Title

Explorations in Stochastic Bandit Problems: Theory and Applications

Author

Kang, Yue

Publication Date

2024

Peer reviewed|Thesis/dissertation

Explorations in Stochastic Bandit Problems: Theory and Applications

By

YUE KANG
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Thomas C. M. Lee, Chair

Cho-Jui Hsieh

Krishnakumar Balasubramanian

Committee in Charge

2024

To the ones I cherish and the ones supporting me.

Contents

Abstract

The stochastic bandit problem (Robbins, 1952) is a type of decision-making problem where an agent must repeatedly choose between multiple arms from a (varying) arm set, where each arm is associated with an unknown and different reward distribution, and the objective is to maximize the cumulative reward over time. This problem gets its name from the analogy of a gambler choosing which arm of a row of slot machines to pull, where each machine provides a different and unknown probability of winning. This problem framework is widely applicable in various areas and several sub-problems of it have been extensively studied during the past few years, e.g. multi-armed bandits (MAB) (Robbins, 1952), linear bandits (Abbasi-Yadkori et al., 2011), Lipschitz bandits (Agrawal, 1995) and so on. However, existing research on bandits faces certain limitations, both theoretical and crucially in practical applications. These challenges have become significant bottlenecks in advancing the field of stochastic bandit problems. To name a few, (1) robustness against adversarial attacks (Chapter 2); (2) auto-hyperparameter tuning (Chapter 3); (3) adaptivity to non-stationary environment (Chapter 3); (4) efficiency under high-dimensional structure with sparsity (Chapter 4); (5) resilience to heavy-tailed payoffs (Chapter 5).

Given that these fundamental issues have rarely been explored in the past, we have committed significant effort to addressing and resolving these challenges both theoretically and practically. In Chapter 1, we present a brief introduction to the bandit problem along with some limitations on the existing literature, which motivates our research. In Chapter 2, we introduce the stochastic Lipschitz bandit problem under the presence of adversarial attacks, and we propose a line of novel algorithms under different types of adversaries even agnostic to the total corruption level C. Subsequently, we study how to dynamically tune the hyperparameters in bandit algorithms with an infinite number of hyperparameter value candidates in Chapter 3. In Chapter 4, we investigate the recently popular low-rank matrix bandit problem and propose two types of algorithms with improved empirical performance and decent regret bounds. Then in Chapter 5, we revisit the low-rank matrix bandit problem but under a more sophisticated scenario: the stochastic payoffs are infused with heavy-tailed noise, and propose a novel framework to handle the heavy-tailedness and sparsity simultaneously. All the algorithms and frameworks we propose are backed by robust theoretical guarantees, with proofs provided in the Appendix.

Acknowledgments

I owe a huge thanks to my advisor, committee members, colleagues, friends and family for being the backbone of my five-year Ph.D. journey. Your guidance and support during my time at University of California, Davis have been priceless. I have no chance to nail it without you all.

To begin with, I'd like to extend my deepest gratitude to my advisors, Dr. Thomas C. M. Lee and Dr. Cho-Jui Hsieh, whose unwavering support and insightful guidance, have been instrumental throughout this journey. Your expertise, patience, and dedication have shaped not only this dissertation but also my academic growth and personal development. Every time I feel frustrated or I meet obstacles in my research, your endless encouragement really pushes my forward and uplifts my spirits regardless of the difficulty. Without your help, I would have no chance to publish any work or find a job in the end.

I am thankful to my committee members, Dr. Krishnakumar Balasubramanian, Dr. Xiucai Ding and Dr. Miles Lopes, for their valuable feedback on my work. Their collective expertise has enriched this work and challenged me to strive for excellence. Dr. Krishnakumar Balasubraminian gave me several constructive suggestions on my first project during my QE, which definitely helps improve the quality of our work. Dr. Xiucai Ding consistently offers me valuable insights into my research endeavors, all while often spending his leisure time discussing soccer games as a close friend, for which I am sincerely grateful. Dr. Miles Lopes's STA 231C is my favorite course during these five years, and I have learned a lot from him as a junior researcher. I'd like to say thank you to all the faculty members and staff in our Department of Statistics who helped me take a deep dive into statistics and gave me much support in the past five years. Furthermore, I'd like to extend my gratitude to several doctorate alumni from our department, to name a few, Dr. Yao Li, Dr. Qin Ding and Dr. Zhenyu Wei. Your endless assistance during my research and my job hunting has been immensely beneficial to me.

My sincere appreciation goes to my colleagues and peers for their support and collaboration. It is very lucky to start my Ph.D. journey with my roommate Yi Han, and we help each other a lot as we explored Davis in the beginning. My second roommate, Prof. Rui Hu, has been a consistent provider of discussions regarding our statistical research and our daily life. Without the companionship and support of my friends and colleagues at Davis, my Ph.D. experience wouldn't have been as vibrant and memorable. I am immensely grateful to all of them for their constant

support and friendship. To name a few (due to the space limit, but not limited to), I would like to thank Doudou Zhou, Zhaoyang Shi, Mingshuo Liu, Yanhao Jin and so on. It is my great pleasure and privilege to have you all during my Ph.D. journey.

Last but definitely not least, I am truly grateful to my family, especially my parents Jiantian Kang and Huimin Wang, and my partner Jing Lyu for their unconditional love, encouragement, and sacrifices. Your unwavering belief in me has been a constant source of strength and motivation all the time. Without their love and guidance, this journey would have been much harder. I am deeply thankful for every moment of peace and encouragement they provided and every sacrifice they made for me, which allowed me to fully dedicate myself to my studies and future aspirations.

CHAPTER 1

Introduction

The stochastic bandit problem is a cornerstone and the most basic framework in the field of sequential decision-making under uncertainty, where at each round an agent follows some policy and pulls an arm from a (varying) arm pool, and then only observes the stochastic reward of the chosen arm. This partial feedback setting presents a classic dilemma of balancing exploration and exploitation tradeoffs. Exploration involves trying different arms to gather more information about their reward distributions, while exploitation means choosing the arm that has so far appeared to offer the best rewards. In this framework, an agent aims to follow an optimal strategy to choose between multiple arms with uncertain rewards in order to maximize its total reward over some time horizon T. The challenge lies in deciding whether to exploit the arm that has historically provided the best estimated reward, or to explore other arms that might yield even greater rewards. This problem is not only a fundamental model for reinforcement learning but also has far-reaching applications in fields such as healthcare (Woodroofe, 1979), finance (Kleinberg & Leighton, 2003), and personalized recommendation (Li et al., 2010). For example, in the randomized controlled trials problem, multi-armed bandits are widely used to correctly identify the best treatment (best arm identification) and to treat patients as effectively as possible during the clinical trial (exploitation) (Villar et al., 2015). And in the e-commerce recommendation system, companies commonly use contextual bandit algorithms with user-item side information to post items that the users are very likely to click into (Zhu & Van Roy, 2023). The inherent complexity and broad applicability make the bandit problem an intriguing and rich study area, inviting innovative approaches and solutions.

Given the wide-ranging applications of bandit algorithms, there has been a substantial body of literature emerging over recent years in various types of bandits. In general, bandit problems can be classified into two categories according to how the rewards are generated. In stochastic bandits (Robbins, 1952), which are the concentration of our research, assumes all rewards of the same arm are generated from the same distribution. The other extreme is adversarial bandits (Auer et al., 2002b), which assume the rewards of arms can be arbitrarily manipulated by an adversary

at each round. Recently, there has been a line of work studying the best-of-both-world bandit that attempts to jointly optimize under both stochastic and adversarial rewards and achieves the nearly optimal regret bound for both settings. We will mainly focus the stochastic bandit problem in the following text, and we would first recall the basic definitions here: the agent receives an arm set denoted by $\mathcal{A}_t$ at round t and each arm $a \in \mathcal{A}_t$ corresponds to an unknown but fixed distribution ν_a. The agent selects some arm $a_t \in \mathcal{A}_t$ and receives a reward $y_{a_t,t}$ drawn from ν_{a_t} independently of historical observations. Denote the expected reward of the optimal arm at round t as

$$\mu_t^* = \arg\max_{a \in \mathcal{A}_t} \mathbb{E}\left(v_a\right).$$

And the quantity of interest we hope to control in time horizon T is the cumulative regret (pseudo regret) defined as $R_T = \sum_{i=1}^{T} \mu_t^* - \mathbb{E}\left(\nu_{a_t}\right)$.

Before we switch to the motivations and details of our work, we'd like to present a series of seminal works on multiple types of stochastic bandit problems. Most existing works on the stochastic bandits follow two key ideas: one is the upper confidence bound (UCB) strategy (Auer et al., 2002a) that optimistically pulls the arm with the highest upper confidence estimate of potential reward to balance the need to exploit well-performing options and explore less certain ones; the other is the Thompson sampling (a.k.a. Bayesian bandit) methodology (Chapelle & Li, 2011; Granmo, 2010) choosing actions based on sampling from posterior distributions of their rewards. The most fundamental type of stochastic bandit should be the multi-armed bandit (MAB) (Auer et al., 2002a), which assumes there are finite and fixed arms at each round. To extend this idea with a more complex arm set, the continuum-armed bandit (Agrawal, 1995) (a.k.a. Lipschitz bandit) has been well studied where the infinite arm set lies in a metric space and the expected reward function is an unknown Lipschitz function. Another popular approach to modelling bandit problems with a continuum arm setting is via the framework of Gaussian processes (Srinivas et al., 2009) (a.k.a. kernel bandit (Chowdhury & Gopalan, 2017)) which assumes the unknown expected reward function is defined on a reproducing kernel Hilbert space with various norms. Some other types of stochastic bandits, such as dueling bandit Yue et al. (2012) where the agent pulls a pair of arms and then observes the winner, and causal bandit that regards each intervention as an arm in a causal structure (Ma et al., 2022b; 2023), has been systematically investigated recently. Parametric bandits are a well-researched area within stochastic decision-making settings as well.

The most popular (generalized) linear contextual (Abbasi-Yadkori et al., 2011; Filippi et al., 2010) bandit assumes each arm is associated with a known, finite-dimensional vector (its feature vector), and the expected reward is presumed to be an unknown (generalized) linear function of this vector. This comprehensive review sets the stage for introducing our own contributions and innovations in tackling some of the unresolved challenges within this field.

1.1. Motivation

Although the bandit framework has been extensively studied under various settings during the past few years, several interesting challenges still remain unexplored in the existing literature:

- **Robustness to adversarial corruptions:** Although there has been extensive research on the adversarial robustness of stochastic bandits, most existing works consider problems with a discrete arm set, such as the traditional MAB (Gupta et al., 2019; Jun et al., 2018; Lykouris et al., 2018) and contextual linear bandit (Bogunovic et al., 2021; He et al., 2022; Li et al., 2019). To the best of our knowledge, the stochastic Lipschitz bandit problem with adversarial corruptions with a wide range of applications has never been explored so far under the weak or strong adversary, and it is intrinsically more challenging due to the complex structure of different metric spaces and the infinite number of arms in the Lipschitz bandit setting.
- **Auto-tuning hyperparameters:** The empirical performance of all bandit algorithms significantly depends on the configuration of hyperparameters, and simply using theoretical optimal values is too conservative and always yields unsatisfactory practical results. This limitation has already become a bottleneck for bandits algorithms in real-world applications, and the few existing methods Bogunovic et al. (2021); Ding et al. (2022b) have apparent flaws such as the lack of theoretical support and the inefficiency of model configurations.
- **Non-stationarity:** As previously mentioned, a crucial assumption in the stochastic bandit setting is that the rewards from the arms remain constant, and the stochastic contextual bandit setting presumes that the relationship between features and rewards is stationary. However, these conditions are often not met in real-world applications. For example,

in the context of news article recommendations—a significant application area for bandit algorithms—users' preferences might lean towards sports news during the Olympics. Although the non-stationary bandits under the *drifting* environment (gradually drifting) and the *switching* environment (abruptly change) have been well studied for MABs (Auer et al., 2002b; 2019; Besbes et al., 2014) and contextual linear bandits (Cheung et al., 2019; Zhao et al., 2020), this intriguing problem has never been explored in the Lipschitz bandit literature.

- **Practical efficiency:** For the high-dimensional bandit problems with sparsity, such as the LASSO bandit (Kim & Paik, 2019) and the low-rank matrix bandit (Jun et al., 2019; Lu et al., 2021), a challenging problem is to propose an efficient algorithm with a decent theoretical guarantee. We focus on the popular low-rank matrix bandit problem, where the contextual information of the arm can be represented by a matrix and the unknown parameter matrix preserves a low-rank property. The existing low-rank matrix bandit algorithms (Lu et al., 2021) are practically inefficient, which dampers their applications in the real-world problems.
- **Resilience to heavy-tailedness:** Another very important assumptions for most stochastic bandit literature is that the random noise follows some sub-Gaussian distribution, but in many real-life applications such as financial markets (Bradley & Taqqu, 2003; Cont & Bouchaud, 2000), there's a notable trend where extreme noise, a.k.a. heavy-tailed noise, in observations occur more frequently than what would be expected under a sub-Gaussian distribution. To address this practical challenge, a line of algorithms has been proposed to handle heavy-tailed noise under the MAB (Bubeck et al., 2013) and linear bandit (Medina & Yang, 2016), but effectively managing heavy-tailed noise under the more complex high-dimensional bandit framework such as the low-rank matrix bandit still remains unexplored.

Overall, the aforementioned difficulties persist within the literature, and our goal in this dissertation is to tackle and resolve these challenges.

1.2. Contribution

Build upon the motivations we present in the earlier section, the detailed contributions of our work can be summarized as follows:

In Chapter 2, we present our paper entitled "Robust Lipschitz bandits to adversarial corruptions" (Kang et al., 2024c). We develop efficient robust Lipschitz bandit algorithms whose regret bounds degrade sub-linearly in terms of the corruption budget C. Under the weak adversary, we extend the idea in Lykouris et al. (2018) and propose an efficient algorithm named Robust Multi-layer Elimination Lipschitz bandit algorithm (RMEL) that is agnostic to C and attains $\tilde{O}(C^{\frac{1}{d_z+2}}T^{\frac{d_z+1}{d_z+2}})$ regret bound. This bound matches the minimax regret bound of Lipschitz bandits (Bubeck et al., 2008; Kleinberg et al., 2019) in the absence of corruptions up to logarithmic terms. Under the strong adversary, we first show that when the budget C is given, a simple modification on the classic Zooming algorithm (Kleinberg et al., 2019) would lead to a robust method, namely, Robust Zooming algorithm, which could obtain a regret bound of order $\tilde{O}(T^{\frac{d_z+1}{d_z+2}}+C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}})$ (d_z is called zooming dimension and is explained in Chapter 2 in detail). We then provide a lower bound to prove the extra $O(C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}})$ regret is unavoidable. Further, inspired by the Bandit-over-Bandit (BoB) model selection idea (Cheung et al., 2019; Ding et al., 2022b; Pacchiano et al., 2020), we design a two-layer framework adapting to the unknown C. Three types of algorithms are discussed and compared in both theory and practice.

In Chapter 3, we present our work entitled "Online continuous hyperparameter optimization for generalized linear contextual bandits" (Kang et al., 2024a). We propose an online continuous hyperparameter optimization framework for contextual bandits called CDT with theoretical guarantees. To the best of our knowledge, CDT is the first hyperparameter tuning method (even model selection method) with continuous candidates in the bandit community. For the top layer of CDT, we propose the Zooming TS algorithm with Restarts for Lipschitz bandits under the *switching* environment. To the best of our knowledge, our work is the first one to consider the Lipschitz bandits under the *switching* environment, and the first one to utilize TS methodology in Lipschitz bandits. Experiments on both synthetic and real datasets with various GLBs validate the efficiency of our method.

In Chapter 4, we present our work named "Efficient frameworks for generalized low-rank matrix bandit problems" (Kang et al., 2022). We propose two efficient methods called G-ESTT and G-ESTS for this problem by modifying two stages of ESTR (Jun et al., 2019) appropriately from different perspectives. To the best of our knowledge, the proposed methods are the first two generalized (contextual) low-rank bandit algorithms that are computationally feasible, and achieve the decent regret bound. The practical superiority of our algorithms are firmly validated based on our experimental results.

In Chapter 5, we introduce our work entitled "Low-rank matrix bandits with heavy-tailed rewards" (Kang et al., 2024b). Inspired by the success of Huber loss (Kang & Kim, 2023; Sun et al., 2020) and nuclear norm penalization (Negahban & Wainwright, 2011), we first introduce a convex-relaxation-based estimator to approximate the low-rank parameter matrix with heavy-tailed noise. As far as we're aware, our work is the first one to solve the trace regression problem under arbitrary heavy-tailed noise with bounded $(1+\delta)$ moment ($\delta \in (0,1)$). Equipped with this estimator, we develop an algorithm named LOTUS for the low-rank matrix bandits under heavy-tailed noise. LOTUS is agnostic to the time horizon T and can work without knowing the rank r, and its practical superiority is then validated in our simulations.

In Chapter 6, we ultimately wrap up our work and touch upon potential future work. The rest of the dissertation includes important technical proof and additional experimental results as Appendix.

Notations: For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its l_2 norm and $\|x\|_A \coloneqq \sqrt{x^\top A x}$ for any positive definite matrix $A \in \mathbb{R}^{d\times d}$. For matrices $X, Y \in \mathbb{R}^{n_1\times n_2}$, we use $\|X\|_{\text{op}}$, $\|X\|_{\text{nuc}}$ and $\|X\|_F$ to define the operator norm, nuclear norm and Frobenious norm of matrix X respectively, and we denote $\langle X, Y\rangle \coloneqq \textbf{trace}(X^\top Y)$ as the inner product between X and Y. We write $f(n) \asymp g(n)$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$, $f(n) \gtrsim g(n)$ if $g(n) = O(f(n))$, and $f(n) \lesssim g(n)$ if $f(n) = O(g(n))$, and these are the common notations used in high-dimensional statistics (Wainwright, 2019). The notation $\tilde{O}(\cdot)$ ignores the polylogarithmic factors. We also denote $[T] = \{1, \dots, T\}$ for $T \in \mathbb{N}^+$.

CHAPTER 2

Robust Lipschitz Bandits to Adversarial Corruptions

2.1. Introduction

Multi-armed Bandit (MAB) (Auer et al., 2002a) is a fundamental and powerful framework in sequential decision-making problems. Given the potential existence of malicious users in real-world scenarios (Chen & Hsieh, 2022), a recent line of works considers the stochastic bandit problem under adversarial corruptions: an agent adaptively updates its policy to choose an arm from the arm set, and an adversary may contaminate the reward generated from the stochastic bandit before the agent could observe it. To robustify bandit learning algorithms under adversarial corruptions, several algorithms have been developed in the setting of traditional MAB (Gupta et al., 2019; Jun et al., 2018; Lykouris et al., 2018) and contextual linear bandits (Bogunovic et al., 2021; Ding et al., 2022a; He et al., 2022; Li et al., 2019; Zhao et al., 2021). These works consider either the weak adversary (Lykouris et al., 2018), which has access to all past data but not the current action before choosing its attack, or the strong adversary (Bogunovic et al., 2021), which is also aware of the current action for contamination. Details of these two adversaries will be elaborated in Section 2.3. In practice, bandits under adversarial corruptions can be used in many real-world problems such as pay-per-click advertising with click fraud and recommendation systems with fake reviews (Lykouris et al., 2018), and it has been empirically validated that stochastic MABs are vulnerable to slight corruption (Ding et al., 2022a; Garcelon et al., 2020; Jun et al., 2018).

