adaptive Oracle and Instance Definition. The non-adaptive oracle algorithm $\mathcal{A}$ selects T measurements to query prior to collecting any data. Let I_t represent the index of the vector $z \in \mathcal{Z}$ chosen at time $t \in \{1, \dots, T\}$. The noise process for the instance under consideration is assumed to be jointly Gaussian and defined by $\gamma_t := \begin{bmatrix}\eta_t & \varepsilon_t\end{bmatrix} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{(d+1)\times(d+1)}$ is an arbitrary positive semidefinite matrix. Defining $\bar{x}_{I_t} := \Gamma^\top z_{I_t}$, $v := \begin{bmatrix}\theta & 1\end{bmatrix} \in \mathbb{R}^{d+1}$ and $\nu_t := v^\top \gamma_t$, the feedback model can be described as follows:

$$\begin{aligned} y_t &= x_t\theta + \varepsilon_t \\ &= (\Gamma^\top z_{I_t})^\top \theta + \eta_t^\top \theta + \varepsilon_t \\ &=: (\Gamma^\top z_{I_t})^\top \theta + v^\top \gamma_t \\ &=: \bar{x}_{I_t}^\top \theta + \nu_t. \end{aligned}$$

Observe that the noise is independent and identically distributed as $\nu_t \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 := v^\top \Sigma v$ since $\gamma_t \sim \mathcal{N}(0, \Sigma)$. Moreover, the noise process is exogeneous with $\mathbb{E}[\nu_t | \bar{x}_{I_t}] = 0$ since $\bar{x}_{I_t}$ is deterministic given the index choice I_t.

Let $\{z_{I_t}\}_{t=1}^T, \{\bar{x}_{I_t}\}_{t=1}^T$, and $\{y_t\}_{t=1}^T$ denote the observations collected by the non-adaptive oracle algorithm $\mathcal{A}$ and define $Z_T \in \mathbb{R}^{T\times d}$, $\bar{X}_T \in \mathbb{R}^{T\times d}$, and $Y_T \in \mathbb{R}^T$ to contain the respective stacked observations. Algorithm $\mathcal{A}$ obtains an estimate $\widehat{\theta}_{\texttt{oracle}}$ by minimizing the sum of

squares as follows:

$$\widehat{\theta}_{\texttt{oracle}} := \arg\min_{\widehat{\theta}\in\mathbb{R}^d} \sum_{t=1}^{T}(y_t - \bar{x}_{I_t}\widehat{\theta})^2 = (\bar{X}_T^\top \bar{X}_T)^{-1}\bar{X}_T^\top Y_T = (Z_T^\top Z_T \Gamma)^{-1} Z_T^\top Y_T.$$

Given $\widehat{\theta}_{\texttt{oracle}}$, the non-adaptive oracle algorithms $\mathcal{A}$ returns a recommendation defined by $\widehat{w} = \arg\max_{w\in\mathcal{W}} w^\top \widehat{\theta}_{\texttt{oracle}}$. Note that since $\widehat{\theta}_{\texttt{oracle}}$ is obtained by least squares with exogeneous, independent and identically distributed mean-zero Gaussian noise, it is straightforward to verify the estimator is distributed as

$$\widehat{\theta}_{\texttt{Oracle}} - \theta \sim \mathcal{N}\Big(0, \sigma^2 \cdot \bar{A}(Z_T, \Gamma)^{-1}\Big), \tag{3.8}$$

where

$$\bar{A}(Z_T, \Gamma) := \left(\sum_{t=1}^{T} \Gamma^\top z_{I_t} z_{I_t} \Gamma\right) = \bar{X}_T^\top \bar{X}_T = \Gamma^\top Z_T^\top Z_T \Gamma.$$

Proof by Contradiction. To begin, recall that

$$\rho^* := \min_{\lambda\in\Delta(\mathcal{Z})} \max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^* - w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^* - w, \theta\rangle^2}.$$

Suppose for the sake of contradiction that the number of samples collected by the non-adaptive oracle algorithm $\mathcal{A}$ is $T \le \sigma^2 \rho^* \log(1/\delta)/2$ and $\mathbb{P}(\widehat{w} \neq w^*) < \delta$ for $\delta \in (0, 0.05]$. To reach a contradiction, we analyze the distribution of $(w - w^*)^\top \widehat{\theta}$ for some $w \neq w^*$ and show that with probability at least δ it is positive. We remark that this proof follows similar techniques to that of the proof of Theorem 3 of Katz-Samuels et al. (2020).

Let $\overline{\lambda} \in \Delta(\mathcal{Z})$ represent the empirical sampling distribution of the algorithm $\mathcal{A}$, which is defined such that $\overline{\lambda}_z = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{z_{I_t} = z\}$ for each $z \in \mathcal{Z}$. Moreover, define

$$\rho^*(\overline{\lambda}) := \max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^* - w\|^2_{A(\overline{\lambda},\Gamma)^{-1}}}{\langle w^* - w, \theta\rangle^2} \quad \text{and} \quad \widetilde{w} \in \arg\max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^* - w\|^2_{A(\overline{\lambda},\Gamma)^{-1}}}{\langle w^* - w, \theta\rangle^2}.$$

Note that $\rho^*(\overline{\lambda}) \ge \rho^*$ and observe by definition,

$$\frac{A(\overline{\lambda}, \Gamma)^{-1}}{T} := \frac{(\sum_{z\in\mathcal{Z}} \overline{\lambda}_z \Gamma^\top z z^\top \Gamma)^{-1}}{T} = \left(\sum_{t=1}^{T} \Gamma^\top z_{I_t} z_{I_t} \Gamma\right)^{-1} := \bar{A}(Z_T, \Gamma)^{-1}.$$

Thus, by Eq. (3.8),

$$\widehat{\theta}_{\texttt{Oracle}} - \theta \sim \mathcal{N}\left(0, \sigma^2 \cdot \frac{A(\overline{\lambda}, \Gamma)^{-1}}{T}\right),$$

and

$$\frac{(\widetilde{w} - w^*)^\top (\widehat{\theta}_{\texttt{Oracle}} - \theta)}{\langle w^* - \widetilde{w}, \theta \rangle} \sim \mathcal{N}\left(0, \sigma^2 \cdot \frac{\|w^* - \widetilde{w}\|^2_{A(\overline{\lambda}, \Gamma)^{-1}}}{T \cdot \langle w^* - \widetilde{w}, \theta \rangle^2}\right).$$

Furthermore, by the definition of $\rho^*(\overline{\lambda})$, the assumption $T \leq \sigma^2 \rho^* \log(1/\delta)/2$, and the fact $\rho^*(\overline{\lambda}) \geq \rho^*$, we obtain

$$\mathbb{V}\left(\frac{(\widetilde{w} - w^*)^\top (\widehat{\theta}_{\texttt{Oracle}} - \theta)}{\langle w^* - \widetilde{w}, \theta \rangle}\right) = \sigma^2 \cdot \frac{\|w^* - \widetilde{w}\|^2_{A(\overline{\lambda}, \Gamma)^{-1}}}{T \cdot \langle w^* - \widetilde{w}, \theta \rangle^2} := \sigma^2 \cdot \frac{\rho^*(\overline{\lambda})}{T} \geq \frac{2}{\log(1/\delta)}. \tag{3.9}$$

Now, consider a random variable $W \sim \mathcal{N}(0, 1)$. Proposition 2.1.2 of Vershynin (2018) gives an anti-concentration result showing that for all $\zeta > 0$,

$$\mathbb{P}(W \geq \zeta) \geq \left(\frac{1}{\zeta} - \frac{1}{\zeta^3}\right) \frac{1}{\sqrt{2\pi}} e^{-\zeta^2/2}. \tag{3.10}$$

We apply this result to the quantity $(\widetilde{w} - w^*)^\top (\widehat{\theta}_{\texttt{Oracle}} - \theta)/\langle w^* - \widetilde{w}, \theta \rangle$ to conclude that $(\widetilde{w} - w^*)^\top \widehat{\theta}_{\texttt{Oracle}} > 0$ with probability at least δ. Toward doing so, let $c \in (1, 1.15]$ be a constant and define

$$\widetilde{W} \sim \mathcal{N}\left(0, \frac{2}{\log(1/\delta)}\right) \quad \text{and} \quad W := \frac{\widetilde{W}}{\sqrt{2/\log(1/\delta)}} \quad \text{and} \quad \gamma := \frac{c}{\sqrt{2/\log(1/\delta)}}.$$

Observe that $W \sim \mathcal{N}(0,1)$. The following analysis holds for $\delta \in (0, 0.05]$ given that $c \in (1, 1.15]$ as assumed:

$$\begin{aligned}
\mathbb{P}\left(\frac{(\widetilde{w}-w^*)^\top(\widehat{\theta}_{\texttt{Oracle}}-\theta)}{\langle w^*-\widetilde{w},\theta\rangle} \geq c\right) &\geq \mathbb{P}(\widetilde{W} \geq c) && \text{(By Eq. 3.9)} \\
&= \mathbb{P}\left(\frac{\widetilde{W}}{\sqrt{2/\log(1/\delta)}} \geq \frac{c}{\sqrt{2/\log(1/\delta)}}\right) \\
&:= \mathbb{P}(W \geq \gamma) \\
&\geq \left(\frac{1}{\gamma}-\frac{1}{\gamma^3}\right)\frac{1}{\sqrt{2\pi}}e^{-\gamma^2/2} && \text{(Proposition 2.1.2 Vershynin 2018)} \\
&= \left(\frac{\sqrt{2}}{c\sqrt{\log(1/\delta)}}-\frac{\sqrt{2}^3}{c^3\sqrt{\log(1/\delta)}^3}\right)\frac{1}{\sqrt{2\pi}}\delta^{c^2/4} \\
&\geq \delta.
\end{aligned}$$

The final inequality can be verified computationally. Thus, with probability at least δ for $\delta \in (0, 0.05]$, we obtain

$$\begin{aligned}
(\widetilde{w}-w^*)^\top\widehat{\theta}_{\texttt{Oracle}} &\geq c(w^*-\widetilde{w})^\top\theta + (\widetilde{w}-w^*)^\top\theta \\
&= c(w^*-\widetilde{w})^\top\theta - (w^*-\widetilde{w})^\top\theta \\
&= (c-1)(w^*-\widetilde{w})^\top\theta \\
&> 0.
\end{aligned}$$

Observe that the final inequality holds since $c > 1$ and $(w^* - \widetilde{w})^\top\theta > 0$ by definition.

This result directly implies that with probability at least δ for $\delta \in (0, 0.05]$, the vector $\widehat{w}$ returned by algorithm $\mathcal{A}$ is not w^*. This is a contradiction, so we conclude that if the non-adaptive oracle algorithm $\mathcal{A}$ acquires $T \leq \sigma^2\rho^*\log(1/\delta)/2$ samples, then $\mathbb{P}(\widehat{w} \neq w^*) \geq \delta$ for $\delta \in (0, 0.05]$. □

3.7.5 Proof of Corollary 18.1

Corollary 18.1. *There exists a problem instance characterized by $\mathcal{W} \subset \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^d$, $\Gamma \in \mathbb{R}^{d\times d}$, and $\theta \in \mathbb{R}^d$ with a noise process satisfying Assumption 1 such that if the non-adaptive oracle*

algorithm acquires $T \leq \max\{d\|\theta\|_2^2, \sqrt{d}\|\theta\|_2\}\rho^* \log(1/\delta)/2$ *samples, then* $\mathbb{P}(\widehat{w} \neq w^*) \geq \delta$ *for* $\delta \in (0, 0.05]$.

Proof. To begin, consider the specifications of Theorem 18 and its result. That is, a problem an arbitrary instance characterized by $\mathcal{W} \subset \mathbb{R}^d, \mathcal{Z} \subset \mathbb{R}^d$, $\Gamma \in \mathbb{R}^{d\times d}$, and $\theta \in \mathbb{R}^d$ where the noise process is jointly Gaussian and defined by $\gamma := \begin{bmatrix} \eta & \varepsilon \end{bmatrix} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{(d+1)\times(d+1)}$ is an arbitrary correlation matrix. Observe that the noise process defined by γ satisfies Assumption 1. The result states that if the non-adaptive oracle algorithm acquires $T \leq \sigma^2\rho^* \log(1/\delta)/2$ samples on the problem instance where $\sigma^2 := v^\top \Sigma v$ and $v := \begin{bmatrix} \theta & 1 \end{bmatrix} \in \mathbb{R}^{d+1}$, then $\mathbb{P}(\widehat{w} \neq w^*) \geq \delta$ for $\delta \in (0, 0.05]$. From this point, we show that there exists a parameter θ and correlation matrix Σ such that $\sigma^2 := v^\top \Sigma v \geq \max\{d\|\theta\|_2^2, \sqrt{d}\|\theta\|_2\}$ in order to reach the stated conclusion.

Notation. Let $\zeta_{\eta_i,\varepsilon} \in [-1, 1]$ denote the correlation between η_i and ε for $i \in \{1, \ldots, d\}$. Similarly, let $\zeta_{\eta_i,\eta_j} = \zeta_{\eta_j,\eta_i} \in [-1, 1]$ denote the correlation between η_i and η_j for $i \neq j \in \{1, \ldots, d\}$. Note that the correlation of η_i with itself for $i \in \{1, \ldots, d\}$ is $\sigma_{\eta_i}^2 = \zeta_{\eta_i,\eta_i} = 1$ and similarly the correlation of ε with itself is $\sigma_\varepsilon^2 = \zeta_{\varepsilon,\varepsilon} = 1$. The correlation matrix Σ is then given by

$$\Sigma = \begin{bmatrix} 1 & \zeta_{\eta_2,\eta_1} & \cdots & \zeta_{\eta_d,\eta_1} & \zeta_{\varepsilon,\eta_1} \\ \zeta_{\eta_1,\eta_2} & 1 & \cdots & \zeta_{\eta_d,\eta_2} & \zeta_{\varepsilon,\eta_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \zeta_{\eta_1,\eta_d} & \zeta_{\eta_2,\eta_d} & \cdots & 1 & \zeta_{\varepsilon,\eta_d} \\ \zeta_{\eta_1,\varepsilon} & \zeta_{\eta_2,\varepsilon} & \cdots & \zeta_{\eta_d,\varepsilon} & 1 \end{bmatrix} := \begin{bmatrix} \Sigma_\theta & \zeta_{\eta,\varepsilon} \\ \zeta_{\eta,\varepsilon}^\top & 1 \end{bmatrix},$$

where

$$\Sigma_\theta = \begin{bmatrix} 1 & \zeta_{\eta_2,\eta_1} & \cdots & \zeta_{\eta_d,\eta_1} \\ \zeta_{\eta_1,\eta_2} & 1 & \cdots & \zeta_{\eta_d,\eta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \zeta_{\eta_1,\eta_d} & \zeta_{\eta_2,\eta_d} & \cdots & 1 \end{bmatrix} \in [-1, 1]^{d\times d} \quad \text{and} \quad \zeta_{\eta,\varepsilon} = \begin{bmatrix} \zeta_{\eta_1,\varepsilon} \\ \zeta_{\eta_2,\varepsilon} \\ \vdots \\ \zeta_{\eta_d,\varepsilon} \end{bmatrix} \in [-1, 1]^d.$$

Moreover, we use the notation $\Sigma_{\theta,i}^\top \in \mathbb{R}^d$ to denote the i–th row of $\Sigma_\theta^\top$, or equivalently the i–th column of Σ_θ, for any $i \in \{1, \ldots, d\}$.

Lower Bounding Noise Variance. Given the above notation, we now work toward lower bounding $\sigma^2 := v^\top \Sigma v$. Observe that by algebraic manipulations,

$$\begin{aligned}
v^\top \Sigma v &:= \begin{bmatrix} \theta \\ 1 \end{bmatrix}^\top \begin{bmatrix} \Sigma_\theta & \zeta_{\eta,\varepsilon} \\ \zeta_{\eta,\varepsilon}^\top & 1 \end{bmatrix} \begin{bmatrix} \theta \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \theta^\top \Sigma_{\theta,1}^\top + \zeta_{\eta_1,\varepsilon} & \theta^\top \Sigma_{\theta,2}^\top + \zeta_{\eta_2,\varepsilon} & \cdots & \theta^\top \Sigma_{\theta,d}^\top + \zeta_{\eta_d,\varepsilon} & \theta^\top \zeta_{\eta,\varepsilon} + 1 \end{bmatrix} \begin{bmatrix} \theta \\ 1 \end{bmatrix} \\
&= \theta^\top \sum_{i=1}^d \Sigma_{\theta,i}^\top \theta_i + \sum_{i=1}^d \zeta_{\eta_i,\varepsilon} \theta_i + \theta^\top \zeta_{\eta,\varepsilon} + 1 \\
&= \theta^\top \Sigma_\theta \theta + 2\theta^\top \zeta_{\eta,\varepsilon} + 1.
\end{aligned}$$

Since Σ_θ is a real symmetric matrix, an eigendecomposition exists such that $\Sigma_\theta = Q \Lambda Q^\top$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d \times d}$ is a diagonal matrix containing the eigenvalues of Σ_θ and $Q \in \mathbb{R}^{d \times d}$ is an orthogonal matrix with columns corresponding to the eigenvectors of Σ_θ. Let $q_i := Q_i^\top$ denote column i of the matrix Q for $i = \{1, \ldots, d\}$, which is equivalently eigenvector i of Σ_θ for $i = \{1, \ldots, d\}$. Without loss of generality, assume that the eigenvectors are of unit length so that $\|q_i\|_2 = 1$ for all $i = \{1, \ldots, d\}$.

Given this information, suppose that the parameter θ in the instance is equal to a scalar multiple of the eigenvector of Σ_θ corresponding to the maximum eigenvalue. Note this is equivalent to the statement that θ is equal to some scalar multiple of the column $q_* \in \mathbb{R}^d$ of the matrix Q where q_* is the eigenvector of Σ_θ corresponding to the maximum eigenvalue λ^*. Thus, we take $\theta = c \cdot q_*$ for some $c \in \mathbb{R}$ and observe that $\|\theta\|_2 = c$. Toward quantifying the value of $v^\top \Sigma v$ for the problem instance, we begin by characterizing $\theta^\top \Sigma_\theta \theta$ for the choice of

θ. Consider the following analysis:

$$\begin{aligned}
\theta^\top \Sigma_\theta \theta &= \theta^\top Q \Lambda Q^\top \theta \\
&:= (cq_*)^\top Q \Lambda Q^\top (cq_*) && (\theta := cq_*) \\
&= \|\theta\|_2^2 q_*^\top \begin{bmatrix} q_1 & \cdots & q_* & \cdots & q_d \end{bmatrix} \Lambda \begin{bmatrix} q_1 & \cdots & q_* & \cdots & q_d \end{bmatrix}^\top q_* \\
&= \|\theta\|_2^2 \begin{bmatrix} 0 & \cdots & \|q_*\|_2^2 & \cdots & 0 \end{bmatrix} \operatorname{diag}(\lambda_1, \ldots, \lambda^*, \ldots, \lambda_1) \begin{bmatrix} 0 & \cdots & \|q_*\|_2^2 & \cdots & 0 \end{bmatrix}^\top \\
& && (q_i^\top q_j = 0 \ \forall i \neq j) \\
&= \|\theta\|_2 \lambda^*. && (\|q_*\|_2 = 1)
\end{aligned}$$

Thus, in general for this choice of θ,

$$v^\top \Sigma_\theta v = \|\theta\|_2^2 \lambda^* + 2\theta^\top \zeta_{\eta,\varepsilon} + 1.$$

To conclude, take $\Sigma_\theta := \mathbf{1}_d \mathbf{1}_d^\top$ where $\mathbf{1}_d$ represents the d-dimensional vector of all ones. Since this is a rank-1 matrix, the maximum eigenvalue is $\lambda^* = \mathbf{1}_d^\top \mathbf{1}_d = d$ and the remainder of the eigenvalues are zero. Observe that $q_* = \mathbf{1}_d / \sqrt{d}$ is an eigenvector corresponding to the maximum eigenvalue since $\mathbf{1}_d \mathbf{1}_d^\top q_* = d q_*$. Thus,

$$\begin{aligned}
v^\top \Sigma_\theta v &= \|\theta\|_2^2 \lambda^* + 2\theta^\top \zeta_{\eta,\varepsilon} + 1 \\
&:= \|\theta\|_2^2 \lambda^* + 2\|\theta\|_2 \mathbf{1}_d^\top \mathbf{1}_d / \sqrt{d} + 1 && (\theta := cq_* := \|\theta\|_2 \mathbf{1}_d / \sqrt{d} = \text{ and } \zeta_{\eta,\varepsilon} := \mathbf{1}_d) \\
&= d\|\theta\|_2^2 + 2\sqrt{d}\|\theta\|_2 + 1 \\
&\geq \max\{d\|\theta\|_2^2, \sqrt{d}\|\theta\|_2\}.
\end{aligned}$$

This completes the proof since we have shown that there exists a parameter θ and correlation matrix Σ such that $\sigma^2 := v^\top \Sigma v \geq \max\{d\|\theta\|_2^2, \sqrt{d}\|\theta\|_2\}$, which by Theorem 18 allows us to make the stated conclusion. □

3.7.6 Proof of Lemma 14.1

The first statement is an immediate consequence of Lemma 18.2 and the second statement is proven in Lemma 18.1.

Lemma 18.1. *In the compliance model, the noise $\eta^\top\theta+\varepsilon$ follows a $\left(8\|\theta\|_2^2+2\right)$-sub-Gaussian distribution.*

Proof. In compliance, we have $z, x \in \{e_1, \cdots, e_d\}$, and

$$\eta = x - \Big(\mathbb{P}\big(e_1 \mid z\big), \cdots, \mathbb{P}\big(e_d \mid z\big)\Big)^\top.$$

Let us figure out the sub-Gaussian parameter of the random vector η. Fix any unit vector a. First, we have $\mathbb{E}[\langle \eta, a\mathring{a}]=0$. Second, we have

$$\begin{aligned}
\left|\eta^\top a\right| &\le \|\eta\|_2 && \text{(Cauchy–Schwarz inequality)}\\
&\le \left\| x - \Big(\mathbb{P}\big(e_1 \mid z\big), \cdots, \mathbb{P}\big(e_d \mid z\big)\Big)^\top \right\|_2 \\
&\le \left(\|x\|_2 + \left\| \Big(\mathbb{P}\big(e_1 \mid z\big), \cdots, \mathbb{P}\big(e_d \mid z\big)\Big)^\top \right\|_2 \right) \\
&\le \left(1 + \left\| \Big(\mathbb{P}\big(e_1 \mid z\big), \cdots, \mathbb{P}\big(e_d \mid z\big)\Big)^\top \right\|_1 \right) && (x \in \{e_1, \cdots, e_d\} \text{ and } \|x\|_2 \le \|x\|_1, \forall x)\\
&= 2.
\end{aligned}$$

Thus, $\eta^\top a$ is bounded and zero-mean and thus 2^2-sub-Gaussian. This implies that

$$\forall \beta, \max_{a:\|a\|\le 1} \mathbb{E}[\exp(\beta\langle \eta, a\mathring{a}\rangle)] \le \exp(\frac{\beta^2 2^2}{2}) \ .$$

and thus η is a 2^2-sub-Gaussan random vector. Then, $\eta^\top\theta$ is $(2\|\theta\|)^2$-sub-Gaussian.

Using Lemma 18.2, we have that $\eta^\top\theta+\varepsilon$ is $2(4\|\theta\|^2+1)$-sub-Gaussian. □

Lemma 18.2. *Let A and B random variables that are each σ_A^2- and σ_B^2-sub-Gaussian but are correlated. Then, $A+B$ is $2(\sigma_A^2+\sigma_B^2)$-sub-Gaussian.*

Proof. By definition of sub-Gaussian, we have for any $\gamma \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(\gamma(A+B)\big)\Big] &= \mathbb{E}\big[\exp(\gamma A)\exp(\gamma B)\big] \\
&\le \sqrt{\mathbb{E}\big[\exp(2\gamma A)\big]}\sqrt{\mathbb{E}\big[\exp(2\gamma B)\big]} && \text{(Cauchy-Schwarz)} \\
&\le \sqrt{\exp\big(2\gamma^2\sigma_A^2\big)}\sqrt{\exp\big(2\gamma^2\sigma_B^2\big)} \\
&\le \exp\big(2\gamma^2(\sigma_A^2+\sigma_B^2)\big).
\end{aligned}$$

□

3.7.7 Proof of Lemma 14.2

Lemma 14.2. *Suppose that T observations are collected non-adaptively from the structural equation model in Eqs. (3.1) and $\Gamma \in \mathbb{R}^{d\times d}$ is known. Then, with probability at least $1-\delta$ for $\delta \in (0,1)$ and $w \in \mathbb{R}^d$,*

$$|w^\top(\widehat{\theta}_{oracle} - \theta)| \le \sqrt{2\sigma_\nu^2 \|w\|_{\bar{A}(Z_T,\Gamma)^{-1}} \log\big(2/\delta\big)}.$$

where σ_ν^2 is the sub-Gaussian parameter of the noise process $\nu := \eta^\top\theta + \varepsilon$ as characterized in Lemma 14.1.

Proof. Given the knowledge of Γ, we have the oracle 2SLS estimator

$$\widehat{\theta}_{\texttt{oracle}} = \left(\sum_{t=1}^{T} z_s\big(\Gamma^\top z_s\big)^\top\right)^{-1}\sum_{t=1}^{T} z_s y_t = \left(\sum_{t=1}^{T} z_s z_s^\top\Gamma\right)^{-1}\sum_{t=1}^{T} z_s y_t.$$

Note that

$$y_t = x_t^\top\theta + \varepsilon_t = \big(\Gamma^\top z_t\big)^\top\theta + \eta_t^\top\theta + \varepsilon_t.$$

Denote $\nu_t := \eta_t^\top \theta + \varepsilon_t$. For any $w \in \mathcal{W}$, we have

$$\begin{aligned}
\left\langle \widehat{\theta}_{\texttt{oracle}} - \theta, w \right\rangle &= \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \sum_{t=1}^{T} z_s y_t - \theta, w \right\rangle \\
&= \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \sum_{t=1}^{T} z_s \left(z_s^\top \Gamma \theta + \nu_t \right) - \theta, w \right\rangle \\
&= \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \theta + \sum_{t=1}^{T} z_s \nu_t \right) - \theta, w \right\rangle \\
&= \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \sum_{t=1}^{T} z_s \nu_t, w \right\rangle \\
&= \sum_{q=1}^{t} \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} z_q, w \right\rangle \nu_q.
\end{aligned}$$

By Lemma 18.1, the noise ν_t is σ_ν^2-sub-Gaussian, we have

$$\left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} z_q, w \right\rangle \nu_q$$

is $\left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} z_q, w \right\rangle^2 \sigma_\nu^2$-sub-Gaussian. Thus

$$\mathbb{P}\left(\left\langle \widehat{\theta}_{\texttt{oracle}} - \theta, w \right\rangle \geq \sqrt{2 \sum_{q=1}^{t} \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} z_q, w \right\rangle^2 \sigma_\nu^2 \log\left(\frac{1}{\delta} \right)} \right) \leq \delta.$$

Concisely,

$$\begin{aligned}
\sum_{q=1}^{t} \left\langle \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} z_q, w \right\rangle^2 &= w^\top \left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \left(\sum_{q=1}^{t} z_q z_q^\top \right) \left(\left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \right)^\top w \\
&= w^\top \Gamma^{-1} \left(\left(\sum_{t=1}^{T} z_s z_s^\top \Gamma \right)^{-1} \right)^\top w \\
&= w^\top \left(\Gamma^\top \left(\sum_{t=1}^{T} z_s z_s^\top \right) \Gamma \right)^{-1} w. \qquad (3.11)
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\left\langle \widehat{\theta}_{\texttt{oracle}} - \theta, w \right\rangle \geq \sqrt{2w^\top \left(\Gamma^\top \left(\sum_{t=1}^{T} z_s z_s^\top\right)\Gamma\right)^{-1} w\sigma_\nu^2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

We can further write it as

$$\mathbb{P}\left(\left\langle \widehat{\theta}_{\texttt{oracle}} - \theta, w \right\rangle \geq \sqrt{2\|w\|^2_{\left(\Gamma^\top\left(\sum_{t=1}^{T} z_s z_s^\top\right)\Gamma\right)^{-1}} \sigma_\nu^2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

By taking a union bound over, we have the confidence interval for the absolute value as in the statement of the lemma. □

3.7.8 Proof of Theorem 15

Theorem 15. *Suppose that $\widehat{\Gamma}$ is estimated through a design matrix $Z_{T_1} \in \mathbb{R}^{T_1 \times d}$ and $\widehat{\theta}_{\texttt{P-2SLS}}$ is estimated through a design matrix $Z_{T_2} \in \mathbb{R}^{T_2 \times d}$. Then, for any $w \in \mathcal{W}$, with probability at least $1-\delta$,*

$$|w^\top(\widehat{\theta}_{\texttt{P-2SLS}} - \theta)| \leq \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4}{\delta}\right)} + \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}} \|\theta\|_2 \sqrt{\sigma_\eta^2 \overline{\log}\left(Z_{T_1}, \delta/4\right)}.$$

where σ_ν^2 is the sub-Gaussian parameter of the noise $\nu := \eta^\top\theta + \varepsilon$, σ_η^2 is the sub-Gaussian parameter of the noise η, and

$$\overline{\log}(Z_T, \delta) := 8d \ln\left(1 + \frac{2TL_z^2}{d(2 \wedge \sigma_{\min}(Z_{T_1}^\top Z_{T_1}))}\right) + 16 \ln\left(\frac{2 \cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(Z_{T_1}^\top Z_{T_1})}\right)\right)$$

Proof. For the pseudo 2SLS estimator, we have

$$\begin{aligned}
\widehat{\theta}_{\texttt{P-2SLS}} - \theta &= \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \sum_{t=1}^{T_2} z_{I_t} y_t - \theta \\
&= \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \sum_{t=1}^{T_2} z_{I_t} \left(z_{I_t}^\top \Gamma\theta + \nu_t\right) - \theta \\
&= \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \Gamma\theta + \sum_{t=1}^{T_2} z_{I_t}\nu_t\right) - \theta \\
&= \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \sum_{t=1}^{T_2} z_{I_t}\nu_t + \left(\widehat{\Gamma}^{-1}\Gamma - I\right)\theta.
\end{aligned}$$

For any $w \in \mathcal{W}$, we have

$$\begin{aligned}
\left\langle \widehat{\theta}_{\texttt{P-2SLS}} - \theta, w \right\rangle &= \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \sum_{t=1}^{T_2} z_{I_t}\nu_t + \left(\widehat{\Gamma}^{-1}\Gamma - I\right)\theta, w \right\rangle \\
&= \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} \sum_{t=1}^{T_2} z_{I_t}\nu_t, w \right\rangle + \left\langle \left(\widehat{\Gamma}^{-1}\Gamma - I\right)\theta, w \right\rangle \\
&= \sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w \right\rangle \nu_q + \left\langle \left(\widehat{\Gamma}^{-1}\Gamma - I\right)\theta, w \right\rangle \\
&= \sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w \right\rangle \nu_q + \left\langle \left(\widehat{\Gamma}^{-1} - \Gamma^{-1}\right)\Gamma\theta, w \right\rangle. \qquad (3.12)
\end{aligned}$$

We upper bound the first term and the second term separately. For the first term, by Lemma 14.1, we know ν_t is σ_ν^2-subGaussian, we have $\sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w \right\rangle \nu_q$ is $\sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_I$ subGaussian. Thus by the concentration inequality of subGaussian random variables, we have

$$\mathbb{P}\left(\sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w \right\rangle \nu_q \geq \sqrt{2 \sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w \right\rangle^2 \sigma_\nu^2 \log\left(\frac{2}{\delta}\right)} \right) \leq \frac{\delta}{2}.$$

By similar calculation as (3.11), we have with probability at least $1-\frac{\delta}{2}$,

$$\sum_{q=1}^{T_2}\left\langle\left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma}\right)^{-1} z_{I_q}, w\right\rangle \nu_q \le \sqrt{2w^\top\left(\widehat{\Gamma}^\top\left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top\right)\widehat{\Gamma}\right)^{-1} w\sigma_\nu^2 \log\left(\frac{2}{\delta}\right)}$$
$$\le \sqrt{2\|w\|^2_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}}\sigma_\nu^2\log\left(\frac{2}{\delta}\right)}. \tag{3.13}$$

Thus with probability at least $1-\frac{\delta}{2}$,

$$\left\langle\widehat{\theta}_{\texttt{P-2SLS}}-\theta, w\right\rangle \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}}\sqrt{2\sigma_\nu^2\log\left(\frac{2}{\delta}\right)} + \left\langle\left(\widehat{\Gamma}^{-1}-\Gamma^{-1}\right)\Gamma\theta, w\right\rangle$$

By Theorem 19, we have with probability at least $1-\frac{\delta}{2}$,

$$\left\langle\left(\widehat{\Gamma}^{-1}-\Gamma^{-1}\right)\Gamma\theta, w\right\rangle \le \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}}\|\theta\|\sqrt{\sigma_\eta^2\overline{\log}\left(Z_{T_1},\delta/2\right)} \tag{3.14}$$

Combining (3.13) and (3.14), we have with probability at least $1-\delta$, for any $w\in\mathcal{W}$,

$$\left\langle\widehat{\theta}_{\texttt{P-2SLS}}-\theta, w\right\rangle \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}}\sqrt{2\sigma_\nu^2\log\left(\frac{2}{\delta}\right)} + \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}}\|\theta\|\sqrt{\sigma_\eta^2\overline{\log}\left(Z_{T_1},\delta/2\right)}.$$

By a union bound, we have the confidence interval for the absolute value of the inner product,

$$|w^\top(\widehat{\theta}_{\texttt{P-2SLS}}-\theta)| \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}}\sqrt{2\sigma_\nu^2\log\left(\frac{4}{\delta}\right)} + \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}}\|\theta\|\sqrt{\sigma_\eta^2\overline{\log}\left(Z_{T_1},\delta/4\right)}.$$

□

Theorem 19. *Suppose that the least square estimator $\widehat{\Gamma}$ is estimated through a design matrix $Z_{T_1}\in\mathbb{R}^{T_1\times d}$, then it satisfies, with probability at least $1-\delta$,*

$$\left\langle\left(\widehat{\Gamma}^{-1}-\Gamma^{-1}\right)\Gamma\theta, w\right\rangle \le \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}}\|\theta\|\sqrt{\sigma_\eta^2\overline{\log}(Z_{T_1},\delta)}$$

for any $w\in\mathcal{W}$.

Proof. Define $V = Z_{T_1}^\top Z_{T_1}$, $S \in \mathbb{R}^{T_1 \times d}$ as the matrix with i-th row being $\eta_i^\top$, the stacked noise, i.e., the data collection process of the design matrix Z_{T_1} is $X = Z_{T_1}\Gamma + S$. Then we have

$$\begin{aligned}
\left\langle \left(\widehat{\Gamma}^{-1} - \Gamma^{-1}\right)\Gamma\theta, w \right\rangle =& w^\top \left(\widehat{\Gamma}^{-1} - \Gamma^{-1}\right)\Gamma\theta \\
=& w^\top \left(\Gamma + V^{-1} Z_{T_1}^\top S\right)^{-1} V^{-1} Z_{T_1}^\top S\Gamma^{-1}\Gamma\theta \qquad \text{(Lemma 21.10)} \\
=& w^\top \widehat{\Gamma}^{-1} V^{-1/2} V^{-1/2} Z_{T_1}^\top S\theta \\
\le& \left\| w^\top \widehat{\Gamma}^{-1} V^{-1/2} \right\| \left\| V^{-1/2} Z_{T_1}^\top S\theta \right\| \\
\le& \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}} \left\| V^{-1/2} Z_{T_1}^\top S \right\|_{\mathrm{op}} \|\theta\| \\
\le& \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}} \|\theta\| \sqrt{\sigma_\eta^2 \overline{\log}(Z_{T_1}, \delta)},
\end{aligned}$$

where the last inequality is due to Lemma 19.1. □

Lemma 19.1. *Suppose we have $z_1, \dots, z_T \in \mathbb{R}^d$ and $\eta_1, \dots, \eta_T \in \mathbb{R}^d$ such that $\eta_T \mid z_1, \eta_1, \dots, z_{T-1}, \eta_{T-1}, z_T$ is σ_η^2-sub-Gaussian vector (defined in Assumption 1). Let $Z, S \in \mathbb{R}^{T\times d}$ be matrices whose t-th row is $z_t^\top$ and $\eta_t^\top$ respectively. Suppose $\|z_t\| \le L_z, \forall t$. Let $V = Z^\top Z$. Then, $\forall \delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\left\| V^{-1/2} Z_T^\top S \right\|_{op} \le \sigma_\eta \sqrt{8d \ln\left(1 + \frac{2TL_z^2}{d(2 \wedge \sigma_{\min}(V))}\right) + 16 \ln\left(\frac{2\cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(V)}\right)\right)}.$$

We abbreviate $\overline{\log}(Z_T, \delta) := 8d \ln\left(1 + \frac{2TL_z^2}{d(2\wedge\sigma_{\min}(V))}\right) + 16 \ln\left(\frac{2\cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V)}\right)\right)$.

Proof. By the definition of operator norm, we have

$$\begin{aligned}
\left\| V^{-1/2} Z_T^\top S \right\|_{\mathrm{op}} &= \sup_{\{x \mid \|x\|_2 = 1\}} \left\| V^{-1/2} Z_T^\top S x \right\|_2 \\
&= \sup_{\{x \mid \|x\|_2 = 1\}} \sqrt{x^\top S^\top Z_T V^{-1} Z_T^\top S x} \\
&= \sup_{\{x \mid \|x\|_2 = 1\}} \left\| Z_T^\top S x \right\|_{V^{-1}}.
\end{aligned}$$

Considering a fixed $w \in \mathbb{R}^d$, by Lemma 19.2, we have with probability at least $1-\delta$,

$$\left\|Z_T^\top S x\right\|_{V^{-1}} \le \sqrt{2}\sigma_\eta \sqrt{d\ln\left(1+\frac{2TL_z^2}{d(2\wedge\sigma_{\min}(V))}\right)+2\ln\left(\frac{2}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V)}\right)\right)}.$$

By Lemma 19.3 and a union bound, for the ε-covering $\mathcal{C}_\varepsilon$, we have the following event happens with probability no more than δ:

$$\mathcal{E} := \left\{\exists x\in\mathcal{C}_\varepsilon, \left\|Z_T^\top S x\right\|_{V^{-1}} \ge \sqrt{2}\sigma_\eta \sqrt{d\ln\left(1+\frac{2TL_z^2}{d(2\wedge\sigma_{\min}(V))}\right)+2\ln\left(\frac{2|\mathcal{C}_\varepsilon|}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V)}\right)\right)}\right\}.$$

We abbreviate $\widehat{\log}(Z_T,\delta) := \sqrt{2}\sqrt{d\ln\left(1+\frac{2TL_z^2}{d(2\wedge\sigma_{\min}(V))}\right)+2\ln\left(\frac{2|\mathcal{C}_\varepsilon|}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V)}\right)\right)}$ When $\mathcal{E}$ does not happen, we have

$$\begin{aligned}
\left\|V^{-1/2}Z^\top S\right\|_{\mathrm{op}} &= \sup_{\{x\mid\|x\|=1\}}\left\|Z_T^\top S x\right\|_{V^{-1}} \\
&= \sup_{\{x\mid\|x\|=1\}}\min_{\{y\mid y\in C_\varepsilon\}}\left\|Z_T^\top S(x-y+y)\right\|_{V^{-1}} \\
&\le \sup_{\{x\mid\|x\|=1\}}\min_{\{y\mid y\in C_\varepsilon\}}\left(\left\|Z_T^\top S(x-y)\right\|_{V^{-1}}+\left\|Z_T^\top S y\right\|_{V^{-1}}\right) \\
&\le \sup_{\{x\mid\|x\|=1\}}\min_{\{y\mid y\in C_\varepsilon\}}\left(\left\|V^{-1/2}Z^\top S\right\|_{\mathrm{op}}\|x-y\|_2+\sigma_\eta\widehat{\log}(Z_T,\delta)\right) \\
&\le \varepsilon\left\|V^{-1/2}Z_T^\top S\right\|_{\mathrm{op}}+\sigma_\eta\widehat{\log}(Z_T,\delta).
\end{aligned}$$

Thus,

$$\left\|V^{-1/2}Z_T^\top S\right\|_{\mathrm{op}} \le \frac{\sigma_\eta}{1-\varepsilon}\widehat{\log}(Z_T,\delta).$$

By choosing $\varepsilon=\frac{1}{2}$ and Lemma 19.3, we have

$$\left\|V^{-1/2}Z_T^\top S\right\|_{\mathrm{op}} \le \sigma_\eta\sqrt{8d\ln\left(1+\frac{2TL_z^2}{d(2\wedge\sigma_{\min}(V))}\right)+16\ln\left(\frac{2\cdot 6^d}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V)}\right)\right)}.$$

□

Lemma 19.2. *[Self-Normalized Bound for Vector-Valued Martingales] Suppose we have* $z_1, \dots, z_t \in \mathbb{R}^d$ *and* $\eta_1, \dots, \eta_t \in \mathbb{R}^d$ *such that* $\eta_t \mid z_1, \eta_1, \dots, z_{t-1}, \eta_{t-1}, z_t$ *is* σ_η^2*-sub-Gaussian vector (defined in Assumption 1). Let* $Z, S \in \mathbb{R}^{t\times d}$ *be matrices whose* s*-th row is* $z_s^\top$ *and* $\eta_s^\top$ *respectively.*

Suppose $\|z_s\| \le L_z, \forall s$*. Let* $V_t = Z^\top Z$*. Then,* $\forall b \in \mathbb{R}^t, \delta \in (0,1)$*, we have, with probability at least* $1-\delta$*,*

$$\forall t \ge 1, \left\|Z^\top Sb\right\|_{V_t^{-1}} \le \sqrt{2}\|b\|_2 \sigma_\eta \sqrt{d \ln\left(1 + \frac{2tL_z^2}{d(2 \wedge \sigma_{\min}(V_t))}\right) + 2\ln\left(\frac{2}{\delta}\cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(V_t)}\right)\right)}.$$

Proof. Since each row of S is a σ_η^2-subGaussian vector, we have that $(Sb)_i/\|b\|$ is σ_η^2-subGaussian. Using Lemma 19.4 with $\varepsilon_s = (Sb)_i/\|b\|$ completes the proof. □

Lemma 19.3. *(Lattimore and Szepesvári, 2020)[Lemma 20.1] There exists a set* $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ *with* $|\mathcal{C}_\varepsilon| \le \left(\frac{3}{\varepsilon}\right)^d$ *such that for any* $x \in \left\{x \mid x \in \mathbb{R}^d, \|x\|_2 = 1\right\}$*, there exists a* $y \in \mathcal{C}_\varepsilon$ *such that* $\|x - y\|_2 \le \varepsilon$*.*

Lemma 19.4. *Let* $z_1, z_2, \dots \in \{z \in \mathbb{R}^d : \|z\|_2 \le L_z\}$ *and* $\varepsilon_1, \varepsilon_2, \dots \in \mathbb{R}$ *be random variables such that* $\varepsilon_k \mid z_1, \varepsilon_1, \dots, z_{t-1}, \varepsilon_{t-1}, z_t$ *is* σ_ε^2*-sub-Gaussian. Let* $V_t = \sum_{s=1}^t z_s z_s^\top$*. Then,*

$$1-\delta \le \mathbb{P}\left(\forall t \ge 1, \left\|\sum_{s=1}^t z_s\varepsilon_s\right\|_{V_t^{-1}} \le \sqrt{2}\sigma_\varepsilon \sqrt{d\ln\left(1 + \frac{2tL_z^2}{d(2\wedge\sigma_{\min}(V_t))}\right) + 2\ln\left(\frac{2}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V_t)}\right)\right)}\right)$$

Proof. Let $Z_t \in \mathbb{R}^{t\times d}$ be the design matrix and define $\varepsilon_t := (\varepsilon_1, \dots, \varepsilon_t)^\top \in \mathbb{R}^t$. Let us omit the subscript t from Z_t, ε_t and V_t. Note that

$$\|Z^\top\varepsilon\|_{V^{-1}} = \|Z^\top\varepsilon\|_{(\frac{1}{2}V+\frac{1}{2}V)^{-1}} \le \|Z^\top\varepsilon\|_{(\frac{1}{2}V+\frac{1}{2}\sigma_{\min}(V)I)^{-1}} \le \sqrt{2}\|Z^\top\varepsilon\|_{(V+\sigma_{\min}(V)I)^{-1}}$$

It remains to bound $\|Z^\top\varepsilon\|_{(V+\sigma_{\min}(V)I)^{-1}}$. We use union bound with the standard self-normalized inequality of Lattimore and Szepesvári (2020)[Theorem 20.4 and Note 20.2].

Specifically, let $\lambda_k = 2^{-k+1}$ for $k \geq 1$. Then,

$$1-\delta \leq \mathbb{P}\left(\mathcal{E} := \left\{\forall k \in \mathbb{N}_+, \forall t \geq 1, \|Z^\top \varepsilon\|_{(V+\lambda_k I)^{-1}} \leq \sigma_\varepsilon \sqrt{d \ln\left(1+\frac{tL^2}{d\lambda_k}\right) + 2\ln\left(\frac{(\pi^2/6)k^2}{\delta}\right)}\right\}\right)$$

Under the event $\mathcal{E}$, we have two cases:

- $\sigma_{\min}(V) \geq 1$: choose $k = 1$. Then,

$$\begin{aligned}\|Z^\top \varepsilon\|_{(V+\sigma_{\min}(V)I)^{-1}} &\leq \|Z^\top \varepsilon\|_{(V+I)^{-1}} \\ &\leq \sigma_\varepsilon \sqrt{d \ln\left(1+\frac{tL^2}{d}\right) + 2\ln\left(\frac{(\pi^2/6)}{\delta}\right)}\end{aligned}$$

- $\sigma_{\min}(V) < 1$: choose $k = \lceil \log_2(2\sigma_{\min}(V)^{-1})\rceil \geq 2$, which is the k that satisfies $\lambda_k \leq \sigma_{\min}(V) < \lambda_{k-1}$. Then,

$$\begin{aligned}\|Z^\top \varepsilon\|_{(V+\sigma_{\min}(V)I)^{-1}} &\leq \|Z^\top \varepsilon\|_{(V+\lambda_k I)^{-1}} \\ &\leq \sigma_\varepsilon \sqrt{d \ln\left(1+\frac{tL^2}{d\lambda_k}\right) + 2\ln\left(\frac{(\pi^2/6)k^2}{\delta}\right)} \\ &\leq \sigma_\varepsilon \sqrt{d \ln\left(1+\frac{tL^2}{d\sigma_{\min}(V)/2}\right) + 2\ln\left(\frac{(\pi^2/6)\lceil \log_2(2\sigma_{\min}(V)^{-1})\rceil^2}{\delta}\right)}. \\ &\qquad (\lambda_k = \tfrac{1}{2}\lambda_{k-1} > \tfrac{1}{2}\sigma_{\min}(V);\ \text{def'n of } k)\end{aligned}$$

Altogether, we have

$$\|Z^\top \varepsilon\|_{(V+\sigma_{\min}(V)I)^{-1}} \leq \sigma_\varepsilon \sqrt{d \ln\left(1+\frac{tL^2(1 \vee 2\sigma_{\min}^{-1}(V))}{d}\right) + 2\ln\left(\frac{(\pi^2/6)(1 \vee \lceil \log_2(2\sigma_{\min}^{-1}(V))\rceil^2)}{\delta}\right)}.$$

We conclude the proof by simplifying the RHS. □

3.7.9 Proof of Theorem 16

Theorem 16. *Algorithm 4 is δ-PAC for the* `CPET-LB` *problem and terminates in at most $c(1+\omega)L_\nu \rho^* \log(1/\delta) + cr(\omega)$ samples, where c hides logarithmic factors of Δ and $|\mathcal{W}|$.*

Proof. **Part 1 δ-PAC**

By the confidence interval in Lemma 14.2, we have, with probability at least $1 - \frac{\delta}{2k^2|\mathcal{W}|}$,

$$\begin{aligned}
\left|\left\langle \widehat{\theta}^k_{\texttt{oracle}} - \theta, w - w^* \right\rangle\right| &\le \|w - w^*\|_{\left(\Gamma^\top\left(\sum_{s=1}^{N_k} z_{I_t} z_{I_t}^\top\right)\Gamma\right)^{-1}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} \\
&\le \frac{\sqrt{1+\omega}\|w - w^*\|_{(\Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}}{\sqrt{N_k}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} \\
&\le \frac{\sqrt{1+\omega}\|w - w^*\|_{(\Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}}{\sqrt{\left\lceil 2(1+\omega)\zeta_k^{-2}\rho(\mathcal{W}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)\right\rceil \vee r(\omega)}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} \\
&\le \zeta_k.
\end{aligned}$$

Define a good event $\mathcal{E}_{k,w}$ for each k and $\mathcal{W}_k$ as

$$\mathcal{E}_{k,w} = \left\{ \left|\left\langle \widehat{\theta}^k_{\texttt{oracle}} - \theta, w - w^* \right\rangle\right| \le \zeta_k \right\}.$$

We claim that with probability at least $1 - \delta$, the event $\bigcap_{k=1}^\infty \bigcap_{w \in \mathcal{W}_k} \mathcal{E}_{k,w}$ holds. It can be proved by a union bound.

$$\begin{aligned}
\mathbb{P}\left(\left(\bigcap_{k=1}^\infty \bigcap_{w\in\mathcal{W}_k} \mathcal{E}_{w,k}\right)^c\right) &\le \sum_{k=1}^\infty \sum_{w\in\mathcal{W}_k} \mathbb{P}\left(\mathcal{E}_{k,w}^c\right) \\
&\le \sum_{k=1}^\infty \sum_{w\in\mathcal{W}_k} \frac{\delta}{2k^2|\mathcal{W}|} \\
&\le \frac{\delta}{2} \sum_{k=1}^\infty \frac{1}{k^2} \\
&\le \delta, \qquad (3.15)
\end{aligned}$$

where the last step is by the fact that $\sum_{k=1}^\infty \frac{1}{k^2} = \frac{\pi^2}{6} < 2$. Under the event $\bigcap_{k=1}^\infty \bigcap_{w\in\mathcal{W}_k} \mathcal{E}_{w,k}$, to show that the best arm is never eliminated, it suffices to show that for any sub-optimal arm

$w \in \mathcal{W}_k$,

$$\begin{aligned}\left\langle w - w^*, \widehat{\theta}^k_{\texttt{oracle}} \right\rangle &= \left\langle w - w^*, \widehat{\theta}^k_{\texttt{oracle}} - \theta \right\rangle + \langle w - w^*, \theta \rangle \\ &\leq \left\langle w - w^*, \widehat{\theta}^k_{\texttt{oracle}} - \theta \right\rangle \\ &\leq \zeta_k.\end{aligned}$$

Thus the best arm w^* never satisfies the elimination condition. Next we show that at the end of stage k, any suboptimal arm w that satisfies

$$\langle \theta, w^* - w \rangle \geq 2\zeta_k$$

is eliminated. To show this, we need to show that w satisfies the elimination condition,

$$\begin{aligned}\max_{w' \in \mathcal{W}_k} \left\langle \widehat{\theta}^k_{\texttt{oracle}}, w' - w \right\rangle \geq & \left\langle \widehat{\theta}^k_{\texttt{oracle}}, w^* - w \right\rangle \\ = & \left\langle \widehat{\theta}^k_{\texttt{oracle}} - \theta, w^* - w \right\rangle + \langle \theta, w^* - w \rangle \\ \geq & -\zeta_k + 2\zeta_k \\ = & \zeta_k.\end{aligned}$$

This implies that with probability at least $1 - \delta$, w^* always survives.

Part 2 sample complexity

Define $S_k = \left\{ w \in \mathcal{W} \mid \langle w^* - w, \theta \rangle \leq 4\zeta_k \right\}$. Thus with probability at least $1 - \delta$, we have $\bigcap_k \{\mathcal{W}_k \subseteq S_k\}$. This implies the following is true with probability at least $1 - \delta$ for all k

$$\begin{aligned}\rho(\mathcal{W}_k) &= \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w, w' \in \mathcal{W}_k} \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \\ &\leq \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w, w' \in S_k} \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \\ &= \rho(S_k).\end{aligned}$$

Define Δ to be the minimum gap between w^* and any other $w \in \mathcal{W}$, i.e., $\Delta := \min_{w \in \mathcal{W} \setminus \{w^*\}} \langle w^* - w, \theta \rangle$. Then for $k \geq \left\lceil \log(4\Delta^{-1}) \right\rceil$, we have $S_k = \{w^*\}$ with probability at least $1 - \delta$. The total sample

complexity is the summation of the number of samples in each round,

$$\begin{aligned}
&\sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} N_k \\
=&\sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} \left(\left\lceil 4(1+\omega)\zeta_k^{-2} \rho(\mathcal{W}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta} \right) \right\rceil \vee r(\omega) \right) \\
\leq&\sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} \left(8(1+\omega)2^{2k} \rho(\mathcal{W}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta} \right) \vee r(\omega) \right) \\
\leq&\sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} \left(8(1+\omega)2^{2k} \rho(\mathcal{S}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta} \right) \vee r(\omega) \right). \qquad (3.16)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\rho^* =& \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w \in \mathcal{W} \setminus \{w^*\}} \frac{\|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}}{\langle w^* - w, \theta \rangle^2} \\
=& \min_{\lambda \in \Delta(\mathcal{Z})} \max_k \max_{w \in S_k} \frac{\|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}}{\langle w^* - w, \theta \rangle^2} \\
\geq& \frac{1}{\lceil \log(4\Delta^{-1}) \rceil} \min_{\lambda \in \Delta(\mathcal{Z})} \sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} \max_{w \in S_k} \frac{\|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}}{\langle w^* - w, \theta \rangle^2} \\
\geq& \frac{1}{16\lceil \log(4\Delta^{-1}) \rceil} \sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} 2^{2k} \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w \in S_k} \|w^* - w\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \\
\geq& \frac{1}{64\lceil \log(4\Delta^{-1}) \rceil} \sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} 2^{2k} \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w, w' \in S_k} \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \\
=& \frac{1}{64\lceil \log(4\Delta^{-1}) \rceil} \sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} 2^{2k} \rho(S_k), \qquad (3.17)
\end{aligned}$$

where the last inequality is by the triangle inequality, i.e.,

$$\begin{aligned}
\max_{w, w' \in S_k} \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \leq& \max_{w, w' \in S_k} \left(\|w - w^*\|_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} + \|w^* - w'\|_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}} \right)^2 \\
\leq& 4 \max_{w \in S_k} \|w - w^*\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}.
\end{aligned}$$

Combining (3.16) and (3.17), we have

$$\sum_{k=1}^{\lceil \log(4\Delta^{-1}) \rceil} N_k \le c(1+\omega) L_\nu \log\left(4\Delta^{-1}\right) \log\left(\frac{\log(4\Delta^{-1})^2 |\mathcal{W}|}{\delta}\right) \rho^* + \log\left(4\Delta^{-1}\right) r(\omega).$$

□

3.7.10 Proofs of sample complexity when given $\widehat{\Gamma}$

Algorithm 8: Confounded pure exploration with known $\widehat{\Gamma}$

Input: $\mathcal{Z}, \mathcal{W}, \widehat{\Gamma}, \delta, \varepsilon, L_\nu \ge \sigma_\nu^2, L_\eta \ge \sigma_\eta^2$

Initialize: $k = 1, \hat{\mathcal{W}}_k = \mathcal{W}$, calculate γ using (3.18);

Define: $f(w, w', \Gamma, \lambda) := \|w - w'\|^2_{(\sum_{z \in \mathcal{Z}} \Gamma^\top \lambda_z z z^\top \Gamma)^{-1}}$;

while $\exists w, w' \in \mathcal{W}_k, s.t., \left\langle w' - w, \widehat{\theta}^k_{oracle} \right\rangle > 4\gamma$ *or* $\zeta_k > \gamma$ **do**

- $\lambda_k = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w, w' \in \mathcal{W}_k} f(w, w', \widehat{\Gamma}, \lambda)$;
- $\rho(\mathcal{W}_k) = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w, w' \in \mathcal{W}_k} f(w, w', \widehat{\Gamma}, \lambda)$;
- $\zeta_k = 2^{-k}$;
- $N_k = \left\lceil 2(1+\omega)\left(\zeta_k^{-2} \wedge \frac{1}{\gamma^2}\right) \rho(\mathcal{W}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) \right\rceil \vee r(\omega)$;
- $Z_{N_k} = \mathrm{ROUND}(\lambda_k, N_k)$;
- Pull arms in Z_{N_k} and observe Y_{N_k};
- Compute $\widehat{\theta}^k_{\texttt{P-2SLS}} = \left(Z_{N_k}^\top Z_{N_k} \widehat{\Gamma}\right)^{-1} Z_{N_k}^\top Y_{N_k}$;
- $\mathcal{W}_{k+1} = \mathcal{W}_k \backslash \left\{ w \in \mathcal{W}_k \mid \exists w' \in \mathcal{W}_k, \text{s.t.}, \left\langle w' - w, \widehat{\theta}^k_{\texttt{P-2SLS}} \right\rangle > \zeta_k + \gamma \right\}$;
- $k \leftarrow k + 1$;

Output: Any $w \in \mathcal{W}_k$

Theorem 20. *Suppose that we have* $\widehat{\Gamma}$ *that is an OLS estimate from an offline dataset* $\{Z_T, X_T\}$ *collected non-adaptively through a fixed design* ξ *and the efficient rounding procedure ROUND,as well as* $T \ge \frac{4\sigma_\eta^2}{\sigma_{\min}(A(\xi,\Gamma))} \overline{\log}(Z_T, \delta/2)$. *Algorithm 8 guarantees that with probability at least* $1-\delta$, *a* 6γ*-good arm is returned, where*

$$\gamma := \max_{w, w' \in \mathcal{W}} \|w - w'\|_{\bar{A}(Z_T, \widehat{\Gamma})^{-1}} \|\theta\|_2 \sqrt{L_\eta \overline{\log}(Z_T, \delta)}. \tag{3.18}$$

Also, the algorithm terminates in at most

$$c(1+\omega)L_\nu \log(1/\delta)\rho^*(\gamma) + c(1+\omega)^2 L_\nu \log(1/\delta)\frac{\sigma_\eta^2 \overline{\log}(Z_T,\delta)}{\sigma_{\min}\big(A(\xi,\Gamma)\big)}\frac{\rho(\xi,\gamma)}{T} \vee cr(\omega)$$

samples, where c hides logarithmic factors of Δ and $|\mathcal{W}|$ and γ.

$$\rho(\lambda,\gamma) \coloneqq \max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2 \vee \gamma^2}$$

and

$$\rho^*(\gamma) = \min_{\lambda\in\Delta(\mathcal{Z})}\max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2 \vee \gamma^2}.$$

Proof. The proof can be divided into four steps:

- The best arm w^* is never eliminated.
- At the end of stage k, any suboptimal arm w that satisfies $\langle \theta, w^*-w\rangle \geq 2\zeta_k + 2\gamma$ is eliminated.
- The stopping condition is met in finite time.
- When the stopping condition is met, there are only 6γ-good arms left.
- The upper bound of the sample complexity.

Step 1: The best arm w^* is never eliminated.

By the confidence interval in Theorem 15, we have, with probability at least $1 - \frac{\delta}{2k^2|\mathcal{W}|}$, for any k and $w \in \mathcal{W}_k$,

$$\begin{aligned}
\left|\left\langle \widehat{\theta}^k_{\texttt{P-2SLS}} - \theta, w - w^* \right\rangle\right| &\le \|w - w^*\|_{\bar{A}(Z_{N_k},\widehat{\Gamma})^{-1}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} + \gamma \\
&\le \frac{\sqrt{1+\omega}\|w - w^*\|_{A(\lambda_k,\widehat{\Gamma})^{-1}}}{\sqrt{N_k}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} + \gamma \\
&\le \frac{\sqrt{1+\omega}\|w - w^*\|_{A(\lambda_k,\widehat{\Gamma})^{-1}}}{\sqrt{\left\lceil 2(1+\omega)\zeta_k^{-2}\rho(\mathcal{W}_k)L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)\right\rceil \vee r(\omega)}} \sqrt{2\sigma_\nu^2 \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)} + \gamma \\
&\le \zeta_k + \gamma.
\end{aligned}$$

Define a good event $\mathcal{E}_{k,w}$ for each k and $w \in \mathcal{W}_k$ as

$$\mathcal{E}_{k,w} = \left\{ \left| \left\langle \widehat{\theta}^k_{\texttt{P-2SLS}} - \theta, w - w^* \right\rangle \right| \le \zeta_k + \gamma \right\}.$$

By the same calculation as (3.15), we claim that with probability at least $1 - \delta$, the event $\bigcap_{k=1}^{\infty} \bigcap_{w \in \mathcal{W}_k} \mathcal{E}_{k,w}$ holds. Under the event $\bigcap_{k=1}^{\infty} \bigcap_{w \in \mathcal{W}_k} \mathcal{E}_{w,k}$, to show that the best arm is never eliminated, it suffices to show that for any sub-optimal arm $w \in \mathcal{W}_k$,

$$\begin{aligned}
\left\langle w - w^*, \widehat{\theta}^k_{\texttt{P-2SLS}} \right\rangle &= \left\langle w - w^*, \widehat{\theta}^k_{\texttt{P-2SLS}} - \theta \right\rangle + \langle w - w^*, \theta \rangle \\
&\le \left\langle w - w^*, \widehat{\theta}^k_{\texttt{P-2SLS}} - \theta \right\rangle \\
&\le \zeta_k + \gamma.
\end{aligned}$$

Thus the best arm w^* never satisfies the elimination condition.

Step 2: At the end of stage k, any suboptimal arm w that satisfies $\langle \theta, w^* - w \rangle \ge 2\zeta_k + 2\gamma$ is eliminated.

To prove this, we show that such arm w must satisfy the elimination condition,

$$\begin{aligned}\max_{w' \in \mathcal{W}_k} \left\langle \widehat{\theta}^k_{\text{P-2SLS}}, w' - w \right\rangle \geq & \left\langle \widehat{\theta}^k_{\text{P-2SLS}}, w^* - w \right\rangle \\ = & \left\langle \widehat{\theta}^k_{\text{P-2SLS}} - \theta, w^* - w \right\rangle + \langle \theta, w^* - w \rangle \\ \geq & -\zeta_k - \gamma + 2\zeta_k + 2\gamma \\ = & \zeta_k + \gamma.\end{aligned}$$

Thus the arm w is eliminated.