Although there has been extensive research on the adversarial robustness of stochastic bandits, most existing works consider problems with a discrete arm set, such as the traditional MAB and contextual linear bandit. In this paper, we investigate robust bandit algorithms against adversarial corruptions in the Lipschitz bandit setting, where a continuously infinite arm set lie in a known metric space with covering dimension d and the expected reward function is an unknown Lipschitz function. Lipschitz bandit can be used to efficiently model many real-world tasks such as dynamic pricing, auction bidding (Slivkins et al., 2019) and hyperparameter tuning (Kang et al., 2024a). The

TABLE 2.1. Comparisons of regret bounds for our proposed robust Lipschitz bandit algorithms.

ALGORITHM	REGRET BOUND	FORMAT	C	ADVERSARY
Robust Zooming	$\tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right)$	HIGH. PROB.	KNOWN	STRONG
RMEL	$\tilde{O}\left((C^{\frac{1}{d_z+2}} + 1) T^{\frac{d_z+1}{d_z+2}}\right)$	HIGH. PROB.	UNKNOWN	WEAK
EXP3.P	$\tilde{O}\left((C^{\frac{1}{d+2}} + 1) T^{\frac{d+2}{d+3}}\right)$	EXPECTED	UNKNOWN	STRONG
CORRAL	$\tilde{O}\left((C^{\frac{1}{d+1}} + 1) T^{\frac{d+1}{d+2}}\right)$	EXPECTED	UNKNOWN	STRONG
BoB Robust Zooming	$\tilde{O}\left(T^{\frac{d+3}{d+4}} + C^{\frac{1}{d+1}} T^{\frac{d+2}{d+3}}\right)$	HIGH. PROB.	UNKNOWN	STRONG

stochastic Lipschitz bandit has been well understood after a large body of literature (Bubeck et al., 2008; Kleinberg et al., 2019; Magureanu et al., 2014), and state-of-the-art algorithms could achieve a cumulative regret bound of order $\tilde{O}(T^{\frac{d_z+1}{d_z+2}})$[1] in time T. However, to the best of our knowledge, the stochastic Lipschitz bandit problem with adversarial corruptions has never been explored, and we believe it is challenging since most of the existing robust MAB algorithms utilized the idea of elimination, which is much more difficult under a continuously infinite arm pool. Furthermore, the complex structure of different metric spaces also poses challenges for defending against adversarial attacks (Solomon et al., 2021) in theory. Therefore, it remains intriguing to design computationally efficient Lipschitz bandits that are robust to adversarial corruptions under both weak and strong adversaries.

We develop efficient robust algorithms whose regret bounds degrade sub-linearly in terms of the corruption budget C. Our contributions can be summarized as follows: (1) Under the weak adversary, we extend the idea in Lykouris et al. (2018) and propose an efficient algorithm named Robust Multi-layer Elimination Lipschitz bandit algorithm (RMEL) that is agnostic to C and attains $\tilde{O}(C^{\frac{1}{d_z+2}} T^{\frac{d_z+1}{d_z+2}})$ regret bound. This bound matches the minimax regret bound of Lipschitz bandits (Bubeck et al., 2008; Kleinberg et al., 2019) in the absence of corruptions up to logarithmic terms. This algorithm consists of multiple parallel sub-layers with different tolerance against the budget C, where each layer adaptively discretizes the action space and eliminates some less promising regions based on its corruption tolerance level in each crafted epoch. Interactions between layers assure the promptness of the elimination process. (2) Under the strong adversary, we first show that when the budget C is given, a simple modification on the classic Zooming algorithm (Kleinberg

[1] d_z is the zooming dimension defined in Section 2.3.

et al., 2019) would lead to a robust method, namely, Robust Zooming algorithm, which could obtain a regret bound of order $\tilde{O}(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}})$. We then provide a lower bound to prove the extra $O(C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}})$ regret is unavoidable. Further, inspired by the Bandit-over-Bandit (BoB) model selection idea (Cheung et al., 2019; Ding et al., 2022b; Pacchiano et al., 2020), we design a two-layer framework adapting to the unknown C where a master algorithm in the top layer dynamically tunes the corruption budget for the Robust Zooming algorithm. Three types of master algorithms are discussed and compared in both theory and practice. Table 2.1 outlines our algorithms as well as their regret bounds under different scenarios.

2.2. Related Work

Stochastic and Adversarial Bandit. Extensive studies have been conducted on MAB and its variations, including linear bandit (Abbasi-Yadkori et al., 2011), matrix bandit (Kang et al., 2022), etc. The majority of literature can be categorized into two types of models (Lattimore & Szepesvári, 2020): stochastic bandit, in which rewards for each arm are independently sampled from a fixed distribution, and adversarial bandit, where rewards are maliciously generated at all time. However, adversarial bandit differs from our problem setting in the sense that rewards are arbitrarily chosen without any budget or distribution constraint. Another line of work aims to obtain "the best of both worlds" guarantee simultaneously (Bubeck & Slivkins, 2012). However, neither of these models is reliable in practice (Cao et al., 2019), since the former one is too ideal, while the latter one remains very pessimistic, assuming a fully unconstrained setting. Therefore, it is more natural to consider the scenario that lies "in between" the two extremes: the stochastic bandit under adversarial corruptions.

Lipschitz Bandit. Most existing works on the stochastic Lipschitz bandit (Agrawal, 1995) follow two key ideas. One is to uniformly discretize the action space into a mesh in the initial phase so that any MAB algorithm could be implemented (Kleinberg, 2004; Magureanu et al., 2014). The other is to adaptively discretize the action space by placing more probes in more promising regions, and then UCB (Bubeck et al., 2008; Kleinberg et al., 2019; Lu et al., 2019), TS (Kang et al., 2024a) or elimination (Feng et al., 2022) method could be utilized to deal with the exploration-exploitation tradeoff. The adversarial Lipschitz bandit was recently introduced and solved in Podimata & Slivkins (2021), where the expected reward Lipschitz function is arbitrarily chosen at each round.

However, as mentioned in the previous paragraph, this fully adversarial setting is quite different from ours. And their algorithm relies on several unspecified hyperparameters and hence is computationally formidable in practice.

Robust Bandit to Adversarial Corruptions. Adversarial attacks were studied in the setting of MAB (Jun et al., 2018) and linear bandits (Garcelon et al., 2020). And we will use two classic attacks for experiments in Section 2.5. To defend against attacks from weak adversaries, Lykouris et al. (2018) proposed the first MAB algorithm robust to corruptions with a regret C times worse than regret in the stochastic setting. An improved algorithm whose regret only contains an additive term on C was then proposed in Gupta et al. (2019). Li et al. (2019) subsequently studied the linear bandits with adversarial corruptions and achieved instance-dependent regret bounds. Lee et al. (2021) also studied the corrupted linear bandit problem while assuming the attacks on reward are linear in action. Recently, a robust VOFUL algorithm achieving regret bound only logarithmically dependent on T was proposed in Wei et al. (2022). Another line of work on the robust bandit problem focuses on a more challenging setting with strong adversaries who could observe current actions before attacking rewards. Bogunovic et al. (2021) considered the corrupted linear bandit when small random perturbations are applied to context vectors, and Ding et al. (2022a); He et al. (2022); Zhao et al. (2021) extended the OFUL algorithm (Abbasi-Yadkori et al., 2011) and achieved improved regret bounds. However, the study of Lipschitz bandits under attacks remains an unaddressed open area.

2.3. Preliminaries

We will introduce the setting of Lipschitz bandits with adversarial corruptions in this section. The Lipschitz bandit is defined on a triplet $(\mathcal{X}, D, \mu)$, where $\mathcal{X}$ is the arm set space equipped with some metric D, and $\mu : \mathcal{X} \to \mathbb{R}$ is an unknown Lipschitz reward function on the metric space $(\mathcal{X}, D)$ with Lipschitz constant 1. W.l.o.g. we assume $\mathcal{X}$ is compact with its diameter no more than 1. Under the stochastic setting, at each round $t \in [T] \coloneqq \{1, 2, \ldots, T\}$, stochastic rewards are sampled for each arm $x \in \mathcal{X}$ from some unknown distribution P_x independently, and then the agent pulls an arm x_t and receives the corresponding stochastic reward $\tilde{y}_t$ such that,

$$\tilde{y}_t = \mu(x_t) + \eta_t, \tag{2.1}$$

where η_t is i.i.d. zero-mean random error with sub-Gaussian parameter σ conditional on the filtration $\mathcal{F}_t = \{x_t, x_{t-1}, \eta_{t-1}, \ldots, x_1, \eta_1\}$. W.l.o.g we assume $\sigma = 1$ for simplicity in the rest of our analysis. At each round $t \in [T]$, the weak adversary observes the payoff function $\mu(\cdot)$, the realizations of P_x for each arm $x \in \mathcal{X}$ and choices of the agent $\{x_i\}_{i=1}^{t-1}$ in previous rounds, and injects an attack $c_t(x_t)$ into the reward before the agent pulls x_t. The agent then receives a corrupted reward $y_t = \tilde{y}_t + c_t(x_t)$. The strong adversary would be omniscient and have complete information about the problem $\mathcal{F}_t$. In addition to the knowledge that a weak adversary possesses, it would also be aware of the current action x_t while contaminating the data, and subsequently decide upon the corrupted reward $y_t = \tilde{y}_t + c_t(x_t)$. Some literature in corrupted bandits (Ding et al., 2022a; Garcelon et al., 2020) also consider attacking on the contexts or arms, i.e. the adversary modifies the true arm x_t in a small region, while in our problem setting it is obvious that attacking contexts is only a sub-case of attacking rewards due to the Lipschitzness of $\mu(\cdot)$, and hence studying the adversarial attacks on rewards alone is sufficient under the Lipschitz bandit setting.

The total corruption budget C of the adversary is defined as $C = \sum_{t=1}^{T} \max_{x \in \mathcal{X}} |c_t(x)|$, which is the sum of maximum perturbation from the adversary at each round across the horizon T. Note the strong adversary may only corrupt the rewards of pulled arms and hence we could equivalently write $C = \sum_{t=1}^{T} |c_t(x_t)|$ in that case as Bogunovic et al. (2021); He et al. (2022). Define the optimal arm $x_* = \arg\max_{x \in \mathcal{X}} \mu(x)$ and the loss of arm x as $\Delta(x) = \mu(x_*) - \mu(x), x \in \mathcal{X}$. W.l.o.g. we assume $C \leq T$ and each instance of attack $|c_t(x)| \leq 1, \forall t \in [T], x \in \mathcal{X}$ as in other robust bandit literature (Gupta et al., 2019; Lykouris et al., 2018) since the adversary could already make any arm $x \in \mathcal{X}$ optimal given that $\Delta(x) \leq 1$. (We can assume $|c_t(x)| \leq u, \forall t \in [T], x \in \mathcal{X}$ for any positive constant u.) Similar to the stochastic case (Kleinberg, 2004), the goal of the agent is to minimize the cumulative regret defined as:

$$Regret_T = T\mu(x_*) - \sum_{t=1}^{T} \mu(x_t). \tag{2.2}$$

An important pair of concepts in Lipschitz bandits defined on $(\mathcal{X}, D, \mu)$ are the covering dimension d and the zooming dimension d_z. Let $\mathcal{B}(x, r)$ denotes a closed ball centered at x with radius r in $\mathcal{X}$, i.e. $\mathcal{B}(x, r) = \{x' \in \mathcal{X} : D(x, x') \leq r\}$, the r-covering number $N_c(r)$ of metric space $(\mathcal{X}, D)$ is defined as the minimal number of balls with radius of no more than r required to cover $\mathcal{X}$. On the contrary, the r-zooming number $N_z(r)$ introduced in Kleinberg et al. (2019) not only depends on the metric

Algorithm 1 Robust Zooming Algorithm

Input: Arm metric space $(\mathcal{X}, D)$, time horizon T, probability rate δ.

Active arm set $J = \{\}$, active space $\mathcal{X}_{act} = \mathcal{X}$.
for $t = 1$ **to** T **do**
  if $f(v) - f(u) \geq r(v) + 2r(u)$ for some pair of active arms $u, v \in J$. **then**
    Set $J = J \setminus \{u\}$ and $\mathcal{X}_{act} = \mathcal{X}_{act} \setminus \mathcal{B}(u, r(u))$. ▷ Removal
  if $\mathcal{X}_{act} \nsubseteq \cup_{v \in J} \mathcal{B}(v, r(v))$ **then**
    Activate and pull some arm $x \notin \cup_{v \in J} \mathcal{B}(v, r(v))$ in $\mathcal{X}_{act}$ such that $x_t = x$, $J = J \cup \{x\}$, and set the components $n(x) = 0$, $f(x) = 0$. ▷ Activation
  else
    Pull $x_t = \arg\max_{v \in J} I(v) = f(v) + 2r(v)$, and break ties arbitrarily. ▷ Selection
  Observe the payoff y_t. And update components associated with x_t in the Robust Zooming Algorithm: $n(x_t) = n(x_t) + 1, f(x_t) = \left(f(x_t)\left(n(x_t) - 1\right) + y_t\right)/n(x_t)$.

space $(\mathcal{X}, D)$ but also the payoff function $\mu(\cdot)$. It describes the minimal number of balls of radius not more than $r/16$ required to cover the r-optimal region defined as $\{x \in \mathcal{X} : \Delta(x) \leq r\}$ (Bubeck et al., 2008)[2]. Next, we define the covering dimension d (zooming dimension d_z) as the smallest $q \geq 0$ such that for every $r \in (0, 1]$ the r-covering number $N_c(r)$ (r-zooming number $N_z(r)$) can be upper bounded by αr^{-q} for some multiplier $\alpha > 0$ that is free of r:

$$d = \min\{q \geq 0 : \exists \alpha > 0, N_c(r) \leq \alpha r^{-q}, \forall r \in (0, 1]\},$$

$$d_z = \min\{q \geq 0 : \exists \alpha > 0, N_z(r) \leq \alpha r^{-q}, \forall r \in (0, 1]\}.$$

It is clear that $0 \leq d_z \leq d$ since the r-optimal region is a subset of $\mathcal{X}$. On the other hand, d_z could be much smaller than d in some benign cases. For example, if the payoff function $\mu(\cdot)$ defined on the metric space $(\mathbb{R}^k, \|\cdot\|_2), k \in \mathbb{N}$ is C^2-smooth and strongly concave in a neighborhood of the optimal arm x_*, then it could be easily verified that $d_z = k/2$ whereas $d = k$. However, d_z is never revealed to the agent as it relies on the underlying function $\mu(\cdot)$, and hence designing an algorithm whose regret bound depends on d_z without knowledge of d_z would be considerably difficult.

2.4. Methods

We will present our main algorithms in this Section.

2.4.1. Known Budgets under Strong Adversaries. To defend against attacks on Lipschitz bandits, we first consider a simpler case where the agent is aware of the corruption budget C. We

[2]We actually use the near-optimality dimension introduced in Bubeck et al. (2008), where the authors imply the equivalence between this definition and the original zooming dimension proposed in Kleinberg et al. (2019).

demonstrate that a slight modification of the classic Zooming algorithm (Kleinberg et al., 2019) can result in a robust Lipschitz bandit algorithm even under the strong adversary, called the Robust Zooming algorithm, which achieves a regret bound of order $\tilde{O}(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}})$.

We first introduce some notations of the algorithm: denote J as the active arm set. For each active arm $x \in J$, let $n(x)$ be the number of times arm x has been pulled, $f(x)$ be the corresponding average sample reward, and $r(x)$ be the confidence radius controlling the deviation of the sample average $f(x)$ from its expectation $\mu(x)$. We also define $\mathcal{B}(x, r(x))$ as the confidence ball of an active arm x. In essence, the Zooming algorithm works by focusing on regions that have the potential for higher rewards and allocating fewer probes to less promising regions. The algorithm consists of two phases: in the activation phase, a new arm gets activated if it is not covered by the confidence balls of all active arms. This allows the algorithm to quickly zoom into the regions where arms are frequently pulled due to their encouraging rewards. In the selection phase, the algorithm chooses an arm with the largest value of $f(v) + 2r(v)$ among J based on the UCB methodology.

Our key idea is to enlarge the confidence radius of active arms to account for the known corruption budget C. Specifically, we could set the value of $r(x)$ as:

$$r(x) = \sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n(x)}} + \frac{C}{n(x)},$$

where the first term accounts for deviation in stochastic rewards and the second term is used to defend the corruptions from the adversary. The robust algorithm is shown in Algorithm 1. In addition to the two phases presented above, our algorithm also conducts a removal procedure at the beginning of each round for better efficiency. This step adaptively removes regions that are likely to yield low rewards with high confidence. Theorem 2.4.1 provides a regret bound for Algorithm 1.

THEOREM 2.4.1. *Given the total corruption budget that is at most C, with probability at least $1-\delta$, the overall regret of Robust Zooming Algorithm (Algorithm 2.1) can be bounded as:*

$$Regret_T = O\left(\ln(T)^{\frac{1}{d_z+2}} T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right) = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right).$$

Furthermore, the following Theorem 2.4.2 implies that our regret bound attains the lower bound and hence is unimprovable. The detailed proof is given in Appendix A.6.1.

THEOREM 2.4.2. *Under the strong adversary with a corruption budget of C, for any zooming dimension $d_z \in \mathbb{Z}^+$, there exists an instance for which any algorithm (even one that is aware of C) must suffer a regret of order $\Omega(C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}})$ with probability at least 0.5.*

In addition to the lower bound provided in Theorem 2.4.2, we further propose another lower bound for the strong adversary particularly in the case that C is unknown in the following Theorem 2.4.3:

THEOREM 2.4.3. *For any algorithm, when there is no corruption, we denote R_T^0 as the upper bound of cumulative regret in T rounds under our problem setting described in Section 2.3, i.e. $Regret_T \leq R_T^0$ with high probability, and it holds that $R_T^0 = o(T)$. Then under the strong adversary and unknown attacking budget C, there exists a problem instance on which this algorithm will incur linear regret $\Omega(T)$ with probability at least 0.5, if $C = \Omega(R_T^0/4^{d_z}) = \Omega(R_T^0)$.*

However, there are also some weaknesses to our Algorithm 1. The first weakness is that the algorithm is too conservative and pessimistic in practice since the second term of $r(x)$ would dominate under a large given value of C. We could set the second term of $r(x)$ as $\min\{1, C/n(x)\}$ to address this issue and the analysis of Theorem 2.4.1 will still hold as shown in Appendix A.1 Remark A.1.1. The second weakness is that it still incurs the same regret bound shown in Theorem 2.4.1 even if there are actually no corruptions applied. To overcome these problems and to further adapt to the unknown corruption budget C, we propose two types of robust algorithms in the following Sections.