Step 3: The stopping condition is met in finite time.

Given the result in Step 2 and the fact that ζ_k is an exponentially decreasing sequence, we know that all of the arms w satisfying $\langle \theta, w^* - w \rangle > 2\gamma$ will be eliminated in finite time. This means that only the arms w satisfying $\langle \theta, w^* - w \rangle \leq 2\gamma$ will remain. We need to show that $\forall w, w' \in \mathcal{W}_k, \left\langle w' - w, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle \leq 4\gamma$ can be achieved in finite time. When $\zeta_k \leq \gamma$, (which will happen in finite time),

$$\begin{aligned}\left\langle w' - w, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle = & \left\langle w' - w + w^* - w^*, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle \\ = & \left\langle w^* - w, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle + \left\langle w' - w^*, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle \\ = & \left\langle w^* - w, \widehat{\theta}^k_{\text{P-2SLS}} - \theta + \theta \right\rangle + \left\langle w' - w^*, \widehat{\theta}^k_{\text{P-2SLS}} - \theta + \theta \right\rangle \\ = & \left\langle w^* - w, \widehat{\theta}^k_{\text{P-2SLS}} - \theta \right\rangle + \langle w^* - w, \theta \rangle + \left\langle w' - w^*, \widehat{\theta}^k_{\text{P-2SLS}} - \theta \right\rangle + \langle w' - w^*, \theta \rangle \\ \leq & \zeta_k + \gamma + 2\gamma + \zeta_k + \gamma \\ = & 4\gamma.\end{aligned}$$

Step 4: When the stopping condition is met, there are only 6γ-good arms left.

For any $w \in \mathcal{W}_k$, we have

$$\begin{aligned}\langle w^* - w, \theta \rangle = & \left\langle w^* - w, \theta - \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle + \left\langle w^* - w, \widehat{\theta}^k_{\text{P-2SLS}} \right\rangle \\ \leq & \zeta_k + \gamma + 4\gamma \\ = & 6\gamma.\end{aligned}$$

Step 5: The upper bound of the sample complexity.

Define $\mathcal{W}(2\gamma)$ as the set of 2γ-good arms, i.e., $\mathcal{W}(2\gamma) := \{w \in \mathcal{W} \mid \langle \theta, w^* - w\rangle \le 2\gamma\}$. Then the best arm in the set $\mathcal{W}\backslash\mathcal{W}(2\gamma)$ has a suboptimality gap $\min_{w\in\mathcal{W}\backslash\mathcal{W}(2\gamma)}\langle\theta, w^* - w\rangle$. We define $\Delta_{\min}(2\gamma) := \min_{w\in\mathcal{W}\backslash\mathcal{W}(2\gamma)}\langle\theta, w^* - w\rangle - 2\gamma$. By the result in Step 2, we know that after $k^* := \left\lceil \log\left(4\Delta_{\min}(2\gamma)^{-1}\right)\right\rceil$ stages, all of the arms in $\mathcal{W}\backslash\mathcal{W}(2\gamma)$ are eliminated. Define $S_k = \{w \in \mathcal{W} \mid \langle w^* - w, \theta\rangle \le 4\zeta_k + 2\gamma\}$. Thus with probability at least $1-\delta$, we have $\bigcap_k \{\mathcal{W}_k \subseteq S_k\}$.

The sample complexity of the algorithm is the total number of samples pulled, which is

$$
\begin{aligned}
&\sum_{k=1}^{k^*} N_k \\
&= \sum_{k=1}^{k^*}\left(\left\lceil 2(1+\omega)\left(\zeta_k^{-2} \wedge \frac{1}{\gamma^2}\right)\rho(\mathcal{W}_k)L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)\right\rceil \vee r(\omega)\right) \\
&\le \sum_{k=1}^{k^*}\left(3(1+\omega)\left(2^{2k} \wedge \frac{1}{\gamma^2}\right)\rho(\mathcal{W}_k)L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) \vee r(\omega)\right) \\
&\le \sum_{k=1}^{k^*}\left(3(1+\omega)\left(2^{2k} \wedge \frac{1}{\gamma^2}\right)\rho(\mathcal{S}_k)L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) \vee r(\omega)\right). \qquad (3.19)
\end{aligned}
$$

Note that the factor of $\min_{\lambda\in\Delta(\mathcal{Z})}\max_{w,w'\in S_k}\|w - w'\|^2_{A(\lambda,\Gamma)^{-1}}$ has the underlying true Γ in it, while the algorithm uses the plugged-in $\widehat{\Gamma}$. We need to relate it to $\rho(S_k) := \min_{\lambda\in\Delta(\mathcal{Z})}\max_{w,w'\in S_k}\|w - w'\|^2_{A(\lambda,\widehat{\Gamma})^{-1}}$. By Lemma 21.7, and defining $\lambda_z^*(S_k)$ as the optimal design for S_k, i.e.,

$$
\lambda_k^* = \arg\min_{\lambda\in\Delta(\mathcal{Z})}\max_{w,w'\in S_k}\|w - w'\|^2_{A(\lambda,\Gamma)^{-1}}.
$$

Define $\lambda_k' = \frac{1}{2}\lambda_k^* + \frac{1}{2}\xi$, we have

$$
\begin{aligned}
&\max_{w,w' \in W_k} \|w - w'\|^2_{A(\lambda_k, \widehat{\Gamma})^{-1}} \\
=& \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in W_k} \|w - w'\|^2_{A(\lambda, \widehat{\Gamma})^{-1}} \\
\le& \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in S_k} \|w - w'\|^2_{A(\lambda, \widehat{\Gamma})^{-1}} \\
\le& \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in S_k} 3\|w - w'\|^2_{A(\lambda, \Gamma)^{-1}} + 2 \max_{w,w' \in S_k} \left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})(w - w')\right\|^2_{A(\lambda, I)^{-1}} \\
\le& \max_{w,w' \in S_k} 3\|w - w'\|^2_{A(\lambda_k', \Gamma)^{-1}} + 2 \max_{w,w' \in S_k} \left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})(w - w')\right\|^2_{A(\lambda_k', I)^{-1}} \\
\le& \max_{w,w' \in S_k} 6\|w - w'\|^2_{A(\lambda_k^*, \Gamma)^{-1}} + \frac{16\sigma_\eta^2 \overline{\log}\big(Z_T, \delta/2\big)}{\sigma_{\min}\big(A(\xi, \Gamma)\big)} \max_{w,w' \in S_k} \|w - w'\|^2_{A(\xi, \Gamma)^{-1}} \frac{1}{T},
\end{aligned} \tag{3.20}
$$

where the last inequality is due to

$$
\begin{aligned}
\|w - w'\|^2_{A(\lambda_k', \Gamma)^{-1}} \le& \|w - w'\|^2_{A\left(\frac{1}{2}\lambda_k^* + \frac{1}{2}\xi, \Gamma\right)^{-1}} \\
\le& \|w - w'\|^2_{A\left(\frac{1}{2}\lambda_k^*, \Gamma\right)^{-1}} \\
=& 2\|w - w'\|^2_{A(\lambda_k^*, \Gamma)^{-1}},
\end{aligned}
$$

as well as, by Lemma 21.9, when $T \ge \frac{4\sigma_\eta^2}{\sigma_{\min}(A(\xi,\Gamma))} \overline{\log}\big(Z_T, \delta/2\big)$, we have, with probability at least $1 - \delta/2$, we have

$$
\begin{aligned}
&\max_{w,w' \in S_k} \left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})(w - w')\right\|^2_{A(\lambda_k', I)^{-1}} \\
\le& \frac{4\sigma_\eta^2 \overline{\log}\big(Z_T, \delta/2\big)}{\sigma_{\min}\big(A(\lambda_k', \Gamma)\big)} \max_{w,w' \in S_k} \|w - w'\|^2_{A(\xi, \Gamma)^{-1}} \frac{1}{T} \\
\le& \frac{4\sigma_\eta^2 \overline{\log}\big(Z_T, \delta/2\big)}{\sigma_{\min}\big(A(\alpha\lambda_k^* + (1-\alpha)\xi, \Gamma)\big)} \max_{w,w' \in S_k} \|w - w'\|^2_{A(\xi, \Gamma)^{-1}} \frac{1}{T} \\
\le& \frac{8\sigma_\eta^2 \overline{\log}\big(Z_T, \delta/2\big)}{\sigma_{\min}\big(A(\xi, \Gamma)\big)} \max_{w,w' \in S_k} \|w - w'\|^2_{A(\xi, \Gamma)^{-1}} \frac{1}{T}.
\end{aligned}
$$

We can lower bound $\rho^*(\gamma)$ by

$$\begin{aligned}
\rho^*(\gamma) &= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2\vee\gamma^2} \\
&= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{k} \max_{w\in S_k\setminus\{w^*\}} \frac{\|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2\vee\gamma^2} \\
&\geq \frac{1}{k^*} \min_{\lambda\in\Delta(\mathcal{Z})} \sum_{k=1}^{k^*} \max_{w\in S_k\setminus\{w^*\}} \frac{\|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2\vee\gamma^2} \\
&\geq \frac{1}{16k^*} \sum_{k=1}^{k^*} \left(2^{2k}\wedge\frac{1}{\gamma^2}\right) \min_{\lambda\in\Delta(\mathcal{Z})} \max_{w\in S_k\setminus\{w^*\}} \|w^*-w\|^2_{A(\lambda,\Gamma)^{-1}} \\
&\geq \frac{1}{64k^*} \sum_{k=1}^{k^*} \left(2^{2k}\wedge\frac{1}{\gamma^2}\right) \min_{\lambda\in\Delta(\mathcal{Z})} \max_{w,w'\in S_k} \|w-w'\|^2_{A(\lambda,\Gamma)^{-1}}. \qquad (3.21)
\end{aligned}$$

Given (3.19), (3.20) and (3.21), we have

$$\sum_{k=1}^{k^*} N_k \leq c(1+\omega)L_\nu \log(1/\delta)\rho^*(\gamma) + c(1+\omega)^2 L_\nu \log(1/\delta)\frac{\sigma_\eta^2\overline{\log}(Z_T,\delta)}{\sigma_{\min}\big(A(\xi,\Gamma)\big)}\frac{\rho(\xi,\gamma)}{T} + cr(\omega),$$

where c hides logarithmic factors of Δ and $|\mathcal{W}|$ and γ. □

3.7.11 Proofs of sample complexity with unknown Γ

Algorithm 9: Optimal design with unknown Γ

Input: $\mathcal{Z}, \mathcal{W}, \delta, L_\nu \geq \sigma_\nu^2, L_\eta \geq \sigma_\eta^2, \omega, \gamma_{\min} \leq \lambda_{\min}(\Gamma), \lambda_E, \kappa_0$

Initialize: $k=1, \mathcal{W}_1=\mathcal{W}, \widehat{\Gamma}_0 = \perp, \zeta_1 = 1$

Define: $f(w,w',\Gamma,\lambda) := \|w-w'\|^2_{(\sum_{z\in\mathcal{Z}}\Gamma^\top\lambda_z zz^\top\Gamma)^{-1}}$, $M := \frac{32L_\eta}{\gamma^2_{\min}\sigma_{\min}(A(\lambda_E,I))}\vee 1$, $\delta_{k,\ell} := \frac{\delta}{4k^2\ell^2}$

while $|\mathcal{W}_k| > 1$ **do**

$\widehat{\Gamma}_k = \Gamma - \texttt{estimator}\Big(\mathcal{W}_k, \widehat{\Gamma}_{k-1}, \zeta_k, \delta, k, \omega, \lambda_E, M, L_\eta\Big)$ `// Step 1: update` $\widehat{\Gamma}$

$\widehat{\theta}^k_{\texttt{P-2SLS}} = \theta - \texttt{estimator}\Big(\mathcal{W}_k, \delta, \zeta_k, \widehat{\Gamma}_k, \omega, L_\nu, k\Big)$ `// Step 2: update` $\widehat{\theta}$

$\mathcal{W}_{k+1} = \mathcal{W}_k \setminus \Big\{w\in\mathcal{W}_k \mid \exists w'\in\mathcal{W}_k, \text{s.t.}, \Big\langle w'-w, \widehat{\theta}^k_{\texttt{P-2SLS}}\Big\rangle > \zeta_k\Big\}$ `// Step 3: elimination`

$k \leftarrow k+1$, $\zeta_k = 2^{-k}$

Output: The single remaining element in $\mathcal{W}_k$

The algorithms of $\Gamma-$`estimator` and $\theta-$`estimator` we present below are slightly different from the one in the main text. In the main text, we omit the phase index k in the algorithm for simplicity.

Algorithm 10: $\Gamma-$`estimator`

Input: $\mathcal{W}_k, \widehat{\Gamma}_{k-1}, \zeta_k, \delta, k, \omega, \lambda_E, M, L_\eta$

Define: $\texttt{Stop}(\mathcal{W}, Z, \Gamma, \delta) := \max_{w,w' \epsilon \mathcal{W}} \|w - w'\|_{\bar{A}(Z,\Gamma)^{-1}} \|\theta\|_2 \sqrt{L_\eta \overline{\log}(Z, \delta)}$

Initialize: $\ell = 1$, $N_{k,0,0} = 0$ `// doubling trick initialization`

if $k = 1$ **then**

  while $\ell = 1$ *or* $\texttt{Stop}\Big(\mathcal{W}_k, Z_{k,0,\ell}, \widehat{\Gamma}_k, \delta_{k,\ell}\Big) > 1$ **do**

    Get $2^{\ell-1}\Big(r(\omega) \vee \frac{2}{\kappa_0}\Big)$ samples denoted as $\big\{Z_{k,0,\ell}, X_{k,0,\ell}, Y_{k,0,\ell}\big\}$ per design λ_E `// via ROUND`

    Update $\widehat{\Gamma}_k$ by `OLS` on $\big\{Z_{k,0,\ell}, X_{k,0,\ell}\big\}$, $\ell \leftarrow \ell + 1$

else

  $\tilde{\lambda}_k = \arg\min_{\lambda \epsilon \Delta(\mathcal{Z})} \max_{w,w' \epsilon \mathcal{W}_k} f(w, w', \widehat{\Gamma}_{k-1}, \lambda)$

  $N' = \left\lfloor 4gdM \ln\big(1 + 2M(d + L_z^2) + 2M2gdM\big) + 8M \ln\left(\frac{2 \cdot 6^d}{\delta_{k,1}}\right) \vee r(\omega) \right\rfloor$

  while $\ell = 1$ *or* $\texttt{Stop}\Big(\mathcal{W}_k, Z_{k,0,\ell} \cup Z_{k,1,\ell}, \widehat{\Gamma}_k, \delta_{k,\ell}\Big) > \zeta_k$ **do**

    $N_{k,1,\ell} = 2^\ell N'$ `// doubling trick update`

    Get $N_{k,1,\ell}$ samples per $\tilde{\lambda}_k$ denoted as $\big\{Z_{k,1,\ell}, X_{k,1,\ell}, Y_{k,1,\ell}\big\}$ `// via ROUND`

    $N_{k,0,\ell} = \left\lceil 2gdM \ln\Big(M\big(d + N_{k,1,\ell} + L_z^2\big)\Big) + 4M \ln\left(\frac{2 \cdot 6^d}{\delta_{k,\ell}}\right) \vee r(\omega) \vee \frac{2}{\kappa_0} \right\rceil$

    Get $(N_{k,0,\ell} - N_{k,0,\ell-1})$ samples per λ_E augmented to $\big\{Z_{k,0,\ell-1}, X_{k,0,\ell-1}\big\}$ and get $\big\{Z_{k,0,\ell}, X_{k,0,\ell}\big\}$

    Update $\widehat{\Gamma}_k$ by `OLS` on $\big\{Z_{k,0,\ell} \cup Z_{k,1,\ell}, X_{k,0,\ell} \cup X_{k,1,\ell}\big\}$, $\ell \leftarrow \ell + 1$

Output: $\widehat{\Gamma}_k$

Algorithm 11: θ – `estimator`

Input: $\mathcal{W}_k, \delta, \zeta_k, \widehat{\Gamma}_k, \omega, L_\nu, k$

$\widehat{\lambda}_k = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} f(w, w', \widehat{\Gamma}_k, \lambda)$

$\rho(\mathcal{W}_k) = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} f(w, w', \widehat{\Gamma}_k, \lambda)$

$N_{k,2} = \left\lceil 2(1+\omega)\zeta_k^{-2}\rho(\mathcal{W}_k)L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right)\right\rceil \vee r(\omega)$

Get $N_{k,2}$ samples per design $\widehat{\lambda}_k$ denoted as $\left\{Z_{k,2}, X_{k,2}, Y_{k,2}\right\}$ `// via ROUND`

Update $\widehat{\theta}^k_{\texttt{P-2SLS}} = (\widehat{\Gamma}_k^\top Z_{k,2}^\top Z_{k,2} \widehat{\Gamma}_k)^{-1} \widehat{\Gamma}_k^\top Z_{k,2}^\top Y_{k,2}$

Output: $\widehat{\theta}^k_{\texttt{P-2SLS}}$

Lemma 20.1. *Algorithm 6 and $N_{k,1,\ell}, N_{k,0,\ell}$ guarantees three properties:*

- *Property 1:* $N_{k,1,\ell} \geq N_{k,0,\ell}$.
- *Property 2:* $\frac{1}{2}\left(N_{k,0,\ell} + N_{k,1,\ell}\right) \leq N_{k,0,\ell-1} + N_{k,1,\ell-1}$.
- *Property 3:* $\frac{N_{k,1,1}}{8d\ln\left(1+\frac{N_{k,1,1}L_z^2}{d}\right)+16\ln\left(\frac{2\cdot 6^d k^2}{\delta}\right)} \leq M\ln(dM)$.

Proof. For Property 1, recall that $N_{k,0,\ell}$ and $N_{k,1,\ell}$ are defined as

$$N_{k,0,\ell} = \left\lceil 2gdM\ln\left(M\left(d + N_{k,1,\ell} + L_z^2\right)\right) + 4M\ln\left(\frac{2\cdot 6^d}{\delta_{k,\ell}}\right) \vee r(\omega)\right\rceil$$

and

$$N_{k,1,\ell} = 2^\ell\left\lceil 4gdM\ln\left(1 + 2M\left(d + L_z^2\right) + 2M2gdM\right) + 8M\ln\left(\frac{2\cdot 6^d}{\delta_{k,1}}\right) \vee r(\omega)\right\rceil.$$

We prove the result by induction. For the result for $\ell = 1$, note that $N_{k,1,1}$ is of the form $N_{k,1,1} = 2a + 2b\ln(1 + 2c + 2bd)$ and $N_{k,0,1}$ is of the form $N_{k,0,1} = a + b\ln(c + dr)$, where

- $a = 4M\ln\left(\frac{2\cdot 6^d}{\delta_{k,1}}\right)$
- $b = 2gdM$
- $c = M(d + L_z^2)$
- $d = M$

- $r = N_{k,1,\ell}$.

By the contraposition of Lemma 23.2, we have

$$r > 2a + 2b\ln(1 + 2c + 2bd) \implies r > a + b\ln(c + dr).$$

Thus we have $N_{k,1,1} \geq N_{k,0,1}$. Now we assume the result holds for ℓ, i.e., $N_{k,1,\ell} \geq N_{k,0,\ell}$ and prove that it holds for $\ell + 1$. We have

$$N_{k,1,\ell+1} = 2N_{k,1,\ell} \geq 2N_{k,0,\ell}.$$

It suffices to prove that

$$2N_{k,0,\ell} \geq N_{k,0,\ell+1}.$$

We have

$$\begin{aligned}
&2N_{k,0,\ell} \\
=&2\left\lceil 2gdM\ln\left(M\left(d + N_{k,1,\ell} + L_z^2\right)\right) + 4M\ln\left(\frac{2\cdot 6^d}{\delta_{k,\ell}}\right) \vee r(\omega)\right\rceil \\
\geq&2gdM\ln\left(M^2\left(d + N_{k,1,\ell} + L_z^2\right)^2\right) + 8M\ln\left(\frac{2\cdot 6^d\ell^2}{\delta_{k,1}}\right) \vee r(\omega) \\
\geq&2gdM\ln\left(M^2\left(d + N_{k,1,\ell} + L_z^2\right)^2\right) + 8M\left(\ln\left(\frac{2\cdot 6^d}{\delta_{k,1}}\right) + 2\ln(\ell)\right) \vee r(\omega) \\
\geq&2gdM\ln\left(M^2\left(d + N_{k,1,\ell} + L_z^2\right)^2\right) + 4M\left(\ln\left(\frac{2\cdot 6^d}{\delta_{k,1}}\right) + 2\ln(\ell + 1)\right) \vee r(\omega) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2\ln\ell \geq \ln(\ell+1) \text{ for } \ell \geq 2) \\
\geq&2gdM\ln\left(M\left(d + 2N_{k,1,\ell} + L_z^2\right)\right) + 4M\ln\left(\frac{2\cdot 6^d(\ell+1)^2}{\delta_{k,1}}\right) \vee r(\omega) \\
=&N_{k,0,\ell+1}.
\end{aligned}$$

For Property 2, it suffices to prove that

$$\frac{1}{2}N_{k,0,\ell} \le N_{k,0,\ell-1}.$$

This is equivalent to prove that

$$\frac{1}{2}\big(a + b\ln(c+2dr)\big) \le a + b\ln(c+dr).$$

We have

$$\begin{aligned}
&\frac{1}{2}\big(a + b\ln(c+2dr)\big) \le a + b\ln(c+dr)\\
\iff &\frac{1}{2}b\ln(c+2dr) \le \frac{1}{2}a + b\ln(c+dr)\\
\iff &b\ln\Big(\sqrt{c+2dr}\Big) \le \frac{1}{2}a + b\ln(c+dr)\\
\impliedby &\sqrt{c+2dr} \le c+dr\\
\iff &c+2dr \le c^2 + 2cdr + d^2r^2\\
\iff &dr(dr+2c-2) + c^2 - c \ge 0.
\end{aligned}$$

The last inequality holds because $c = M > 1$.

For Property 3, we have,

$$\begin{aligned}
&\frac{4dM\ln\big(1+2M(d+L_z^2)+2MdM\big) + 8M\ln\Big(\frac{2\cdot 6^d k^2}{\delta}\Big)}{8d\ln\Big(1+\frac{N_{k,1,1}L_z^2}{d}\Big) + 16\ln\Big(\frac{2\cdot 6^d k^2}{\delta}\Big)}\\
\le &\frac{4dM\ln\big(1+2M(d+L_z^2)+2MdM\big)}{8d\ln\Big(1+\frac{N_{k,1,1}L_z^2}{d}\Big) + 16\ln\Big(\frac{2\cdot 6^d k^2}{\delta}\Big)} + \frac{8M\ln\Big(\frac{2\cdot 6^d k^2}{\delta}\Big)}{8d\ln\Big(1+\frac{N_{k,1,1}L_z^2}{d}\Big) + 16\ln\Big(\frac{2\cdot 6^d k^2}{\delta}\Big)}\\
\le &\frac{M\ln\big(1+2M(d+L_z^2)+2MdM\big)}{2\ln\big(1+ML_z^2\big)} + \frac{M}{2} \qquad \text{(loosely apply } N_{k,1,1} \ge dM\text{)}\\
\le &M\ln(dM).
\end{aligned}$$

□

Theorem 21. *Algorithm 6 guarantees that with probability at least $1-\delta$, the best arm is returned, and the algorithm terminates in at most*

$$(1+\omega)\Big(\Big(L_\nu \log(1/\delta) + L_\eta \|\theta\|_2^2\big(d+\log(1/\delta)\big)\Big)\rho^* + \big(d+\log(1/\delta)\big)\Big(L_\eta\|\theta\|_2^2\rho_0 + M\Big)\Big)$$

pulls, ignoring both of the additive and multiplicative logarithms of $\Delta, |\mathcal{W}|, \rho^*, \rho_0, M$, where

$$\rho^* = \min_{\lambda\in\Delta(\mathcal{Z})}\max_{w\in\mathcal{W}\setminus\{w^*\}} \frac{\|w^*-w\|^2_{(\sum_{z\in\mathcal{Z}}\lambda_z\Gamma^\top zz^\top\Gamma)^{-1}}}{\langle w^*-w,\theta\rangle^2},$$

and

$$\rho_0 = \max_{w\in\mathcal{W}\setminus\{w^*\}} \|w^*-w\|^2_{(\sum_{z\in\mathcal{Z}}\lambda_{E,z}\Gamma^\top zz^\top\Gamma)^{-1}},$$

and

$$M = \frac{32L_\eta}{\gamma^2_{\min}\sigma_{\min}\big(A(\lambda_E, I)\big)} \vee 1.$$

Note that ρ_0 does not get hurt by $\langle w^*-w,\theta\rangle$. It comes from the fact that in the first phase, we initialize that algorithm with E-optimal design.

Proof. **Part 1: correctness of the algorithm**

The idea of the proof is similar to the proof of Theorem 16.

Recall that the confidence interval of `P-2SLS` can be break down into two terms (3.12).

$$\Big\langle \widehat{\theta}_{\texttt{P-2SLS}} - \theta, w\Big\rangle = \sum_{q=1}^{T_2}\Big\langle\Big(\sum_{t=1}^{T_2} z_{I_t}z_{I_t}^\top\widehat{\Gamma}\Big)^{-1} z_{I_q}, w\Big\rangle\nu_q + \Big\langle\big(\widehat{\Gamma}^{-1}-\Gamma^{-1}\big)\Gamma\theta, w\Big\rangle.$$

Given $\widehat{\Gamma}$, for a $w\in\mathcal{W}$, with probability at least $1-\frac{\delta}{2}$ the first term satisfies

$$\sum_{q=1}^{T_2}\Big\langle\Big(\sum_{t=1}^{T_2} z_{I_t}z_{I_t}^\top\widehat{\Gamma}\Big)^{-1} z_{I_q}, w\Big\rangle\nu_q \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}}\sqrt{2\sigma_\nu^2\log\Big(\frac{4}{\delta}\Big)}.$$

For any $w \in \mathcal{W}$, with probability at least $1 - \frac{\delta}{2}$, the second term satisfies

$$\left\langle \left(\widehat{\Gamma}^{-1} - \Gamma^{-1}\right)\Gamma\theta, w \right\rangle \le \|w\|_{\bar{A}(Z_{T_1},\widehat{\Gamma})^{-1}} \|\theta\|_2 \sqrt{\sigma_\eta^2 \overline{\log}\left(Z_{T_1}, \delta/4\right)}.$$

Note that by Lemma 19.1, the above inequality holds for all $w \in \mathcal{W}$, and the RHS is essentially a result of

$$\left\|V^{-1/2} Z_T^\top S\right\|_{\text{op}} \le \sqrt{\sigma_\eta^2 \overline{\log}\left(Z_T, \delta/4\right)}.$$

In the vanilla form of the confidence of `P-2SLS`, we can define good events as

- for the first term, for any $w \in \mathcal{W}$,

$$\sum_{q=1}^{T_2} \left\langle \left(\sum_{t=1}^{T_2} z_{I_t} z_{I_t}^\top \widehat{\Gamma} \right)^{-1} z_{I_q}, w \right\rangle \nu_q \le \|w\|_{\bar{A}(Z_{T_2},\widehat{\Gamma})^{-1}} \sqrt{2\sigma_\nu^2 \log\left(16k^2|\mathcal{W}|/\delta\right)}.$$

- for the second term, $\left\|V^{-1/2} Z_T^\top S\right\|_{\text{op}} \le \sqrt{\sigma_\eta^2 \overline{\log}\left(Z_T, \delta/4\right)}$.

For our algorithm design, since we use the doubling trick for the first sub-phase, we need to define the good event for the first sub-phase as the samples from each doubling trick iteration satisfies the self-normalized concentration inequality of Lemma 19.1.

We define the good event for ℓ-th doubling trick iteration in the first sub-phase of phase k as

$$\mathcal{E}_{k,\ell}^1 = \left\{ \left\| V_{k,\ell}^{-1/2} Z_{k,\ell}^\top S_{k,\ell} \right\|_{\text{op}}^2 \le \sigma_\eta^2 \overline{\log}\left(Z_{k,\ell}, \frac{\delta}{4k^2(\ell+1)^2} \right) \right\},$$

where $Z_{k,\ell} = Z_{k,0,\ell} \cup Z_{k,1,\ell}$, $V_{k,\ell} = Z_{k,\ell}^\top Z_{k,\ell}$ and $S_{k,\ell}$ is stacked noise matrix during collecting samples $Z_{k,\ell}$ per the model $X = Z\Gamma + S$. By a union bound, we have

$$\mathbb{P}\left(\left(\bigcap_{k=1}^{\infty} \bigcap_{\ell=1}^{\infty} \mathcal{E}_{k,\ell}^1 \right)^c \right) \le \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \mathbb{P}\left(\mathcal{E}_{k,\ell}^1 \right) \le \delta/2.$$

For the second sub-phase in phase k, we define the good event for the second sub-phase in phase k and $w \in \mathcal{W}$ as

$$\mathcal{E}^2_{k,w} = \left\{ \sum_{q=1}^{N_{k,2}} \left\langle \left(\sum_{t=1}^{N_{k,2}} z_{I_t} z_{I_t}^\top \widehat{\Gamma} \right)^{-1} z_{I_q}, w \right\rangle \nu_q \le \|w\|_{\bar{A}(Z_{k,2},\widehat{\Gamma})^{-1}} \sqrt{2\sigma_\nu^2 \log\left(16k^2|\mathcal{W}|/\delta\right)} \right\}.$$

By a union bound, we have

$$\mathbb{P}\left(\left(\bigcap_{k=1}^{\infty} \bigcap_{w \in \mathcal{W}_k} \mathcal{E}^2_{k,w} \right)^c \right) \le \sum_{k=1}^{\infty} \sum_{w \in \mathcal{W}} \mathbb{P}\left(\mathcal{E}^2_{k,w}\right) \le \delta/2.$$

Under the good event $\left(\bigcap_{k=1}^{\infty} \bigcap_{\ell=1}^{\infty} \mathcal{E}^1_{k,\ell}\right) \bigcap \left(\bigcap_{k=1}^{\infty} \bigcap_{w \in \mathcal{W}_k} \mathcal{E}^2_{k,w} \right)$, we have with probability at least $1-\delta$, for all k and $w \in \mathcal{W}$,

$$\left| \left\langle \widehat{\theta}^k_{\texttt{P-2SLS}} - \theta, w - w^* \right\rangle \right| \le \zeta_k.$$

The rest of proof is same as the proof of Theorem 16.

Part 2: sample complexity of algorithm

Sample complexity for first sub-phase

Recall that $\lambda_E = \arg\max_{\lambda \in \Delta(\mathcal{Z})} \sigma_{\min}(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)$ is the E-optimal design to maximize the minimum singular value of $\sum_{z \in \mathcal{Z}} \lambda_z z z^\top$ and $\kappa_0 = \max_\lambda \sigma_{\min}(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)$ is the maximum minimum singular value of $\sum_{z \in \mathcal{Z}} \lambda_z z z^\top$. At the beginning of first sub-phase in phase k, the algorithm first samples $N_{k,0,0}$ arms according to λ_E.