2.4.2. Unknown Budgets under Weak Adversaries. The weak adversary is unaware of the agent's current action before contaminating the stochastic rewards. We introduce an efficient algorithm called Robust Multi-layer Elimination Lipschitz bandit algorithm (RMEL) that is summarized in Algorithm 2. Four core steps are introduced as follows.

Multi-layer Parallel Running: Our algorithm consists of multiple sub-layers running in parallel, each with a different tolerance level against corruptions. As shown in Algorithm 2, there are l^* layers and the tolerance level of each layer, denoted as v_l, increases geometrically with a ratio of B (a hyperparameter). At each round, a layer l is sampled with probability $1/v_l$, meaning that layers that are more resilient to attacks are less likely to be chosen and thus may make slower progress. This sampling scheme helps mitigate adversarial perturbations across layers by limiting the amount of corruptions distributed to layers whose tolerance levels exceed the unknown budget C to at most

Algorithm 2 Robust Multi-layer Elimination Lipschitz Bandit Algorithm (RMEL)

Input: Arm metric space $(\mathcal{X}, D)$, time horizon T, probability rate δ, base parameter B.

1: Tolerance level $v_l = \ln(4T/\delta)B^{l-1}, m_l = 1, n_l = 0, \mathcal{A}_l = 1/2$-covering of $\mathcal{X}$, $f_{l,A} = n_{l,A} = 0$ for all $A \in \mathcal{A}_l, l \in [l^*]$ where $l^* := \min\{l \in \mathbb{N} : \ln(4T/\delta)B^{l-1} \geq T\}$.
2: **for** $t = 1$ **to** T **do**
3: Sample layer $l \in [l^*]$ with probability $1/v_l$, with the remaining probability sampling $l = 1$. Find the minimum layer index $l_t \geq l$ such that $\mathcal{A}_{l_t} \neq \emptyset$. ▷ Layer sampling
4: Choose $A_t = \arg\min_{A \in \mathcal{A}_{l_t}} n_{l_t,A}$, break ties arbitrary.
5: Randomly pull an arm $x_t \in A_t$, and observe the payoff y_t.
6: Set $n_{l_t} = n_{l_t} + 1$, $n_{l_t,A_t} = n_{l_t,A_t} + 1$, and $f_{l_t,A_t} = \left(f_{l_t,A_t}(n_{l_t,A_t} - 1) + y_t\right)/n_{l_t,A_t}$.
7: **if** $n_{l_t} = 6\ln(4T/\delta) \cdot 4^{m_l} \times |\mathcal{A}_{l_t}|$ **then**
8: Obtain $f_{l_t,*} = \max_{A \in \mathcal{A}_{l_t}} f_{l_t,A}$.
9: For each $A \in \mathcal{A}_{l_t}$, if $f_{l_t,*} - f_{l_t,A} > 4/2^{m_{l_t}}$, then we eliminate A from $\mathcal{A}_{l_t}$ and all active regions A' from $\mathcal{A}_{l'}$ in the case that $A' \subseteq A, A' \in \mathcal{A}_{l'}, l' < l$. ▷ Removal
10: Find $1/2^{m_l+1}$-covering of each remaining $A \in \mathcal{A}_{l_t}$ in the same way as A was partitioned in other layers. Then reload the active region set $\mathcal{A}_{l_t}$ as the collection of these coverings.
11: Set $n_{l_t} = 0$, $m_{l_t} = m_{l_t} + 1$. And renew $n_{l_t,A} = f_{l_t,A} = 0, \forall A \in \mathcal{A}_{l_t}$. ▷ Refresh

$O(\ln(T/\delta))$. For the other low-tolerance layers which may suffer from high volume of attacks, we use the techniques introduced below to rectify them in the guidance of the elimination procedure on robust layers. While we build on the multi-layer idea introduced in Lykouris et al. (2018), our work introduces significant refinements and novelty by extending this approach to continuous and infinitely large arm sets, as demonstrated below.

Active Region Mechanism: For each layer ℓ, our algorithm proceeds in epochs: we initialize the epoch index $m_l = 1$ and construct a $1/2^{m_l}$-covering of $\mathcal{X}$ as the active region set $\mathcal{A}_l$. In addition, we denote n_l as the number of times that layer l has been chosen, and for each active region $A \in \mathcal{A}_l$ we define $n_{l,A}, f_{l,A}$ as the number of times A has been chosen as well as its corresponding average empirical reward respectively. Assume layer l_t is selected at time t, then only one active region (denoted as A_t) in $\mathcal{A}_{l_t}$ would be played where we arbitrarily pull an arm $x_t \in A_t$ and collect the stochastic payoff y_t. For any layer l, if each active region in $\mathcal{A}_l$ is played for $6\ln(4T/\delta) \cdot 4^m$ times (i.e. line 6 of Algorithm 2), it will progress to the next epoch after an elimination process that is described below. All components mentioned above that are associated with the layer l will subsequently be refreshed (i.e. line 10 of Algorithm 2).

Within-layer Region Elimination and Discretization: For any layer $l \in [l^*]$, the within-layer elimination occurs at the end of each epoch as stated above. We obtain the average empirical reward $f_{l,A}$ for all $A \in \mathcal{A}_l$ and then discard regions with unpromising payoffs compared with the

optimal one with the maximum estimated reward (i.e. $f_{l,*}$ defined in line 7 of Algorithm 2). We further "zoom in" on the remaining regions of the layer l that yield satisfactory rewards: we divide them into $1/2^{m_l+1}$-covering and then reload A_l as the collection of these new partitions for the next epoch (line 9 of Algorithm 2) for the layer l. In consequence, only regions with nearly optimal rewards would remain and be adaptively discretized in the long run.

Cross-layer Region Elimination: While layers are running in parallel, it is essential to facilitate communication among them to prevent less reliable layers from getting trapped in suboptimal regions. In our Algorithm 2, if an active region $A \in \mathcal{A}_l$ is eliminated based on the aforementioned rule, then A will also be discarded in all layers $l' \leq l$. This is because the lower layers are faster whereas more vulnerable and less resilient to malicious attacks, and hence they should learn from the upper trustworthy layers whose tolerance levels surpass C by imitating their elimination decisions. A tradeoff lies in the selection of the hyperparameter B, which controls the ratio of tolerance levels between adjacent layers. With a larger value of B, only fewer layers are required, and hence more samples could be assigned to each layer for better efficiency. But the cumulative regret bound would deteriorate since it's associated with B sub-linearly. The cumulative regret bound is presented in the following Theorem 2.4.4, with its detailed proof in Appendix A.2.

THEOREM 2.4.4. *If the underlying corruption budget is C, then with probability at least $1-\delta$, the overall regret of our RMEL algorithm (Algorithm 2) could be bounded as:*

$$Regret_T = \tilde{O}\left(\left((BC)^{\frac{1}{d_z+2}}+1\right)T^{\frac{d_z+1}{d_z+2}}\right) = \tilde{O}\left(\left(C^{\frac{1}{d_z+2}}+1\right)T^{\frac{d_z+1}{d_z+2}}\right).$$

Note that if no corruption is actually applied (i.e. $C = 0$), our RMEL algorithm could attain a regret bound of order $\tilde{O}(T^{\frac{d_z+1}{d_z+2}})$ which coincides with the lower bound of stochastic Lipschitz bandits up to logarithmic terms. We further prove a regret lower bound of order $\Omega(C)$ under the weak adversary in Theorem 2.4.5 with its detailed proof in Appendix A.6.2. Therefore, a compelling open problem is to narrow the regret gap by proposing an algorithm whose regret bound depends on C in another additive term free of T under the weak adversary, like Gupta et al. (2019) for MABs and He et al. (2022) for linear bandits.

THEOREM 2.4.5. *Under the weak adversary with corruption budget C, for any zooming dimension d_z, there exists an instance such that any algorithm (even is aware of C) must suffer from the regret of order $\Omega(C)$ with probability at least* 0.5.

2.4.3. Unknown Budgets under Strong Adversaries. In Section 2.4.1, we propose the Robust Zooming algorithm to handle the strong adversary given the knowledge of budget C and prove that it achieves the optimal regret bound. However, compared with the known budget C case, defending against strong adversaries naturally becomes more challenging when the agent is unaware of the budget C. Motivated by the literature on model selection in bandits, we extend our Robust Zooming algorithm by combining it with different master algorithms to learn and adapt to the unknown C on the fly. We consider two approaches along this line: the first approach uses the master algorithms EXP3.P and CORRAL with the smoothing transformation (Pacchiano et al., 2020) to deal with unknown C, which leads to a promising regret bound but a high computational cost. We then equip Robust Zooming algorithm with the efficient bandit-over-bandit (BoB) idea (Cheung et al., 2019) to adapt to the unknown C, leading to a more efficient algorithm with a slightly worse regret bound.

Model Selection: When an upper bound on C is known, we propose the Robust Zooming algorithm with regret bound $\tilde{O}(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}})$ against strong adversaries in Section 2.4.1. Therefore, it is natural to consider a decent master algorithm that selects between $\lceil \log_2(T) \rceil$ base algorithms where the i-th base algorithm is the Robust Zooming algorithm with corruptions at most 2^i. As $C \leq T$, there must exist a base algorithm that is at most $2C$-corrupted. Here we choose the stochastic EXP3.P and CORRAL with smooth transformation proposed in Pacchiano et al. (2020) as the master algorithm due to the following two reasons with respect to theoretical analysis: (1). our action set $\mathcal{A}$ is fixed and the expected payoff is a function of the chosen arm, which satisfies the restrictive assumptions of this master algorithm (Section 2, (Pacchiano et al., 2020)); (2). the analysis in Pacchiano et al. (2020) still works even the regret bounds of base algorithms contain unknown values, and note the regret bound of our Zooming Robust algorithm depends on the unknown C. Based on Theorem 3.2 in Pacchiano et al. (2020), the expected cumulative regret of our Robust Zooming algorithm with these two types of master algorithms could be bounded as follows:

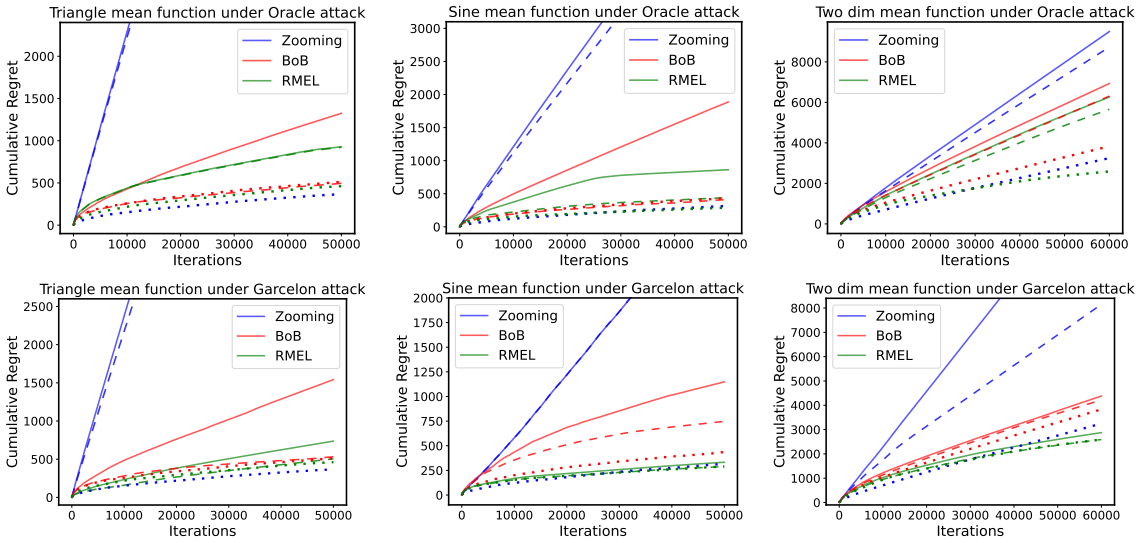


FIGURE 2.1. Plots of regrets of Zooming algorithm (blue), RMEL (green) and BoB Robust Zooming algorithm (red) under different settings with three levels of corruptions: (1) dotted line: no corruption; (2) dashed line: moderate corruptions; (3) solid line: strong corruptions. Numerical values of final cumulative regrets in our experiments are also displayed in Table A.2 in Appendix A.7.

THEOREM 2.4.6. *When the corruption budget C is unknown, by using our Algorithm 1 with $\{2^i\}_{i=1}^{\lceil \log_2(T) \rceil}$ corruptions as base algorithms and the EXP3.P and CORRAL with smooth transformation (Pacchiano et al., 2020) as the master algorithm, the expected regret could be upper bounded by*

$$\mathbb{E}(Regret_T) = \begin{cases} \tilde{O}\left((C^{\frac{1}{d+2}}+1)T^{\frac{d+2}{d+3}}\right) & EXP3.P, \\ \tilde{O}\left((C^{\frac{1}{d+1}}+1)T^{\frac{d+1}{d+2}}\right) & CORRAL. \end{cases}$$

We can observe that the regret bounds given in Theorem 2.4.6 are consistent with the lower bounds presented in Theorem 2.4.3. And CORRAL is better under small corruption budgets C (i.e. $C = \tilde{O}(T^{\frac{d+1}{d+3}})$) whereas EXP3.P is superior otherwise. Note that the order of regret relies on d instead of d_z since the unknown d_z couldn't be used as a parameter in practice, and both regret bounds are worse than the lower bound given in Theorem 2.4.2 for the strong adversary. Another drawback of the above method is that a two-step smoothing procedure is required at each round, which is computationally expensive. Therefore, for better practical efficiency, we propose a simple BoB-based method as follows:

BoB Robust Zooming: The BoB idea (Cheung et al., 2019) is a special case of model selection in bandits and aims to adjust some unspecified hyperparameters dynamically in batches. Here

we use $\lceil \log_2(T) \rceil$ Robust Zooming algorithms with different corruption levels shown above as base algorithms in the bottom layer and the classic EXP3.P (Auer et al., 2002b) as the top layer. Our method, named BoB Robust Zooming, divides T into H batches of the same length, and in one batch keeps using the same base algorithm that is selected from the top layer at the beginning of this batch. When a batch ends, we refresh the base algorithm and use the normalized accumulated rewards of this batch to update the top layer EXP3.P since the EXP3.P algorithm (Auer et al., 2002b) requires the magnitude of rewards should at most be 1 in default. Specifically, we normalize the cumulative reward at the end of each batch by dividing it with $(2H + \sqrt{2H \log(12T/H\delta)})$ due to the fact that the magnitude of the cumulative reward at each batch would at most be this value with high probability as shown in Lemma A.4.0.1 in Appendix A.4. Note that this method is highly efficient since a single update of the EXP3.P algorithm only requires $O(1)$ time complexity, and hence the additional computation from updating EXP3.P is only $O(H)$. Due to space limit, we defer Algorithm 10 to Appendix A.5, and the regret bound is given as follows:

THEOREM 2.4.7. *When the corruption budget C is unknown, with probability at least $1-\delta$, the regret of our BoB Robust Zooming algorithm with $H = T^{(d+2)/(d+4)}$ could be bounded as:*

$$Regret_T = \tilde{O}\left(T^{\frac{d+3}{d+4}} + C^{\frac{1}{d+1}} T^{\frac{d+2}{d+3}}\right).$$

Although we could deduce the more challenging high-probability regret bound for this algorithm, its order is strictly worse than those given in Theorem 2.4.6. In summary, the BoB Robust Zooming algorithm is more efficient and easier to use in practice, while yielding worse regret bound in theory. However, due to its practical applicability, we will implement this BoB Robust Zooming algorithm in the experiments. It is also noteworthy that we can attain a better regret bound with Algorithm 2 under the weak adversary as shown in Theorem 2.4.4, which aligns with our expectation since the strong adversary considered here is more malicious and difficult to defend against.

2.5. Experimental Results

In this section, we show by simulations that our proposed RMEL and BoB Robust Zooming algorithm outperform the classic Zooming algorithm in the presence of adversarial corruptions. To firmly validate the robustness of our proposed methods, we use three types of models and two sorts of attacks with different corruption levels. We first consider the metric space $([0,1], |\cdot|)$

with two expected reward functions that behave differently around their maximum: (1). $\mu(x) = 0.9 - 0.95|x - 1/3|$ (triangle) and (2). $\mu(x) = 2/(3\pi) \cdot \sin(3\pi x/2)$ (sine). We then utilize a more complicated metric space $([0,1]^2, \|\cdot\|_\infty)$ with the expected reward function (3). $\mu(x) = 1 - 0.8\|x - (0.75, 0.75)\|_2 - 0.4\|x - (0,1)\|_2$ (two dim). We set the time horizon $T = 50,000\,(60,000)$ for the metric space with $d = 1\,(2)$ and the false probability rate $\delta = 0.01$. The random noise at each round is sampled IID from $N(0, 0.01)$. Average cumulative regrets over 20 repetitions are reported in Figure 2.1.

Since adversarial attacks designed for stochastic Lipschitz bandits have never been studied, we extend two types of classic attacks, named Oracle (Jun et al., 2018) for the MAB and Garcelon (Garcelon et al., 2020) for the linear bandit, to our setting. The details of these two attacks are summarized as follows:

- **Oracle:** This attack (Jun et al., 2018) was proposed for the traditional MAB, and it pushes the rewards of "good arms" to the very bottom. Specifically, we call an arm is benign if the distance between it and the optimal arm is no larger than 0.2. And we inject this attack by pushing the expected reward of any benign arm below that of the worst arm with an additional margin of 0.1 with probability 0.5.
- **Garcelon:** We modify this type of attack studied in Garcelon et al. (2020) for linear bandit framework, which replaces expected rewards of arms outside some targeted region with IID Gaussian noise. For $d = 1$, since the optimal arm is set to be $1/3$ for both triangle and sine payoff functions, we set the targeted arm interval as $[0.5, 1]$. For $d = 2$, since the optimal arm is close to $(0.75, 0.75)$, we set the targeted region as $[0, 0.5]^2$. Here we contaminate the stochastic reward if the pulled arm is not inside the target region by modifying it into a random Gaussian noise $N(0, 0.01)$ with probability 0.5.

We consider the strong adversary in experiments as both types of attack are injected only if the pulled arms lie in some specific regions. Note although we originally propose RMEL algorithm for the weak adversary in theory, empirically we find it works exceptionally well (Figure 2.1) across all settings here. We also conduct simulations based on the weak adversary and defer their settings and results to Appendix A.7 due to the limited space. The first Oracle attack is considered to be more malicious in the sense that it specifically focuses on the arms with good rewards, while the

second Garcelon attack could corrupt rewards generated from broader regions, which may contain some "bad arms" as well.

Since there is no existing robust Lipschitz bandit algorithm, we use the classic Zooming algorithm (Kleinberg et al., 2019) as the baseline. As shown in Figure 2.1, we consider three levels of quantities of corruptions applied on each case to show how attacks progressively disturb different methods. Specifically, we set $C = 0$ for the non-corrupted case, $C = 3,000$ for the moderate-corrupted case and $C = 4,500$ for the strong-corrupted case. Due to space limit, we defer detailed settings of algorithms to Appendix A.7.

From the plots in Figure 2.1, we observe that our proposed algorithms consistently outperform the Zooming algorithm and achieve sub-linear cumulative regrets under both types of attacks, whereas the Zooming algorithm becomes incompetent and suffers from linear regrets even under a moderate volume of corruption. This fact also implies that the two types of adversarial corruptions used here are severely detrimental to the performance of stochastic Lipschitz bandit algorithms. And it is evident our proposed RMEL yields the most robust results under various scenarios with different volumes of attacks. It is also worth noting that the Zooming algorithm attains promising regrets under a purely stochastic setting, while it experiences a huge increase in regrets after the corruptions emerge. This phenomenon aligns with our expectation and highlights the fact that our proposed algorithms balance the tradeoff between accuracy and robustness in a much smoother fashion.

CHAPTER 3

Online Continuous Hyperparameter Optimization for Generalized Linear Contextual Bandits

3.1. Introduction

Generalized linear bandit (GLB) was first proposed in Filippi et al. (2010) and has been extensively studied under various settings over the recent years (Jun et al., 2017; Kang et al., 2022), where the stochastic payoff of an arm follows a generalized linear model (GLM) of its associated feature vector and some fixed, but initially unknown parameter θ^*. Note that GLB extends the linear bandit (Abbasi-Yadkori et al., 2011) in representation power and has greater applicability in real-world applications, e.g. logistic bandit algorithms (Zhang et al., 2016) can achieve improvement over linear bandit when the rewards are binary. Upper Confidence Bound (UCB) (Auer et al., 2002a; Filippi et al., 2010; Li et al., 2010) and Thompson Sampling (TS) (Agrawal & Goyal, 2012; 2013) are the two most popular ideas to solve the GLB problem. Both of these methods could achieve the optimal regret bound of order $\tilde{O}(\sqrt{T})$ under some mild conditions, where T stands for the total number of rounds (Agrawal & Goyal, 2013).

However, the empirical performance of these bandit algorithms significantly depends on the configuration of hyperparameters, and simply using theoretical optimal values often yields unsatisfactory practical results, not to mention some of them are unspecified and need to be learned in reality. For example, in both LinUCB (Li et al., 2010) and LinTS (Abeille & Lazaric, 2017; Agrawal & Goyal, 2013) algorithms, there are hyperparameters called exploration rates that govern the tradeoff and hence the learning process. But it has been empirically verified that the best exploration rate to use is always instance-dependent and may vary at different iterations (Bouneffouf & Claeys, 2020; Ding et al., 2022b). Note it is inherently impossible to use any state-of-the-art offline hyperparameter tuning methods such as cross validation (Stone, 1974) or Bayesian optimization (Frazier, 2018) since decisions in bandits should be made in real time. To choose the best hyperparameters, some previous works use grid search in their experiments (Ding et al., 2021; Jun et al., 2019), but

obviously, this approach is infeasible when it comes to reality, and how to manually discretize the hyperparameter space is also unclear. Conclusively, this limitation has already become a bottleneck for bandit algorithms in real-world applications, but unfortunately, it has rarely been studied in the previous literature.

The problem of hyperparameter optimization for contextual bandits was first studied in Bouneffouf & Claeys (2020), where the authors proposed two methods named OPLINUCB and DOPLINUCB to learn the practically optimal exploration rate of LinUCB in a finite candidate set by viewing each candidate as an arm and then using multi-armed bandit to pull the best one. However, 1) the authors did not provide any theoretical support, and 2) we believe the best exploration parameter in practice would vary during iterations – more exploration may be preferred at the beginning due to the lack of observations, while more exploitation would be favorable in the long run when the model estimate becomes more accurate. Furthermore, 3) they only consider tuning one single hyperparameter. To tackle these issues, Ding et al. (2022b) proposed TL and Syndicated framework by using a non-stationary multi-armed bandit for the hyperparameter set. However, their approach still requires a pre-defined set of hyperparameter candidates. In practice, choosing the candidates requires domain knowledge and plays a crucial role in the performance. Also, using a piecewise-stationary setting instead of a complete adversarial bandit (e.g. EXP3) for hyperparameter tuning is more efficient since we expect a fixed hyperparameter setting would yield indistinguishable results in a period of time. Conclusively, it would be more efficient to use a continuous space for bandit hyperparameter tuning.

We propose an efficient bandit-over-bandit (BOB) framework (Cheung et al., 2019) named Continuous Dynamic Tuning (CDT) framework for bandit hyperparameter tuning in the continuous hyperparameter space, without requiring a pre-defined set of hyperparameter candidate configurations. For the top layer bandit we formulate the online hyperparameter tuning as a non-stationary Lipschitz continuum-arm bandit problem with noise where each arm represents a hyperparameter configuration and the corresponding reward is the performance of the GLB, and the expected reward is a time-dependent Lipschitz function of the arm with some biased noise. Here the bias depends on the previous observations since the history could also affect the update of bandit algorithms. It is also reasonable to assume the Lipschitz functions are piecewise stationary since we believe the expected reward would be stationary with the same hyperparameter configuration over

a period of time (i.e. *switching* environment). Specifically, for the top layer of our CDT framework, we propose the Zooming TS algorithm with Restarts, and the key idea is to adaptively refine the hyperparameter space and zoom into the regions with more promising reward (Kleinberg et al., 2019) by using the TS methodology (Chapelle & Li, 2011). Moreover, we demonstrate that a simple restart trick could handle the piecewise changes of the bandit environments in both theory and practice. To sum up, we summarize our contributions as follows:

1) We propose an online continuous hyperparameter optimization framework for contextual bandits called CDT that handles all aforementioned issues of previous methods with theoretical guarantees. To the best of our knowledge, CDT is the first hyperparameter tuning method (even model selection method) with continuous candidates in the bandit community. 2) For the top layer of CDT, we propose the Zooming TS algorithm with Restarts for Lipschitz bandits under the *switching* environment. To the best of our knowledge, our work is the first one to consider the Lipschitz bandits under the *switching* environment, and the first one to utilize TS methodology in Lipschitz bandits. 3) Experiments on both synthetic and real datasets with various GLBs validate the efficiency of our method.

3.2. Related Work

There has been extensive literature on contextual bandit algorithms, and most of them are based on the UCB or TS techniques. For example, several UCB-type algorithms have been proposed for GLB, such as GLM-UCB (Filippi et al., 2010) and UCB-GLM (Li et al., 2017) that achieve the optimal $\tilde{O}(\sqrt{T})$ regret bound. Another rich line of work on GLBs follows the TS idea, including Laplace-TS (Chapelle & Li, 2011), SGD-TS (Ding et al., 2021), etc. In this paper, we focus on the hyperparameter tuning of contextual bandits, which is a practical but under-explored problem. For related work, Sharaf & Daumé III (2019) first studied how to learn the exploration parameters in contextual bandits via a meta-learning method. However, this algorithm fails to adjust the learning process based on previous observations and hence can be unstable in practice. Bouneffouf & Claeys (2020) then proposed OPLINUCB and DOPLINUCB to choose the exploration rate of LinUCB from a candidate set, and moreover Ding et al. (2022b) formulates the hyperparameter tuning problem as a non-stochastic multi-armed bandit and utilizes the classic EXP3 algorithm. However, as we mentioned in Section 3.1, both works have several limitations that could be decently fixed. Note that

hyperparameter tuning could be regarded as a branch of model selection in bandit algorithms. To name a few for this general problem, Agarwal et al. (2017) proposed a master algorithm that could combine multiple bandit algorithms, while Foster et al. (2019) initiated the study of model selection tradeoff in contextual bandits and proposed the first model selection algorithm for contextual linear bandits. However, these general model selection methods may fail for the bandit hyperparameter tuning task. To clarify this point, we take the state-of-the-art corralling idea (Agarwal et al., 2017) as an example: in theory, it has regret bound or order $O(\sqrt{MT} + MR_{\max})$ where M is the number of base models (number of hyperparameter combinations in our setting) and $R_{\max}$ is the regret of the worst candidate model in the tuning set. Therefore, on the one hand, M is infinitely large in our problem setting with a continuous candidate set, which means the regret bound would also be infinitely large. On the other hand, in order to achieve sub-linear regret in hyperparameter tuning, the corralling idea requires that all hyperparameter candidates yield sub-linear regret in theory, which is a very unrealistic assumption. On the contrary, our work only assumes the existence of a hyperparameter candidate in the tuning set which yields good theoretical regret in theory. In experiments, it is also costly to use since it requires updating all base models at each round, and we have infinitely many base models under our setting. (Ding et al., 2022b) includes the corralling idea in their experiments, and we can observe that it achieves almost linear regret in each setting since it has no sub-linear regret guarantee for the bandit hyperparameter tuning problem. In conclusion, the only existing methods that focus on hyperparameter tuning of bandits are OP and TL (Syndicated), and we use both of them in our paper as baselines. And we propose the first continuous hyperparameter tuning framework for contextual bandits, which doesn't require a pre-defined set of candidates. Note it is doable to finely discretize the continuous space and then implement an algorithm with discrete candidate sets (e.g. Syndicated) in methodology, but we highlight the inefficiency of this idea on both the empirical and theoretical side in Appendix B.1.4. We also briefly review the literature on Lipschitz bandits that follows two key ideas. One is uniformly discretizing the action space into a mesh (Kleinberg, 2004; Magureanu et al., 2014) so that any learning process like UCB could be directly utilized. Another more popular idea is adaptive discretization on the action space by placing more probes in more encouraging regions Bubeck et al. (2008); Kleinberg et al. (2019); Lu et al. (2019); Valko et al. (2013), and UCB could be used for exploration. Furthermore, the Lipschitz bandit under adversarial corruption was recently studied

in Kang et al. (2024c). In addition, Podimata & Slivkins (2021) proposed the first fully adversarial Lipschitz bandit in an adaptive refinement manner and derived instance-dependent regret bounds, but their algorithm relies on some unspecified hyperparameters and is computationally infeasible. Since the expected reward function for hyperparameters would not drastically change every time, it is also inefficient to use a fully adversarial algorithm here. Therefore, we introduce a new problem of Lipschitz bandits under the *switching* environment, and propose the Zooming TS algorithm with a restart trick to deal with the "almost stationary" nature of the bandit hyperparameter tuning problem.

3.3. Preliminaries

We first review the problem setting of contextual bandit algorithms. Denote T as the total number of rounds and K as the number of arms we could choose at each round, where K could be infinite. At each round $t \in [T] \coloneqq \{1, \dots, T\}$, the player is given K arms represented by a set of feature vectors $\mathcal{X}_t = \{x_{t,a} \,|\, a \in [K]\} \subseteq \mathbb{R}^d$ drawn from some unknown distribution, where $x_{t,a}$ is a d-dimensional vector containing information of arm a at round t. The player selects an action $a_t \in [K]$ based on the current $\mathcal{X}_t$ and previous observations, and only receives the payoff of the pulled arm a_t. Denote $x_t \coloneqq x_{t,a_t}$ as the feature vector of the chosen arm a_t and y_t as the corresponding reward. We assume the reward y_t follows a canonical exponential family with minimal representation, a.k.a. generalized linear bandits (GLB) with some mean function $\mu(\cdot)$. In addition, one can represent this model by $y_t = \mu(x_t^\top \theta^*) + \epsilon_t$, where ϵ_t follows a sub-Gaussian distribution with parameter σ^2 independent with the information filtration $\mathcal{F}_t = \sigma(\{a_s, \mathcal{X}_s, y_s\}_{s=1}^{t-1})$ and $\sigma(\mathcal{X}_t)$ up to round t, and θ^* is some unknown coefficient. Denote $a_{t,*} \coloneqq \arg\max_{a \in [K]} \mu(x_{t,a}^\top \theta^*)$ as the optimal arm at round t and $x_{t,*}$ as its corresponding feature vector. The goal is to minimize the expected cumulative regret defined as:

$$R(T) = \sum_{t=1}^{T} \left[\mu({x_{t,*}}^\top \theta^*) - \mathbb{E}\left(\mu(x_t^\top \theta^*) \right) \right]. \tag{3.1}$$

Note that all state-of-the-art contextual GLB algorithms depend on at least one hyperparameter to balance the well-known exploration-exploitation tradeoff. For example, LinUCB (Li et al., 2010),

the most popular UCB linear bandit, uses the following rule for arm selection at round t:

$$\text{(LinUCB)} \qquad a_t = \arg\max_{a\in[K]} x_{t,a}^\top \hat{\theta}_t + \alpha_1(t) \left\|x_{t,a}\right\|_{V_t^{-1}} .$$

Here the model parameter $\hat{\theta}_t$ is estimated at each round t via ridge regression, i.e. $\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s y_s$ where $V_t = \lambda I_r + \sum_{s=1}^{t-1} x_s x_s^\top$. And it considers the standard deviation of each arm with an exploration parameter $\alpha_1(t)$, where with a larger value of $\alpha_1(t)$ the algorithm will be more likely to explore uncertain arms. Note that the regularization parameter λ is only used to ensure V_t is invertible and hence its value is not crucial and commonly set to 1. In theory we can choose the value of $\alpha_1(t)$ as $\alpha_1(t) = \sigma\sqrt{r\log\left((1+t/\lambda)/\delta\right)} + \|\theta^*\|\sqrt{\lambda}$, to achieve the optimal $\widetilde{O}(\sqrt{T})$ bound of regret: However, in practice, the values of σ and $\|\theta^*\|$ are unspecified, and hence this theoretical value of $\alpha_1(t)$ is inaccessible. Furthermore, it has been shown that this is a very conservative choice that would lead to unsatisfactory practical performance, and the practically optimal hyperparameter values to use are distinct and far from the theoretical ones under different algorithms or settings. We also conduct a series of simulations with several state-of-the-art GLB algorithms to validate this fact, which is deferred to Appendix B.1.1. Conclusively, the best exploration parameter to use in practice should always be chosen dynamically based on the specific scenario and past observations. In addition, many GLB algorithms depend on some other hyperparameters, which may also affect the performance. For example, SGD-TS also involves a stepsize parameter for the stochastic gradient descent besides the exploration rate, and it is well known that a decent stepsize could remarkably accelerate the convergence (Loizou et al., 2021). To handle all these cases, we propose a general framework that can be used to automatically tune multiple continuous hyperparameters for a contextual bandit.

For a certain contextual bandit, assume there are p different hyperparameters $\alpha(t) = \{\alpha_i(t)\}_{i=1}^p$, and each hyperparameter $\alpha_i(t)$ could take values in an interval $[a_i, b_i]$, $\forall t$. Denote the parameter space $A = \bigotimes_{i=1}^p [a_i, b_i]$, and the theoretical optimal values as $\alpha^*(t)$. Given the observations $\mathcal{F}_t$ up to round t, we write $a_t(\alpha(t)|\mathcal{F}_t)$ as the arm we pulled when the hyperparameters are set to $\alpha(t)$, and $x_t(\alpha(t)|\mathcal{F}_t)$ as the corresponding feature vector.

Motivated by the success of Bayesian optimization (Frazier, 2018) on the hyperparameter tuning of the offline machine learning algorithms, the main idea of our algorithm is to formulate the hyperparameter optimization as a (another layer of) non-stationary Lipschitz bandit in the continuous

space $A \subseteq \mathbb{R}^p$, i.e. the agent chooses an arm (hyperparameter combination) $\alpha \in A$ in round $t \in [T]$, and then we decompose $\mu(x_t(\alpha|\mathcal{F}_t)^\top \theta^*)$ as

$$\mu(x_t(\alpha|\mathcal{F}_t)^\top \theta^*) = g_t(\alpha) + \eta_{\mathcal{F}_t,\alpha}. \tag{3.2}$$

Here g_t is some time-dependent Lipschitz function that formulates the performance of the bandit algorithm under the hyperparameter combination α at round t, since the bandit algorithm tends to pull similar arms if the chosen values of hyperparameters are close at round t. In other words, we expect close hyperparameter values to yield similar results with other conditions fixed, as in Bayesian optimization on offline hyperparameter tuning. To demonstrate that our Lipschitz assumption w.r.t. the hyperparameter values in Eqn. (3.3) is reasonable, we conduct simulations with LinUCB and LinTS, and defer it to Appendix B.1 due to the space limit. Moreover, $(\eta_{\mathcal{F}_t,\alpha} - \mathbb{E}[\eta_{\mathcal{F}_t,\alpha}])$ is IID sub-Gaussian with parameter τ^2, and to be fair we assume $\mathbb{E}[\eta_{\mathcal{F}_t,\alpha}]$ could also depend on the history $\mathcal{F}_t$ since past observations and action sets would explicitly influence the model parameter estimation and hence the decision making at each round. In addition to Lipschitzness, we also suppose g_t follows a *switching* environment: g_t is piecewise stationary with some change points, i.e.

$$|g_t(\alpha_1) - g_t(\alpha_2)| \leq \|\alpha_1 - \alpha_2\|, \ \forall \alpha_1, \alpha_2 \in A; \tag{3.3}$$

$$\sum_{t=1}^{T-1} \mathbf{1}[\exists \alpha \in A : \ g_t(\alpha) \neq g_{t+1}(\alpha)] = c(T), \ c(T) \in \mathbb{N}. \tag{3.4}$$

Since after sufficient exploration, the expected reward should be stable with the same hyperparameter setting, we could assume that $c(T) = \tilde{O}(1)$. Detailed justification on this piecewise Lipschitz assumption is deferred to Remark B.2.1 in Appendix B.2. Although numerous research works have considered the *switching* environment (a.k.a. *abruptly-changing* environment) for multi-armed or linear bandits (Auer et al., 2002b; Wei et al., 2016), our work is the first to introduce this setting into the continuum-armed bandits. The *switching* environment is extended from the well-known classic change-point detection problem in statistics, which has great applications in a wide range of fields such as climatology (Reeves et al., 2007) and neuroscience (Ma et al., 2022a). In Section 3.4.1, we will show that by combining our proposed Zooming TS algorithm for Lipschitz bandits with a simple restarted strategy, a decent regret bound could be achieved under the *switching* environment.