Before we proceed to the main proof of the sample complexity, we first address a minor technique issue to avoid cumbersomeness. For the logarithmic term that appears in the algorithm and confidence interval,

$$\overline{\log}(Z_T, \delta) := 8d \ln\left(1 + \frac{2TL_z^2}{d(2 \wedge \sigma_{\min}(Z_T^\top Z_T))}\right) + 16 \ln\left(\frac{2 \cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(Z_T^\top Z_T)}\right)\right),$$

when $2 \wedge \sigma_{\min}(V) = 2$, it is equivalent to

$$\widetilde{\log}(T, \delta) := 8d \ln\left(1 + \frac{TL_z^2}{d}\right) + 16 \ln\left(\frac{2{\cdot}6^d}{\delta}\right).$$

Our algorithm design guarantees that $2 \wedge \sigma_{\min}(V) = 2$ is always true whenever we need to use the logarithmic term $\overline{\log}(Z_T, \delta)$, given that the samples of our interest always includes the E-optimal design samples $Z_{k,0,\ell}$ and the number of samples from the E-optimal design $\left|Z_{k,0,\ell}\right|$ is always larger than $\frac{2}{\kappa_0}$. So for the remaining part of the proof, we will use $\widetilde{\log}(N, \delta)$ instead of $\overline{\log}(Z_T, \delta)$.

Denote the samples of E-optimal design that mixed into the samples from ℓ-th doubling trick iteration in the first sub-phase of phase k as $Z_{k,0,\ell}$ and $\left|Z_{k,0,\ell}\right| = N_{k,0,\ell}$. By Lemma 21.1, our choice of $N_{k,0,\ell}$

$$N_{k,0,\ell} \geq \frac{64gd\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \ln\left(\frac{32g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)}\big(d + N_{k,1,\ell}L_z^2 + L_z^2\big)\right) + \frac{128g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \ln\left(\frac{2{\cdot}6^d}{\delta}\right),$$

is a sufficient condition to guarantee

$$N_{k,0,\ell} \geq \frac{4g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big)}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \vee r(\omega). \tag{3.22}$$

Multiply both sides of (3.22) by $\frac{N_{k,0,\ell}+N_{k,1,\ell}}{N_{k,0,\ell}}$ and we have

$$N_{k,0,\ell} + N_{k,1,\ell} \geq \left(\frac{4g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big)}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \vee r(\omega)\right) \frac{N_{k,0,\ell} + N_{k,1,\ell}}{N_{k,0,\ell}}.$$

By Property 1 of Lemma 20.1, we have $\alpha_{k,\ell} \coloneqq \frac{N_{k,0,\ell}}{N_{k,0,\ell}+N_{k,1,\ell}} < 1/2$, then

$$\begin{aligned} N_{k,0,\ell} + N_{k,1,\ell} &\geq \frac{1}{\alpha_{k,\ell}} \left(\frac{4g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big)}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \vee r(\omega) \right) \\ &= \frac{4g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big)}{\sigma_{\min}\big(A(\alpha_{k,\ell}\lambda_E, \Gamma)\big)} \vee \frac{r(\omega)}{\alpha_{k,\ell}} \\ &\geq \frac{4g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big)}{\sigma_{\min}\Big(A(\alpha_{k,\ell}\lambda_E + (1-\alpha_{k,\ell})\tilde{\lambda}_k, \Gamma)\Big)} \vee r(\omega). \end{aligned} \tag{3.23}$$

These condition on $N_{k,0,\ell}$ and $N_{k,0,\ell} + N_{k,1,\ell}$ are needed for the proof.

Denote the total number of doubling trick iterations as L_k for phase k. In the case of $L_k = 1$, the samples from the first doubling trick iteration satisfies stopping condition of the first sub-phase already, and the algorithm will not enter the second doubling trick iteration. Thus the total number of samples for the first sub-phase is

$$\begin{aligned} N_{k,0,1} + N_{k,1,1} \leq & 2N_{k,1,1} && \text{(Property 1 of Lemma 20.1)} \\ \leq & 8gdM \ln\Big(1 + 2M\big(d + L_z^2\big) + 2M2gdM\Big) + 16M \ln\left(\frac{2{\cdot}6^d}{\delta_{k,1}}\right) \vee r(\omega). \end{aligned}$$

In the case of $L_k > 1$, for $\ell \in \{1, \cdots, L_k\}$, denote $\widehat{\Gamma}^\ell$ as the estimate of Γ at the end of the ℓ-th doubling trick iteration. With these notations, we have

- At the end of L_k-th doubling trick iteration, the stopping condition is satisfied, i.e.,

$$\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{\bar{A}(Z_{k,0,L_k} \cup Z_{k,1,L_k}, \widehat{\Gamma}^{L_k})^{-1}} \|\theta\|_2^2 L_\eta \widetilde{\log}\big(N_{k,0,L_k} + N_{k,1,L_k}, \delta_{k,L_k}\big) \leq \zeta_k^2. \tag{3.24}$$

  We short $\widetilde{\log}\big(N_{k,0,\ell} + N_{k,1,\ell}, \delta_{k,\ell}\big) = \widetilde{\log}_{k,\ell}$.

- At the end of $(L_k - 1)$-th doubling trick iteration, the stopping condition is not satisfied, i.e.,

$$\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{\bar{A}(Z_{k,0,L_k-1} \cup Z_{k,1,L_k-1}, \widehat{\Gamma}^{L_k-1})^{-1}} \|\theta\|_2^2 L_\eta \widetilde{\log}_{k,L_k-1} > \zeta_k^2. \tag{3.25}$$

Denote ξ_{L_k} as the empirical distribution of $Z_{k,0,L_k} \cup Z_{k,1,L_k}$. Then above two conditions imply that the number of samples for L_k-th and $(L_k - 1)$-th doubling trick iterations respectively satisfy

$$N_{k,0,L_k} + N_{k,1,L_k} \geq \frac{\|\theta\|_2^2 L_\eta \widetilde{\log}_{k,L_k}}{\zeta_k^2} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_k}, \widehat{\Gamma}^{L_k})^{-1}}. \tag{3.26}$$

and

$$N_{k,0,L_k-1} + N_{k,1,L_k-1} < \frac{\|\theta\|_2^2 L_\eta \widetilde{\log}_{k,L_k-1}}{\zeta_k^2} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_k-1}, \widehat{\Gamma}^{L_k-1})^{-1}}. \tag{3.27}$$

Note that by Property 2 of Lemma 20.1,

$$\frac{1}{2}\big(N_{k,0,L_k} + N_{k,1,L_k}\big) \leq N_{k,0,L_k-1} + N_{k,1,L_k-1}.$$

Thus

$$N_{k,0,L_k} + N_{k,1,L_k} < \frac{2\|\theta\|_2^2 L_\eta \widetilde{\log}_{k,L_k-1}}{\zeta_k^2} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_k-1}, \widehat{\Gamma}^{L_k-1})^{-1}}. \tag{3.28}$$

For any $\ell \in \{1, \ldots, L_k\}$, by Lemma 21.5, the factor of $\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^\ell)^{-1}}$ can be upper bounded by

$$\begin{aligned}
&\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^\ell)^{-1}} \\
\leq& 3 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^{L_{k-1}})^{-1}} + 2 \max_{w,w' \in \mathcal{W}_k} \Big\|((\widehat{\Gamma}^\ell)^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w')\Big\|^2_{(\sum_z \xi_\ell z z^\top)^{-1}} \\
\leq& 3 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^{L_{k-1}})^{-1}} + 2 \max_{w,w' \in \mathcal{W}_k} \Big\|(\Gamma^{-\top} - (\widehat{\Gamma}^\ell)^{-\top})(w - w')\Big\|^2_{(\sum_z \xi_\ell z z^\top)^{-1}} \\
&+ 2 \max_{w,w' \in \mathcal{W}_k} \Big\|(\Gamma^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w')\Big\|^2_{(\sum_z \xi_\ell z z^\top)^{-1}}.
\end{aligned} \tag{3.29}$$

We will upper bound the three terms in the RHS of (3.29) separately.

For the first term, by Lemma 21.4,

$$\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^{L_{k-1}})^{-1}} \leq 4 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\tilde{\lambda}_k, \widehat{\Gamma}^{L_{k-1}})^{-1}}, \tag{3.30}$$

where $\tilde{\lambda}_k$ is the optimal design for $\mathcal{W}_k$ in the first sub-phase of phase $k-1$ (based on the last doubling trick iteration), i.e.,

$$\tilde{\lambda}_k = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda, \widehat{\Gamma}^{L_{k-1}})^{-1}}.$$

Then for any λ,

$$\begin{aligned}
&\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\tilde{\lambda}_k, \widehat{\Gamma}^{L_{k-1}})^{-1}} \\
\leq & 3 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda,\Gamma)^{-1}} + 2 \max_{w,w' \in \mathcal{W}_k} \left\| (\Gamma^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w') \right\|^2_{A(\lambda, I)^{-1}} && \text{(Lemma 21.5)} \\
\overset{b_1}{\leq} & 3 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda'_k,\Gamma)^{-1}} + 2 \max_{w,w' \in \mathcal{W}_k} \left\| (\Gamma^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w') \right\|^2_{(\sum_z \lambda'_k zz^\top)^{-1}} \\
\overset{b_2}{\leq} & 6 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda^*_k,\Gamma)^{-1}} + 2 \max_{w,w' \in \mathcal{W}_k} \left\| (\Gamma^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w') \right\|^2_{(\sum_z \lambda'_k zz^\top)^{-1}}. && (3.31)
\end{aligned}$$

where for (b_1), we plug in $\lambda'_k := \alpha^*_k \lambda_E + (1 - \alpha^*_k)\lambda^*_k$, with $\alpha^*_k \leq 1/2$ will be defined later and λ^*_k is the optimal design for $\mathcal{W}_k$ given Γ, i.e.,

$$\lambda^*_k = \arg\min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda,\Gamma)^{-1}}.$$

(b_2) is due to the fact that $\alpha^*_k \leq 1/2$. For the second term in the RHS of (3.31), given the condition (3.23), i.e.,

$$N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}} \geq \frac{4gL_\eta \widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\xi_{L_{k-1}}, \Gamma)\big)}$$

by Lemma 21.9 we have,

$$
\begin{aligned}
&\left\|(\Gamma^{-\top}-(\widehat{\Gamma}^{L_{k-1}})^{-\top})(w-w')\right\|^2_{(\sum_z \lambda'_k zz^\top)^{-1}}\\
\leq&\frac{2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\lambda'_k,\Gamma)\big)}\frac{1}{N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}}}\|w-w'\|^2_{A(\xi_{L_{k-1}},\Gamma)^{-1}}\\
\leq&\frac{1}{N_{k-1,0,L_{k-1}}}\frac{2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\lambda'_k,\Gamma)\big)}\frac{N_{k-1,0,L_{k-1}}}{N_{k-1,0,L_{k-1}+N_{k-1,1,L_{k-1}}}}\|w-w'\|^2_{A(\xi_{L_{k-1}},\Gamma)^{-1}}\\
\leq&\frac{1}{N_{k-1,0,L_{k-1}}}\frac{2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\alpha^*_k\lambda_E,\Gamma)\big)}\frac{N_{k-1,0,L_{k-1}}}{N_{k-1,0,L_{k-1}+N_{k-1,1,L_{k-1}}}}\|w-w'\|^2_{A(\xi_{L_{k-1}},\Gamma)^{-1}}\\
\leq&\frac{1}{N_{k-1,0,L_{k-1}}}\frac{2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\|w-w'\|^2_{A(\xi_{L_{k-1}},\Gamma)^{-1}} \qquad (\text{set } \alpha^*_k=\alpha_{k-1})\\
\overset{b_1}{\leq}&\frac{1}{4g}\|w-w'\|^2_{A(\xi_{L_{k-1}},\Gamma)^{-1}}\\
\overset{b_2}{\leq}&\frac{6}{4g}\|w-w'\|^2_{A(\xi_{L_{k-1}},\widehat{\Gamma}^{L_{k-1}})^{-1}}. \qquad (3.32)
\end{aligned}
$$

where for (b_1), we use the condition (3.22) on $N_{k-1,0,L_{k-1}}$, and (b_2) is due to Lemma 21.3. Plug (3.32) into (3.31), we have

$$
\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\tilde{\lambda}_k,\widehat{\Gamma}^{L_{k-1}})^{-1}}\leq 6\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda^*_k,\Gamma)^{-1}}+\frac{3}{g}\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_{L_{k-1}},\widehat{\Gamma}^{L_{k-1}})^{-1}}. \qquad (3.33)
$$

Plug (3.33) into (3.30), we have the first term in the RHS of (3.29) can be upper bounded by

$$
\begin{aligned}
&\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_\ell,\widehat{\Gamma}^{L_{k-1}})^{-1}}\\
\leq&\,24\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda^*_k,\Gamma)^{-1}}+\frac{12}{g}\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_{L_{k-1}},\widehat{\Gamma}^{L_{k-1}})^{-1}}\\
\leq&\,24\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda^*_k,\Gamma)^{-1}}+\frac{12}{g}\frac{\zeta^2_{k-1}}{\|\theta\|_2^2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}(N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}}), \qquad (3.34)
\end{aligned}
$$

where for the last inequality we use the fact that the stopping condition is satisfied at the end of (L_{k-1})-th doubling trick iteration for phase $k-1$ per (3.26).

For the second term in the RHS of (3.29), by Lemma 21.8, when the condition (3.23) is satisfied, i.e., when

$$N_{k,0,\ell} + N_{k,1,\ell} \geq \frac{4gL_\eta \widetilde{\log}_{k,\ell}}{\sigma_{\min}\big(A(\xi_\ell, \Gamma)\big)},$$

we have

$$\begin{aligned}\max_{w,w' \in \mathcal{W}_k} \Big\| (\Gamma^{-\top} - (\widehat{\Gamma}^{\ell})^{-\top})(w - w') \Big\|^2_{(\sum_z \xi_\ell z z^\top)^{-1}} \leq & \frac{1}{g} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \Gamma)^{-1}} \\ \leq & \frac{6}{g} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_\ell, \widehat{\Gamma}^{\ell})^{-1}}, \end{aligned} \tag{3.35}$$

where the last inequality is due to Lemma 21.3.

For the third term in the RHS of (3.29), by Lemma 21.9, when the condition (3.23) is satisfied, i.e., when

$$N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}} \geq \frac{4gL_\eta \widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\xi_{L_{k-1}}, \Gamma)\big)}$$

we have

$$\begin{aligned} & \max_{w,w' \in \mathcal{W}_k} \Big\| (\Gamma^{-\top} - (\widehat{\Gamma}^{L_{k-1}})^{-\top})(w - w') \Big\|^2_{(\sum_z \xi_\ell z z^\top)^{-1}} \\ \leq & \frac{2L_\eta \widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\xi_\ell, \Gamma)\big)} \frac{1}{N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}}} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_{k-1}}, \Gamma)^{-1}} \\ \leq & \frac{2L_\eta \widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\alpha_\ell \lambda_E, \Gamma)\big)} \frac{1}{N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}}} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_{k-1}}, \widehat{\Gamma}^{L_{k-1}})^{-1}} \\ \overset{b_1}{\leq} & \frac{N_{k,0,\ell} + N_{k,1,\ell}}{N_{k,0,\ell}} \frac{1}{N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}}} \frac{2\widetilde{\log}_{k-1,L_{k-1}}}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} (N_{k-1,0,L_{k-1}} + N_{k-1,1,L_{k-1}}) \frac{\zeta_{k-1}^2}{\|\theta\|_2^2 \widetilde{\log}_{k-1,L_{k-1}}} \\ \leq & \frac{N_{k,0,\ell} + N_{k,1,\ell}}{N_{k,0,\ell}} \frac{2}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \frac{\zeta_{k-1}^2}{\|\theta\|_2^2} \\ \overset{b_2}{\leq} & \frac{N_{k,0,\ell} + N_{k,1,\ell}}{4gL_\eta \widetilde{\log}_{k,\ell}} \frac{\zeta_{k-1}^2}{\|\theta\|_2^2} \end{aligned} \tag{3.36}$$

where (b_1) is due to (3.26), (b_2) is due to (3.22). Plug (3.34), (3.35) and (3.36) into (3.29), we have

$$\begin{aligned}
&\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_\ell,\widehat{\Gamma}^\ell)^{-1}}\\
\leq&72\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}}+\frac{36}{g}\frac{\zeta_{k-1}^2}{\|\theta\|_2^2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}(N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}})\\
&+\frac{12}{g}\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_\ell,\widehat{\Gamma}^\ell)^{-1}}+\frac{2}{g}\frac{N_{k,0,\ell+1}+N_{k,1,\ell+1}}{L_\eta\widetilde{\log}_{k,\ell}}\frac{\zeta_{k-1}^2}{\|\theta\|_2^2}.
\end{aligned}\tag{3.37}$$

With $\ell=L_k-1$ and the fact that $g>24$ whose exact value will be set later, we can rearrange (3.37) as

$$\begin{aligned}
&\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_{L_{k^{-1}}},\widehat{\Gamma}^{L_{k^{-1}}})^{-1}}\\
\leq&144\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}}+\frac{72}{g}\frac{\zeta_{k-1}^2}{\|\theta\|_2^2L_\eta\widetilde{\log}_{k-1,L_{k-1}}}(N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}})\\
&+\frac{1}{g}\frac{N_{k,0,L_k}+N_{k,1,L_k}}{L_\eta\widetilde{\log}_{k,L_k-1}}\frac{\zeta_{k-1}^2}{\|\theta\|_2^2}.
\end{aligned}\tag{3.38}$$

Note that the LHS of above can be lower bounded by (3.28),

$$\frac{\zeta_k^2}{2\|\theta\|_2^2L_\eta\widetilde{\log}_{k,L_k-1}}(N_{k,0,L_k}+N_{k,1,L_k})<\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\xi_{L_{k^{-1}}},\widehat{\Gamma}^{L_{k^{-1}}})^{-1}}.\tag{3.39}$$

Rearrange the terms in (3.39) and (3.38) and setting g to be larger enough and using the fact that $\zeta_k=\zeta_{k-1}/2$, we have

$$N_{k,0,L_k}+N_{k,1,L_k}\leq\frac{576}{\zeta_k^2}\|\theta\|_2^2L_\eta\widetilde{\log}_{k,L_k-1}\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}}+\frac{72}{g}\frac{\widetilde{\log}_{k,L_k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}(N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}}).$$

Note that by definition $\widetilde{\log}_{k,L_k-1}<\widetilde{\log}_{k,L_k}$, thus

$$N_{k,0,L_k}+N_{k,1,L_k}\leq\frac{576}{\zeta_k^2}\|\theta\|_2^2L_\eta\widetilde{\log}_{k,L_k}\max_{w,w'\in\mathcal{W}_k}\|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}}+\frac{72}{g}\frac{\widetilde{\log}_{k,L_k}}{\widetilde{\log}_{k-1,L_{k-1}}}(N_{k-1,0,L_{k-1}}+N_{k-1,1,L_{k-1}}).$$

Denote $D_k \coloneqq N_{k,0,L_k} + N_{k,1,L_k}$ and divide both sides of above by $\widetilde{\log}_{k,L_k}$, we have

$$\frac{D_k}{\widetilde{\log}_{k,L_k}} \leq \frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + \frac{72}{g}\frac{D_{k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}.$$

In the case of $L_k = 1$, by Property 3 of Lemma 20.1, we have

$$\frac{D_k}{\widetilde{\log}_{k,L_k}} = \frac{N_{k,1,1}}{8d\ln\left(1 + \frac{N_{k,1,1}L_z^2}{d}\right) + 16\ln\left(\frac{2\cdot 6^d k^2}{\delta}\right)} \leq M\ln(dM).$$

Thereby we have

$$\frac{D_k}{\widetilde{\log}_{k,L_k}} \leq \max\left\{M\ln(dM), \frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + \frac{72}{g}\frac{D_{k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}\right\}. \tag{3.40}$$

Taking a summation over k on both sides of (3.40), we have

$$\sum_{k=1}^{K^*}\frac{D_k}{\widetilde{\log}_{k,L_k}} = \frac{D_1}{\widetilde{\log}_{1,L_1}} + \sum_{k=2}^{K^*}\frac{D_k}{\widetilde{\log}_{k,L_k}} \tag{3.41}$$

$$\leq \frac{D_1}{\widetilde{\log}_{1,L_1}} + \sum_{k=2}^{K^*}\max\left\{M\ln(dM), \frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + \frac{72}{g}\frac{D_{k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}\right\} \tag{3.42}$$

$$\leq \frac{D_1}{\widetilde{\log}_{1,L_1}} + \sum_{k=2}^{K^*}M\ln(dM) + \sum_{k=2}^{K^*}\left(\frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + \frac{72}{g}\frac{D_{k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}\right) \tag{3.43}$$

$$\leq \frac{D_1}{\widetilde{\log}_{1,L_1}} + \sum_{k=2}^{K^*}M\ln(dM) + \sum_{k=2}^{K^*}\frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + \sum_{k=2}^{K^*}\frac{72}{g}\frac{D_{k-1}}{\widetilde{\log}_{k-1,L_{k-1}}}. \tag{3.44}$$

Thus by setting $g = 72 \times 2$ and rearranging the terms, we have

$$\sum_{k=1}^{K^*}\frac{D_k}{\widetilde{\log}_{k,L_k}} \leq \frac{2D_1}{\widetilde{\log}_{1,L_1}} + 2\sum_{k=2}^{K^*}M\ln(dM) + 2\sum_{k=2}^{K^*}\frac{576}{\zeta_k^2}\|\theta\|_2^2 L_\eta \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}}. \tag{3.45}$$

For D_1, which corresponds to the first sub-phase where we use E-optimal design with doubling trick, we have the stopping condition,

$$\max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{\bar{A}(Z_{1,0,L_1}, \widehat{\Gamma}^{L_1})^{-1}} \|\theta\|_2^2 L_\eta \widetilde{\log}_{1,L_1} \leq \zeta_1^2.$$

This implies that when the stopping condition is met

$$N_{1,0,L_1} \geq \frac{\|\theta\|_2^2 L_\eta \widetilde{\log}_{1,L_1}}{\zeta_1^2} \max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{A(\xi_{L_1}, \widehat{\Gamma}^{L_1})^{-1}}. \tag{3.46}$$

and

$$N_{1,0,L_1-1} < \frac{2\|\theta\|_2^2 L_\eta \widetilde{\log}_{1,L_1-1}}{\zeta_1^2} \max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{A(\xi_{L_1-1}, \widehat{\Gamma}^{L_1-1})^{-1}}. \tag{3.47}$$

Since we use E-optimal design for the first phase, the factor of $\|w - w'\|^2_{A(\xi_{L_1-1}, \widehat{\Gamma}^{L_1-1})^{-1}}$ can be upper bounded by

$$\begin{aligned}
&\|w - w'\|^2_{A(\xi_{L_1-1}, \widehat{\Gamma}^{L_1-1})^{-1}} \\
\leq & 3\|w - w'\|^2_{A(\xi_{L_1-1}, \Gamma)^{-1}} + 2\left\|(\Gamma^{-\top} - (\widehat{\Gamma}^{L_1-1})^{-\top})(w - w')\right\|^2_{\left(\sum_z \xi_{L_1-1} z z^\top\right)^{-1}} && \text{(Lemma 21.5)} \\
\leq & 3\|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}} + 2\|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}} && \text{(Lemma 21.8)} \\
\leq & 6\|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}}. && (3.48)
\end{aligned}$$

Plug (3.48) into (3.47), and use the fact that $2N_{1,0,L_1-1} = N_{1,0,L_1}, \zeta_1 = 1$, we have

$$N_{1,0,L_1} \leq 24\|\theta\|_2^2 L_\eta \widetilde{\log}_{1,L_1-1} \max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}} \leq 24\|\theta\|_2^2 L_\eta \widetilde{\log}_{1,L_1} \max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}}. \tag{3.49}$$

Thus we have,

$$\frac{D_1}{\widetilde{\log}_{1,L_1}} \leq \frac{2N_{1,0,L_1}}{\widetilde{\log}_{1,L_1}} \leq 48\|\theta\|_2^2 L_\eta \max_{w,w' \epsilon \mathcal{W}_1} \|w - w'\|^2_{A(\lambda_E, \Gamma)^{-1}} =: \rho_1.$$

By the same calculation as (3.17) and (3.19), we have

$$\|\theta\|_2^2 L_\eta \sum_{k=2}^{K^*} \frac{1}{\zeta_k^2} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} \le c\|\theta\|_2^2 L_\eta K^* \rho^* =: \rho_2,$$

where c is an absolute constant. Next we lower bound the left hand side of (3.45). To do this, we first upper bound $\widetilde{\log}_{k,L_k}$ as

$$\begin{aligned}
\widetilde{\log}_{k,L_k} =& 8d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d k^2 L_k^2}{\delta}\right) \\
\le& 8d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 32 \ln(L_k) + 16 \ln\left(\frac{2 \cdot 6^d k^2}{\delta}\right) \\
\le& 8d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 32 \ln\left(\log_2\left(\frac{D_k}{d}\right)\right) + 16 \ln\left(\frac{2 \cdot 6^d k^2}{\delta}\right) \\
\le& 32d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d k^2}{\delta}\right),
\end{aligned} \tag{3.50}$$

where the inequality above uses the fact that L_k is the index of the last doubling trick iteration for phase k, by the design of the doubling trick, we have $L_k \le \log_2\left(\frac{D_k}{d}\right)$.

$$\begin{aligned}
\sum_{k=1}^{K^*} \frac{D_k}{\widetilde{\log}_{k,L_k}} =& \sum_{k=1}^{K^*} \frac{D_k}{8d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d k^2 L_k^2}{\delta}\right)} \\
\ge& \sum_{k=1}^{K^*} \frac{D_k}{32d \ln\left(1 + \frac{D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d K^{*2}}{\delta}\right)} && \text{(due to (3.50))} \\
\ge& \sum_{k=1}^{K^*} \frac{D_k}{32d \ln\left(1 + \frac{\sum_{k=1}^{K^*} D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d K^{*2}}{\delta}\right)} \\
=& \frac{1}{32d \ln\left(1 + \frac{\sum_{k=1}^{K^*} D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d K^{*2}}{\delta}\right)} \sum_{k=1}^{K^*} D_k.
\end{aligned}$$

Denote $\rho_3 := K^* M \ln(dM)$, looking back at (3.44), we have

$$\sum_{k=1}^{K^*} D_k \le \left(32d \ln\left(1 + \frac{\sum_{k=1}^{K^*} D_k L_z^2}{d}\right) + 16 \ln\left(\frac{2 \cdot 6^d K^{*2}}{\delta}\right)\right)(\rho_1 + \rho_2 + \rho_3).$$

By Lemma 23.2, we have

$$\sum_{k=1}^{K^*} D_k \le 32(\rho_1+\rho_2+\rho_3)\ln\left(\frac{2\cdot 6^d K^{*2}}{\delta}\right) + 64d(\rho_1+\rho_2+\rho_3)\ln\left(3+(\rho_1+\rho_2+\rho_3)\frac{2L_z^2}{d}\right). \quad (3.51)$$

Sample complexity for second sub-phase

The design for the second sub-phase of phase k is based on $\widehat{\Gamma}^{L_k}$, the estimate of Γ at the end of the L_k-th doubling trick iteration in the first sub-phase of phase k,

$$\hat{\lambda}_k = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\lambda,\widehat{\Gamma}^{L_k})^{-1}}.$$

Then for any λ, by Lemma 21.5 we have

$$\begin{aligned}
&\max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\hat{\lambda}_k,\widehat{\Gamma}^{L_k})^{-1}} \\
\le& 3\max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\lambda,\Gamma)^{-1}} + 2\max_{w,w'\in\mathcal{W}_k} \left\|(\Gamma^{-\top}-(\widehat{\Gamma}^{L_k})^{-\top})(w-w')\right\|^2_{A(\lambda,I)^{-1}}
\end{aligned}$$

We plug in $\lambda_k'' := \alpha_k^*\lambda_E + (1-\alpha_k^*)\lambda_k^*$, with $\alpha_k^* \le 1/2$ will be defined later and λ_k^* is the optimal design for $\mathcal{W}_k$,

$$\begin{aligned}
&\max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\hat{\lambda}_k,\widehat{\Gamma}^{L_k})^{-1}} \\
\le& 3\max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\lambda_k'',\Gamma)^{-1}} + 2\max_{w,w'\in\mathcal{W}_k} \left\|(\Gamma^{-\top}-(\widehat{\Gamma}^{L_k})^{-\top})(w-w')\right\|^2_{(\sum_z \lambda_k'' zz^\top)^{-1}} \\
\le& 6\max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} + 2\max_{w,w'\in\mathcal{W}_k} \left\|(\Gamma^{-\top}-(\widehat{\Gamma}^{L_k})^{-\top})(w-w')\right\|^2_{(\sum_z \lambda_k'' zz^\top)^{-1}}, \quad (3.52)
\end{aligned}$$

where the last inequality is due to the fact that $\alpha_k^* \le 1/2$. For the second term in the RHS of above, by Lemma 21.9 we have,

$$
\begin{aligned}
&\left\|(\Gamma^{-\top} - (\widehat{\Gamma}^{L_k})^{-\top})(w - w')\right\|^2_{(\sum_z \lambda_k'' zz^\top)^{-1}} \\
\le& \frac{4L_\eta \widetilde{\log}_{k,L_k}}{\sigma_{\min}\big(A(\lambda_k'', \Gamma)\big)} \|w - w'\|^2_{\bar{A}(Z_{k,0} \cup Z^{L_k}, \Gamma)^{-1}} \\
\le& \frac{1}{N_{k,0,L_k}} \frac{4L_\eta \widetilde{\log}_k}{\sigma_{\min}\big(A(\lambda_k'', \Gamma)\big)} \frac{N_{k,0,L_k}}{N_{k,0,L_k} + N_{k,1L_k}} \|w - w'\|^2_{A(\xi_{L_k}, \Gamma)^{-1}} \\
\le& \frac{1}{N_{k,0,L_k}} \frac{4L_\eta \widetilde{\log}_{k,L_k}}{\sigma_{\min}\big(A(\alpha_k^* \lambda_E, \Gamma)\big)} \frac{N_{k,0,L_k}}{N_{k,0,L_k} + N_{k,1,L_k}} \|w - w'\|^2_{A(\xi_{L_k}, \Gamma)^{-1}} \\
\le& \frac{1}{N_{k,0,L_k}} \frac{4L_\eta \widetilde{\log}_{k,L_k}}{\sigma_{\min}\big(A(\lambda_E, \Gamma)\big)} \|w - w'\|^2_{A(\xi_{L_k}, \Gamma)^{-1}} && (\text{set } \alpha_k^* = \alpha_k) \\
\overset{b_1}{\le}& \frac{1}{g} \|w - w'\|^2_{A(\xi_{L_k}, \Gamma)^{-1}} \\
\overset{b_2}{\le}& \frac{1}{g} \|w - w'\|^2_{A(\xi_{L_k}, \widehat{\Gamma}^{L_k})^{-1}}, && (3.53)
\end{aligned}
$$

where for (b_1), we have $N_{k,0,L_k} \ge \frac{4gL_\eta \widetilde{\log}_{k,L_k}}{\sigma_{\min}(A(\lambda_E, \Gamma))}$, and (b_2) is due to Lemma 21.3. Plug (3.53) into (3.52), we have

$$
\begin{aligned}
\max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\hat{\lambda}_k, \widehat{\Gamma}^{L_k})^{-1}} \le& 6 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*, \Gamma)^{-1}} + \frac{2}{g} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\xi_{L_k}, \widehat{\Gamma}^{L_k})^{-1}} \\
\le& 6 \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*, \Gamma)^{-1}} + \frac{2}{g} \frac{\zeta_k^2}{\|\theta\|_2^2 L_\eta \widetilde{\log}_{k,L_k}} (N_{k,0,L_k} + N_{k,1,L_k}), \qquad (3.54)
\end{aligned}
$$

where the last inequality is due to (3.26). According to the algorithm design, the number of samples in the second sub-phase of phase k is defined as

$$
N_{k,2} = \left\lceil 2(1+\omega)\zeta_k^{-2} \rho(\mathcal{W}_k) L_\nu \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) \right\rceil \vee r(\omega),
$$

with $\rho(\mathcal{W}_k) = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda, \widehat{\Gamma}^{L_k})^{-1}}$. Then we have, by setting $g \ge 4$,

$$
N_{k,2} \lessapprox (1+\omega)\zeta_k^{-2} L_\nu \max_{w,w' \in \mathcal{W}_k} \|w - w'\|^2_{A(\lambda_k^*, \Gamma)^{-1}} \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) + (1+\omega)(N_{k,0,L_k} + N_{k,1,L_k}),
$$

where '$\lessapprox$' hides logarithmic factors of $|\mathcal{W}|$ for the second term and constants for simplicity. Plug (3.54) into the above inequality. Also note that by Lemma 14.1, we can always set $L_\nu = 2(\|\theta\|_2^2 L_\eta + 1)$.