Algorithm 3 Zooming TS algorithm with Restarts

Input: Time horizon T, space A, epoch size H.

for $t = 1$ **to** T **do**
  if $t \in \{\tau H + 1 : \tau = 0, 1, \dots\}$ **then**
    Initialize the total candidate space $A_0 = A$ and the active set $J \subseteq A_0$ s.t. $A_0 \subseteq \cup_{v \in J} B(v, r_1(v))$ and $n_1(v) \leftarrow 1, \forall v \in J$. ▷ Restart
  else if $\hat{f}_t(v) - \hat{f}_t(u) > r_t(v) + 2r_t(u)$ for some pair of $u, v \in J$ **then**
    Set $J = J \backslash \{u\}$ and $A_0 = A_0 \backslash B(u, r_t(u))$. ▷ Removal
  if $A_0 \nsubseteq \cup_{v \in J} B(v, r_t(v))$ **then** ▷ Activation
    Activate and pull some point $v \in A_0$ that has not been covered: $J = J \cup \{v\}, v_t = v$.
  else
    $v_t = \arg\max_{v \in J} I_t(v)$, break ties arbitrarily. ▷ Selection
  Observe the reward $\tilde{y}_{t+1}$, and then update components in the Zooming TS algorithm: $n_{t+1}(v), \hat{f}_{t+1}(v), r_{t+1}(v), s_{t+1}(v)$ for the chosen $v_t \in J$:

$$n_{t+1}(v_t) = n_t(v_t) + 1, \hat{f}_{t+1}(v_t) = (\hat{f}_t(v_t) n_t(v_t) + \tilde{y}_{t+1}) / n_{t+1}(v_t).$$

3.4. Methods

3.4.1. Lipschitz Bandits under the Switching Environment. For simplicity and consistency, we will reload and introduce a new system of notations in this subsection. Consider the non-stationary Lipschitz bandit problem on a compact space A under some metric $\text{Dist}(\cdot, \cdot) \geq 0$, where the covering dimension is denoted by p_c. The learner pulls an arm $v_t \in A$ at round $t \in [T]$ and subsequently receives a reward $\tilde{y}_t$ sampled independently of $\mathbb{P}_{v_t}$ as $\tilde{y}_t = f_t(v_t) + \eta_v$, where $t = 1, \dots, T$ and η_v is IID zero-mean error with sub-Guassian parameter τ_0^2, and f_t is the expected reward function at round t and is Lipschitz with respect to $\text{Dist}(\cdot, \cdot)$. The *switching* environment assumes the time horizon T is partitioned into $c(T) + 1$ intervals, and the bandit stays stationary within each interval, i.e.

$$|f_t(m) - f_t(n)| \leq \text{Dist}(m, n), \quad m, n \in A; \text{ and } \sum_{t=1}^{T-1} \mathbf{1}[\exists m \in A : \, f_t(m) \neq f_{t+1}(m)] = c(T).$$

Here in this section $c(T) = o(T)$ could be any integer. The goal of the learner is to minimize the expected (dynamic) regret that is defined as:

$$R_L(T) = \sum_{t=1}^{T} \max_{v \in A} f_t(v) - \sum\nolimits_{t=1}^{T} \mathbb{E}\left(f_t(v_t)\right).$$

Since we consider the non-stationary Lipschitz bandit different from the setting in Chapter 2, we will reload some notations here: at each round t, $v_t^* := \arg\max_{v \in A} f_t(v)$ denotes the maximal point

(w.l.o.g. assume it's unique), and $\Delta_t(v) = f_t(v^*) - f_t(v)$ is the "badness" of each arm v. We also denote $A_{r,t}$ as the r-optimal region at the scale $r \in (0, 1]$, i.e. $A_{r,t} = \{v \in A : r/2 < \Delta_t(v) \leq r\}$ at time t. Then the r-zooming number $N_{z,t}(r)$ of (A, f_t) is defined as the minimal number of balls of radius no more than r required to cover $A_{r,t}$. (Note the subscript z stands for zooming here.) Next, we define the zooming dimension $p_{z,t}$ at time t as the smallest $q \geq 0$ such that for every $r \in (0, 1]$ the r-zooming number can be upper bounded by cr^{-q} for some multiplier $c > 0$ free of r:

$$p_{z,t} = \min\{q \geq 0 : \exists c > 0, N_{z,t}(r) \leq cr^{-q}, \forall r \in (0, 1]\}.$$

It's obvious that $0 \leq p_{z,t} \leq p_c$, $\forall t \in [T]$. (Note $p_{z,t}$ is fixed under the stationary environment.) On the other hand, the zooming dimension could be much smaller than p_c under some mild conditions. For example, if the payoff function f_t defined on $\mathbb{R}^{p_c}$ is greater than $\|v_t^* - v\|^\beta$ in scale for some $\beta \geq 1$ around v^* in the space A, i.e. $f_t(v_t^*) - f_t(v) = \Omega(\|v_t^* - v\|^\beta)$, then it holds that $p_{z,t} \leq (1 - 1/\beta)p_c$. Note that we have $\beta = 2$ (i.e. $p_{z,t} \leq p_c/2$) when $f_t(\cdot)$ is C^2-smooth and strongly concave in a neighborhood of v^*. More details are presented in Appendix B.3. Since the expected reward Lipschitz function $f_t(\cdot)$ is fixed in each time interval under the *switching* environment, the zooming number and zooming dimension $p_{z,t}$ would also stay identical. And we also write $p_{z,*} = \max_{t\in[T]} p_{z,t} \leq p_c$.

Our proposed Algorithm 3 extends the classic Zooming algorithm (Kleinberg et al., 2019), which was used under the stationary Lipschitz bandit environment, by adding several new ingredients for better efficiency and adaptivity to non-stationary environment: on the one hand, we employ the TS methodology and propose a novel removal step. Here we utilize TS since it was shown that TS is more robust than UCB in practice (Chapelle & Li, 2011; Wang & Chen, 2018), and the removal procedure in line 5 of Algorithm 3 could adaptively subtract regions that are prone to yield low rewards. Both of these two ideas could enhance the algorithmic efficiency, which coincides with the practical orientation of our work. On the other hand, the restarted strategy proceeds our proposed Zooming TS in epochs and refreshes the algorithm after every H rounds. The epoch size H is fixed through the total time horizon and controls the tradeoff between non-stationarity and stability. Note that H in our algorithm does not need to match the actual length of stationary intervals of the environment, and we would discuss its selection later. At each epoch, we maintain a time-varying active arm set $S_t \subseteq A$, which is initially empty and updated every time. For each

arm $v \in A$ and time t, denote $n_t(v)$ as the number of times arm v has been played before time t since the last restart, and $\hat{f}_t(v)$ as the corresponding average sample reward. We let $\hat{f}_t(v) = 0$ when $n_t(v) = 0$. Define the confidence radius and the TS standard deviation of active arm v at time t respectively as

$$r_t(v) = \sqrt{\frac{13\tau_0^2 \ln T}{2n_t(v)}}, \quad s_t(v) = s_0\sqrt{\frac{1}{n_t(v)}}, \tag{3.5}$$

where $s_0 = \sqrt{52\pi\tau_0^2 \ln(T)}$. We call $B(v, r_t(v)) = \{u \in \mathbb{R}^p : \text{Dist}(u, v) \leq r_t(v)\}$ as the confidence ball of arm v at time $t \in [T]$. We construct a randomized algorithm by choosing the best active arm according to the perturbed estimate mean $I_t(\cdot)$:

$$I_t(v) = \hat{f}_t(v) + s_t(v)Z_{t,v}, \tag{3.6}$$

where $Z_{t,v}$ is i.i.d. drawn from the clipped standard normal distribution: we first sample $\tilde{Z}_{t,v}$ from the standard normal distribution and then set $Z_{t,v} = \max\{1/\sqrt{2\pi}, \tilde{Z}_{t,v}\}$. This truncation was also used in TS multi-armed bandits (Jin et al., 2021), and our algorithm clips the posterior samples with a lower threshold to avoid underestimation of good arms. Moreover, the explanations of the TS update is deferred to Appendix B.4 due to the space limit.

The regret analysis of Algorithm 3 is very challenging since the active arm set is constantly changing and the optimal arm v^* cannot be exactly recovered under the Lipschitz bandit setting. Thus, existing theory on multi-armed bandits with TS is not applicable here. We overcome these difficulties with some innovative use of metric entropy theory, and the regret bound of Algorithm 3 is given as follows.

THEOREM 3.4.1. *With $H = \Theta\left((T/c(T))^{(p_{z,*}+2)/(p_{z,*}+3)}]\right)$, the total regret of our Zooming TS algorithm with Restarts under the switching environment over time T is bounded as*

$$R_L(T) \leq \tilde{O}\left((c(T))^{1/(p_{z,*}+3)}\, T^{(p_{z,*}+2)/(p_{z,*}+3)}\right),$$

when $c(T) > 0$. In addition, if the environment is stationary (i.e. $c(T) = 0, f_t = f, p_{z,t} = p_{z,*} \coloneqq p_z, \forall t \in [T]$), then by using $H = T$ (i.e. no restart), our Zooming TS algorithm could achieve the

Algorithm 4 Continuous Dynamic Tuning (CDT)

Input: $T_1, T_2, \{\mathcal{X}_t\}_{t=1}^{T}, A = \bigotimes_{i=1}^{p}[a_i, b_i]$.

1: Randomly choose $a_t \in [K]$ and observe x_t, y_t, $t \leq T_1$.
2: Initialize the hyperparameter active set J s.t. $A \subseteq \cup_{v \in J} B(v, r_1(v))$ where $n_{T_1}(v) \leftarrow 1, \forall v \in J$.
3: **for** $t = (T_1 + 1)$ **to** T **do**
4: Run the t-th iteration of Algorithm 3 with initial input horizon $T - T_1$, input space A and restarting epoch length T_2. Denote the pulled arm at round t as $\alpha(i_t) \in A$. ▷ Top
5: Run the contextual bandit algorithm with hyperparameter $\alpha(i_t)$ to pull an arm a_t. ▷ Bottom
6: Obtain y_t and update components in the contextual bandit algorithm. ▷ Bottom Update
7: Update components in Algorithm 1 by treating y_t as the reward of arm $\alpha(i_t)$ ▷ Top Update

optimal regret bound for Lipschitz bandits up to logarithmic factors:

$$R_L(T) \leq \tilde{O}\left(T^{(p_z+1)/(p_z+2)}\right).$$

We also present empirical studies to further evaluate the performance of our Algorithm 3 compared with stochastic Lipschitz bandit algorithms in Appendix B.1.3. A potential drawback of Theorem 3.4.1 is that the optimal epoch size H under *switching* environment relies on the value of $c(T)$ and $p_{z,*}$, which are unspecified in reality. However, this problem could be solved in theory by using the BOB idea (Cheung et al., 2019; Zhao et al., 2020) to adaptively choose the optimal epoch size with a meta algorithm (e.g. EXP3 (Auer et al., 2002b)) in real time. In this case, we prove the expected regret can be bounded by the order of $\tilde{O}\left(T^{\frac{p_c+2}{p_c+3}} \cdot \max\left\{c(T)^{\frac{1}{p_c+3}}, T^{\frac{1}{(p_c+3)(p_c+4)}}\right\}\right)$ in general, and some better regret bounds in problem-dependent cases. More details are presented in Theorem B.6.1 with its proof in Appendix B.6. However, in the following Section 3.4.2 we could simply set $H = T^{(2+p)/(3+p)}$ in our CDT framework where p is the number of hyperparameters to be tuned after assuming $c(T) = \tilde{O}(1)$ is of constant scale up to logarithmic terms. The value of τ_0 can be determined by assuring the observed rewards are bounded. Note our work introduces a new problem on Lipschitz bandits under the *switching* environment. One potential limitation of our work is how to deduce a regret lower bound under this problem setting is unclear, and we leave it as a future work.

3.4.2. Continuous Hyperparameter Tuning (CDT). Based on the proposed algorithm in the previous subsection, we introduce our online double-layer Continuous Dynamic Tuning (CDT) framework for hyperparameter optimization of contextual bandit algorithms. We assume the arm to be pulled follows a fixed distribution given the hyperparameters to be used and the history at

each round. The detailed algorithm is shown in Algorithm 4. Our method extends the bandit-over-bandit (BOB) idea that was first proposed for non-stationary stochastic bandit problems (Cheung et al., 2019), where it adjusts the sliding-window size dynamically based on the changing model. In our work, for the top layer we use our proposed Algorithm 3 to tune the best hyperparameter values from the admissible space, where each arm represents a hyperparameter configuration and the corresponding reward is the algorithmic result. T_2 is the length of each epoch (i.e. H in Algorithm 3), and we would refresh our Zooming TS Lipschitz bandit after every T_2 rounds as shown in Line 5 of Algorithm 4 due to the non-stationarity. The bottom layer is the primary contextual bandit and would run with the hyperparameter values $\alpha(i_t)$ chosen from the top layer at each round t. We also include a warming-up period of length T_1 in the beginning to guarantee sufficient exploration as in Ding et al. (2021); Li et al. (2017). Despite the focus of our CDT framework is on the practical aspect, we also present a novel theoretical analysis in the following for the completeness of our work.

Although there has been a rich line of work on regret analysis of UCB and TS GLB algorithms, most literature certainly requires that some hyperparameters, e.g. exploration rate, always take their theoretical values. It is challenging to study the regret bound of GLB algorithms when their hyperparameters are synchronously tuned in real time, since the chosen hyperparameter values may be far from the theoretical ones in practice, not to mention that previous decisions would also affect the current update cumulatively. Moreover, there is currently no existing literature and regret analysis on hyperparameter tuning (or model selection) for bandit algorithms with an infinite number of candidates in a continuous space. Recall that we denote $\mathcal{F}_t = \sigma\left(\{a_s, \mathcal{X}_s, y_s\}_{s=1}^{t-1}\right)$ as the past information before round t under our CDT framework, and a_t, x_t are the chosen arm and its corresponding feature vector at time t, which implies that $a_t = a_t(\alpha(i_t)|\mathcal{F}_t), x_t = x_t(\alpha(i_t)|\mathcal{F}_t)$. Furthermore, we denote $\alpha^*(t)$ as the theoretical optimal value at round t and $\mathcal{F}_t^*$ as the past information filtration by always using the theoretical optimal $\alpha^*(t)$. Since the decision at each round t also depends on the history observed by time t, the pulled arm with the same hyperparameter $\alpha(t)$ might be different under $\mathcal{F}_t$ or $\mathcal{F}_t^*$. To analyze the cumulative regret $R(T)$ of our Algorithm 4, we first decompose it into four quantities:

$$R(T) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T_1}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t{}^\top\theta^*)\right)\right]}_{\text{Quantity (A)}} + \underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*)\right)\right]}_{\text{Quantity (B)}}$$

$$+\underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*\right) - \mu(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*)\right)\right]}_{\text{Quantity (C)}}$$

$$+\underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*\right) - \mu(x_t(\alpha(i_t)|\mathcal{F}_t)^\top\theta^*)\right)\right]}_{\text{Quantity (D)}}.$$

Intuitively, Quantity (A) is the regret paid for pure exploration during the warming-up period and could be controlled by the order $O(T_1)$. Quantity (B) is the regret of the contextual bandit algorithm that runs with the theoretical optimal hyperparameters $\alpha^*(t)$ all the time, and hence it could be easily bounded by the optimal scale $\tilde{O}(\sqrt{T})$ based on the literature. Quantity (C) is the difference of cumulative reward with the same $\alpha^*(t)$ under two separate lines of history. Quantity (D) is the extra regret paid to tune the hyperparameters on the fly. By using the same line of history $\mathcal{F}_t$ in Quantity (D), the regret of our Zooming TS algorithm with Restarts in Theorem 3.4.1 can be directly used to bound Quantity (D). Conclusively, we deduce the following theorem for the regret bound:

THEOREM 3.4.2. *Under our problem setting in Section 3.3, for UCB and TS GLB algorithms with exploration hyperparameters (e.g. LinUCB, UCB-GLM, GLM-UCB, LinTS), by taking* $T_1 = O(T^{2/(p+3)}), T_2 = O(T^{(p+2)/(p+3)})$ *where* p *is the number of hyperparameters, and let the theoretically optimal hyperparameter combination* $\alpha^*(T) \in A$*, it holds that*

$$\mathbb{E}[R(T)] \leq \tilde{O}(T^{(p+2)/(p+3)}).$$

The detailed proof of Theorem 3.4.2 is presented in Appendix B.7. Note that this regret bound could be further improved to $\tilde{O}(T^{(p_0+2)/(p_0+3)})$ where p_0 is any constant that is no smaller than the zooming dimension of $(A, g_t), \forall t$. For example, from Figure B.1 in Appendix B.1 we can

observe that in practice g_t would be C^2-smooth and strongly concave, which implies that $\mathbb{E}[R(T)] \leq \tilde{O}(T^{(p+4)/(p+6)})$.

Note our work is the first one to consider model selection for bandits with a continuous candidate set, and the regret analysis for online model selection in the bandit setting (Foster et al., 2019) is intrinsically more difficult compared with the offline model selection (Han & Lee, 2022; Zhao & Yu, 2006). For example, regret bounds of the algorithm CORRAL (Agarwal et al., 2017) for model selection and Syndicated (Ding et al., 2022b) for bandit hyperparameter tuning are (sub)linearly dependent on the number of candidates, which would be infinitely large and futile in our case. Furthermore, given the fact that Syndicated in Ding et al. (2022b) fails to recover the optimal $O(\sqrt{T})$ bound of regret without stringent assumptions under the easier setting with finite hyperparameter candidates, it would be substantially difficult to deduce a feasible regret bound under our more complicated problem setting. Moreover, the non-stationarity under the *switching* environment would further deteriorate the optimal order of cumulative regret (Cheung et al., 2019). And it is intrinsically more difficult to consider the continuum-armed bandit over the multi-armed bandit. Therefore, we believe our theoretical result is non-trivial and significant. Our work stands as the first seminal attempt in bandit hyperparameter tuning (or even bandit model selection) with an infinite number of candidates. An extensive study on this new problem will be an interesting future direction.