Thus,

$$\begin{aligned}\sum_{k=1}^{K^*} N_{k,2} &\lessapprox (1+\omega)\sum_{k=1}^{K^*} \zeta_k^{-2} L_\nu \max_{w,w'\in\mathcal{W}_k} \|w-w'\|^2_{A(\lambda_k^*,\Gamma)^{-1}} \log\left(\frac{4k^2|\mathcal{W}|}{\delta}\right) + (1+\omega)\sum_{k=1}^{K^*}(N_{k,0,L_k}+N_{k,1,L_k}) \\ &\lessapprox (1+\omega)K^* L_\nu \rho^* \log\left(\frac{4K^{*2}|\mathcal{W}|}{\delta}\right) + (1+\omega)\sum_{k=1}^{K^*}(N_{k,0,L_k}+N_{k,1,L_k}). \end{aligned} \tag{3.55}$$

This essentially means that the sample complexity for the second sub-phase $\sum N_{k,2}$ can be upper bounded by summation of the sample complexity we pay for Algorithm 4 and the sample complexity of the first sub-phase $\sum_{k=1}^{K^*}(N_{k,0,L_k}+N_{k,1,L_k})$. Combine (3.51) and (3.55), we conclude the result.

□

Lemma 21.1. *Denote*

$$\widetilde{\log}\big(N_{k,0},\delta_{k,0}\big) = 8d\ln\left(1+\frac{N_{k,0}L_z^2}{d}\right) + 16\ln\left(\frac{2{\cdot}6^d}{\delta}\right).$$

A sufficient condition for

$$N_{k,0,\ell} \geq \frac{g\sigma_\eta^2 \widetilde{\log}\big(N_{k,0,\ell}+N_{k,1,\ell},\delta_{k,\ell}\big)}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)} \vee r(\omega). \tag{3.56}$$

to hold is

$$N_{k,0,\ell} \geq \frac{16gd\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\ln\left(\frac{8g}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\big(d+N_{k,1,\ell}L_z^2+L_z^2\big)\right) + \frac{32g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\ln\left(\frac{2{\cdot}6^d}{\delta}\right).$$

Proof. Given

$$\widetilde{\log}\big(N_{k,0,\ell}+N_{k,1,\ell},\delta_{k,\ell}\big) = 8d\ln\left(1+\frac{(N_{k,0,\ell}+N_{k,1,\ell})L_z^2}{d}\right) + 16\ln\left(\frac{2{\cdot}6^d}{\delta}\right).$$

By Lemma 21.2, for the formula

$$X \geq A \ln (D + BX) + C$$

we have

- $A = \frac{8gd\sigma_\eta^2}{\sigma_{\min}(A(\lambda_E,\Gamma))}$,
- $B = \frac{L_z^2}{d}$,
- $C = \frac{16g\sigma_\eta^2}{\sigma_{\min}(A(\lambda_E,\Gamma))} \ln\left(\frac{2\cdot 6^d}{\delta}\right)$,
- $D = 1 + \frac{N_{k,1,\ell}L_z^2}{d}$.

Thus a sufficient condition for the inequality to hold is

$$\begin{aligned} N_{k,0,\ell} \geq & \frac{16gd\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)} \ln\left(\frac{8gd\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\left(1 + \frac{N_{k,1,\ell}L_z^2}{d} + \frac{L_z^2}{d}\right)\right) + \frac{32g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)} \ln\left(\frac{2\cdot 6^d}{\delta}\right) \\ = & \frac{16gd\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)} \ln\left(\frac{8g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)}\big(d + N_{k,1,\ell}L_z^2 + L_z^2\big)\right) + \frac{32g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda_E,\Gamma)\big)} \ln\left(\frac{2\cdot 6^d}{\delta}\right). \end{aligned}$$

□

Lemma 21.2. *Let $X \geq 1, A, B \geq 0$, then a sufficient condition for $X \geq A \ln (D + BX) + C$ is*

$$X \geq 2A \ln(AD + AB) + 2C.$$

Proof. The proof is motivated by Gales et al. (2022). Let $f \in (0,1)$, then

$$
\begin{aligned}
& X \geq A \ln (D + BX) + C \\
\Leftrightarrow & X \geq A \ln\left(\frac{fX}{A}\right) + A \ln\left(\frac{A}{f}\left(\frac{D}{X} + B\right)\right) + C \\
\Leftarrow & X \geq A\left(\frac{fX}{A} - 1\right) + A \ln\left(\frac{A}{f}\left(\frac{D}{X} + B\right)\right) + C \qquad \text{(since } \ln(x) \leq x - 1\text{)} \\
\Leftarrow & X(1-f) \geq A \ln\left(\frac{A}{2f}\left(\frac{D}{X} + B\right)\right) + C \\
\Leftrightarrow & X \geq \frac{1}{1-f} A \ln\left(\frac{A}{2f}\left(\frac{D}{X} + B\right)\right) + \frac{c}{1-f}.
\end{aligned}
$$

Set $f = 1/2$ and by the fact $X \geq 1$, we have

$$X \geq 2A \ln(AD + AB) + 2C.$$

□

Lemma 21.3. *Suppose that we have a data set $\{Z_T, X_T\}$. Denote the empirical distribution of Z_T as ξ. The number of samples satisfies*

$$T \geq \frac{8\sigma_\eta^2}{\sigma_{\min}\big(A(\xi,\Gamma)\big)} \overline{\log}(Z_T, \delta) \vee r(\omega).$$

$\widehat{\Gamma}$ is the OLS estimate of Γ based on $\{Z_T, X_T\}$. Then

$$\|w\|^2_{A(\xi,\Gamma)^{-1}} \leq 6\|w\|^2_{A(\xi,\widehat{\Gamma})^{-1}}.$$

Proof. By Lemma 21.5, we have

$$
\begin{aligned}
\|w\|^2_{A(\xi,\Gamma)^{-1}} \leq & 3\|w\|^2_{A(\xi,\widehat{\Gamma})^{-1}} + 2\left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})w\right\|^2_{A(\xi,I)^{-1}} \\
\leq & 3\|w\|^2_{A(\xi,\widehat{\Gamma})^{-1}} + \frac{1}{2}\|w\|^2_{A(\xi,\Gamma)^{-1}},
\end{aligned}
$$

where the last inequality is due to Lemma 21.8 with $g = 2$. Rearranging the terms, we have

$$\|w\|^2_{A(\xi,\Gamma)^{-1}} \leq 6\|w\|^2_{A(\xi,\widehat{\Gamma})^{-1}}.$$

□

Lemma 21.4. *Suppose that we use ROUND to sample N_0 arms according to λ_E denoted as Z_0, and N_1 arms according to λ_1 denoted as Z_1, with $N_1 \geq N_0$. Denote the empirical distribution of all the collected samples as ξ, then*

$$\|w\|^2_{A(\xi,\Gamma)} \leq 4\|w\|^2_{A(\lambda_1,\Gamma)^{-1}}.$$

Proof. Denote the empirical distribution of Z_0 as ξ_0, and the empirical distribution of Z_1 as ξ_1, then we have

$$\begin{aligned}
\|w\|^2_{\bar{A}(Z_0\cup Z_1,\Gamma)^{-1}} =& w^\top \left(\sum_{z\in Z_0\cup Z_1} \Gamma^\top z z^\top \Gamma\right)^{-1} w \\
=& w^\top \left(\sum_{z\in Z_0} \Gamma^\top z z^\top \Gamma + \sum_{z\in Z_1} \Gamma^\top z z^\top \Gamma\right)^{-1} w \\
=& w^\top \left(N_0 \sum_{z\in\mathcal{Z}} \xi_{z,0}\Gamma^\top z z^\top \Gamma + N_1 \sum_{z\in\mathcal{Z}} \xi_{z,1}\Gamma^\top z z^\top \Gamma\right)^{-1} w \\
=& \frac{1}{N_0+N_1} w^\top \left(\frac{N_0}{N_0+N_1} \sum_{z\in\mathcal{Z}} \xi_{z,0}\Gamma^\top z z^\top \Gamma + \frac{N_1}{N_0+N_1} \sum_{z\in\mathcal{Z}} \xi_{z,1}\Gamma^\top z z^\top \Gamma\right)^{-1} w \\
\leq& \frac{1}{N_0+N_1} w^\top \left(\frac{N_1}{N_0+N_1} \sum_{z\in\mathcal{Z}} \xi_{z,1}\Gamma^\top z z^\top \Gamma\right)^{-1} w \\
\leq& \frac{2}{N_0+N_1} w^\top \left(\sum_{z\in\mathcal{Z}} \xi_{z,1}\Gamma^\top z z^\top \Gamma\right)^{-1} w \qquad (N_1 \geq N_0) \\
\leq& \frac{4}{N_0+N_1} w^\top \left(\sum_{z\in\mathcal{Z}} \lambda_1\Gamma^\top z z^\top \Gamma\right)^{-1} w \\
=& \frac{4}{N_0+N_1} \|w\|^2_{A(\lambda_1,\Gamma)^{-1}}.
\end{aligned}$$

The result follows by noting that

$$\|w\|^2_{\bar{A}(Z_0\cup Z_1,\Gamma)^{-1}} = \frac{1}{N_0+N_1}\|w\|^2_{A(\xi,\Gamma)}.$$

□

Lemma 21.5. *Suppose that we have an estimate $\widehat{\Gamma}$ that is invertible, for any $x \in \mathbb{R}^d$ and covariance matrix V, we have*

$$\|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} \le 3\|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + 2\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}}. \tag{3.57}$$

Proof. Suppose we have an estimate $\widehat{\Gamma}$. Then,

$$\|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} = \|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + \|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} - \|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}}.$$

Note that

$$\begin{aligned}(\Gamma^\top V\Gamma)^{-1} - (\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1} &= \Gamma^{-1}V^{-1}\Gamma^{-\top} - \widehat{\Gamma}^{-1}V^{-1}\widehat{\Gamma}^{-\top}\\ &= \Gamma^{-1}V^{-1}\Gamma^{-\top} - \widehat{\Gamma}^{-1}V^{-1}\widehat{\Gamma}^{-\top} + \Gamma^{-1}V^{-1}\widehat{\Gamma}^{-\top} - \Gamma^{-1}V^{-1}\widehat{\Gamma}^{-\top}\\ &= \Gamma^{-1}V^{-1}(\Gamma^{-\top}-\widehat{\Gamma}^{-\top}) + (\Gamma^{-1}-\widehat{\Gamma}^{-1})V^{-1}\widehat{\Gamma}^{-\top}.\end{aligned}$$

Thus,

$$\begin{aligned}\|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} &= \|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + x^\top\Gamma^{-1}V^{-1}(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x + x^\top(\Gamma^{-1}-\widehat{\Gamma}^{-1})V^{-1}\widehat{\Gamma}^{-\top}x\\ &\le \|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + \|x\|_{(\Gamma^\top V\Gamma)^{-1}}\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|_{V^{-1}} + \left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|_{V^{-1}}\|x\|_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}}\\ &\le \|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + \frac{1}{2}\|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} + \frac{1}{2}\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}} + \frac{1}{2}\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}} + \frac{1}{2}\|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}}.\end{aligned}$$

(AM-GM)

This implies that

$$\|x\|^2_{(\Gamma^\top V\Gamma)^{-1}} \le 3\|x\|^2_{(\widehat{\Gamma}^\top V\widehat{\Gamma})^{-1}} + 2\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}}.$$

□

Lemma 21.6. *Suppose* $\lambda_f = \arg\min f(\lambda)$ *and* $\lambda_g = \arg\min g(\lambda)$ *and* $f(\lambda) \le g(\lambda) + h(\lambda)$, *then*

$$f(\lambda_f) \le g(\lambda_g) + h(\lambda_g).$$

Proof.

$$f(\lambda_f) \le \min_{\lambda} \big(g(\lambda) + h(\lambda)\big) \le g(\lambda_g) + h(\lambda_g).$$

□

Lemma 21.7. *Define* λ_z^* *and* $\tilde{\lambda}_z$

$$\lambda_z^* := \arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}} \|w - w'\|^2_{A(\lambda_z,\Gamma)^{-1}},$$

and

$$\tilde{\lambda}_z := \arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{w,w' \in \mathcal{W}} \|w - w'\|^2_{A(\lambda_z,\widehat{\Gamma})^{-1}}$$

as the optimal design regarding Γ *and that regarding its estimate* $\widehat{\Gamma}$ *respectively. Then, we have*

$$\max_{w,w' \in \mathcal{W}} \|w - w'\|^2_{A(\tilde{\lambda}_z,\widehat{\Gamma})^{-1}} \le \max_{w,w' \in \mathcal{W}} 3\|w - w'\|^2_{A(\lambda_z^*,\Gamma)^{-1}} + \max_{w,w' \in \mathcal{W}} 2\left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})(w - w')\right\|^2_{(\sum_z \lambda_z^* z z^\top)^{-1}}.$$

Proof. By Lemma 21.5, for any $w, w' \in \mathcal{W}$ and $\lambda_z \in \Delta_{\mathcal{Z}}$,

$$\|w - w'\|^2_{A(\lambda_z,\widehat{\Gamma})^{-1}} \le 3\|w - w'\|^2_{A(\lambda_z,\Gamma)^{-1}} + 2\left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top})(w - w')\right\|^2_{(\sum_z \lambda_z z z^\top)^{-1}}.$$

Thus

$$\max_{w,w' \in \mathcal{W}} \|w - w'\|^2_{A(\lambda_z,\widehat{\Gamma})^{-1}} \le \max_{w,w' \in \mathcal{W}} 3\|w - w'\|^2_{A(\lambda_z,\Gamma)^{-1}} + \max_{w,w' \in \mathcal{W}} 2\left\|(\Gamma^{-\top} - \widehat{\Gamma}_k^{-\top})(w - w')\right\|^2_{(\sum_z \lambda_z z z^\top)^{-1}}.$$

By Lemma 21.6,

$$\max_{w,w'\in\mathcal{W}}\|w-w'\|^2_{A(\tilde{\lambda}_z,\widehat{\Gamma})^{-1}} \le \max_{w,w'\in\mathcal{W}} 3\|w-w'\|^2_{A(\lambda^*_z,\Gamma)^{-1}} + \max_{w,w'\in\mathcal{W}} 2\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})(w-w')\right\|^2_{\left(\sum_z \lambda^*_z zz^\top\right)^{-1}}.$$

□

Lemma 21.8. *Suppose that we have $\widehat{\Gamma}$ that is an OLS estimate from an offline dataset $\{Z_{T_1}, X_{T_1}\}$ collected non-adaptively through a fixed design λ and the efficient rounding procedure ROUND. Let $V = Z_{T_1}^\top Z_{T_1}$. Then, for any $x \in \mathbb{R}^d$ and $g \ge 1$, we have, with probability $1-\delta$,*

$$\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}} \le \frac{1}{g}\|x\|^2_{\bar{A}(Z_{T_1},\Gamma)^{-1}},$$

when

$$T_1 \ge \frac{4g\sigma_\eta^2}{\sigma_{\min}\big(A(\lambda,\Gamma)\big)}\overline{\log}(Z_{T_1},\delta)\vee 2p, \tag{3.58}$$

where p is the cardinality of support of λ.

Proof. We first show that

$$\begin{aligned}
&\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|^2_{V^{-1}} \\
=&\left\|\Gamma^{-\top}S^\top Z_{T_1}V^{-1}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|^2_{V^{-1}} && \text{(Lemma 21.10)}\\
\overset{a_1}{=}&\left\|V^{-\frac12}\Gamma^{-\top}S^\top Z_{T_1}V^{-1}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|^2_2 && (\|Ax\|^2_V = \|V^{\frac12}Ax\|^2_2)\\
\le&\left\|V^{-\frac12}\Gamma^{-\top}S^\top Z_{T_1}V^{-\frac12}\right\|^2_{\mathrm{op}}\left\|V^{-\frac12}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|^2_2 && (\|Ax\|\le\|A\|_{\mathrm{op}}\|x\|)\\
\le&\left\|V^{-\frac12}\Gamma^{-\top}\right\|^2_{\mathrm{op}}\left\|S^\top Z_{T_1}V^{-\frac12}\right\|^2_{\mathrm{op}}\left\|V^{-\frac12}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|^2_2 && (\|AB\|_{\mathrm{op}}\le\|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}})\\
=&\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|^2_{\mathrm{op}}\left\|V^{-1/2}\big(\Gamma+V^{-1}Z_{T_1}^\top S\big)^{-\top}x\right\|^2_2 && (\|A^\top A\|_{\mathrm{op}} = \|A\|^2_{\mathrm{op}})\\
=&\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|^2_{\mathrm{op}}\left\|V^{-1/2}\Gamma^{-\top}x - V^{-1/2}\Gamma^{-\top}\big(V^{-1}Z_{T_1}^\top S\big)^\top\big(\Gamma+V^{-1}Z_{T_1}^\top S\big)^{-\top}x\right\|^2_2 \\
& && \text{(Lemma 21.11: } (A+B)^{-1} = A^{-1}-(A+B)^{-1}BA^{-1})\\
\overset{a_2}{\le}&\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|^2_{\mathrm{op}}\left(\left\|V^{-1/2}\Gamma^{-\top}x\right\|^2_2 + \left\|V^{-1/2}\Gamma^{-\top}\big(V^{-1}Z_{T_1}^\top S\big)^\top\big(\Gamma+V^{-1}Z_{T_1}^\top S\big)^{-\top}x\right\|^2_2\right).
\end{aligned}$$

We can upper bound the the term $\left\|V^{-1/2}\Gamma^{-\top}\left(V^{-1}Z_{T_1}^\top S\right)^\top\left(\Gamma+V^{-1}Z_{T_1}^\top S\right)^{-\top}x\right\|_2^2 =: \mathfrak{V}$ by noticing that it appears in both of a_1, a_2 above. Thus we have the inequality

$$\mathfrak{V} \le \left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\text{op}}^2\left(\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2+\mathfrak{V}\right).$$

By rearranging the terms, we have

$$\mathfrak{V} \le \frac{1}{1-\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\text{op}}^2}\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\text{op}}^2\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2.$$

By Lemma 19.1, with probability $1-\delta$, we have

$$\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\text{op}}^2 \le \sigma_\eta^2\overline{\log}(Z_{T_1},\delta).$$

Thus, with probability $1-\delta$, we have

$$\mathfrak{V} \le \frac{1}{1-\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\sigma_\eta^2\overline{\log}(Z_{T_1},\delta)}\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\sigma_\eta^2\overline{\log}(Z_{T_1},\delta)\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2.$$

To further upper bound $\mathfrak{V}$, we first find a sufficient condition on T_1 such that $\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\sigma_\eta^2\overline{\log}(Z_{T_1},\delta) \le \frac{1}{2g}$, $g \ge 1$. By Lemma 21.12, when $T_1 \ge 2p$,

$$\left\|\Gamma^{-1}V^{-1}\Gamma^{-\top}\right\|_{\text{op}}\sigma_\eta^2\overline{\log}(Z_{T_1},\delta) \le \frac{2\sigma_\eta^2}{T_1\sigma_{\min}(\sum_{z\in\mathcal{Z}}\lambda_z\Gamma^\top zz^\top\Gamma)}\overline{\log}(Z_{T_1},\delta).$$

To upper bound the right hand side by $\frac{1}{2g}$, we need

$$T_1 \ge \frac{4g\sigma_\eta^2}{\sigma_{\min}(\sum_{z\in\mathcal{Z}}\lambda_z\Gamma^\top zz^\top\Gamma)}\overline{\log}(Z_{T_1},\delta). \tag{3.59}$$

With this condition (3.59), we have with probability $1-\delta$,

$$\begin{aligned}\mathfrak{V} \leq & \frac{1}{1-\left\|\Gamma^{-1} V^{-1} \Gamma^{-\top}\right\|_{\mathrm{op}} \sigma_\eta^2 \overline{\log}\left(Z_{T_1}, \delta\right)}\left\|\Gamma^{-1} V^{-1} \Gamma^{-\top}\right\|_{\mathrm{op}} \sigma_\eta^2 \overline{\log}\left(Z_{T_1}, \delta\right)\left\|V^{-1 / 2} \Gamma^{-\top} x\right\|_2^2 \\ \leq & \frac{1}{1-\frac{1}{2 g}} \frac{1}{2 g}\left\|V^{-1 / 2} \Gamma^{-\top} x\right\|_2^2 \\ \leq & \frac{1}{g}\left\|V^{-1 / 2} \Gamma^{-\top} x\right\|_2^2 . \end{aligned} \tag{3.60}$$

□

Lemma 21.9. *Suppose that we have $\widehat{\Gamma}$ that is an OLS estimate from an offline dataset $\{Z_{T_1}, X_{T_1}\}$ collected non-adaptively through a fixed design λ and the efficient rounding procedure ROUND. Let $\dot{V}$ be any positive definite matrix. Then, for any $x \in \mathbb{R}^d$, we have, with probability $1-\delta$,*

$$\left\|(\Gamma^{-\top} - \widehat{\Gamma}^{-\top}) x\right\|_{\dot{V}^{-1}}^2 \leq 2\left\|\Gamma^{-1} \dot{V}^{-1} \Gamma^{-\top}\right\|_{\mathrm{op}} \sigma_\eta^2 \overline{\log}\left(Z_{T_1}, \delta\right)\|x\|_{\bar{A}\left(Z_{T_1}, \Gamma\right)^{-1}}^2,$$

when

$$T_1 \geq \frac{4 \sigma_\eta^2}{\sigma_{\min}\big(A(\lambda, \Gamma)\big)} \overline{\log}\left(Z_{T_1}, \delta\right) \vee 2p,$$

where p is the cardinality of support of λ.

Proof.

$$
\begin{aligned}
&\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|_{\dot{V}^{-1}}^2 \\
=&\left\|\Gamma^{-\top}S^\top Z_{T_1}V^{-1}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|_{\dot{V}^{-1}}^2 && \text{(Lemma 21.10)}\\
\overset{a_1}{=}&\left\|\dot{V}^{-\frac{1}{2}}\Gamma^{-\top}S^\top Z_{T_1}V^{-1}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|_2^2 && (\|Ax\|_V^2=\left\|V^{\frac{1}{2}}Ax\right\|_2^2)\\
\leq&\left\|\dot{V}^{-\frac{1}{2}}\Gamma^{-\top}S^\top Z_{T_1}V^{-\frac{1}{2}}\right\|_{\mathrm{op}}^2\left\|V^{-\frac{1}{2}}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|_2^2 && (\|Ax\|\leq\|A\|_{\mathrm{op}}\|x\|)\\
\leq&\left\|\dot{V}^{-\frac{1}{2}}\Gamma^{-\top}\right\|_{\mathrm{op}}^2\left\|S^\top Z_{T_1}V^{-\frac{1}{2}}\right\|_{\mathrm{op}}^2\left\|V^{-\frac{1}{2}}(\Gamma+V^{-1}Z_{T_1}^\top S)^{-\top}x\right\|_2^2 && (\|AB\|_{\mathrm{op}}\leq\|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}})\\
=&\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\mathrm{op}}^2\left\|V^{-1/2}\left(\Gamma+V^{-1}Z_{T_1}^\top S\right)^{-\top}x\right\|_2^2 && (\|A^\top A\|_{\mathrm{op}}=\|A\|_{\mathrm{op}}^2)\\
=&\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\mathrm{op}}^2\left\|V^{-1/2}\Gamma^{-\top}x-V^{-1/2}\Gamma^{-\top}\left(V^{-1}Z_{T_1}^\top S\right)^\top\left(\Gamma+V^{-1}Z_{T_1}^\top S\right)^{-\top}x\right\|_2^2 \\
& && \text{(Lemma 21.11: } (A+B)^{-1}=A^{-1}-(A+B)^{-1}BA^{-1})\\
\overset{a_2}{\leq}&\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\mathrm{op}}^2\left(\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2+\left\|V^{-1/2}\Gamma^{-\top}\left(V^{-1}Z_{T_1}^\top S\right)^\top\left(\Gamma+V^{-1}Z_{T_1}^\top S\right)^{-\top}x\right\|_2^2\right).
\end{aligned}
$$

Given the condition (3.59) holds, we have with probability $1-\delta$,

$$
\begin{aligned}
&\left\|(\Gamma^{-\top}-\widehat{\Gamma}^{-\top})x\right\|_{\dot{V}^{-1}}^2 \\
\overset{a_1}{\leq}&\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\mathrm{op}}^2\left(\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2+\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2\right)\\
=&2\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\left\|V^{-1/2}Z_{T_1}^\top S\right\|_{\mathrm{op}}^2\left\|V^{-1/2}\Gamma^{-\top}x\right\|_2^2\\
\leq&2\left\|\Gamma^{-1}\dot{V}^{-1}\Gamma^{-\top}\right\|_{\mathrm{op}}\sigma_\eta^2\overline{\log}(Z_{T_1},\delta)\|x\|_{(\Gamma^\top V\Gamma)^{-1}}^2,
\end{aligned}
$$

where (a_1) is due to (3.60) and setting $g=1$. □

Lemma 21.10. *For a least square estimate $\widehat{\Gamma}$ that is estimated through a design matrix Z and $\widehat{\Gamma}$ is invertible, we have*

$$
\widehat{\Gamma}^{-1}-\Gamma^{-1}=-\left(\Gamma+V^{-1}Z^\top S\right)^{-1}V^{-1}Z^\top S\Gamma^{-1}.
$$

Proof. Since $\widehat{\Gamma}$ is a least square estimator, we have

$$
\widehat{\Gamma}=\Gamma+V^{-1}Z^\top S.
$$

By Lemma 21.11, we have

$$\begin{aligned}\widehat{\Gamma}^{-1} - \Gamma^{-1} =& \left(\Gamma + V^{-1}Z^\top S\right)^{-1} - \Gamma^{-1} \\ =& -\left(\Gamma + V^{-1}Z^\top S\right)^{-1} V^{-1} Z^\top S \Gamma^{-1}.\end{aligned}$$

□

Lemma 21.11. *For two invertible matrixes $A, B \in \mathbb{R}^{d\times d}$, we have*

$$(A+B)^{-1} = A^{-1} - (A+B)^{-1} B A^{-1}.$$

Proof. We have

$$\begin{aligned}(A+B)^{-1} &= A^{-1} + (A+B)^{-1} - A^{-1} \\ &= A^{-1} + \left((A+B)^{-1}A - I\right)A^{-1} \\ &= A^{-1} + (A+B)^{-1}\big(A - (A+B)\big)A^{-1} \\ &= A^{-1} - (A+B)^{-1} B A^{-1}.\end{aligned}$$

□

Lemma 21.12. *Suppose that we have a design matrix Z_T that is sampled from a distribution $\lambda \in \Delta_{\mathcal{Z}}$, with the efficient rounding procedure ROUND. Let p represent the cardinality of support of λ. We have, if $T \geq 2p$,*

$$\left\|\left(\Gamma^\top Z_T^\top Z_T \Gamma\right)^{-1}\right\|_{\text{op}} \leq \frac{2}{T\sigma_{\min}(\sum_{z\in\mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)}.$$

where $\sigma_{\min}(\cdot)$ is the smallest singular value of a matrix.

Proof. Suppose that each arm $z \in \mathcal{Z}$ is sampled t_z times, the empirical distribution of Z_T is $\xi := \left(\frac{t_z}{T}\right)_{z\in\mathcal{Z}}$. Thus, we have

$$\begin{aligned}\left\|\left(\Gamma^\top Z_T^\top Z_T \Gamma\right)^{-1}\right\|_{\text{op}} &= \frac{1}{\sigma_{\min}\left(\Gamma^\top Z_T^\top Z_T \Gamma\right)} \\ &= \frac{1}{\sigma_{\min}(T\sum_{z\in\mathcal{Z}} \xi_z \Gamma^\top z z^\top \Gamma)} \\ &= \frac{1}{T\sigma_{\min}(\sum_{z\in\mathcal{Z}} \xi_z \Gamma^\top z z^\top \Gamma)}.\end{aligned}$$

By Fiez et al. (2019, Proposition 2), we have

$$\sum_{z\in\mathcal{Z}} \xi_z \Gamma^\top z z^\top \Gamma \succeq \alpha \sum_{z\in\mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma,$$

where $\alpha \in [\frac{T}{T+2p}, 1]$ when $T \geq 2p$. Given the fact that both of $\sum_{z\in\mathcal{Z}} \xi_z \Gamma^\top z z^\top \Gamma$ and $\sum_{z\in\mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma$ are positive definite, we have $\sigma_{\min}(\sum_{z\in\mathcal{Z}} \xi_z \Gamma^\top z z^\top \Gamma) \geq \alpha\sigma_{\min}(\sum_{z\in\mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)$. Thus, we have

$$\left\|\left(\Gamma^\top Z_T^\top Z_T \Gamma\right)^{-1}\right\|_{\text{op}} \leq \frac{1}{\alpha T\sigma_{\min}(\sum_{z\in\mathcal{Z}} \lambda_z \Gamma^\top z z^\top \Gamma)}.$$

When $T \geq 2p$, we have $\alpha \geq 1/2$, which implies the result. □

3.7.12 Estimating $\lambda_{\min}(\Gamma)$

In this section, we introduce a simple adaptive procedure that finds a high probability lower bound on $\gamma^*_{\min} := \lambda_{\min}(\Gamma)$ that is sufficiently accurate (i.e., within a constant factor of $\gamma^*_{\min}$). For simplicity, we assume $\|z_t\| \leq 1, \forall t$ in this section.

Our algorithm leverages confidence bounds to adaptively determine how many samples we like to take. Let $\hat{\Gamma}_t := (Z^\top Z)^{-1} Z^\top X$ be the least square estimate of Γ after sampling t times to obtain $\{(z_s, x_s)\}_{s=1}^t$ where $Z \in \mathbb{R}^{t\times d}$ and $X \in \mathbb{R}^{t\times d}$ are the design matrices. Let $V_t := \sum_{s=1}^t z_s z_s^\top$. We define the lower and upper confidence bound for $\gamma^*_{\min}$ as follows:

$$\text{LCB}(t) := \lambda_{\min}(\hat{\Gamma}_t) - \sqrt{\frac{\psi_t}{t}}, \qquad \text{UCB}(t) := \lambda_{\min}(\hat{\Gamma}_t) + \sqrt{\frac{\psi_t}{t}}$$

where

$$\psi_t = \sigma_{\min}^{-1}\left(\frac{1}{t}V_t\right)\cdot\left(8d\ln\left(1+\frac{2t}{d(2\wedge\sigma_{\min}(V_t))}\right)+16\ln\left(\frac{2\cdot 6^d}{\delta}\cdot\log_2^2\left(\frac{4}{2\wedge\sigma_{\min}(V_t)}\right)\right)\right).$$

and $\mathrm{LCB}(0) := -\infty$ and $\mathrm{UCB}(0) := \infty$.

The following lemma shows that $\mathrm{LCB}(t)$ and $\mathrm{UCB}(t)$ form a valid anytime confidence bound for $\gamma^*_{\min}$.

Lemma 21.13. *(Correctness of the confidence bounds)*

$$1-\delta \le \mathbb{P}(\forall t \ge 1, \mathrm{LCB}(t) \le \gamma^*_{\min} \le \mathrm{UCB}(t))$$

Equipped with the confidence bounds, we are now ready to describe our algorithm for learning $\gamma^*_{\min}$ (see Algorithm 12). Since the tightness of the confidence bounds depends on the smallest eigenvalue of V_t, it is natural to use the E-optimal design as defined in Section 3.3.2. Recall that the solution of the E-optimal design is λ^*_E and κ_0 is the smallest singular value achieved by λ^*_E. We take in a rounding procedure for the E-optimal design $\mathrm{ROUND}_E(\lambda, t)$ that takes in t samples and design λ and outputs integer sample count assignments $\{N_z\}_{z\in\mathcal{Z}}$ so that if we sample according to these counts then we have

$$\sigma_{\min}^{-1}(\frac{1}{t}V_t) \le (1+\omega)\kappa_0^{-1} \tag{3.61}$$

After determining the base sample counts $\{m_z\}_{z\in\mathcal{Z}}$ by $\mathrm{ROUND}(\lambda^*_E, \lceil r(\omega)\rceil, \omega)$, we start doubling the sample size until we satisfy the condition $\mathrm{LCB}(t) > \frac{1}{2}\mathrm{UCB}(t)$. Note that the sampling scheme in the while loop is designed such that the total number of samples collected up to (and including) j-th iteration is $2^{j-1}\lceil r(\omega)\rceil$. Once the loop stops, we return $\mathrm{LCB}(t)$ as the claimed lower approximation of the $\gamma^*_{\min}$.

Let N_w be the total number of samples we used in Algorithm 12. Then, the next theorem shows that the estimate returned by our algorithm is both a valid lower bound to $\gamma^*_{\min}$ and sufficiently accurate.

Algorithm 12: Learning $\lambda_{\min}(\Gamma)$

Input: Arm set $\mathcal{Z}$, rounding procedure ROUND_E for the E-optimal design, rounding accuracy ω

Initialize: $j = 1$, $t = 0$.

Compute the E-optimal design λ_E^* for $\mathcal{Z}$.

Compute $\{m_z\}_{z\in\mathcal{Z}}$ by $\mathrm{ROUND}_E(\lambda_E^*, \lceil r(\omega)\rceil)$.

while $\mathrm{LCB}(t) \le \frac{1}{2}\mathrm{UCB}(t)$ **do**

  $t \leftarrow 2^{j-1}\lceil r(\omega)\rceil$

  For each arm $z \in \mathcal{Z}$, sample $2^{j-1}m_z - \mathbb{1}\{j \neq 1\}\, 2^{j-2}m_z$ times

  Estimate $\hat{\Gamma}_t$ using all samples collected so far (total t data points)

  $j \leftarrow j + 1$

Output: $\mathrm{LCB}(t)$

Theorem 22. *(Correctness of Algorithm 12) The total number of samples denoted by N_w used in Algorithm 12 satisfies that, with probability at least $1-\delta$,*

$$\frac{\gamma_{\min}^*}{2} < \mathrm{LCB}(N_w) \le \gamma_{\min}^*$$

We next analyze the sample complexity of the algorithm, which essentially shows the scaling of $\frac{1}{(\gamma_{\min}^*)^2\kappa_0}$ even if the algorithm does not need knowledge of $\gamma_{\min}^*$.

Theorem 23. *(Sample complexity of Algorithm 12) Then, with $\omega = 1$, we have, with probability at least $1-\delta$,*

$$N_w = O\left(r(1) + (\gamma_{\min}^*)^{-2}\kappa_0^{-1}\cdot\left(d\cdot\mathrm{polylog}((\gamma_{\min}^*)^{-2}, \kappa_0^{-1}, d) + \ln(2/\delta)\right)\right)$$

We remark that Allen-Zhu et al. (2021) provides a rounding procedure with $r(\varepsilon) = O(d/\varepsilon^2)$.

Proof of Lemma 21.13. Note that $\hat{\Gamma} = \Gamma + V_t^{-1}Z^\top S$ where $Z, S \in \mathbb{R}^{t\times d}$ is the design matrices with s-th row being $z_s^\top$ and $\eta_s^\top$ respectively. Using Lemma 19.1, we have, with probability

at least $1-\delta$,

$$\begin{aligned}
\forall t \geq 1, \left\|\hat{\Gamma}_t - \Gamma\right\|_{\mathrm{op}}^2 &= \left\|V_t^{-1} Z^\top S\right\|_{\mathrm{op}}^2 \\
&\leq \left\|V_t^{-1/2}\right\|_{\mathrm{op}}^2 \left\|V_t^{-1/2} Z^\top S\right\|_{\mathrm{op}}^2 \\
&= \frac{1}{t\sigma_{\min}(\frac{1}{t}V_t)} \left\|V_t^{-1/2} Z^\top S\right\|_{\mathrm{op}}^2 \\
&\leq \frac{1}{t\sigma_{\min}(\frac{1}{t}V_t)} \cdot \left(8d \ln\left(1 + \frac{2t}{d(2 \wedge \sigma_{\min}(V_t))}\right) + 16 \ln\left(\frac{2 \cdot 6^d}{\delta} \cdot \log_2^2\left(\frac{4}{2 \wedge \sigma_{\min}(V_t)}\right)\right)\right) \\
&\qquad \text{(Lemma 19.1)} \\
&= \frac{1}{t}\psi_t
\end{aligned}$$

The well-known Weyl's theorem implies that $\max_k |\lambda_k(\hat{\Gamma}_t) - \lambda_k(\Gamma)| \leq \left\|\hat{\Gamma}_t - \Gamma\right\|_{\mathrm{op}}$ where $\lambda_k(A)$ is the k-th largest singular value of the matrix A. Choosing $k = d$ and combining it with the display above conclude the proof. □

Proof of Theorem 22. We assume $\forall t \geq 1, \mathrm{LCB}(t) \leq \gamma_{\min}^* \leq \mathrm{UCB}(t)$, which happens with probability at least $1-\delta$. Then, it is trivial to see that $\mathrm{LCB}(N_w) \leq \gamma_{\min}^*$.

For the other inequality, we use the fact that the stopping condition was satisfied with N_w:

$$\mathrm{LCB}(N_w) > \frac{1}{2}\mathrm{UCB}(N_w) \geq \frac{\gamma_{\min}^*}{2} \ .$$

This concludes the proof. □

Proof of Theorem 23. We assume $\forall t \geq 1, \mathrm{LCB}(t) \leq \lambda_{\min}(\Gamma) \leq \mathrm{UCB}(t)$, which happens with probability at least $1-\delta$. In this proof, we let $\hat{\gamma}_{\min} := \lambda_{\min}(\hat{\Gamma}_{N_w})$ and $\gamma_{\min}^* := \lambda_{\min}(\Gamma)$ for brevity.

If $N_w = \lceil r(1) \rceil$, then there is nothing to prove. Otherwise, the loop in the algorithm was iterated more than once. Then, since the stopping condition was satisfied with N_w, we have that in the previous iteration where $t = N_w/2$ the stopping condition was not satisfied. Thus,

$$2\mathrm{LCB}(N_w/2) \leq \mathrm{UCB}(N_w/2) \ .$$

Using the following two inequalities:

$$2\text{LCB}(N_w/2) \geq 2\left(\hat{\gamma}_{\min} - \sqrt{\frac{\psi_{N_w/2}}{N_w/2}}\right) \geq 2\left(\gamma^*_{\min} - 2\sqrt{\frac{\psi_{N_w/2}}{N_w/2}}\right)$$
$$\text{UCB}(N_w/2) \leq \hat{\gamma}_{\min} + \sqrt{\frac{\psi_{N_w/2}}{N_w/2}} \leq \gamma^*_{\min} + 2\sqrt{\frac{\psi_{N_w/2}}{N_w/2}} \ ,$$

we have

$$\gamma^*_{\min} \leq 6\sqrt{\frac{\psi_{N_w/2}}{N_w/2}} \implies N_w \leq \frac{72}{(\gamma^*_{\min})^2}\psi_{N_w/2} \ .$$

On the other hand, with the rounding procedure, we have

$$\sigma^{-1}_{\min}\left(\sum_{t=1}^{N} z_t z_t^\top\right) = \frac{1}{N}\sigma^{-1}_{\min}\left(\frac{1}{N}\sum_{t=1}^{N} z_t z_t^\top\right) = \frac{1}{N}\sigma_{\max}\left(\left(\frac{1}{N}\sum_{t=1}^{N} z_t z_t^\top\right)^{-1}\right) \leq \frac{1}{N}(1+\omega)\kappa_0^{-1} = \frac{2}{N}\kappa_0^{-1}$$

since $\omega = 1$. Using this and the fact that $N_w \geq d$, it is easy to see that there exists an absolute constant c_1 such that

$$\psi_{N_w/2} \leq c_1\kappa_0^{-1}\left(d\ln\left(1+N_w+\kappa_0^{-1}\right) + \ln\left(\frac{2}{\delta}\right)\right) \ .$$

Then, there exists an absolute constant c_2 such that

$$N_w \leq (\gamma^*_{\min})^{-2}\cdot c_2\kappa_0^{-1}\left(d\ln\left(1+N_w+\kappa_0^{-1}\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

We have N_w on both sides. We invoke Lemma 23.2 with $r = 1 + N_w$ to obtain

$$N_w \leq 1 + 2c_2(\gamma^*_{\min})^{-2}\kappa_0^{-1}\left(d\ln(1+2\kappa_0^{-1}(1+c_2 d(\gamma^*_{\min})^{-2})) + \ln\left(\frac{2}{\delta}\right)\right) \ .$$

□

Lemma 23.1. *Let $A, \Gamma \in \mathbb{R}^{d\times d}$ where A is symmetric positive semi-definite. Then,*

$$\sigma_{\min}\left(\Gamma^\top A\Gamma\right) \geq \sigma_{\min}(\Gamma)^2\sigma_{\min}(A)$$

Proof.

$$\begin{aligned}
(\sigma_{\min}(\Gamma^\top A\Gamma))^{-1} &= \|\Gamma^{-1}A^{-1}\Gamma^{-T}\|_{\text{op}} \\
&\leq \|\Gamma^{-1}\|_{\text{op}}\|A^{-1}\|_{\text{op}}\|\Gamma^{-T}\|_{\text{op}} \qquad \text{(submultiplicity of the operator norm)} \\
&= \sigma_{\min}(\Gamma)^{-2}\sigma_{\min}(A)^{-1}
\end{aligned}$$

Taking the inverse on both sides concludes the proof. □

3.7.13 Proof of Lemma 16.1

Define $\Lambda = \sum_{i=1}^d \lambda_i e_i e_i^\top$, i.e. the diagonal matrix with λ on the diagonal.

Note that

$$\min_\lambda \max_{i,j}(e_i^\top - e_j)^\top(\Gamma\Lambda\Gamma^\top)^{-1}(e_i - e_j) = \min_\lambda \max_{i,j}\sum_{k=1}^d \frac{(\Gamma_{k,i}^{-1} - \Gamma_{k,j}^{-1})^2}{\lambda_k}$$

So for an upper bound, we consider $\lambda = 1/d1$,

$$\min_\lambda \max_{i,j}(e_i^\top - e_j)^\top(\Gamma\Lambda\Gamma^\top)^{-1}(e_i - e_j) \leq d\max_{i,j}\|\Gamma^{-1}(e_i - e_j)\|_2^2$$

The result about ρ^* follows immediately.

When $\Gamma = (1-\varepsilon)/\mathbf{1}\mathbf{1}^\top + \varepsilon I$, a computation using Sherman-Morrison shows that $\Gamma^{-1} = 1/\varepsilon[I - (1-\varepsilon)/\mathbf{1}\mathbf{1}^\top]$. Thus

$$\begin{aligned}
\min_\lambda \max_{i,j}(e_i^\top - e_j)^\top(\Gamma\Lambda\Gamma^\top)^{-1}(e_i - e_j) &= \varepsilon^{-2}\min_\lambda \max_{i,j}(e_i - e_j)^\top\Lambda^{-1}(e_i - e_j) \\
&= \varepsilon^{-2}\min_\lambda \max_{i,j} e_i\top\Lambda^{-1}e_i + e_j\Lambda^{-1}e_j \\
&= \varepsilon^{-2}\min_\lambda \max_{i,j} e_i\top\Lambda^{-1}e_i + e_j\Lambda^{-1}e_j \\
&\geq \varepsilon^{-2}\min_\lambda \max_{i} e_i\top\Lambda^{-1}e_i \qquad\qquad = \varepsilon^{-2}d
\end{aligned}$$

where the last line follows from the Kiefer-Wolfowitz Theorem (Lattimore and Szepesvári, 2020).

3.7.14 lemma for solving x less than ln(x)

Lemma 23.2. *Let $a, b, c, d > 0$. Then, for every r,*

$$r \le a + b\ln(c + dr) \implies r \le 2a + 2b\ln(1 + 2c + 2bd)$$

Proof. If $dr \le c$, then

$$r \le a + b\ln(2c) \le a + b\ln(1 + 2c)$$

If $dr > c$, then,

$$\begin{aligned} r &\le a + b\ln(2dr) \\ &= a + b\ln(2d\frac{r}{2b}\cdot 2b) \\ &\le a + b\left(\frac{r}{2b} - 1 + \ln(4bd)\right) && (\ln(x) \le x - 1) \\ \implies r &\le 2a + 2b\ln\left(\frac{4}{e}bd\right) \\ &\le 2a + 2b\ln(2bd) \end{aligned}$$

Either case, we have

$$r \le 2a + 2b\ln((1 + 2c) \vee 2bd) \le 2a + 2b\ln(1 + 2c + 2bd)$$

□

Lemma 23.3. *Let $\alpha, \beta > 0$. Then, for any r,*

$$r \le \alpha\ln(1 + r) + \beta \implies r \le 2\alpha\ln(e + \alpha) + 2\beta$$

Proof.

$$
\begin{aligned}
& r \le \alpha \ln(r(\frac{1}{r}+1)) + \beta \\
& \le \alpha \ln(\frac{r}{2\alpha} \cdot 2\alpha(\frac{1}{r}+1)) + \beta \\
& \le \alpha \left(\frac{r}{2\alpha} - 1 + \ln(2\alpha(\frac{1}{r}+1)) \right) + \beta && (\forall x, \ln(x) \le x - 1) \\
\implies & r \le 2\alpha \ln(\frac{2}{e}\alpha(\frac{1}{r}+1)) + \beta
\end{aligned}
$$

If $r \le 2\alpha$, then there is nothing to prove. If $r > 2\alpha$, then $r \le 2\alpha \ln(1+\alpha) + 2\beta$. Either case, we have $r \le 2\alpha \ln(e+\alpha) + 2\beta$. □

Chapter 4

Nealy Optimal Active Preference Learning and Its Application to LLM Alignment

Aligning Large Language Models (LLMs) critically relies on high quality datasets of human preference labels, which are expensive to collect. While various active learning methods have been proposed to improve sample efficiency over passive collection, many existing works rely on off-the-shelf experimental design criteria, such as G or D-optimality. These objectives are not specifically tailored to the structure of preference learning, leaving an open problem of designing more targeted, problem-specific algorithms. In this work, we identify a simple intuition unique to the preference learning problem that challenges the common choice of these design objectives. Motivated by this, we propose two novel active learning algorithms. The first achieves the first known instance-dependent label complexity guarantee for this problem, while the second is a simple and practical greedy method. We validate our algorithms on real-world preference datasets, demonstrating improvements in sample efficiency over existing methods.

4.1 Introduction

Reinforcement Learning from Human Feedback (RLHF) is a pivotal post-training technique for aligning Large Language Models (LLMs) with human preferences, moving beyond the limitations of standard supervised fine-tuning. A core component of RLHF is learning a reward model that accurately reflects human judgments on the quality of model-generated responses. This reward model then guides the optimization of the LLM, often using reinforcement learning algorithms like Proximal Policy Optimization (PPO) (Ouyang et al., 2022).

The standard paradigm for learning a reward model is pairwise preference learning. In this setup, human labelers are presented with pairs of LLM-generated responses to a given prompt and are asked to indicate their preference. This collected preference data is then used to train a reward model, such as the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952). This approach is motivated by the cognitive observation that humans often find it easier to make relative preference judgments between two options than to assign absolute scores to individual items.

However, collecting human preference labels is expensive and time-consuming. This challenge is magnified when high-quality labels are required, such as in safety-critical applications like medical assistance, where each label may demand expert knowledge and careful consideration. These factors make the data collection process difficult to scale and can lead to issues with data quality (Zhang et al., 2023). Consequently, developing data-efficient methods that minimize the number of queries to human labelers while ensuring reward model quality is a critical problem.

To address this, we formulate preference learning as an active learning problem, where an algorithm actively selects the most informative pairs to query, since *not all pairs of answers are equally informative*. While recent works have applied active learning to this problem using well studied tools of G-optimal and D-optimal experimental design (Das et al., 2024; Mehta et al., 2023; Scheid et al., 2024; Mukherjee et al., 2024), they predominantly focus on worst-case performance guarantees, meaning that either their analysis is not tight enough or the algorithm design does not adapt to the true difficulty of the problem instance. Such results may not fully exploit the underlying structure of a given problem instance, potentially leading to suboptimal sampling strategies.

In this work, we argue for an instance-dependent approach. Our method starts from a simple intuition: pairs of answers that are close to the decision boundary (i.e., those with small reward differences) are the most ambiguous and thus provide the most information when queried. Based on this, we develop two distinct but complementary algorithms: the first, based on a novel experimental design formulation, achieves a nearly optimal instance-dependent label complexity; the second is a practical and easily implemented greedy algorithm based on a new uncertainty sampling heuristic.

We summarize our contributions as follows:

- We propose a novel experimental design objective for active preference learning that, to the best of our knowledge, is the first one to yield an instance-dependent label complexity guarantee for the active preference learning problem.
- We design a simple and practical greedy algorithm based on a new uncertainty sampling heuristic, which is easy to implement and does not require solving complex optimization problems.
- We provide empirical validation on real-world preference datasets. Our results demonstrate that our proposed methods notably reduce the number of required queries while achieving high reward model accuracy compared to existing methods.

Notations: Thoughout the paper, we use the following notations, $[n] = \{1, 2, \ldots, n\}$; $\triangle(\mathcal{Z}) = \{\lambda \in \mathbb{R}^{|\mathcal{Z}|} \mid \lambda_{z \in \mathcal{Z}} \geq 0, \sum_{z \in \mathcal{Z}} \lambda_z = 1\}$ is the probability simplex over the set $\mathcal{Z}$; $\|z\|_A = \sqrt{z^\top A z}$ is the norm of z with respect to a positive definite matrix A; $H(\lambda, \theta) = \sum_{z \in \mathcal{Z}} \lambda_z \dot{\mu}(z^\top \theta) z z^\top$, where $\dot{\mu}$ is the derivative of the link function μ.

4.2 Related Work

The Reinforcement Learning from Human Feedback (RLHF) pipeline was popularized by InstructGPT (Ouyang et al., 2022), though the core concept dates back to earlier work of Christiano et al. (2017). The now-dominant optimization algorithm, Proximal Policy Optimization (PPO), was introduced in Schulman et al. (2017). While alternative pipelines like Direct Preference Optimization (DPO) have since been proposed in Rafailov et al. (2023), all of these alignment techniques necessitate a high-quality dataset of preference labels, motivating the need for sample-efficient data collection.

A recent line of work has focused on improving the sample efficiency through active learning for preference collection (Das et al., 2024; Mehta et al., 2023; Scheid et al., 2024; Zhang et al., 2024; Dwaracherla et al., 2024; Muldrew et al., 2024; Mukherjee et al., 2024; Liu et al., 2024; Schlaginhaufen et al., 2025; Kveton et al., 2025). Mehta et al. (2023) was among the first to formally bring active learning to RLHF, proposing an active contextual dueling bandit

method with a worst-case suboptimality gap guarantee. Das et al. (2024) later relaxed the underlying assumptions to better fit the Bradley-Terry-Luce (BTL) model, and Liu et al. (2024) further extended this to incorporate human labelers with varying levels of expertise. Other works have framed the problem differently, for instance, through the lens of best arm identification (Scheid et al., 2024) or for online, rather than offline, alignment (Zhang et al., 2024). Notably, most of these works provide worst-case performance guarantees only. More recently, active learning has also been tailored specifically for the DPO pipeline (Muldrew et al., 2024; Kveton et al., 2025).

Many of these methods are grounded in the experimental design framework, with D-optimal experimental design being a particularly popular approach (Mukherjee et al., 2024; Liu et al., 2024; Schlaginhaufen et al., 2025). The seminal work for this approach dates back to Kiefer and Wolfowitz (1960). In recent works, Mukherjee et al. (2024) proposed D-optimal experiment design based algorithms for various feedback models and evaluated performance using ranking loss functions, distinct from the common use of classification accuracy, however their algorithms are not adaptive but fixed allocation algorithms. Liu et al. (2024) proposed a dual active learning approach using D-optimal experimental design that also accounts for labeler expertise. There are also another work with methods combining randomized exploration with D-optimal experimental design (Schlaginhaufen et al., 2025). D-optimal experimental design has also been adapted for DPO, with worst case theoretical guarantees on the logit errors (Kveton et al., 2025).

Beyond preference learning for alignment, the principles of active learning or broadly speaking, interactive machine learning are also being applied to other areas of LLM, such as in-context learning (Margatina et al., 2023) and prompt engineering (Diao et al., 2024; Shi et al., 2024).

4.3 Preliminaries

Consider a set of prompts $\mathcal{X}$, $|\mathcal{X}| = n$, where each prompt $x \in \mathcal{X}$ has two associated answers, a_1 and a_2. The unlabeled dataset is structured as triplets (x, a_1, a_2). These answer pairs are typically generated either by using different settings in the same LLM or by using two different LLMs. Borrowing terminology from the bandit literature, we refer to each triplet as

an arm. Our focus is more on the embedding space. Let d be the embedding dimension, and $z \in \mathcal{Z} \subset \mathbb{R}^d$ denotes the embedding of the triplet (x, a_1, a_2), where the order of the two answers is preserved. We assume $\|z\| \leq 1$ for all $z \in \mathcal{Z}$ as Das et al. (2024). The preference label is denoted as $y \in \{1, 0\}$, where $y = 1$ indicates that the first answer is preferred, otherwise the second. We use the notation $a_1 \succeq a_2$ to denote that a_1 is preferred over a_2.

We consider a realizable setting, that is to say, there exist a true parameter $\theta^* \in \mathbb{R}^d$ such that the preference model can be expressed as, for any $i \in [n]$,

$$p(y = 1|x, a_1, a_2) = \psi\big(z^\top \theta^*\big), \tag{4.1}$$

where ψ is a monotonic link function that maps the inner product $z^\top \theta^*$ to a probability. This can be instantiated into different common models, such as logistic model: $\psi(x) = \frac{1}{1+e^{-x}}$, linear model: $\psi(x) = \frac{x+1}{2}$, with $x \in [-1, 1]$, and Gaussian CDF, $\psi(x) = \Phi(x)$, where Φ is the cumulative distribution function of the standard Gaussian distribution. The logistic model is often called Bradley-Terry-Luce (BTL) model in the human preference learning problem, and will be the main focus of this paper.

We formulate the pairwise preference learning task as an active classification problem for the sign of $z^\top \theta^*$. The objective is to train a classifier that correctly classifies the preference for all of the answer pairs, by actively and adaptively selecting arm at each round and querying human annotators for preference labels.

At every round t, based on all the data collected up to time t, the learner selects an arm and queries the human annotator for its preference label. The learner then receives the labels that is generated by Equation (4.1). This can be seen as a special case of a batched setting, which is more realistic in large scale label collection where the learner can select a batch of arms and query for their preference labels at once (Scheid et al., 2024). The labeled preference dataset at time t thus can be represented as $\mathcal{D}_t = \big\{(z_s, y_s) \mid s \in [t]\big\}$. A classifier $\hat{\theta}_t$ can be trained based on $\mathcal{D}_t$ by solving the following maximum likelihood estimation (MLE) problem,

$$\hat{\theta}_t = \arg\max_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} y_s \log \psi(z_s^\top \theta) + (1 - y_s) \log(1 - \psi(z_s^\top \theta)). \tag{4.2}$$

An arm z is said to be classified correctly by the classifier $\hat{\theta}_t$ if $\text{sign}(z^\top \hat{\theta}_t) = \text{sign}(z^\top \theta^*)$. The goal of the learner is to design a strategy that can classify all the arms correctly with high probability. We define the following notion of probably approximately correct (PAC) strategy.

Definition 3. *(δ-PAC) A strategy is said to be δ-PAC if, for any problem instance $(\mathcal{Z}, \theta^*)$ and any $\delta > 0$, it can be executed in a finite number of steps and with probability at least $1-\delta$, it outputs a classifier that can classify all the arms correctly.*

Our goal is to design a δ-PAC strategy, while minimizing the number of queries made to human annotators.

4.3.1 Confidence Interval

Our approach critically relies on two ingredients: the confidence width of the MLE estimator from Equation (4.2) and the relative position of the confidence interval, which will be clear in the next section. Here we present the confidence interval we use for rest of the paper, but our approach is rather general and not sensitive to the specific choice of the confidence interval.

Theorem 24. *(Theorem 1 of Jun et al. (2021)) Let $\delta \le e^{-1}$. Let $\hat{\theta}_t$ be the solution of Equation (4.2) where, for every $s \in [t]$, y_s is conditionally independent from $\{z_i\}_{i=1}^t \setminus \{z_s\}$ given z_s (i.e., the z_s's are a fixed design). Fix $z \in \mathbb{R}^d$ with $\|z\| \le 1$. Let t_{eff} be the number of distinct vectors in $\{z_s\}_{s=1}^t$. Define $\gamma(d) = 64(d\log(6) + \log((2+t_{eff})/\delta))$. Define the event*

$$\mathcal{E}_{var} = \left\{ \forall z', \ \frac{1}{\sqrt{2.2}} \|z'\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \le \|z'\|_{H'(\{z_s, s\in[t]\},\theta_t)^{-1}} \le \sqrt{2.2}\|z'\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \right\}.$$

If $\xi_t^2 := \max_{s\in[t]} \|z_s\|^2_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \le \frac{1}{\gamma(d)}$, then

$$\mathbb{P}\left(\left| z^\top(\hat{\theta}_t - \theta^*) \right| > D_{\delta,t}(z) := 2.4\|z\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \sqrt{\log \frac{2(2+t_{eff})}{\delta}}, \mathcal{E}_{var} \right) \le \delta,$$

where $H'(\mathcal{D}, \theta) = \sum_{z\in\mathcal{D}} \dot{\mu}(z^\top\theta) zz^\top$ is the Fisher information matrix at θ.

4.3.2 Intuition

The core intuition of our approach is to leverage the positions of the confidence intervals for $z^\top \theta^*$ relative to zero for each arm $z \in \mathcal{Z}$. This contrasts with many existing works that primarily focus on reducing the overall uncertainty of the parameter θ^*, for example, by minimizing the volume of the confidence ellipsoid of θ^* (D-optimal experimental design) or the maximum prediction variance over all arms (G-optimal design). In other words, these methods consider only the width of the confidence intervals for $z^\top \theta^*$, while neglecting their positions.

We illustrate our intuition in Figure 4.1 using four arms with different confidence intervals on $z^\top \theta^*$. The varying widths of these intervals reflect different levels of uncertainty at a given time t during the learning process. Algorithms that are designed to minimize uncertainty of each individual arm, such as G-optimal experimental design (Das et al., 2024), would preferentially pull arms with high uncertainty (i.e., wide confidence intervals), like arms 1 and 4 in the figure.

Although this strategy may seem reasonable from a general uncertainty reduction perspective, and often be used in many heuristics, from a sign classification perspective, this strategy can be suboptimal. Arm 1, for example, is already easy to classify, as its entire confidence interval locates above zero, despite its width. In contrast, the confidence interval for arm 2 overlaps with zero. This indicates that under the current estimate $\hat{\theta}$, its sign remains uncertain, even though its associated uncertainty width is smaller. Therefore, intuitively it is more beneficial to pull arm 2 to reduce its uncertainty until its confidence interval is fully separated from zero.

In the next two sections, we formalize the above intuition into two algorithm designs.

4.4 Active Preference Learning via Experimental Design

In this section, we develop a δ-PAC algorithm for the active preference learning problem based on the experimental design framework.

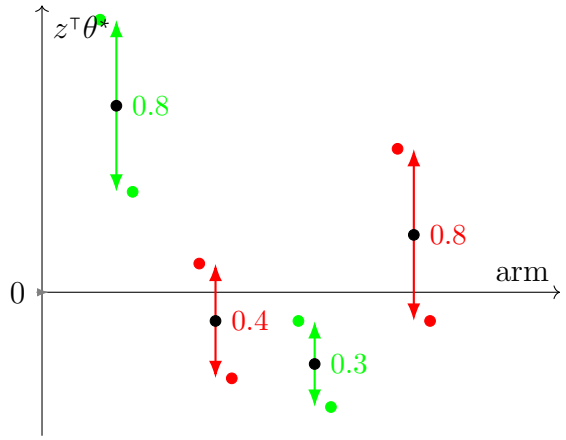


Figure 4.1: Intuition for our proposed method. An arm's sign is considered confidently determined (green) if its confidence interval does not contain zero, and uncertain (red) otherwise. The numbers denote the width of each confidence interval.

4.4.1 Optimal Allocation

To motivate our algorithm, we first analyze the optimal query allocation in an oracle setting where the true parameter θ^* is known. This allocation represents the ideal distribution of preference queries across pairs of arms. Subsequently, we will introduce an adaptive algorithm that approximates this optimal allocation without knowledge of θ^*.