3.5. Experimental Results

In this section, we show by experiments that our hyperparameter tuning framework outperforms the theoretical hyperparameter setting and other tuning methods with various (generalized) linear bandit algorithms. We utilize seven state-of-the-art bandit algorithms: two of them (LinUCB (Li et al., 2010), LinTS (Agrawal & Goyal, 2013)) are linear bandits, and the other five algorithms (UCB-GLM (Li et al., 2017), GLM-TSL (Kveton et al., 2020), Laplace-TS (Chapelle & Li, 2011), GLOC (Jun et al., 2017), SGD-TS (Ding et al., 2021)) are GLBs. Note that all these bandit algorithms except Laplace-TS contain an exploration rate hyperparameter, while GLOC and SGD-TS further require an additional learning parameter. And Laplace-TS only depends on one stepsize hyperparameter for a gradient descent optimizer. We compare our CDT framework with the theoretical setting, OP (Bouneffouf & Claeys, 2020) and TL (Ding et al., 2022b) (one hyperparameter)

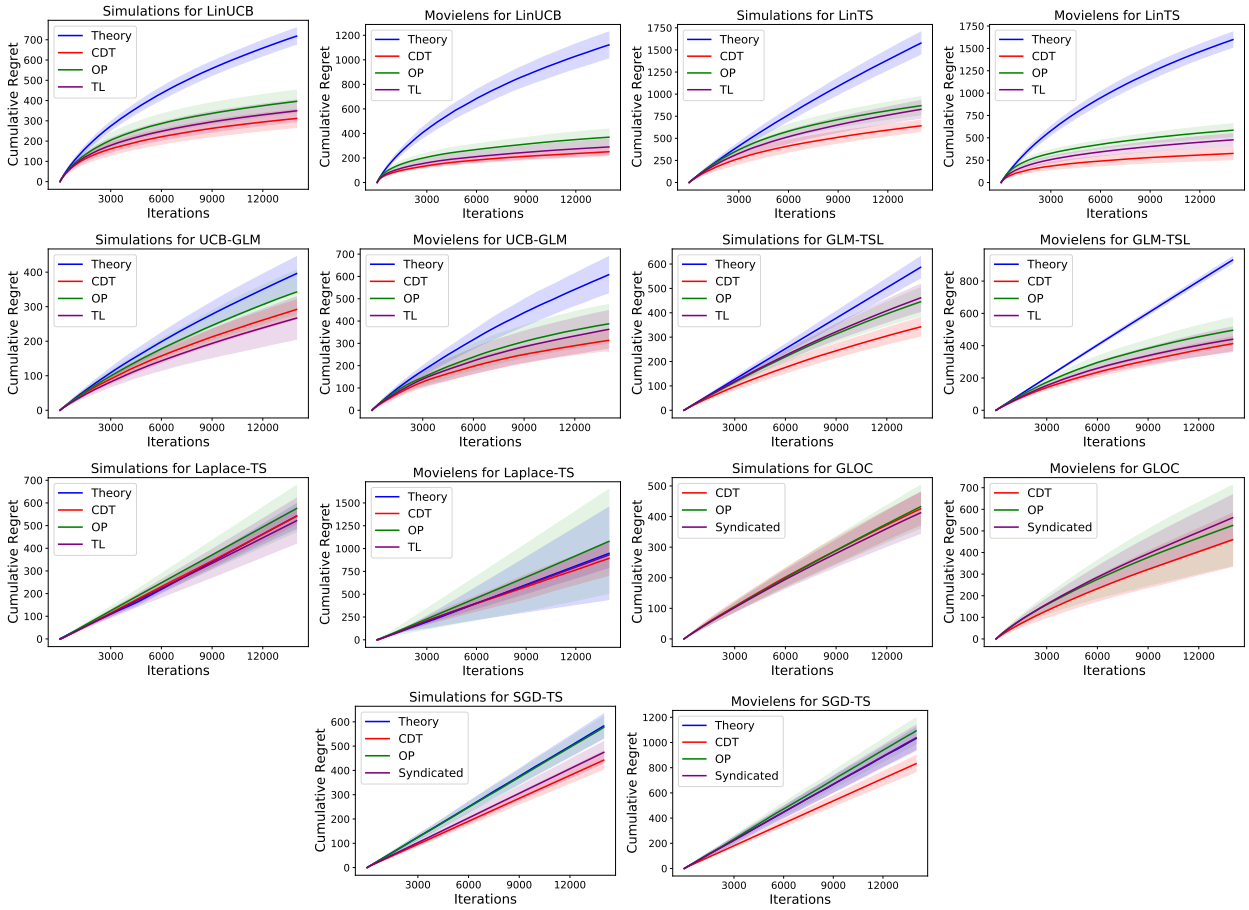


FIGURE 3.1. Cumulative regret curves of our CDT framework compared with existing hyperparameter selection methods under multiple (generalized) linear bandit algorithms on the simulations and Movielens dataset.

and Syndicated (Ding et al., 2022b) (multiple hyperparameters) algorithms. Their details are given as follows:

(1) **Theoretical setting**: We implement the theoretical exploration rate and stepsize for each algorithm. For the stepsize of gradient descent used in SGD-TS and Laplace-TS, we set it as 1 instead. (We observe the algorithmic performance is not sensitive to this stepsize.)

(2) **OP**: Bouneffouf & Claeys (2020) proposes OPLINUCB to tune the exploration rate of LinUCB. Here we modify it so that it could be used in other bandit algorithms. Note that OP is only applicable to algorithms with one hyperparameter, and hence we fix the learning parameter of GLOC and SGD-TS as their theoretical values instead, and only tune the exploration rates.

(3) **TL** (Ding et al., 2022b) (one hyperparameter): For algorithms with only one hyperparameter, TL is used.

(4) **Syndicated** (Ding et al., 2022b) (multiple hyperparameters): For GLOC and SGD-TS (two hyperparameters), the Syndicated framework is utilized for comparison.

We run comprehensive experiments on both simulations and real-world datasets. Specifically, for the real data, we use the benchmark Movielens 100K dataset along with the Yahoo News dataset:

(1) **Simulation**: In each repetition, we simulate all the feature vectors $\{x_{t,a}\}$ and the model parameter θ^* according to Uniform($-1/\sqrt{r}$,$1/\sqrt{r}$) elementwisely, and hence we have $\|x_{t,a}\| \leq 1$. We set d =25, K =120 and T =14,000. For linear model, the expected reward of arm a is formulated as $x_{t,a}^\top \theta^*$ and random noise is sampled from $N(0, 0.25)$; for Logistic model, the mean reward of arm a is defined as $p = 1/(1+\exp(-x_{t,a}^\top \theta^*))$, and the output is drawn from a Bernoulli distribution.

(2) **Movielens 100K dataset**: This dataset contains 100K ratings from 943 users on 1,682 movies. For data pre-processing, we utilize LIBPMF (Yu et al., 2014) to perform matrix factorization and obtain the feature matrices for both users and movies with d =20, and then normalize all feature vectors into unit r-dimensional ball. In each repetition, the model parameter θ^* is defined as the average of 300 randomly chosen users' feature vectors. And for each time t, we randomly choose $K = 300$ movies from 1,682 available feature vectors as arms $\{x_{t,a}\}_{a=1}^{300}$. The time horizon T is set to 14,000. For linear models, the expected reward of arm a is formulated as $x_{t,a}^\top \theta^*$ and random noise is sampled from $N(0, 0.5)$; for Logistic model, the output of arm a is drawn from the Bernoulli distribution with $p = 1/(1 + \exp(-x_{t,a}^\top \theta^*))$.

(3) **Yahoo News dataset:** We downloaded the Yahoo Recommendation dataset R6A, which contains Yahoo data from May 1 to May 10, 2009 with $T = 2881$ timestamps. For each user's visit, the module will select one article from a pool of 20 articles for the user, and then the user will decide whether to click. We transform the contextual information into a 6-dimensional vector based on the processing in Chu et al. (2009). We build a Logistic bandit on this data, and the observed reward is simulated from a Bernoulli distribution with a probability of success equal to its click-through rate at each time.

We first present the results on simulations and Movielens datasets: since all the existing tuning algorithms require a user-defined candidate set, we design the tuning set for all potential hyperparameters as $\{0.1, 1, 2, 3, 4, 5\}$. And for our CDT framework, which is the first algorithm for tuning hyperparameters in an interval, we simply set the interval as $[0.1, 5]$ for all hyperparameters. Each

experiment is repeated for 20 times, and the average regret curves with standard deviation are displayed in Figure 3.1. We further explore the existing methods after enlarging the hyperparameter candidate set to fairly validate the superiority of our proposed CDT in Appendix B.1.4.1. The results in Appendix B.1.4.1 further lead to discussion on why it is inefficient to first discretize the continuous space and then implement an algorithm (e.g. Syndicated) with discrete candidate sets. We believe a large value of warm-up period T_1 may abandon some useful information in practice, and hence we use $T_1 = T^{2/(p+3)}$ according to Theorem 3.4.2 in experiments. And we would restart our hyperparameter tuning layer after every $T_2 = 3T^{(p+2)/(p+3)}$ rounds. An ablation study on the role of T_1, T_2 in our CDT framework is also conducted and deferred to Appendix B.1.4.2, where we demonstrate that the performance of CDT is pretty robust to the choice of T_1, T_2 in practice.

From Figure 3.1, we observe that our CDT framework outperforms all existing hyperparameter tuning methods for most contextual bandit algorithms. It is also clear that CDT performs stably and soundly with the smallest standard deviation across most datasets (e.g. experiments for LinTS, UCB-GLM), indicating that our method is highly flexible and robustly adaptive to different datasets. Moreover, when tuning multiple hyperparameters (GLOC, SGD-TS), we can see that the advantage of our CDT is also evident since our method is intrinsically designed for any hyperparameter space. It is also verified that the theoretical hyperparameter values are too conservative and would lead to terrible performance (e.g. LinUCB, LinTS). Note that all tuning methods exhibit similar results when applied to Laplace-TS. We believe it is because Laplace-TS only relies on an insensitive hyperparameter that controls the stepsize in gradient descent loops, which mostly affects the convergence speed.

For the Yahoo News Recommendation dataset, since it is a logistic bandit, we only output the cumulative rewards of GLBs in Table 3.1. From the table, we can observe that our proposed CDT also performs the best overall. Specifically, it is only slightly worse than TL for GLM-TSL and GLOC, and yields the best results among all hyperparameter tuning frameworks for UCB-GLM, GLM-TSL, and SGD-TS. And the theoretical hyperparameter setting is very unstable again as in Figure 3.1. Conclusively, our proposed CDT yields uniformly the best performances compared with existing baselines in both large-scale and mild-scale experiments with multiple contextual bandit algorithms. This fact also validates the rationality of Lipschitz continuity assumption on the bandit hyperparameter tuning problem in Section 3.3.

Method	UCB-GLM	GLM-TSL	Laplace-TS	GLOC	SGD-TS
Theory	221.51	214.67	217.38		206.73
CDT	221.69	218.27	217.05	217.95	218.35
OP	217.25	217.08	213.95	216.28	215.58
TL/Syndicated	218.95	219.36	214.42	218.19	215.02

TABLE 3.1. Comparisons of cumulative rewards from different algorithms on Yahoo dataset.

CHAPTER 4

Efficient Frameworks for Low-rank Matrix Bandits

4.1. Introduction

The contextual bandit has proven to be a powerful framework for sequential decision-making problems, with great applications to clinical trials (Woodroofe, 1979), recommendation system (Li et al., 2010), and personalized medicine (Bastani & Bayati, 2020). This class of problems evaluates how an agent should choose an action from the potential action set at each round based on an updating policy on-the-fly so as to maximize the cumulative reward or minimize the overall regret. With high dimensional sparse data becoming ubiquitous in various fields nowadays (Zhao & Yu, 2006; Zhu et al., 2019), the most fundamental (generalized) linear bandit framework, although has been extensively studied, becomes inefficient in practice. This fact consequently leads to a line of work on stochastic high dimensional bandit problems with low dimensional structures (Johnson et al., 2016; Li et al., 2022), such as the LASSO bandit and low-rank matrix bandit.

In this work, we investigate on the generalized low-rank matrix bandit problem firstly studied in Lu et al. (2021): at round $t = 1, \ldots, T$, the algorithm selects an action represented by a d_1 by d_2 matrix X_t from the admissible action set $\mathcal{X}_t$ ($\mathcal{X}_t$ may be fixed), and receives its associated noisy reward $y_t = \mu(\langle \Theta^*, X_t \rangle) + \eta_t$ where $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is some unknown low-rank matrix with rank $r \ll \{d_1, d_2\}$ and $\mu(\cdot)$ is the inverse link function. More details about this setting are deferred to Section 4.3. This problem has vast applicability in real world applications. On the one hand, matrix inputs are appropriate when dealing with paired contexts which are omnipresent in practice. For instance, to design a personalized movie recommendation system, we can formulate each user as m d_1-dimensional feature vectors $(x_1, \ldots, x_m \in \mathbb{R}^{d_1})$ and each movie as m d_2-dimensional feature vectors $(y_1, \ldots, y_m \in \mathbb{R}^{d_2})$. A user-item pair can then be naturally represented by a feature matrix defined as the summation of the outer products $\sum_{k=1}^{m} x_k y_k^\top \in \mathbb{R}^{d_1 \times d_2}$, which will become the contextual feature observed by the bandit algorithm. Other applications involve interaction features between two groups, such as flight-hotel bundles (Lu et al., 2021) and dating service (Jun et al., 2019) can

also be similarly established. Besides, low-rank models have gained tremendous success in various areas (Candès & Recht, 2009). In particular, our problem can be regarded as an extension of the inductive matrix factorization problem (Jain & Dhillon, 2013; Zhong et al., 2015), which estimates low-rank matrices with contextual information, under the online learning scenario.

Our study is inspired by a line of work on stochastic contextual low-rank matrix bandit (Jang et al., 2021; Jun et al., 2019; Lu et al., 2021). To design an algorithm for matrix bandit problems, a naïve approach is to flatten the d_1 by d_2 feature matrices into vectors and then apply any (generalized) linear bandit algorithms, which, however, would be inefficient when $d_1 d_2$ is large. To take advantage of the low-rank structure, Jun et al. (2019) have introduced the bilinear low-rank bandit problem and proposed a two-stage algorithm named ESTR which could achieve a regret bound of $\tilde{O}((d_1+d_2)^{3/2}\sqrt{rT}/D_{rr})$[1]. Subsequently, Jang et al. (2021) constructed a new algorithm called ϵ-FALB for bilinear bandits and achieved a better regret of $\tilde{O}(\sqrt{d_1 d_2 (d_1+d_2)T})$. However, they only studied the linear reward framework and also restricted the feature matrix as a rank-one matrix. As a follow-up work, Lu et al. (2021) further released the rank-one restriction on the action feature matrices, and they introduced an algorithm LowGLOC based on the online-to-confidence-set conversion (Abbasi-Yadkori et al., 2012) for generalized low-rank matrix bandits with $\tilde{O}(\sqrt{(d_1+d_2)^3 rT})$ regret bound. However, this method can't handle the contextual setting since the arm set is assumed fixed at each round. This algorithm is also computationally prohibitive since it requires to calculate the weights of a self-constructed covering of the admissible parameter space at each iteration. And how to find this covering for low-rank matrices is also unclear.

In this work, we propose two efficient methods called G-ESTT and G-ESTS for this problem by modifying two stages of ESTR appropriately from different perspectives. To the best of our knowledge, the proposed methods are the first two generalized (contextual) low-rank bandit algorithms that are computationally feasible, and achieve the decent regret bound of $\tilde{O}(\sqrt{(d_1+d_2)^3 rT}/D_{rr})$ and $\tilde{O}((d_1+d_2)^{7/4} r^{3/4} T/D_{rr})$ on low-rank bandits. The main contributions of this paper can be summarized as: **1)** we propose two novel two-stage frameworks G-ESTT and G-ESTS under some mild assumptions. Compared with ESTR in Jun et al. (2019), ϵ-FALB in Jang et al. (2021) and LowESTR in Lu et al. (2021), our algorithms are proposed for the nonlinear reward framework with arbitrary action matrices. Compared with LowGLOC in Lu et al. (2021), our algorithms are

[1] $\tilde{O}$ ignores the polylogarithmic factors.

computationally feasible in practice. **2)** For G-ESTT, we extend the GLM-UCB algorithms (Filippi et al., 2010) via a novel regularization technique. **3)** Our proposed G-ESTS is simple and could be easily implemented based on any state-of-the-art misspecified linear bandit algorithms to achieve the regret bound of order $\tilde{O}(d_1+d_2)^{7/4}r^{3/4}T/D_{rr})$. In practice, it can be used with any generalized linear bandit to achieve high efficiency. Particularly, when we combine G-ESTS with some efficient algorithms (e.g. SGD-TS (Ding et al., 2021)), the total time complexity after a warm-up stage scales as $O(Tr(d_1+d_2))$. **4)** The practical superiority of our algorithms are firmly validated based on our experimental results.

4.2. Related Work

In this section, we briefly discuss some previous algorithms on low-rank matrix bandit problems. Besides the works we have discussed in the former section, Katariya et al. (2017a); Trinh et al. (2020) considered the rank-one bandit problems where the expected reward forms a rank-one matrix and the player selects an element from this matrix as the expected reward at each round. In addition, Katariya et al. (2017b) also studied the rank-one matrix bandit via an elimination-based algorithm. Alternatively, Gopalan et al. (2016); Kveton et al. (2017); Lu et al. (2018) considered the general low-rank matrix bandit, and furthermore Hao et al. (2020) considered a stochastic low-rank tensor bandit. However, for all these works the feature matrix of an action could be flattened into a one-hot basis vector, and our work yields a more general structure.

Additionally, Li et al. (2022) extended some previous works (Johnson et al., 2016) and presented a unified algorithm based on a greedy search for high-dimensional bandit problems. But it's non-trivial to extend the framework to the matrix bandit problem. For example, they assume that the minimum eigenvalue of the covariance matrix could be strictly lower bounded, but this lower bound would mostly depend on the size of feature matrices, and hence would affect the regret bound consequently.

4.3. Preliminaries

In this section we review our problem setting and introduce the assumptions for our theoretical analysis. Let T be the total number of rounds and $\mathcal{X}_t$ be the action set ($\mathcal{X}_t$ could be fixed or not). Throughout this paper, we denote the action set $\mathcal{X}_t = \mathcal{X}$ as fixed for notation simplicity, while our frameworks also work with the same regret bound when $\mathcal{X}_t$ varies over time (see Appendix C.8.3

for more details.) Algorithms along with theory could be identically obtained when the action set varies (Appendix C.8.3). At each round $t \in [T]$, The agent selects an action $X_t \in \mathcal{X}_t$ and gets the payoff y_t which is conditionally independent of the past payoffs and choices. For the generalized low-rank matrix bandits, we assume the payoff y_t follows a canonical exponential family such that:

$$p_{\Theta^*}(y_t|X_t) = \exp\left(\frac{y_t\beta - b(\beta)}{\phi} + c(y_t,\phi)\right), \text{ where } \beta = \mathrm{vec}(X_t)^\top \mathrm{vec}(\Theta^*) \coloneqq \langle X_t, \Theta^*\rangle, \tag{4.1}$$
$$\mathbb{E}_{\Theta^*}(y_t|X_t) = b'(\langle X_t, \Theta^*\rangle) \coloneqq \mu(\langle X_t, \Theta^*\rangle),$$

where $\Theta^* \subseteq \Theta$ is a fixed but unknown matrix with rank $r \ll \{d_1, d_2\}$ and Θ is some admissible compact subset of $\mathbb{R}^{d_1\times d_2}$ (w.l.o.g. $d_1 = \Theta(d_2)$). We also call $\mu(\langle X_t, \Theta^*\rangle)$ the reward of action X_t. In addition, one can represent model (4.1) in the following Eqn. (4.2). Note that if we relax the definition of $\mu(\cdot)$ to any real univariate function with some centered exogenous random noise η_t, the model shown in Eqn. (4.2) generalizes our problem setting to a single index model (SIM) matrix bandit, and the generalized low-rank matrix bandit problem is a special case of this model.

$$y_t = \mu(\langle X_t, \Theta^*\rangle) + \eta_t. \tag{4.2}$$

Here, η_t follows the sub-Gaussian property with some constant parameter σ_0 conditional on the filtration $\mathcal{F}_t = \{X_t, X_{t-1}, \eta_{t-1}, \ldots, X_1, \eta_1\}$. We also denote $d = \max\{d_1, d_2\}$. And it is natural to evaluate the agent's strategy based on the regret (Audibert et al., 2009), defined as the difference between the total reward of optimal policy and the agent's total reward in practice:

$$Regret_t = \sum_{i=1}^{t} \max_{X\in\mathcal{X}} \mu(\langle X, \Theta^*\rangle) - \mu(\langle X_i, \Theta^*\rangle).$$

We also present the following two definitions to facilitate further analysis via Stein's method:

Definition 4.3.1. *Let $p : \mathbb{R} \to \mathbb{R}$ be a univariate probability density function defined on $\mathbb{R}$. The score function $S^p : \mathbb{R} \to \mathbb{R}$ regarding density $p(\cdot)$ is defined as:*

$$S^p(x) = -\nabla_x log(p(x)) = -\nabla_x p(x)/p(x), \qquad x \in \mathbb{R}.$$

In particular, for a random matrix with its entrywise probability density $\mathbf{p} = (p_{ij}) : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$, we define its score function $S^{\mathbf{p}} = (S^{\mathbf{p}}_{ij}) : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ as $S^{\mathbf{p}}_{ij}(x) = S^{p_{ij}}(x)$ by applying the univariate score function to each entry of $\mathbf{p}$ independently.