We begin by defining $\mathcal{C}_\theta$ as the set of all parameters θ that correctly classify all arms,

$$\mathcal{C}_\theta := \left\{\theta \in \mathbb{R}^d \mid \forall z \in \mathcal{Z}, \operatorname{sign}\left(z^\top\theta\right) = \operatorname{sign}\left(z^\top\theta^*\right)\right\}.$$

Intuitively, the algorithm should be designed to return a classifier $\hat{\theta}$ that belongs to the set $\mathcal{C}_\theta$ with high probability and minimal label complexity. This requires an optimal allocation over the arms $\mathcal{Z}$ that efficiently exploits the structure of $\mathcal{C}_\theta$, ensuring the classifier $\hat{\theta}$ converges to this set quickly. We define

$$\mathcal{Z}^+_{\theta^*} := \left\{z \in \mathcal{Z} \mid z^\top\theta^* > 0\right\}, \quad \mathcal{Z}^-_{\theta^*} := \left\{z \in \mathcal{Z} \mid z^\top\theta^* < 0\right\}.$$

By Theorem 24, we have

$$z^\top\theta^* - z^\top\hat{\theta}_t \le \mathcal{C} \cdot \|z\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \sqrt{\log\left(t/\delta\right)}. \tag{4.3}$$

For any arm $z \in \mathcal{Z}^+_{\theta^*}$, the requirement $z^\top \hat{\theta}_t > 0$ is equivalent to:

$$z^\top \theta^* - z^\top \hat{\theta}_t < z^\top \theta^*, \tag{4.4}$$

Given the confidence bound in Equation (4.3), a sufficient condition for Equation (4.4) to hold is

$$\mathcal{C} \cdot \|z\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \sqrt{\log\big(t/\delta\big)} < z^\top \theta^*. \tag{4.5}$$

To ensure $z^\top \hat{\theta}_t > 0$ for all $z \in \mathcal{Z}^+_{\theta^*}$, this must hold for the worst case, leading to the sufficient condition:

$$\frac{1}{\mathcal{C} \cdot \sqrt{\log\big(t/\delta\big)}} \geq \max_{z\in\mathcal{Z}} \frac{\|z\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}}}{z^\top \theta^*} = \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(t\sum_{z\in\mathcal{Z}} \lambda_{z,t} \dot{\mu}(z^\top\theta^*) zz^\top\right)^{-1}}}{z^\top \theta^*}$$

where $\lambda_{z,t} = \sum_{s=1}^{t} \mathbb{1}\{z_s = z\}/t$ is the empirical frequency of arm z till time t. This motivates choosing a design that satisfies this sufficient condition as quickly as possible,

$$\lambda^*_+ = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top\theta^*) zz^\top\right)^{-1}}}{z^\top \theta^*}. \tag{4.6}$$

Similarly, for an arm $z \in \mathcal{Z}^-_{\theta^*}$, a parallel argument leads to the design:

$$\lambda^*_- = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top\theta^*) zz^\top\right)^{-1}}}{-z^\top \theta^*}. \tag{4.7}$$

Combining the objectives from Equation (4.6) and Equation (4.7) yields a unified design that ensures the parameter estimate converges to the true parameter θ^* at a fast rate,

$$\lambda^* = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top\theta^*) zz^\top\right)^{-1}}}{\left|z^\top \theta^*\right|}.$$

Comparison to G-Optimal and XY-Optimal Designs. Readers familiar with experimental design methods, particularly in multi-armed bandits, may notice a similarity between our design objective and the G-optimal or XY-optimal design objective as frequently used

in best arm identification problems (Soare et al., 2014; Zhao et al., 2024; Fiez et al., 2019; Yang and Tan, 2022). However, there are crucial differences. The G-optimal design aims to minimize the maximum prediction variance,

$$\arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{Z}} \|z\|_{(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)^{-1}}.$$

The core distinction in our objective is the instance-dependent denominator $|z^\top \theta^*|$, which is absent in the G-optimal design objective. As explained in Section 4.3.2, this term augments the G-optimal design objective by considering the confidence interval's position relative to the decision boundary. This augmentation is the key to achieving our instance-dependent results in the next section, moving beyond worst-case guarantees (Das et al., 2024; Mukherjee et al., 2024; Liu et al., 2024).

Our objective also differs from the XY-optimal design (Fiez et al., 2019; Zhao et al., 2024), which is defined as,

$$\arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z_1, z_2 \in \mathcal{Z}} \frac{\|z_1 - z_2\|_{(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)^{-1}}}{|(z_1 - z_2)^\top \theta^*|}.$$

While the XY-optimal design also has an instance-dependent denominator, it is tailored specifically for best arm identification, as the allocation focuses on reducing uncertainty on the particular direction of $(z_1 - z_2)$ for all $z_1, z_2 \in \mathcal{Z}$. This is suitable for identifying a single winner among all arms. In contrast, our objective is designed to classify all arms, focusing on resolving the uncertainty of each individual arm z with respect to the decision boundary. There is also other difference between our problem and the best arm identification problem, which we explain in the remark of next section.

4.4.2 Algorithm Design with Unknown θ^*

Guided by the oracle allocation derived in the previous section, we now present an adaptive algorithm that approximates this allocation without knowing the true parameter θ^*. Our approach is inspired by the sequential experimental design framework (Fiez et al., 2019; Jun

et al., 2021; Zhao et al., 2024), which has been successfully applied to various multi-armed bandit problems with linear structures.

The key idea is that to classify all arms with high confidence, we need to shrink the confidence interval for each $z^\top\theta^*$ until it no longer contains zero. Once an interval is separated from zero, the sign of the corresponding $z^\top\theta^*$ is determined with high confidence Equation (4.4)-Equation (4.5). The number of samples required for each arm $z \in \mathcal{Z}$ depends on the magnitude of $z^\top\theta^*$. The oracle design from the previous section provides an allocation that achieves this separation efficiently by adapting to the specific structure of the problem instance.

The full algorithm is presented in Algorithm 14. It proceeds in two phases: a warm-up procedure to obtain an initial estimate of θ^*, and a main sequential experimental design procedure. The warm-up phase, detailed in Algorithm 13, collects an initial dataset by sampling according to a G-optimal design.

The main phase of the algorithm operates in rounds. At each round ℓ, we maintain a set of "active" arms, $\mathcal{Z}_\ell$, which includes all arms whose signs have not yet been confidently classified by the previous estimate, $\hat{\theta}_{\ell-1}$. We start with $\mathcal{Z}_1 = \mathcal{Z}$. The algorithm also uses an exponentially decreasing precision parameter, ε_ℓ, that controls the effective distance to the decision boundary for the active set. In each round, we compute a new design, λ_ℓ, based on the estimate $\hat{\theta}_{\ell-1}$ and the active set $\mathcal{Z}_\ell$. Arms that are far from the decision boundary (i.e., where $|z^\top\theta^*|$ is large) are expected to be classified with high confidence in early rounds and are thus removed from the active set.

We note that the warm-up procedure requires knowledge of a lower bound on the minimum derivative of the link function at the true parameter.

Algorithm 13: Warm-up procedure

Input: $\mathcal{Z}, \delta, \omega, \kappa_0 = \min_{z\in\mathcal{Z}} \dot{\mu}(z^\top\theta^*)$

Initialize: $\lambda_0 = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \|z\|^2_{(\sum_{z\in\mathcal{Z}} \lambda_z zz^\top)^{-1}}$

$N_0 = \left\lceil 3(1+\omega)\kappa_0^{-1} d\gamma(d) \log\Big(2|\mathcal{Z}|\big(2+|\mathcal{Z}|\big)/\delta\Big)\right\rceil$

Pull arms in $Z_{N_0} = \text{ROUND}(\lambda_0, N_0)$ and observe Y_{N_0}

Output: MLE $\hat{\theta}_0$ by Equation (4.2)

Algorithm 14: Experimental design algorithm

Input: $\mathcal{Z}, \delta, \omega$

Initialize: $\ell = 1, \mathcal{Z}_1 = \mathcal{Z}, \varepsilon_\ell = 2^{-\ell+1}, Q = \varnothing$

$\hat{\theta}_0$ = Algorithm 13 (warm-up procedure)

Set:

$f_1(\hat{\theta}_{\ell-1}, \lambda) := \gamma(d) \max_{z \in \mathcal{Z}} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}}$

$f_2(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda) := 2.4^2 \varepsilon_\ell^{-2} \max_{z \in \mathcal{Z}_\ell} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}}$

$f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda) := f_1(\hat{\theta}_{\ell-1}, \lambda) \vee f_2(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$

while $|\mathcal{Z}_\ell| \geq 1$ **do**

$\lambda_\ell = \arg\min_{\lambda \in \Delta(\mathcal{Z})} f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$

$\rho(\mathcal{Z}_\ell) = \min_{\lambda \in \Delta(\mathcal{Z})} f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$

$N_\ell = \left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell) \log\left(\frac{2\ell^2 |\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)\right\rceil \vee r(\omega)$

$Z_\ell = \text{ROUND}(\lambda_\ell, N_\ell)$

$Q \leftarrow Q \cup Z_\ell$

Pull arms in Z_ℓ and observe Y_ℓ

Compute MLE $\hat{\theta}_\ell$ with data $\{Z_\ell, Y_\ell\}$ by Equation (4.2)

$\mathcal{Z}_{\ell+1} = \mathcal{Z}_\ell \backslash \left\{ z \in \mathcal{Z}_\ell \middle| \left| z^\top \hat{\theta}_\ell \right| > \varepsilon_\ell \right\}$

$\ell \leftarrow \ell + 1$

Pull arms in Q and observe Y

Compute MLE $\hat{\theta}$ with data $\{Q, Y\}$ by Equation (4.2)

Output: $\hat{\theta}$

Remark 7. *The final, non-adaptive sampling phase in Algorithm 14 (lines 19-20) is a critical step. It is necessitated by a technical requirement of the confidence bound in Theorem 24 and same for many other confidence bounds (Abbasi-Yadkori et al., 2011), which requires that data is collected non-adaptively. That is, the selection of arms cannot depend on previous observations. Consequently, the confidence guarantee would not hold if we were to compute the final MLE using only the data gathered adaptively during the main rounds. This requirement distinguishes our setting from many standard sequential design problems (Fiez et al., 2019; Jun et al., 2021; Zhao et al., 2024), where the goal is to identify a single best arm. In contrast, our objective is to output a reward model, which necessitates a valid confidence bound on the final parameter estimate itself.*

We show the following two instance dependent quantities that essentially capture the complexity of the problem instance,

$$\rho^* = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{Z}} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{|z^\top \theta^*|^2},$$

and

$$\rho_0 = \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{Z}} 3\gamma(d)\|z\|^2_{H(\lambda,\theta^*)^{-1}}.$$

We define the following shorts for simplicity before we are ready to present an upper bound for the label complexity of Algorithm 14, $\ell^* = \lceil \log(4\Delta^{-1}) \rceil$, $\Delta = \min_{z \in \mathcal{Z}} |z^\top \theta^*|$, $\overline{\log} = \log\left(\frac{2\ell^{*2}|\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)$.

Theorem 25. *(Label complexity) With a probability of at least $1-\delta$, Algorithm 14 will stop after at most $c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega)$ rounds, where c is an absolute constant.*

Theorem 26. *(Correctness) With a probability of at least $1-\delta$, the reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, i.e.,*

$$\operatorname{sign}(z^\top \hat{\theta}) = \operatorname{sign}(z^\top \theta^*) \quad \text{for } \forall z \in \mathcal{Z}.$$

Our result stands in stark contrast to existing work on active learning for RLHF. Prior works (Das et al., 2024; Kveton et al., 2025; Mehta et al., 2023; Scheid et al., 2024), for instance, have analyzed various performance metrics, such as suboptimality gap, simple regret, and logit error, but have only provided worst-case guarantees. Their convergence rates, typically on the order of $O(d/\sqrt{T})$, lack dependence on the problem instance and therefore fail to capture the true complexity of the learning problem.

In contrast, our results are instance-dependent. This is because our approach identifies an experimental design objective that is intrinsically tied to the problem structure. Previous works have relied on standard, off-the-shelf objectives like G-optimality and D-optimality. While these are powerful general purpose tools, they are not specifically tailored to the active learning problem in RLHF, which limits the tightness of their theoretical guarantees.

The instance-dependent quantity $\rho^\star$ is analogous to complexity terms that arise in many best-arm identification problems with linear bandits (Jun et al., 2021; Fiez et al., 2019; Zhao et al., 2024). This term captures the problem's complexity through the geometry of the arm set and the true parameter θ^*. For linear bandits, Fiez et al. (2019); Zhao et al. (2024) have shown that such a term characterizes the optimal sample complexity via a minimax lower bound. However, extending this analysis to the logistic model is not straightforward. As shown by Jun et al. (2021), establishing a similar minimax lower bound for logistic bandits is highly non-trivial. The primary difficulty is that the KL-divergence between two Bernoulli distributions is not a simple function of the squared difference of their means. To the best of our knowledge, a minimax lower bound for best-arm identification in logistic bandits has not yet been established. We therefore take a different approach and provide an information-theoretic lower bound for this problem.

Theorem 27. *Define a cone*

$$\mathcal{C}_{\theta^*,\mathcal{Z}} := \left\{\theta \in \mathbb{R}^d : \exists\, z \in \mathcal{Z},\ \operatorname{sign}(z^\top \theta^*) \neq \operatorname{sign}(\theta^\top z)\right\}.$$

Then for any δ–PAC algorithm to the preference learning problem, the expected label complexity $\mathbb{E}_{\theta^*}[\tau]$ must satisfy

$$\mathbb{E}_{\theta^*}[\tau] \geq \log \frac{1}{2.4\,\delta} \min_{\lambda \in \Delta_n} \max_{\theta \in C_{\theta^*,Z}} \frac{1}{\sum_{i=1}^{n} \lambda_i \, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big)},$$

where $P_{\theta,i}$ is the distribution of the preference feedback when pulling arm z_i under parameter θ.

4.5 Greedy Algorithm

In this section, we propose a simple greedy algorithm for active preference learning problem that is directly motivated by the intuition from Section 4.3.2. In contrast to the method in Section 4.4, this approach avoids solving a complex optimization problem and do not require any knowledge on the minimum derivative of the link function. The key idea is to use a greedy objective that incorporates the position of each confidence interval relative to the decision boundary. As argued in Section 4.3.2, relying solely on confidence width can lead to inefficient sampling. Our objective therefore uses the interval's position to augment its width.

Specifically, for each arm z, we define its *remaining uncertainty* as the minimum remaining amount its confidence interval must shrink to no longer contain the decision boundary (zero). This value captures the uncertainty that must be resolved to classify the arm confidently. Formally, the remaining uncertainty of arm z at round t is defined as:

$$\mathsf{RU}_t(z) = \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\},$$

where $\mathsf{LCB}_t(z)$ and $\mathsf{UCB}_t(z)$ are the lower and upper confidence bounds of $z^\top\theta^*$ at round t respectively, i.e., $\mathsf{LCB}_t(z) = z^\top\hat{\theta}_t - D_{\delta,t}(z)$ and $\mathsf{UCB}_t(z) = z^\top\hat{\theta}_t + D_{\delta,t}(z)$. This concept is illustrated in Figure 4.2, which depicts four possible scenarios for the confidence interval positions.

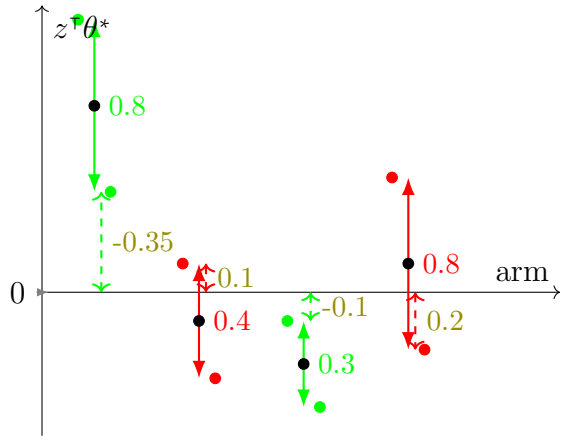


Figure 4.2: Illustration of remaining uncertainty (in olive). The remaining uncertainty is positive only when a confidence interval contains the decision boundary (zero), indicating that the arm's sign is not yet confidently determined.

Based on this concept, we present our algorithm in Algorithm 15. The algorithm follows a simple greedy strategy: in each round, it pulls the arm with the maximum remaining uncertainty.

Algorithm 15: Greedy Algorithm for Active Preference Learning

Input: $\mathcal{Z}$, δ, labeling budget T
Initialize: $\mathcal{D} \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  Compute $\hat{\theta}_t$ and $(\mathsf{LCB}_t(z), \mathsf{UCB}_t(z))$ for all $z \in \mathcal{Z}$
  $z_t \leftarrow \arg\max_{z \in \mathcal{Z}} \min\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\}$ with ties broken arbitrarily
  Query z_t to observe preference y_t
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(z_t, y_t)\}$
Output: MLE $\hat{\theta}$ on $\mathcal{D}$ using Equation (4.2)

The greedy objective in Algorithm 15 can be interpreted as a refined measure of uncertainty for this problem. It refines the standard confidence width by also incorporating the confidence interval's position relative to the decision boundary, as shown in the following theorem.

Theorem 28. *The greedy objective in Algorithm 15 can be equivalently written as*

$$\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} = \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}.$$

To demonstrate the advantage of our algorithm over a G-optimal based method e.g., Das et al. (2024), we show a simple canonical instance where their respective label complexities

differ significantly. Consider the following instance. Let ε be a small constant, e.g., $\varepsilon \in (0, 1/4)$. The arms are the canonical basis vectors, $z_i = e_i$ for $i = 1, \ldots, d$. The true parameter is set to $\theta^* = (1, \ldots, 1, \varepsilon)^\top$. Using the stopping rule from Lemma 29.5, we have the following result.

Theorem 29. *For the canonical instance $z_i = e_i$ for $i = 1, \ldots, d$ and $\theta^* = (1, \ldots, 1, \varepsilon)^\top$ for $\varepsilon \in (0, 1/4)$, the greedy selection objective $\arg\max_{i \in [d]} \|e_i\|_{H_t'^{-1}}$ has label complexity scaling as $\tilde{O}\big(\frac{d}{\varepsilon^2}\big)$, while with Algorithm 15, the label complexity scales as $\tilde{O}\big(d + \frac{1}{\varepsilon^2}\big)$.*

4.6 Experiments

We evaluate our proposed active preference learning algorithm by comparing it against several baselines on multiple datasets. We consider the following five algorithms for comparison with our algorithm in Algorithm 15:

- **Random Sampling:** Arms are selected uniformly at random at each round.
- **Uncertainty Sampling:** This method selects arms closest to the decision boundary based on the current estimate, i.e., those with the smallest absolute estimated reward: $\arg\min_z \left| z^\top \hat{\theta}_t \right|$. A simple and widely adopted heuristic in active learning (Lewis, 1995; Tong and Koller, 2001; Mussmann and Liang, 2018; Raj and Bach, 2022), it has proven effective in various settings.
- **Selective Sampling:** A popular streaming setting active learning framework where the learner decides whether to query an human labeler as each arm arrives. We adapt it to our setting by iterating through the arms in a shuffled order and selecting the first arm z whose lower confidence bound is below a threshold, e.g., $\mathsf{LCB}_t(z) < 0.1$. Note that all true rewards $z^\top \theta^*$ are positive in our setup, since we construct each z as the difference between the chosen and rejected answer.
- **APO (Das et al., 2024):** This method pulls the arm that is most uncertain under the current model, measured by the width of the confidence ellipsoid in its direction: $\arg\max_z \|z\|_{H'(\{z_s, s \in [t]\}, \hat{\theta}_t)^{-1}}$. The same approach is also used in Li et al. (2025).

- **D-optimal experimental design:** This common experimental design criterion, used in Liu et al. (2024); Kveton et al. (2025), selects the arm that greedily maximizes the determinant of the information matrix i.e., $\arg\max_z \det\left(\sum_{s=1}^{t} \dot{\mu}(z_s^\intercal \hat{\theta}_t) z_s z_s^\intercal\right)$. Note that Kveton et al. (2025) propose an equivalent formulation that avoids computing the determinant directly.

We evaluate all methods on three widely-used RLHF preference datasets: the Anthropic helpful and harmless dataset (Bai et al., 2022), the Nectar dataset (Zhu et al., 2024), and the ultrafeedback-binarized-preferences-cleaned dataset from Bartolome et al. (2023). For all datasets, we use Gemma2 (Team, 2024) to generate text embeddings for the content.

Following practical considerations for real-world applications, we implement all algorithms in a batched setting, as a fully sequential, round-by-round implementation is often infeasible due to labeling latency (Scheid et al., 2024). In each epoch, we sample a batch of 50 arms to query, collect their labels, and then update the model. All reported results are averaged over 50 independent runs.

Note that any given dataset represents a single realization of the underlying label noise, meaning the chosen and rejected labels for each pair are fixed. The randomness in our experimental results, therefore, comes from the random train-test split performed at the start of each independent run.

We report the classification accuracy of the estimated reward model on the test set. The results are presented in Figure 4.3. Our proposed method (Algorithm 15) consistently outperforms all baselines on the Anthropic helpful and harmless and the ultrafeedback-binarized-preferences-cleaned datasets. On the Nectar dataset, our algorithm exhibits a significant advantage in the early stages, while Uncertainty Sampling achieves comparable performance in the later stages.

4.7 Conclusion

In this work, we addressed the problem of active preference learning for RLHF by proposing a novel, tailored experimental design objective. Based on this objective, we developed an efficient algorithm that achieves an instance-dependent label complexity bound, improving

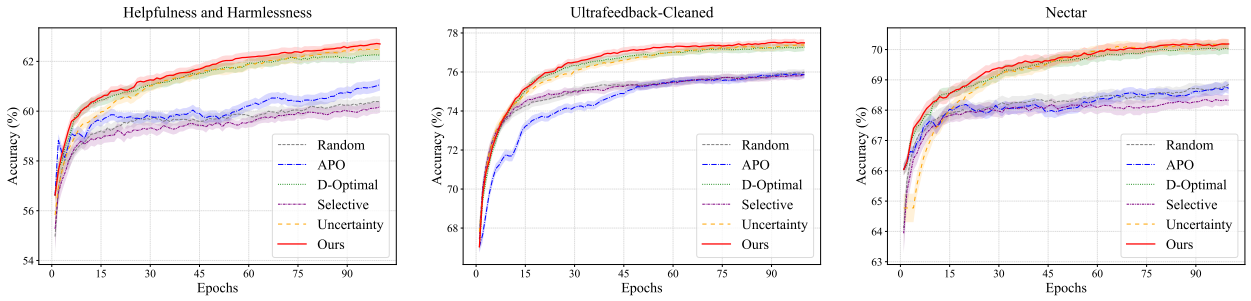


Figure 4.3: Comparison of different active preference learning methods on three datasets.

upon existing worst-case results. We also introduced a simple and practical greedy algorithm as an alternative. Consistent with prior work in this area, we use classification accuracy as a proxy for the performance of the learned reward model. However, a key open question remains: how does the reward model's accuracy impact the final performance of the aligned LLM? Investigating this connection, as well as developing a more precise experimental design objective that directly targets the overall RLHF goal, are compelling directions for future work.

4.8 Proofs

4.8.1 Proof of Theorem 25

Theorem 25. *(Label complexity) With a probability of at least* $1-\delta$*, Algorithm 14 will stop after at most* $c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega)$ *rounds, where* c *is an absolute constant.*

Proof. We prove an upper bound for the label complexity in the following.

Define $\mathcal{S}_\ell := \left\{z \in \mathcal{Z} \middle| |z^\top \theta^*| \le 2\varepsilon_\ell\right\}$. Thus, by Lemma 29.1, with a probability at least $1-\delta$, we have $\bigcap_\ell \{\mathcal{Z}_\ell \subseteq S_\ell\}$ hold. This implies the following is true with a probability at least $1-\delta$

for all ℓ

$$
\begin{aligned}
\rho(\mathcal{Z}_\ell) &= \min_{\lambda \in \Delta(\mathcal{Z})} \max \left\{ \gamma(d) \max_{z \in \mathcal{Z}} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}}, 2.4^2 \varepsilon_\ell^{-2} \max_{z \in \mathcal{Z}_\ell} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}} \right\} \\
&\le \min_{\lambda \in \Delta(\mathcal{Z})} \max \left\{ \gamma(d) \max_{z \in \mathcal{Z}} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}}, 2.4^2 \varepsilon_\ell^{-2} \max_{z \in \mathcal{S}_\ell} \|z\|^2_{H(\lambda, \hat{\theta}_{\ell-1})^{-1}} \right\} && (\mathcal{Z}_\ell \subseteq S_\ell) \\
&\le \min_{\lambda \in \Delta(\mathcal{Z})} \max \left\{ 3\gamma(d) \max_{z \in \mathcal{Z}} \|z\|^2_{H(\lambda, \theta^*)^{-1}}, 3 \cdot 2.4^2 \varepsilon_\ell^{-2} \max_{z \in \mathcal{S}_\ell} \|z\|^2_{H(\lambda, \theta^*)^{-1}} \right\} && \text{(Lemma 29.3)} \\
&\le \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{Z}} 3\gamma(d) \|z\|^2_{H(\lambda, \theta^*)^{-1}} + \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{S}_\ell} 3 \cdot 2.4^2 \varepsilon_\ell^{-2} \|z\|^2_{H(\lambda, \theta^*)^{-1}} \\
&=: \rho_0 + \rho_\ell. && (4.8)
\end{aligned}
$$

By Lemma 29.1, with a probability at least $1 - \delta$, we have $S_\ell = \varnothing$ for $\ell \ge \ell^* = \left\lceil \log(4\Delta^{-1}) \right\rceil$. The total label complexity of Algorithm 14 is the summation of the number of samples in each round, which can be upper bounded as

$$
\begin{aligned}
&\sum_{\ell=1}^{\ell^*} N_\ell \\
&= \sum_{\ell=1}^{\ell^*} \left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell) \log\left(\frac{2\ell^2 |\mathcal{Z}| \left(2 + |\mathcal{Z}|\right)}{\delta} \right) \right\rceil \vee r(\omega) \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega)\rho(\mathcal{Z}_\ell) \overline{\log} \vee r(\omega) \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega)(\rho_0 + \rho_\ell) \overline{\log} \vee r(\omega) && \text{(due to (4.8))} \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega) \overline{\log} \rho_\ell + 4(1+\omega)\ell^* \overline{\log} \rho_0 + \ell^* r(\omega). && (4.9)
\end{aligned}
$$

On the other hand, we have

$$\begin{aligned}
\rho^* &= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{|z^\top\theta^*|^2} \\
&= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{\ell} \max_{z\in S_\ell} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{|z^\top\theta^*|^2} \\
&\geq \frac{1}{\ell^*} \min_{\lambda\in\Delta(\mathcal{Z})} \sum_{\ell=1}^{\ell^*} \max_{z\in S_\ell} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{|z^\top\theta^*|^2} \\
&\geq \frac{1}{16\ell^*} \sum_{\ell=1}^{\ell^*} 2^{2\ell} \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in S_\ell} \|z\|^2_{H(\lambda,\theta^*)^{-1}}. \qquad (4.10)
\end{aligned}$$

Combining (4.9), (4.10), and N_0 for the warm-up, we have

$$\sum_{\ell=0}^{\ell^*} N_\ell \leq c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega),$$

where c is an absolute constant. □

4.8.2 Proof of Theorem 26

Theorem 26. *(Correctness) With a probability of at least $1-\delta$, the reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, i.e.,*

$$\operatorname{sign}(z^\top\hat{\theta}) = \operatorname{sign}(z^\top\theta^*) \quad \text{for } \forall z \in \mathcal{Z}.$$

Proof. The proof is similar to Lemma 29.1 and Lemma 29.2. Readers are recommended to read the proof of Lemma 29.1 and Lemma 29.2 first. The difference is that here we need to show that the final reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, while in Lemma 29.2, each reward model $\hat{\theta}_\ell$ is only shown to classify the arms in $\mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$ correctly. Note that since $\mathcal{Z}_\ell$ will be eventually empty, for any $z \in \mathcal{Z}$, there exists ℓ such that $z \in \mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$. So what we need to show is that for all $z \in \mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$, the reward model $\hat{\theta}$ will classify z correctly.

With a probability of at least $1-\frac{\delta}{2\ell^2|\mathcal{Z}|}$, we have for all $z \in \mathcal{Z}_\ell$,

$$
\begin{aligned}
\left|\left(\hat{\theta}-\theta^*\right)^\top z\right| \overset{a_1}{\leq} & 2.4\|z\|_{H'(Q,\theta^*)^{-1}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\leq & 2.4\|z\|_{H'(\mathcal{Z}_\ell,\theta^*)^{-1}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \qquad (H'(\mathcal{Z}_\ell,\theta^*) \preceq H'(Q,\theta^*)) \\
\overset{a_2}{\leq} & \frac{2.4\sqrt{1+\omega}\|z\|_{H(\lambda_\ell,\theta^*)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\overset{a_3}{\leq} & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
= & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{2.4\sqrt{\left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)\right\rceil \vee r(\omega)}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\leq & \varepsilon_\ell, \qquad (4.11)
\end{aligned}
$$

where b_1 is due to Theorem 24, b_2 is due to Lemma 29.4, b_3 is due to Lemma 29.3. Consider an arm $z \in \mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$, we have $\left|z^\top \hat{\theta}_\ell\right| > \varepsilon_\ell$ by the elimination condition in Algorithm 14. Thus by Equation (4.11) and Lemma 29.5 we know that $\hat{\theta}_\ell$ will classify z correctly. By a union bound same as Equation (4.13), we know that with a probability at least $1-\delta$, all of arms in $\mathcal{Z}$ will be classified correctly by $\hat{\theta}$. □

Lemma 29.1. *With a probability of at least $1-\delta$, for all $z \in \mathcal{Z}_\ell$, we have $|z^\top \theta^*| \leq 2\varepsilon_\ell$, for all $\ell \geq 1$.*

Proof. With a probability of at least $1-\frac{\delta}{2\ell^2|\mathcal{Z}|}$, we have for all $z \in \mathcal{Z}_\ell$,

$$\begin{aligned}
\left|z^\top\left(\hat{\theta}_\ell - \theta^*\right)\right| \overset{a_1}{\leq} & 2.4\|z\|_{H'(\mathcal{Z}_\ell,\theta^*)^{-1}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\overset{a_2}{\leq} & \frac{2.4\sqrt{1+\omega}\|z\|_{H(\lambda_\ell,\theta^*)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\overset{a_3}{\leq} & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
= & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{2.4\sqrt{\left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)\right\rceil \vee r(\omega)}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\leq & \varepsilon_\ell, \qquad (4.12)
\end{aligned}$$

where b_1 is due to Theorem 24, b_2 is due to Lemma 29.4, b_3 is due to Lemma 29.3.

Define a good event $\mathcal{E}_{\ell,z}$ for each ℓ and $z \in \mathcal{Z}_\ell$ as

$$\mathcal{E}_{\ell,z} = \left\{\left|z^\top\left(\hat{\theta}_\ell - \theta^*\right)\right| \leq \varepsilon_\ell\right\}.$$

We claim that with a probability at least $1-\delta$, the event $\bigcap_{\ell=1}^{\infty}\bigcap_{z\in\mathcal{Z}_\ell}\mathcal{E}_{\ell,z}$ holds. It can be proved by a union bound as follows,

$$\begin{aligned}
\mathbb{P}\left(\left(\bigcap_{\ell=1}^{\infty}\bigcap_{z\in\mathcal{Z}_\ell}\mathcal{E}_{\ell,z}\right)^c\right) \leq & \sum_{\ell=1}^{\infty}\sum_{z\in\mathcal{Z}_\ell}\mathbb{P}\left(\mathcal{E}_{\ell,z}^c\right) \\
\leq & \sum_{\ell=1}^{\infty}\sum_{z\in\mathcal{Z}_\ell}\frac{\delta}{2\ell^2|\mathcal{Z}|} \\
\leq & \frac{\delta}{2}\sum_{\ell=1}^{\infty}\frac{1}{\ell^2} \\
\leq & \delta, \qquad (4.13)
\end{aligned}$$

where the last step is by the fact that $\sum_{\ell=1}^{\infty}\frac{1}{\ell^2} = \frac{\pi^2}{6} < 2$. Under the event $\bigcap_{\ell=1}^{\infty}\bigcap_{z\in\mathcal{Z}_\ell}\mathcal{E}_{\ell,z}$, we show the following.