DEFINITION 4.3.2. (Fact 2.6, (Minsker, 2018)) *Given a rectangular matrix $A \in \mathbb{R}^{d_1 \times d_2}$, the (Hermitian) dilation $\mathcal{H} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$ is defined as:*

$$\mathcal{H}(A) = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}.$$

We would omit the subscript x of ∇ and the superscript p of S when the underlying distribution is clear. With these definitions, we make the following mild assumptions:

ASSUMPTION 4.3.3. (Finite second-moment score) There exists a sampling distribution $\mathcal{D}$ over $\mathcal{X}$ such that for the random matrix X drawn from $\mathcal{D}$ with its associated density $\mathbf{p} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$, we have $\mathbb{E}[(S^{\mathbf{p}}(X))_{ij}^2] \leq M, \forall i, j$. And the columns or rows of random matrix X are pairwisely independent.

ASSUMPTION 4.3.4. The norm of true parameter Θ^* and feature matrices in $\mathcal{X}$ is bounded: there exists $S \in \mathbb{R}^+$ such that for all arms $X \in \mathcal{X}$, $\|X\|_F, \|\Theta^*\|_F \leq S_0$.

ASSUMPTION 4.3.5. The inverse link function $\mu(\cdot)$ in GLM is continuously differentiable and there exist two constants c_μ, k_μ such that $0 < c_\mu \leq \mu'(x) \leq k_\mu$ for all $|x| \leq S_0$.

Assumption 4.3.3 is commonly used in Stein's method (Chen et al., 2010), and easily satisfied by a wide range of distributions that are non-zero-mean or even non sub-Gaussian thereby allowing us to work with cases not previously possible. For example, to find $\mathcal{D}$ we only need the convex hull of $\mathcal{X}$ contains a ball with radius R, and then we can use p_{ij} as centered normal p.d.f. with variance $R^2/(d_i d_j)$. This choice works well in our experiments and please refer to Appendix C.9 for more details. Furthermore, Assumption 4.3.4 and 4.3.5 are also standard in contextual generalized bandit literature, and they explicitly imply that we have an upper bound as $|\mu(\langle X, \Theta \rangle)| \leq |\mu(0)| + k_\mu S_0 := S_f$.

4.4. Methods

In this section, we present our novel two-stage frameworks, named Generalized Explore Subspace Then Transform (G-ESTT) and Generalized Explore Subspace Then Subtract (G-ESTS) respectively. These two algorithms are inspired by the two-stage algorithm ESTR proposed in Jun et al. (2019). ESTR estimates the row and column subspaces for the true parameter Θ^* in stage 1. In

Algorithm 5 Generalized Explore Subspace Then Transform (G-ESTT)

Input: $\mathcal{X}, T, T_1, \mathcal{D}$, the probability rate δ, parameters for Stage 2: $\lambda, \lambda_\perp$.

Stage 1: Subspace Estimation

1: **for** $t = 1$ **to** T_1 **do**

2: Pull arm $X_t \in \mathcal{X}$ according to $\mathcal{D}$, observe payoff y_t.

3: Obtain $\widehat{\Theta}$ based on Eqn. (4.6).

4: Obtain the full SVD of $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\, \widehat{D}\, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U} \in \mathbb{R}^{d_1 \times r}, \widehat{V} \in \mathbb{R}^{d_2 \times r}$.

Stage 2: Sparse Generalized Linear Bandits

5: Rotate the arm feature set: $\mathcal{X}' := [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X} [\widehat{V}, \widehat{V}_\perp]$ and the admissible parameter space: $\Theta' := [\widehat{U}, \widehat{U}_\perp]^\top \Theta [\widehat{V}, \widehat{V}_\perp]$.

6: Define the vectorized arm set so that the last $(d_1 - r) \cdot (d_2 - r)$ components are negligible:

$$\mathcal{X}_0 := \{\mathrm{vec}(\mathcal{X}'_{1:r,1:r}), \mathrm{vec}(\mathcal{X}'_{r+1:d_1,1:r}), \mathrm{vec}(\mathcal{X}'_{1:r,r+1:d_2}), \mathrm{vec}(\mathcal{X}'_{r+1:d_1,r+1:d_2})\}, \tag{4.3}$$

and similarly define the parameter set:

$$\Theta_0 := \{\mathrm{vec}(\Theta'_{1:r,1:r}), \mathrm{vec}(\Theta'_{r+1:d_1,1:r}), \mathrm{vec}(\Theta'_{1:r,r+1:d_2}), \mathrm{vec}(\Theta'_{r+1:d_1,r+1:d_2})\}. \tag{4.4}$$

7: For $T_2 = T - T_1$ rounds, invoke (P)LowGLM-UCB with $\mathcal{X}_0, \Theta_0, k = (d_1 + d_2)r - r^2, (\lambda_0, \lambda_\perp)$.

stage 2, it exploits the estimated subspaces and transforms the original matrix bandits into linear bandits with sparsity, and then invoke a penalized approach called LowOFUL.

4.4.1. Subspace Exploration. For any real-value function $f(\cdot)$ defined on $\mathbb{R}$, and symmetric matrix $A \in \mathbb{R}^{d \times d}$ with its SVD decomposition as $A = UDU^\top$, we define

$$f(A) := U\, \mathrm{diag}(f(D_{11}), \ldots, f(D_{dd}))\, U^\top.$$

To explore the valid subspace of the parameter matrix Θ^*, we firstly define a function $\psi : \mathbb{R} \to \mathbb{R}$ (Minsker, 2018) in Eqn. (4.5) and subsequently we define $\tilde{\psi}_\nu : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ as $\tilde{\psi}_\nu(A) = \psi(\nu \mathcal{H}(A))_{1:d_1,(d_1+1):(d_1+d_2)}/\nu$ for some parameter $\nu \in \mathbb{R}^+$.

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0; \\ -\log(1 - x + x^2/2), & x < 0. \end{cases} \tag{4.5}$$

We consider the following well-defined regularized minimization problem with nuclear norm penalty:

$$\widehat{\Theta} = \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} L_{T_1}(\Theta) + \lambda_{T_1} \|\Theta\|_{\mathrm{nuc}}, \quad L_{T_1}(\Theta) = \langle \Theta, \Theta \rangle - \frac{2}{T_1} \sum_{i=1}^{T_1} \langle \tilde{\psi}_\nu(y_i \cdot S(X_i)), \Theta \rangle. \tag{4.6}$$

An interesting fact is that our estimator is invariant under different choices of function $\mu(\cdot)$, and we could present the following oracle inequality regarding the estimation error $\left\|\widehat{\Theta} - \mu^*\Theta^*\right\|_F$ for some nonzero constant μ^* by adapting generalized Stein's Method (Chen et al., 2010).

THEOREM 4.4.1. (Bounds for GLM) *For any low-rank generalized linear model with samples $X_1 \ldots, X_{T_1}$ drawn from $\mathcal{X}$ according to $\mathcal{D}$ in Assumption 4.3.3, and assume Assumption 4.3.4 and 4.3.5 hold, then for the optimal solution to the nuclear norm regularization problem (4.6) with $\nu = \sqrt{2\log(2(d_1+d_2)/\delta)/((4\sigma_0^2 + S_f^2)MT_1(d_1+d_2))}$ and*

$$\lambda_{T_1} = 4\sqrt{\frac{2(4\sigma_0^2 + S_f^2)M(d_1+d_2)\log(2(d_1+d_2)/\delta)}{T_1}},$$

with probability at least $1-\delta$ it holds that:

$$\left\|\widehat{\Theta} - \mu^*\Theta^*\right\|_F^2 \leq \frac{C_1 M(d_1+d_2)r\log(\frac{2(d_1+d_2)}{\delta})}{T_1}, \tag{4.7}$$

for $C_1 = 36(4\sigma_0^2 + S_f^2)$ and some nonzero constant μ^*.

The proof of Theorem 4.4.1 is based on a novel adaptation of Stein-typed Lemmas and is deferred to Appendix C.2. We believe this oracle bound is non-trivial since the rate of convergence is no worse than that deduced from the restricted strong convexity (details in Appendix C.10) given $M = O((d_1+d_2)^2)$, even without the regular sub-Gaussian assumption. For completeness, we also present detailed proof of the matrix recovery rate with the restricted strong convexity in Appendix C.10 which may be of separate interest. And its proof is highly different from the one used in the simple linear case (Lu et al., 2021). We also present an intuitive explanation on why our Stein-type method works well under our problem setting in Appendix C.10.2 even without the sub-Gaussian assumption. In addition, this bound also holds under a more general SIM in Eqn. (4.2) other than just GLM. Furthermore, although there exists a non-zero constant μ^* in the error term, it will not affect the singular vectors and subspace estimation of Θ^* at all.

After acquiring the estimated $\widehat{\Theta}$ in stage 1, we can obtain the corresponding SVD as $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\,\widehat{D}\,[\widehat{V}, \widehat{V}_\perp]^\top$, where $\widehat{U} \in \mathbb{R}^{d_1\times r}, \widehat{U}_\perp \in \mathbb{R}^{d_1\times(d_1-r)}, \widehat{V} \in \mathbb{R}^{d_2\times r}$ and $\widehat{V}_\perp \in \mathbb{R}^{d_2\times(d_2-r)}$. And we assume the SVD of the matrix Θ^* can be represented as $\Theta^* = UDV^\top$ where $U \in \mathbb{R}^{d_1\times r}$ and $V \in \mathbb{R}^{d_2\times r}$. To transform the original generalized matrix bandits into generalized linear bandit problems, we follow the works in Jun et al. (2019) and penalize those covariates that are

Algorithm 6 LowGLM-UCB

Input: $T_2, k, \mathcal{X}_0$, the probability rate δ, penalization parameters $(\lambda_0, \lambda_\perp)$.
Initialize $M_1(c_\mu) = \sum_{i=1}^{T_1} x_{s_1,i}\, x_{s_1,i}^\top + \Lambda/c_\mu$.
for $t \geq 1$ **do**
  Estimate $\hat{\theta}_t$ according to (4.10).
  Choose the arm $x_t = \arg\max_{x \in \mathcal{X}_0} \{\mu(x^\top \hat{\theta}_t) + \rho_t(\delta) \|x\|_{M_t^{-1}(c_\mu)}\}$, receive y_t,
  Update $M_{t+1}(c_\mu) \longleftarrow M_t(c_\mu) + x_t x_t^\top$.

complementary to $\widehat{U}$ and $\widehat{V}$. Specifically, we could orthogonally rotate the parameter space Θ and the action set $\mathcal{X}$ as:

$$\Theta' = [\widehat{U}, \widehat{U}_\perp]^\top \Theta [\widehat{V}, \widehat{V}_\perp], \quad \mathcal{X}' = [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X} [\widehat{V}, \widehat{V}_\perp],$$

Define the total dimension $p := d_1 d_2$, the effective dimension $k := d_1 d_2 - (d_1 - r)(d_2 - r)$ and the r-th largest singular value for Θ^* as D_{rr}, and vectorize the new arm space $\mathcal{X}'$ and admissible parameter space as shown in Eqn. (4.3) and (4.4). Then for the true parameter θ^* after transformation, we know that $\theta^*_{k+1:p} = \mathrm{vec}(\Theta^{*,\prime}_{r+1:d_1, r+1:d_2})$ is almost null based on results in Stewart (1990) and Theorem 4.4.1:

(4.8)
$$\left\|\theta^*_{k+1:p}\right\|_2 = \left\|\widehat{U}_\perp^\top U D V^\top \widehat{V}_\perp\right\|_F \leq \left\|\widehat{U}_\perp^\top U\right\|_F \left\|\widehat{V}_\perp^\top V\right\|_F \cdot \|D\|_{\mathrm{op}} \lesssim \frac{(d_1 + d_2) M r}{T_1 D_{rr}^2} \log\left(\frac{d_1 + d_2}{\delta}\right) := S_\perp.$$

Therefore, this problem degenerates to an equivalent $d_1 d_2-$dimensional generalized linear bandit with a sparse structure (i.e. last $p - k$ entries of θ^* are almost null according to Eqn. (4.8)). To reload the notation we define $\mathcal{X}_0, \Theta_0$ as the new feature set and parameter space as shown in Algorithm 5.

Remark. Note the magnitude of D_{rr} would be free of d since Θ^* contains only r nonzero singular values, and hence we assume that $D_{rr} = \Theta(1/\sqrt{r})$ under Assumption 4.3.4. This issue has been ignored in all previous analysis of explore-then-commit-type algorithms (e.g. ESTR (Jun et al., 2019), LowESTR (Lu et al., 2021)), where the final regret bound of them should be of order $\tilde{O}(d^{3/2} r\sqrt{T})$ instead of the originally-used $\tilde{O}(d^{3/2}\sqrt{rT})$ because of the existence of D_{rr}.

4.4.2. G-ESTT. After reducing the original generalized matrix bandit problem into an identical p-dimensional generalized linear bandit problem in stage 2, we can reformulate the problem in the following way: at each round t, the agent chooses a vector x_t of dimension p from the

transformed action set $\mathcal{X}_0$, and observes a noisy reward $y_t = \mu({x_t}^\top \theta^*) + \eta_t$. To make use of our additional knowledge shown in Eqn. (4.8), we propose LowGLM-UCB as an extension of the standard generalized linear bandit algorithm GLM-UCB (Filippi et al., 2010) combined with self-normalized martingale technique (Abbasi-Yadkori et al., 2011). Specifically, we consider the following maximum quasi-likelihood estimation problem shown in Eqn. (4.9) for each round with a weighted regularizer, where the regularizer is $\|\theta\|_\Lambda^2/2 = \theta^\top \Lambda \theta/2$ for some positive definite diagonal matrix $\Lambda = \mathrm{diag}(\lambda_0, \ldots, \lambda_0, \lambda_\perp, \ldots, \lambda_\perp)$ with λ_0 only applied to the first k diagonal entries. By enlarging $\lambda_\perp$, we ensure more penalization forced on the last $p-k$ element of θ^* as desired.

$$\hat{\theta}_t = \arg\max_\theta \widetilde{L}_t^\Lambda(\theta),$$

$$\widetilde{L}_t^\Lambda(\theta) = \sum_{i=1}^{T_1} \Big[y_{s_1,i} {x_{s_1,i}}^\top \theta - b({x_{s_1,i}}^\top \theta) \Big] + \sum_{i=1}^{t-1} \Big[y_i {x_i}^\top \theta - b({x_i}^\top \theta) \Big] - \frac{1}{2} \|\theta\|_\Lambda^2 . \tag{4.9}$$

Here $x_{s_1,i}$ in Eqn. (4.9) is the special vectorization shown in Eqn. (4.3) of $[\widehat{U}, \widehat{U}_\perp]^\top X_i [\widehat{V}, \widehat{V}_\perp]$ where X_i is the arm we randomly pull at i-th step in stage 1, and $y_{s_1,i}$ is the corresponding payoff we observe. x_i in the second summation of Eqn. (4.9) refers to the arm we pull at i-th step in stage 2. Since $\widetilde{L}_t^\Lambda(\theta)$ is a strictly concave function of θ, we have its gradient equal to 0 at the maximum $\hat{\theta}_t$, i.e. $\nabla_\theta \widetilde{L}_t^\Lambda(\theta)|_{\hat{\theta}_t} = 0$. In what follows, for $t \geq 2, \theta \in \mathbb{R}^p$ we define the function $g_t(\theta)$ and have that

$$\nabla_\theta \widetilde{L}_t^\Lambda(\theta) = \sum_{i=1}^{T_1} y_{s_1,i}\, x_{s_1,i} + \sum_{i=1}^{t-1} y_i\, x_i - \Bigg(\underbrace{\sum_{i=1}^{T_1} \mu({x_{s_1,i}}^\top \theta) x_{s_1,i} + \sum_{i=1}^{t-1} \mu({x_i}^\top \theta) x_i + \Lambda\theta}_{\coloneqq\, g_t(\theta)} \Bigg),$$

$$\nabla_\theta \widetilde{L}_t^\Lambda(\theta)|_{\hat{\theta}_t} = 0 \quad \Longrightarrow \quad g_t(\hat{\theta}_t) = \sum_{i=1}^{T_1} y_{s_1,i}\, x_{s_1,i} + \sum_{i=1}^{t-1} y_i\, x_i. \tag{4.10}$$

We also define a matrix function $M_t(s) = \sum_{i=1}^{T_1} x_{s_1,i} x_{s_1,i}^\top + \sum_{k=1}^{t-1} x_k x_k^\top + \Lambda/s$ for $s \in \mathbb{R}^+$ and denote $V_t \coloneqq M_t(1)$. Furthermore, a remarkable benefit of reusing the actions $\{X_i\}_{i=1}^{T_1}$ we randomly pull in stage 1 is that they contain more randomness and are preferable to the ones we select based on some strategy in stage 2 regarding the parameter estimation because most vector recovery theory requires sufficient randomness during sampling. More inspiring, the projection step in the tradition GLM-UCB (Filippi et al., 2010), which might be nonconvex and hence hard to solve, is no longer required due the consistency of $\hat{\theta}_t$ after reutilizing $\{X_i\}_{i=1}^{T_1}$. Specifically, if we assume the true parameter

θ^* lies in the interior of Θ_0 and the sampling distribution $\mathcal{D}$ satisfies sub-Gaussian property with parameter σ, and Assumption 4.3.4, 4.3.5 held, then we can show that $\left\|\hat{\theta}_t - \theta^*\right\|_2 \leq 1$ holds with probability at least $1-\delta$ as long as $T_1 \geq ((\hat{C}_1\sqrt{p} + \hat{C}_2\sqrt{\log(1/\delta)})/\sigma^2)^2 + 2B/\sigma^2$ holds for some absolute constants $\hat{C}_1, \hat{C}_2$ with the definition $B \coloneqq 16\sigma_0^2(p+\log(1/\delta))/c_\mu^2$. An intuitive explanation along with a rigorous proof are deferred to Appendix C.4 due to the space limit. The proposed LowGLM-UCB is shown in Algorithm 6, and its regret analysis is presented in Theorem C.3.1 in Appendix.