For all $z \in \mathcal{Z}_\ell$ satisfying $|z^\top \theta^*| > 2\varepsilon_\ell$, we have, if $z^\top \theta^* > 2\varepsilon_\ell$,

$$\begin{aligned} z^\top \hat{\theta}_\ell =& z^\top \left(\hat{\theta}_\ell - \theta^* \right) + z^\top \theta^* \\ >& -\varepsilon_\ell + 2\varepsilon_\ell \\ >& \varepsilon_\ell. \end{aligned} \tag{4.14}$$

If $z^\top \theta^* < -2\varepsilon_\ell$,

$$\begin{aligned} z^\top \hat{\theta}_\ell =& z^\top \left(\hat{\theta}_\ell - \theta^* \right) + z^\top \theta^* \\ <& \varepsilon_\ell - 2\varepsilon_\ell \\ <& -\varepsilon_\ell. \end{aligned} \tag{4.15}$$

Combine Equation (4.14) and Equation (4.15), we have $\left| z^\top \hat{\theta}_\ell \right| > \varepsilon_\ell$, which means z satisfying $|z^\top \theta^*| > 2\varepsilon_\ell$ will be removed from $\mathcal{Z}_\ell$ in the end of round ℓ according to the elimination condition in Algorithm 14. □

Lemma 29.2. *For all $\ell \geq 1$, with a probability at least $1 - \delta$, the arms in $\mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$ are all classified correctly by $\hat{\theta}_\ell$.*

Proof. This result directly follows from Equation (4.12), Lemma 29.5, and the elimination condition in Algorithm 14.

By Equation (4.12), we know that the confidence width of $z^\top \theta^*$ is at most ε_ℓ for all $z \in \mathcal{Z}_\ell$. By the elimination condition in Algorithm 14, we have $\left| z^\top \hat{\theta}_\ell \right| > \varepsilon_\ell$ for all $z \in \mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$, thus the sufficient condition in Lemma 29.5 is satisfied, which means $\hat{\theta}_\ell$ will classify all arms in $\mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$ correctly with a probability at least $1 - \delta$. □

Lemma 29.3. *(Lemma 7 of Jun et al. (2021)) For each $\ell \geq 0$, define the event,*

$$\mathcal{R}_\ell = \left\{ \frac{1}{3} H(\lambda, \theta^*) \leq H(\lambda, \hat{\theta}_\ell) \leq 3H(\lambda, \theta^*), \forall \lambda \in \Delta(\mathcal{Z}) \right\}.$$

We have that

- $\mathbb{P}(\mathcal{R}_0) \geq 1 - 2\delta$ *and* $\mathbb{P}\left(\mathcal{R}_\ell \mid \mathcal{R}_{\ell-1}, \cdots, \mathcal{R}_0\right) \geq 1 - 2\delta$ *for all* $\ell \geq 1$*;*

- $\max_{z\in\mathcal{Z}} \|z\|^2_{H'(\mathcal{Z}_\ell,\theta^*)^{-1}} \le 1/\gamma(d)$.

Lemma 29.4. *(Lemma 13 of Jun et al. (2021)) Assume that $\lambda \in \Delta_{\mathcal{X}}$, and that we have sampled $x_1,\dots,x_n \sim ROUND(\lambda,n,\varepsilon)$ with $n \ge r(\varepsilon) = \frac{(d(d+1)+2)}{\varepsilon}$, and $\varepsilon \le 1$. Then, for any θ,*

$$\sum_{s=1}^{n} \dot{\mu}(x_s^\top\theta)x_s x_s^\top \succeq \frac{n}{1+\varepsilon}\sum_{x\in\mathcal{X}} \lambda_x \dot{\mu}(x^\top\theta)xx^\top.$$

This in particular implies

- *For any z,*
$$\|z\|^2_{\left(\sum_{s=1}^n \dot{\mu}(x_s^\top\theta)x_s x_s^\top\right)^{-1}} \le \frac{(1+\varepsilon)}{n}\|z\|^2_{\left(\sum_{x\in\mathcal{X}} \lambda_x\dot{\mu}(x^\top\theta)xx^\top\right)^{-1}}$$
- $$\lambda_{\min}\left(\sum_{s=1}^{n} \dot{\mu}(x_s^\top\theta)x_s x_s^\top\right) \ge \frac{n}{1+\varepsilon}\lambda_{\min}\left(\sum_{x\in\mathcal{X}} \lambda_x\dot{\mu}(x^\top\theta)xx^\top\right)$$

Lemma 29.5. *Let $D_\delta(z)$ be any valid confidence width for the estimated model $\hat{\theta}$ on arm z with a probability of at least $1-\delta$, i.e., with a probability of at least $1-\delta$, we have*

$$z^\top\hat{\theta} - D_\delta(z) < z^\top\theta < z^\top\hat{\theta} + D_\delta(z).$$

Then the following condition

$$D_\delta(z) < \left|z^\top\hat{\theta}\right|$$

is a sufficient condition to guarantee that the reward model $\hat{\theta}$ makes a correct classification on arm z with a probability of at least $1-\delta$ i.e.,

$$\text{sign}(z^\top\hat{\theta}) = \text{sign}(z^\top\theta).$$

Proof. Suppose that $D_\delta(z) < \left|z^\top\hat{\theta}\right|$. We now show that $z^\top\hat{\theta}$ and $z^\top\theta$ share the same sign.

Case 1: $z^\top \hat{\theta} > 0$.

Since $D_\delta(z) < z^\top \hat{\theta}$, we have

$$z^\top \hat{\theta} - D_\delta(z) > 0.$$

Hence, with a probability of at least $1 - \delta$,

$$z^\top \theta > z^\top \hat{\theta} - D_\delta(z) > 0,$$

so $z^\top \theta > 0$ and the classification is correct.

Case 2: $z^\top \hat{\theta} < 0$.

Here $-z^\top \hat{\theta} > 0$ and $D_\delta(z) < -z^\top \hat{\theta}$, so

$$z^\top \hat{\theta} + D_\delta(z) < 0.$$

Hence, with a probability of at least $1 - \delta$,

$$z^\top \theta < z^\top \hat{\theta} + D_\delta(z) < 0,$$

so $z^\top \theta < 0$ and again the classification is correct. In both cases the sign of $z^\top \hat{\theta}$ matches that of $z^\top \theta$. □

4.8.3 Proof of Theorem 27

Theorem 27. *Define a cone*

$$\mathcal{C}_{\theta^*,\mathcal{Z}} := \left\{ \theta \in \mathbb{R}^d : \exists\, z \in \mathcal{Z},\ \operatorname{sign}(z^\top \theta^*) \neq \operatorname{sign}(\theta^\top z) \right\}.$$

Then for any δ–PAC algorithm to the preference learning problem, the expected label complexity $\mathbb{E}_{\theta^*}[\tau]$ must satisfy

$$\mathbb{E}_{\theta^*}[\tau] \geq \log \frac{1}{2.4\,\delta} \min_{\lambda \in \Delta_n} \max_{\theta \in C_{\theta^*,Z}} \frac{1}{\sum_{i=1}^n \lambda_i\, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big)},$$

where $P_{\theta,i}$ is the distribution of the preference feedback when pulling arm z_i under parameter θ.

Proof. Let $\mathcal{Z} = \{z_1, \ldots, z_n\} \subset \mathbb{R}^d$, and θ^* be the true parameter. Let T_i denote the random variable, which is the number of times arm i is pulled. Under any δ–PAC algorithm, by transportation lemma of Kaufmann et al. (2016), for any $\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}$, we have

$$\sum_{i=1}^{n} \mathbb{E}[T_i]\, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \log \frac{1}{2.4\,\delta},$$

where $P_{\theta,i}$ is the distribution of the preference feedback when pulling arm z_i under parameter θ. In our case, we have

$$P_{\theta,i} = \mathrm{Bern}\big(p_i(\theta)\big), \quad p_i(\theta) = \sigma(z_i^\top \theta), \quad \sigma(z) = \frac{1}{1+e^{-z}}.$$

For any feasible solution $\mathbf{t} = (t_1, \ldots, t_n)$ to the following optimization problem,

$$\begin{aligned} &\min_{t} \sum_{i=1}^{n} t_i \\ \text{s.t.} \quad &\min_{\theta \in \mathcal{C}_{\theta^*,Z}} \sum_{i=1}^{n} t_i\, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \log \frac{1}{2.4\,\delta}. \end{aligned}$$

we have

$$\sum_{i=1}^{n} \mathbb{E}[T_i] \;\geq\; \sum_{i=1}^{n} t_i.$$

In particular, taking $\mathbf{t}^*$ to be an optimal solution to this problem,

$$\min_{\theta \in C_{\theta^*,Z}} \sum_{i=1}^{n} \frac{t_i^*}{\sum_{j=1}^{n} t_j^*}\, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \frac{\log\big(1/2.4\delta\big)}{\sum_{j=1}^{n} t_j^*} \;\geq\; \frac{\log\big(1/2.4\delta\big)}{\sum_{j=1}^{n} \mathbb{E}[T_j]}.$$

Since $\sum_{i=1}^{n} \frac{t_i^*}{\sum_{j=1}^{n} t_j^*} = 1$, we conclude

$$\max_{\lambda \in \Delta_n} \min_{\theta \in C_{\theta^*,Z}} \sum_{i=1}^{n} \lambda_i\, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \frac{\log\big(1/2.4\delta\big)}{\sum_{i=1}^{n} \mathbb{E}[T_i]}.$$

Rearranging the terms, we have

$$\mathbb{E}_{\theta^*}[\tau] = \sum_{i=1}^{n} \mathbb{E}[T_i] \geq \log \frac{1}{2.4\,\delta} \min_{\lambda \in \Delta_n} \max_{\theta \in C_{\theta^*, Z}} \frac{1}{\sum_{i=1}^{n} \lambda_i \, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big)}.$$

□

4.8.4 Proof of Theorem 28

Theorem 28. *The greedy objective in Algorithm 15 can be equivalently written as*

$$\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} = \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}.$$

Proof.

$$\begin{aligned}
&\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} \\
&= \max_z \min\left\{-(\hat{\theta}_t^\top z - D_{\delta,t}(z)), \hat{\theta}_t^\top z + D_{\delta,t}(z)\right\} \\
&= \max_z \min\left\{-\hat{\theta}_t^\top z + D_{\delta,t}(z), \hat{\theta}_t^\top z + D_{\delta,t}(z)\right\} \\
&= \max_z \left\{\min\left\{-\hat{\theta}_t^\top z, \hat{\theta}_t^\top z\right\} + D_{\delta,t}(z)\right\} \\
&= \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}.
\end{aligned}$$

□

4.8.5 Proof of Theorem 29

Theorem 29. *For the canonical instance $z_i = e_i$ for $i = 1, \ldots, d$ and $\theta^* = (1, \ldots, 1, \varepsilon)^\top$ for $\varepsilon \in (0, 1/4)$, the greedy selection objective $\arg\max_{i \in [d]} \|e_i\|_{H_t'^{-1}}$ has label complexity scaling as $\tilde{O}\big(\frac{d}{\varepsilon^2}\big)$, while with Algorithm 15, the label complexity scales as $\tilde{O}\big(d + \frac{1}{\varepsilon^2}\big)$.*

Proof. We begin by establishing a sufficient condition for the algorithms to stop. Our analysis considers the canonical instance where the arm set $\mathcal{Z}$ consists of the standard basis vectors $\{e_1, \ldots, e_d\}$. In this setting, the two algorithms greedily pull arms based on their respective selection rules:

- **G-optimal based method (Das et al., 2024):** At each round t, select arm $\arg\max_{i\in[d]} \|e_i\|_{H_t'^{-1}}$, where $H_t' = \sum_{s=1}^{t} \dot{\mu}(z_s^\top \hat{\theta}_t) z_s z_s^\top$.
- **Our method:** At each round t, select arm $\arg\max_{i\in[d]} \{D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t|\}$.

From Lemma 29.5, the suggested stopping rule is

$$D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t| < 0, \quad \text{for all } i \in [d], \tag{4.16}$$

where $D_{\delta,t}(e_i)$ is the confidence width for arm e_i at round t. Since this rule involves the empirical estimate $\hat{\theta}_t$, we first derive a simpler sufficient condition that depends only on the true parameter θ^*.

We show that the condition

$$D_{\delta,t}(e_i) < \frac{1}{2}|e_i^\top \theta^*|, \quad \text{for all } i \in [d], \tag{4.17}$$

is sufficient for the stopping rule in Equation (4.16) to be satisfied. By definition of the confidence width, we know that with probability at least $1-\delta$, $|e_i^\top \hat{\theta}_t - e_i^\top \theta^*| \le D_{\delta,t}(e_i)$. Using the triangle inequality, we have:

$$\begin{aligned} |e_i^\top \hat{\theta}_t| &\ge |e_i^\top \theta^*| - |e_i^\top \hat{\theta}_t - e_i^\top \theta^*| \\ &\ge |e_i^\top \theta^*| - D_{\delta,t}(e_i). \end{aligned}$$

If we assume Equation (4.17) holds, then $|e_i^\top \theta^*| > 2D_{\delta,t}(e_i)$. Substituting this into the inequality above gives:

$$|e_i^\top \hat{\theta}_t| > 2D_{\delta,t}(e_i) - D_{\delta,t}(e_i) = D_{\delta,t}(e_i).$$

This directly implies that $D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t| < 0$, satisfying the stopping rule.

Now, we use this sufficient condition to determine the number of pulls t_i required for each arm e_i. For the canonical instance, Lemma 29.6 allows us to express the condition in

Equation (4.17) as:

$$2.4\sqrt{\frac{1}{\dot{\mu}(\theta_i^*)t_i}}\sqrt{\log\frac{2(2+|\mathcal{Z}|)}{\delta}} < \frac{1}{2}|\theta_i^*|, \quad \text{for all } i \in [d]. \tag{4.18}$$

Solving for t_i, we find that each arm i must be pulled a number of times satisfying:

$$t_i > \frac{\mathcal{C}}{\dot{\mu}(\theta_i^*)(\theta_i^*)^2}\log\frac{2(2+|\mathcal{Z}|)}{\delta},$$

where $\mathcal{C}$ is a numerical constant.

We can now analyze the total sample complexity for each algorithm.

G-optimal Method As established in Lemma 29.7, the G-optimal method behaves as a round-robin for the canonical instance, meaning $t_i \approx T/d$ for all i after T rounds. To satisfy the stopping condition for all arms, the number of pulls for the arm with the smallest $|\theta_d^*| = \varepsilon$ must meet the requirement. This leads to a total sample complexity T of:

$$T \approx d \cdot t_d > d \cdot \frac{\mathcal{C}}{\varepsilon^2}\log\left(\frac{|\mathcal{Z}|}{\delta}\right) \implies T = \tilde{O}\left(\frac{d}{\varepsilon^2}\right).$$

The algorithm continues to pull all arms in a round-robin fashion until the condition is met for every arm, including those that may have satisfied it much earlier.

Our Method A fundamental difference in our algorithm is that its selection rule, $\arg\max_i\{D_{\delta,t}(e_i) - |e_i^\top\hat{\theta}_t|\}$, is directly aligned with the stopping rule. Once an arm e_i satisfies the stopping condition $D_{\delta,t}(e_i) - |e_i^\top\hat{\theta}_t| < 0$, its selection metric becomes negative and it will not be pulled again, as the algorithm prioritizes arms with larger, positive values. The algorithm thus focuses its budget on arms that have not yet met the condition.

The total sample complexity is the sum of the required pulls for each arm. Let $|\theta_d^*| = \varepsilon$. The complexity is dominated by the pulls required for this arm, while other arms require fewer pulls.

$$T = \sum_{i=1}^{d} t_i = \sum_{i=1}^{d}\tilde{O}\left(\frac{1}{(\theta_i^*)^2}\right) = \tilde{O}\left(\sum_{i=1}^{d-1}\frac{1}{(\theta_i^*)^2} + \frac{1}{\varepsilon^2}\right) = \tilde{O}\left(d + \frac{1}{\varepsilon^2}\right).$$

This demonstrates a significant improvement over the G-optimal method when ε is small. □

Lemma 29.6. *For the canonical instance where $z_i = e_i$ for $i = 1, \cdots, d$, the confidence interval in Theorem 24 can reduce to*

$$\left|\hat{\theta}_{t,i} - \theta_i^*\right| < 2.4\sqrt{\frac{1}{\dot{\mu}(\theta_i^*)t_i}}\sqrt{\log\frac{2(2+|\mathcal{Z}|)}{\delta}}.$$

Proof. The result is straightforward by noting that for the canonical instance and the confidence interval in Theorem 24 □

Lemma 29.7. *For the canonical instance where $z_i = e_i$ for $i = 1, \cdots, d$, and $\theta^* = (1, \ldots, 1, \varepsilon)^\top$, the algorithm in Das et al. (2024) pulls arms almost uniformly, i.e.,*

$$t_1 = \cdots = t_{d-1} \quad \text{and} \quad t_d < t_1 < 2t_d.$$

Proof. For the algorithm of Das et al. (2024), the selection rule is

$$\arg\max_i \|z_i\|_{H'(\{z_s, s\in[t]\}, \theta^*)^{-1}} = \arg\max_i \frac{1}{\dot{\mu}(\theta_i^*)t_i} = \arg\min_i \dot{\mu}(\theta_i^*)t_i.$$

Note that we have

$$\dot{\mu}(\theta_i^*) = \dot{\mu}(1) \text{ for } i = 1, \cdots, d-1, \text{ and } \dot{\mu}(\theta_d^*) = \dot{\mu}(\varepsilon).$$

Also note that for a small ε, e.g., $\varepsilon \in (0, 1/4)$,

$$\dot{\mu}(1) < \dot{\mu}(\varepsilon) < 2\dot{\mu}(1).$$

This implies that the algorithm will pull the arms almost uniformly. Specifically, if t_i is the number of times arm i is pulled, then we have

$$t_1 = \cdots = t_{d-1} \quad \text{and} \quad t_d < t_1 < 2t_d.$$

□

Chapter 5

Conclusion and Future Work

In this dissertation, we studied three problems within the bandit and interactive machine learning framework, with contributions ranging from theory to practice. In Chapter 2, we revisited the simple regret minimization problem in multi-armed bandits. Our results solved an open problem mentioned in Lattimore and Szepesvári (2020) regarding the output distribution of pure exploration algorithms and provided a new algorithm that achieves superior performance, even in the challenging data-poor regime. In Chapter 3, we studied the confounded pure exploration problem, where the experimenter cannot directly intervene on the action variable. We brought a tool from econometrics that is less common in machine learning literature, the instrumental variable, to solve the problem and discovered its power in certain real-world pure exploration settings. We were able to solve the challenge of deriving a non-asymptotic confidence bound for a modified version of the two-stage least squares estimator, which eventually led to a nearly optimal algorithm for the problem. In Chapter 4, we studied the problem of preference learning for Large Language Model alignment. We proposed a new algorithm that actively and adaptively queries a human expert for preference data. Compared to existing methods that use off-the-shelf experimental design objectives, our method is developed from a new objective that is more suitable for the problem, enabling us to show the first instance-dependent guarantee for this task.

Throughout this dissertation, a common theme is the development of strong, instance-dependent performance guarantees for problems arising from the area of bandit and interactive machine learning, especially in the pure exploration setting. Such guarantees are superior to worst-case guarantees, which can be significantly loose for many specific problem instances. Another common theme is that all of the works in this dissertation can be sum-

marized as addressing the problem of efficient data collection. This is important, as in past years we have observed significant progress in machine learning and artificial intelligence from the perspective of algorithm computation, while much less attention has been given to the intelligence of the data collection step. We believe that efficient data collection will play an important role in the future of machine learning and artificial intelligence.

While each chapter has summarized its own directions for future work on its specific problem, there are also common themes for future exploration that could involve combining the different problems mentioned in this dissertation. One interesting direction of potentially high practical value, for example, is to apply the theoretical insights from Chapter 2. This chapter demonstrated a significant acceleration effect with the right algorithm and showed the great potential of pure exploration even in the data-poor regime. These findings could be extremely useful in practical problems like those studied in Chapter 3 and Chapter 4, as these are the settings that practitioners often face and care about. The work of Chapter 2 is currently of primarily theoretical value; it would be very interesting to see how these insights can be translated into more practical problems.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Z. Allen-Zhu, Y. Li, A. Singh, and Y. Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming*, 186:439–478, 2021.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

R. Arratia and L. Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 41–66, 2010.

J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Computational Learning Theory (COLT)*, volume 7, pages 1–122, 2009.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-Armed Bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

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

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 263–272. PMLR, 2017.

M. Aziz, J. Anderton, E. Kaufmann, and J. Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pages 3–24. PMLR, 2018.

M. Aziz, J. Anderton, K. Jamieson, A. Wang, H. Bouchard, and J. Aslam. Identifying New Podcasts with High General Appeal Using a Pure Exploration Infinitely-Armed Bandit Strategy. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 134–144, 2022.

T. Baharav and D. Tse. Ultra fast medoid identification via correlated sequential halving. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019a.

T. Baharav and D. Tse. Ultra fast medoid identification via correlated sequential halving. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019b.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

E. Bakshy, D. Eckles, and M. S. Bernstein. Designing and deploying online field experiments. In *International Conference on World Wide Web*, pages 283–292, 2014.

K. Balagopalan*, N. T. Nguyen*, Y. Zhao*, and K.-S. Jun. Fixing the loose brake: Exponential-tailed stopping time in best arm identification. In *International Conference on Machine Learning*. PMLR, 2025. *Equal contribution.

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

A. Bartolome, G. Martin, and D. Vila. Notus. https://github.com/argilla-io/notus, 2023.

R. E. Bechhofer. A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, and Its Use with Various Experimental Designs. *Biometrics*, 14(3):408–429, 1958.

B. Bilodeau, L. Wang, and D. Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

E. Bradlow. Encouragement designs: an approach to self-selected samples in an experimental design. *Marketing Letters*, 9:383–391, 1998.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, apr 2011. URL https://hal-hec.archives-ouvertes.fr/hal-00609550.

R. Camilleri, K. Jamieson, and J. Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning (ICML)*, pages 1227–1237. PMLR, 2021.

A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. *Journal of Machine Learning Research*, 49(June):590–604, 2016.

A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning (ICML)*, pages 1133–1141. PMLR, 2015.

Y. Chandak, S. Shankar, V. Syrgkanis, and E. Brunskill. Adaptive instrument design for indirect experiments. *arXiv preprint arXiv:2312.02438*, 2023.

A. R. Chaudhuri and S. Kalyanakrishnan. PAC identification of a bandit arm relative to a reward quantile. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1777–1783, 2017.

A. R. Chaudhuri and S. Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 991–1000. PMLR, 2019.

L. Chen, J. Li, and M. Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 101–110. PMLR, 2017a.

L. Chen, J. Li, and M. Qiao. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory (COLT)*, pages 535–592. PMLR, 2017b.

H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30 (3):755–770, 1959.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

R. Combes, M. S. Talebi Mazraeh Shahi, A. Proutiere, et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

I. Danihelka, A. Guez, J. Schrittwieser, and D. Silver. Policy improvement by planning with gumbel. In *International Conference on Learning Representations*, 2022.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

N. Das, S. Chakraborty, A. Pacchiano, and S. R. Chowdhury. Active preference optimization for sample efficient rlhf. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024.

R. De Heide, J. Cheshire, P. Ménard, and A. Carpentier. Bandits with many optimal arms. *Advances in Neural Information Processing Systems*, 34:22457–22469, 2021.

R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

R. Degenne, P. Ménard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

R. Della Vecchia and D. Basu. Online instrumental variable regression: Regret analysis and bandit feedback. *arXiv preprint arXiv:2302.09357*, 2023.

S. Diao, P. Wang, Y. Lin, R. Pan, X. Liu, and T. Zhang. Active prompting with chain-of-thought for large language models. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 1330–1350. Association for Computational Linguistics (ACL), 2024.

V. Dwaracherla, S. M. Asghari, B. Hao, and B. Van Roy. Efficient exploration for llms. *arXiv preprint arXiv:2402.00396*, 2024.

S. Ehrenfeld. On the efficiency of experimental designs. *The annals of mathematical statistics*, 26(2):247–255, 1955.

B. Elbers. Encouragement designs and instrumental variables for a/b testing, August 2023. URL https://engineering.atspotify.com/2023/08/encouragement-designs-and-instrumental-variables-for-a-b-testing/. Accessed: 2024-05-19.

S. Engineering. Encouragement designs and instrumental variables for a/b testing, 2023. Accessed: December 26, 2023.

E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

D. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International conference on machine learning*, pages 3199–3210. PMLR, 2020.

S. B. Gales, S. Sethuraman, and K.-S. Jun. Norm-agnostic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 73–91. PMLR, 2022.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 998–1027, 2016a.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 998–1027, 2016b.

X. Gong and J. Zhang. Dual instrumental method for confounded kernelized bandits. *arXiv preprint arXiv:2209.03224*, 2022.

W. H. Greene. *Econometric analysis*. Pearson Education India, 2003.

A. Hassidim, R. Kupfer, and Y. Singer. An optimal elimination algorithm for learning a best arm. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:10788–10798, 2020.

A. Inoue and G. Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

L. Jain, K. Jamieson, R. Mankoff, R. Nowak, and S. Sievert. The new yorker cartoon caption contest dataset. 2020. URL https://nextml.github.io/caption-contest-data/.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. On finding the largest mean among many. *arXiv preprint arXiv:1306.3917*, 2013.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 423–439, 2014.

K.-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 974–982, 2016. ISBN 9781510829008.

K.-S. Jun and C. Zhang. Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality. *ICML Workshop on Theoretical Foundations of Reinforcement Learning (arXiv:2006.08754)*, 2020.

K. S. Jun and C. Zhang. PopArt: Efficient Sparse Regression and Experimental Design for Optimal Sparse Linear Bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022-December, 2022.

K.-S. Jun, K. Jamieson, R. Nowak, and X. Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

K.-S. Jun, L. Jain, B. Mason, and H. Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pages 5148–5157. PMLR, 2021.

N. Kallus. Instrument-armed bandits. In *Algorithmic Learning Theory*, pages 529–546. PMLR, 2018.

S. Kalyanakrishnan and P. Stone. Efficient selection of multiple bandit arms: Theory and practice. In *International Conference on Machine Learning (ICML)*, 2010.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, volume 12, pages 655–662, 2012.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 1238–1246. PMLR, 2013.

J. Katz-Samuels and K. Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1781–1791, 2020.

J. Katz-Samuels, L. Jain, Z. Karnin, and K. G. Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 10371–10382, 2020.

E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory (COLT)*, pages 228–251. PMLR, 2013.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

E. Kaufmann, P. Ménard, O. D. Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

J. Kiefer and J. Wolfowitz. Optimum designs in regression problems. *The annals of mathematical statistics*, 30(2):271–294, 1959.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

R. Kohavi, D. Tang, and Y. Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

B. Kveton, Y. Liu, J. M. Kruijssen, and Y. Nie. Non-compliant bandits. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1138–1147, 2023.

B. Kveton, X. Li, J. McAuley, R. Rossi, J. Shang, J. Wu, and T. Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016a.

F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016b.

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

D. D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018a.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018b.

L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-tzur, M. Hardt, B. Recht, and A. Talwalkar. A System for Massively Parallel Hyperparameter Tuning. In *Proceedings of Machine Learning and Systems (MLSys)*, pages 230–246. 2020.

L.-F. Li, Y.-Y. Qian, P. Zhao, and Z.-H. Zhou. Provably efficient online rlhf with one-pass reward modeling. *arXiv preprint arXiv:2502.07193*, 2025.

Z. Li, K. Jamieson, and L. Jain. Optimal exploration is no harder than thompson sampling. *arXiv preprint arXiv:2310.06069*, 2023.

P. Liu, C. Shi, and W. W. Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

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

A. Malek, V. Aglietti, and S. Chiappa. Additive causal bandits with unknown graph. In *International Conference on Machine Learning*, 2023.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

K. Margatina, T. Schick, N. Aletras, and J. Dwivedi-Yu. Active learning principles for in-context learning with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

V. Mehta, V. Das, O. Neopane, Y. Dai, I. Bogunovic, J. Schneider, and W. Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*, 2023.

S. Mukherjee, A. Lalitha, K. Kalantari, A. A. Deshmukh, G. Liu, Y. Ma, and B. Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

W. Muldrew, P. Hayes, M. Zhang, and D. Barber. Active preference learning for large language models. In *International Conference on Machine Learning*, pages 36577–36590. PMLR, 2024.

S. Mummalaneni, H. Yoganarasimhan, and V. V. Pathak. Producer and consumer engagement on social media platforms. *SSRN 4173537*, 2022.

S. Mussmann and P. S. Liang. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. *Advances in Neural Information Processing Systems*, 31, 2018.

F. Orabona and D. Pal. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 577–585, 2016.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

J. Pearl. *Causality*. Cambridge university press, 2009.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

A. Raj and F. Bach. Convergence of uncertainty sampling for active learning. In *International Conference on Machine Learning*, pages 18310–18331. PMLR, 2022.

A. Scheid, E. Boursier, A. Durmus, M. I. Jordan, P. Ménard, E. Moulines, and M. Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*, 2024.

A. Schlaginhaufen, R. Ouhamma, and M. Kamgarpour. Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *arXiv preprint arXiv:2506.09508*, 2025.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

C. Shi, K. Yang, Z. Chen, J. Li, J. Yang, and C. Shen. Efficient prompt optimization through the lens of best arm identification. *Advances in Neural Information Processing Systems*, 37:99646–99685, 2024.

M. Simchowitz, K. Jamieson, and B. Recht. The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime. *Proceedings of Machine Learning Research*, 65: 1–41, 2017.

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

E. Tanczos, R. Nowak, and B. Mankoff. A KL-LUCB algorithm for Large-Scale Crowdsourcing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5894–5903. 2017.

G. Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

W. 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.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

J. Weltz, T. Fiez, A. Volfovsky, E. Laber, B. Mason, H. Nassif, and L. Jain. Experimental designs for heteroskedastic variance. *arXiv preprint arXiv:2310.04390*, 2023.

J. Yang and V. Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, G. Chen, J. Li, X. Wu, Z. Zhiyi, Q. Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, 2023.

J. Zhang, Y. Chen, and A. Singh. Causal bandits: Online decision-making in endogenous settings. *arXiv preprint arXiv:2211.08649*, 2022.

S. Zhang, D. Yu, H. Sharma, H. Zhong, Z. Liu, Z. Yang, S. Wang, H. Hassan, and Z. Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Y. Zhao, C. Stephens, C. Szepesvári, and K.-S. Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.

Y. Zhao, K.-S. Jun, T. Fiez, and L. Jain. Adaptive experimentation when you can't experiment. *Advances in neural information processing systems*, 37, 2024.

Y. Zhou, X. Chen, and J. Li. Optimal PAC multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225. PMLR, 2014.

B. Zhu, E. Frick, T. Wu, H. Zhu, K. Ganesan, W.-L. Chiang, J. Zhang, and J. Jiao. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*, 2024.

Y. Zhu, J. Katz-Samuels, and R. Nowak. Near instance optimal model selection for pure exploration linear bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6735–6769. PMLR, 2022.