Notice that we can simply replace $M_t(c_\mu)$ by V_t in Algorithm 6, and the regret bound would increase at most up to a constant factor (Appendix C.7). A potential drawback of Algorithm 6 is that in each iteration we have to calculate $\hat{\theta}_t$, which might be computationally expensive. We could resolve this problem by only recomputing $\hat{\theta}_t$ whenever $|M_t(c_\mu)|$ increases significantly, i.e. by a constant factor $C > 1$ in scale. And consequently we only need to solve the Eqn. (4.10) for $O(\log(T_2))$ times up to the horizon T_2, which remarkably saves the computation. Meanwhile, the bound of the regret would only increase by a constant multiplier $\sqrt{C}$. We call this modified algorithm as PLowGLM-UCB with the initial letter "P" standing for "Parsimonious". Its pseudo-code and regret analysis are given in Appendix C.8.1. Equipped with LowGLM-UCB in stage 2, we deduce the overall regret of G-ESTT in the following text.

To quantify the performance of our algorithm, we first define $\alpha_t^x(\cdot)$ and $\beta_t^x(\cdot)$ as

$$\alpha_t(\delta) \coloneqq \frac{k_\mu}{c_\mu}\left(\sigma_0\sqrt{k\log(1+\frac{c_\mu S_0^2 t}{k\lambda_0}) + \frac{c_\mu S_0^2 t}{\lambda_\perp} - \log(\delta^2)} + \sqrt{c_\mu}(\sqrt{\lambda_0}S_0 + \sqrt{\lambda_\perp}S_\perp)\right), \tag{4.11}$$

$$\beta_t^x(\delta) \coloneqq \alpha_t(\delta)\left\|x\right\|_{M_t^{-1}(c_\mu)}. \tag{4.12}$$

And the following Theorem 4.4.2 exhibits the overall regret bound for G-ESTT.

THEOREM 4.4.2. (Regret of G-ESTT) *Suppose we set* $T_1 \asymp \sqrt{M(d_1+d_2)rT\log((d_1+d_2)/\delta)}/D_{rr}$, *and we invoke LowGLM-UCB (or PLowGLM-UCB) in stage 2 with* $\rho_t(\delta) = \alpha_{t+T_1}(\delta/2), p = d_1d_2, k = (d_1+d_2)r - r^2$, $\lambda_\perp = c_\mu S_0^2 T/(k\log(1+c_\mu S_0^2 T/(k\lambda_0)))$, *and the rotated arm sets* $\mathcal{X}_0$ *and available parameter space* Θ_0. *With* $M = O((d_1+d_2)^2)$, *the overall regret of G-ESTT is, with probability at least* $1-\delta$,

$$Regret_T = \tilde{O}\left((\frac{\sqrt{r(d_1+d_2)M}}{D_{rr}} + k)\sqrt{T}\right)$$

Algorithm 7 Generalized Explore Subspace Then Subtract (G-ESTS)

Input: $\mathcal{X}, T, T_1, \mathcal{D}$, the probability rate δ, parameters for Stage 2: $\lambda, \lambda_\perp$.

Stage 1: Subspace Estimation

1: Randomly choose $X_t \in \mathcal{X}$ according to $\mathcal{D}$ and record X_t, Y_t for $t = 1, \dots T_1$.

2: Obtain $\widehat{\Theta}$ from Eqn. (4.6), and calculate its full SVD as $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\, \widehat{D}\, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U} \in \mathbb{R}^{d_1 \times r}, \widehat{V} \in \mathbb{R}^{d_2 \times r}$.

Stage 2: Low Dimensional Bandits

3: Rotate the arm feature set: $\mathcal{X}' := [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X} [\widehat{V}, \widehat{V}_\perp]$ and the admissible parameter space: $\Theta' := [\widehat{U}, \widehat{U}_\perp]^\top \Theta [\widehat{V}, \widehat{V}_\perp]$.

4: Define the vectorized arm set so that the last $(d_1 - r) \cdot (d_2 - r)$ components are negligible, and then **drop** them:

$$\mathcal{X}_{0,sub} := \{\mathrm{vec}(\mathcal{X}'_{1:r,1:r}), \mathrm{vec}(\mathcal{X}'_{r+1:d_1,1:r}), \mathrm{vec}(\mathcal{X}'_{1:r,r+1:d_2})\}, \tag{4.13}$$

and also refine the parameter set accordingly:

$$\Theta_{0,sub} := \{\mathrm{vec}(\Theta'_{1:r,1:r}), \mathrm{vec}(\Theta'_{r+1:d_1,1:r}), \mathrm{vec}(\Theta'_{1:r,r+1:d_2})\}. \tag{4.14}$$

5: For $T_2 = T - T_1$ rounds, invoke any misspecified generalized linear bandit algorithm with $\mathcal{X}_{0,sub}, \Theta_{0,sub}, k = (d_1 + d_2)r - r^2$.

Specifically, with $M = O((d_1 + d_2)^2)$, the regret bound becomes $\tilde{O}\left((\sqrt{r(d_1 + d_2)^3}/D_{rr} + k)\sqrt{T}\right)$.

4.4.3. G-ESTS. Although G-ESTT is more efficient than all existing algorithms on our problem setting, it still needs to calculate the MLE in high dimensional space which might be increasingly formidable with large sizes of feature matrices. Note this computational issue remains ubiquitous among most bandit algorithms on high dimensional problems with sparsity, not to mention these algorithms rely on multiple unspecified hyperparameters. Therefore, to handle this practical issue, we propose another fast and efficient framework called G-ESTS in this section.

Inspired by the success of dimension reduction in machine learning (Van Der Maaten et al., 2009), we propose G-ESTS as shown in Algorithm 7. And we summarize the core idea of G-ESTS as: After rearranging the vectorization of the action set $\mathcal{X}'$ and the unknown Θ'^* as we have shown in Eqn. (4.3) and (4.4) for G-ESTT, we can simply exclude, rather than penalize, the subspaces that are complementary to the rows and columns of $\widehat{\Theta}$. In other words, we could remove the last $p - k$ entries directly, i.e. Eqn. (4.13) and (4.14). Intriguingly, not only can we get a low-dimensional (k) generalized linear bandit problem in stage 2, where redundant dimensions are excluded and hence any state-of-the-art algorithms could be readily invoked. Specifically, by utilizing any misspecified generalized linear bandit algorithm, we could validate the following Theorem 4.4.3.

THEOREM 4.4.3. (Regret of G-ESTS) *Suppose we set* $T_1 \asymp \sqrt{M(d_1+d_2)^{3/2}r^{3/2}T\log((d_1+d_2)/\delta)}/D_{rr}$, *and we invoke any efficient misspecified generalized linear bandit algorithm with regret bound* $\tilde{O}(\epsilon\sqrt{k}T)$ [2] *in stage 2 with* $p = d_1d_2, k = (d_1+d_2)r - r^2$, *and the reduced arm sets* $\mathcal{X}_{0,sub}$ *and available parameter space* $\Theta_{0,sub}$. *The overall regret of G-ESTS is, with probability at least* $1-\delta$,

$$Regret_T = \widetilde{O}\left((\frac{\sqrt{r^{3/2}(d_1+d_2)^{3/2}M}}{D_{rr}} + k)\sqrt{T}\right).$$

Specifically, with $M = O((d_1+d_2)^2)$, *the regret bound becomes* $\widetilde{O}\left(\sqrt{r^{3/2}(d_1+d_2)^{7/2}T}/D_{rr}\right)$.

Although our G-ESTS could achieve decent theoretical regret bound only equipped with misspecified generalized linear bandit algorithms, we showcase in practice it can work well with any state-of-the-art generalized linear bandit algorithm: In the following experiments, We will implement the SGD-TS algorithm (Ding et al., 2021) in stage 2 of G-ESTS since SGD-TS could efficiently proceed with only $O(dT)$ complexity for $d-$dimensional features over T rounds. Therefore, the total computational complexity of stage 2 is at most $O(T_2(d_1+d_2)r)$, which is significantly less than that of other methods for low-rank matrix bandits (e.g. LowESTR (Lu et al., 2021)). And the total time complexity of G-ESTS would only scale $O(T_1d_1d_2/\epsilon^2 + T_2(d_1+d_2)r)$ where ϵ is the accuracy for subgradient methods in stage 1. This fact also firmly validates the practical superiority of our G-ESTS approach. We naturally believe that this G-ESTS framework can be easily implemented in the linear setting as a special case of GLM, where in stage 2 one can utilize any linear bandit algorithm accordingly. In addition, we can easily modify our approaches for the contextual setting by merely transforming the action sets at each iteration with the same regret bound. More details with pseudo-codes for the contextual case are in Appendix C.8.3.

4.5. Experimental Results

In this section, we show by simulation experiments that our proposed G-ESTT (with LowGLM-UCB), G-ESTS (with SGD-TS) outperform existing algorithms for the generalized low-rank matrix bandit problems. Since we are the first to propose a practical algorithm for this problem, currently there is no existing literature for comparison. In order to validate the advantage of utilizing low-rank structure and generalized reward functions, we compare with the original SGD-TS after naïvely

[2] Modern misspecified generalized linear bandit algorithms can achieve $\tilde{O}(\epsilon\sqrt{k}T)$ bound of regret where ϵ is the misspecified rate.

TABLE 4.1. Time in minutes required to make decisions all over round T in simulations (480 arms).

d	10		12	
r	1	2	1	2
G-ESTS	39.46	45.41	41.28	48.52
G-ESTT	516.14	531.95	520.25	539.83
SGD-TS	99.57	101.34	101.82	104.42
LowESTR	401.88	419.15	410.31	425.92

flattening the d_1 by d_2 matrices without using the low-rank structure, and LowESTR (Lu et al., 2021), which works well for linear low-rank matrix bandits.

We simulate a dataset with $d_1 = d_2 = 10\,(12)$ and $r = 1\,(2)$: when $r = 1$, we set the diagonal matrix Θ^* as $\mathrm{diag}(\Theta^*) = (0.8, 0, \cdots, 0)$. When $r = 2$, we set $\Theta^* = v_1 v_1^\top + v_2 v_2^\top$ for two random orthogonal vectors v_1, v_2 with $\|v_1\|_2 = \|v_2\|_2 = 3$. For arms we draw $480\,(1000)$ random matrices from $\{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \leq 1\}$, and we build a logistic model where the payoff y_t is drawn from a Bernoulli distribution with mean $\mu(X_t^\top \theta^*)$. More details on the hyper-parameter tuning are in Appendix C.9. Each experiment is repeated 100 times for credibility and the average regret, along with standard deviation, is displayed in Figure 4.1. Note that our experiments are more comprehensive than those in Lu et al. (2021). And due to the expensive time complexity of UCB-based baselines (Table 4.1), it is formidable for us to increase d here.

From the plots, we observe that our algorithms G-ESTT and G-ESTS always achieve less regret compared with LowESTR and SGD-TS in all four scenarios consistently. Intriguingly, in the warm-up period SGD-TS incurs less regret compared with our methods due to the sacrifice of random sampling in stage 1, but our proposed framework quickly overtakes SGD-TS after utilizing the low-rank structure as desired. This phenomenon exactly coincides with our theory. Notice that G-ESTT is slightly better than G-ESTS in the case for $r = 2$ especially in the very beginning of stage 2, and we believe it is because that our G-ESTT could reutilize the actions in stage 1 and hence could yield more robust performance when switching to stage 2. However, G-ESTS would gradually catch up with G-ESTT in the long run as expected. Besides, it costs G-ESTS extremely less running time than other existing methods to update the decisions due to its dimensional reduction as shown in Table 4.1. We also observe that the cumulative regret of G-ESTS tends to become better eventually if we increase T_1 decently. (Further investigation and plots for 1000 arms

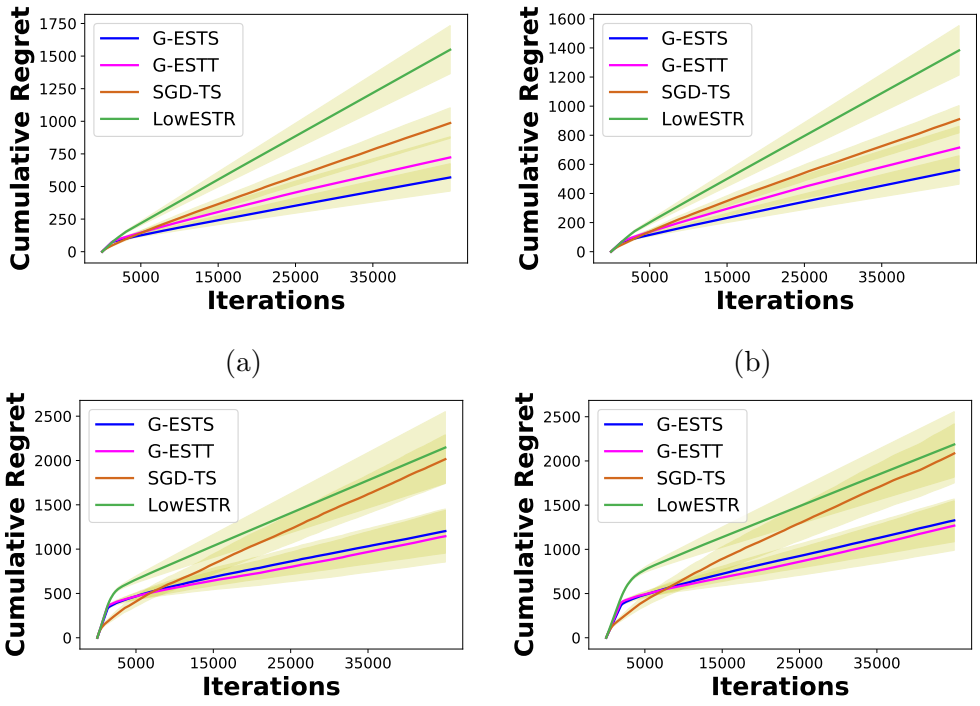


FIGURE 4.1. Plots of regret curves of algorithm G-ESTS, G-ESTT, SGD-TS and LowESTR under four settings (480 arms). (a): diagonal Θ^* $d_1 = d_2 = 10, r = 1$; (b): diagonal Θ^* $d_1 = d_2 = 12, r = 1$; (c): non-diagonal Θ^* $d_1 = d_2 = 10, r = 2$; (d): non-diagonal Θ^* $d_1 = d_2 = 12, r = 2$.

are in Appendix C.9.) Moreover, to pre-check the efficiency of our Stein's lemma-based method for subspace estimation shown in Eqn. (4.6), we also tried some other low-rank subspace detection algorithms for comparison. The details are also deferred to Appendix C.9.4 due to the space limit.

CHAPTER 5

Low-rank Matrix Bandits under Heavy-tailed Rewards

5.1. Introduction

The Multi-armed Bandit (MAB) has proven to be a powerful framework to model various decision-making problems with great applications to medical trials (Villar et al., 2015), personalized recommendation (Li et al., 2010), and hyperparameter learning (Ding et al., 2022b; Kang et al., 2024a), etc. To leverage the side information (contexts) of arms in real-world scenarios, the most important variant of MAB, named stochastic linear bandit (SLB), has been extensively investigated. However, the rise of high-dimensional sparse data in modern applications has revealed the inefficiencies of the traditional SLB, particularly in its failure to account for sparsity. To address this limitation, the stochastic high-dimensional bandit with low-dimensional structures has emerged as the pioneering model, such as the LASSO bandit (Bastani & Bayati, 2020) and the low-rank matrix bandit (Jun et al., 2019). In this work, we investigate the stochastic low-rank matrix bandit, where at each round t the agent first observes the arm set $\mathcal{X}_t \subseteq \mathbb{R}^{d_1 \times d_2}$ composing of context matrices ($\mathcal{X}_t$ can be infinite and changing over time). Then the agent pulls an arm $X_t \in \mathcal{X}_t$ and only obtains its associated noisy reward $y_t = \langle X_t, \Theta^* \rangle + \eta_t$ with some inherent low-rank parameter Θ^* and zero-mean white noise η_t. This bandit problem is broadly applicable in recommendation systems with pair contexts, like dating service and combined flight-hotel promotion (Kang et al., 2022).

In all existing literature on low-rank matrix bandit, a default assumption is that the noise η_t is sub-Gaussian conditioned on historical observations (Jun et al., 2019). However, in various real-world scenarios such as financial markets (Bradley & Taqqu, 2003; Cont & Bouchaud, 2000), there's a notable trend where extreme noise, a.k.a. heavy-tailed noise, in observations occur more frequently than what would be expected under a sub-Gaussian distribution, in which case previous studies would become futile. These heavy-tailed observations do not exhibit exponential decay and may crucially affect the estimation. To address this challenge, a line of algorithms has been proposed to handle heavy-tailed noise under MAB (Bubeck et al., 2013) and SLB (Medina & Yang,

2016). However, to the best of our knowledge, effectively managing heavy-tailed noise under the more complex and efficient low-rank matrix bandit framework remains unexplored. In this study, we examine this crucial problem: low-rank matrix bandit with heavy-tailed rewards (LowHTR). Specifically, to keep consistent with the heavy-tailed studies under MAB and SLB, we assume that the noise has finite $(1+\delta)$ moment for some $\delta \in (0,1]$. We first propose an efficient algorithm named LOTUS when T is unrevealed to the agent. Then we demonstrate it attains a regret lower bound of LowHTR for the order of T ignoring logarithmic factors. Our LOTUS can be further improved to be agnostic to rank r with slightly worse regret bound.

The detailed contributions of our work can be summarized as follows: (1) inspired by the success of Huber loss (Kang & Kim, 2023; Sun et al., 2020) and nuclear norm penalization (Negahban & Wainwright, 2011), we first introduce a convex-relaxation-based estimator to approximate the low-rank parameter matrix with heavy-tailed noise. As far as we're aware, our work is the first one to solve the trace regression problem under arbitrary heavy-tailed noise with bounded $(1+\delta)$ moment ($\delta \in (0,1)$), which is highly non-trivial and stands as a noteworthy advancement on its own merits. (2) Equipped with the aforementioned estimator, we develop an algorithm named LOTUS for LowHTR. LOTUS exploits the estimated subspace by proposing a sub-method called LowTO that extends from the TOFU algorithm (Shao et al., 2018) designed for SLB with heavy-tailed noise. Our LowTO truncates the rewards to mitigate the heavy-tailed effect and penalizes the redundant features within the sparsity structure. When the total horizon T is unrevealed, our algorithm could adaptively switch between exploration and exploitation to achieve the $\tilde{O}(d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}} / \tilde{D}_{rr})$[1] regret bound. (3) We further provide a lower bound for LowHTR of order $\Omega(d^{\frac{\delta}{1+\delta}} r^{\frac{\delta}{1+\delta}} T^{\frac{1}{1+\delta}})$, which indicates that our LOTUS is nearly optimal in the scale of T. (4) While all existing works on low-rank matrix bandits require a priori knowledge of the rank r, we further improve our LOTUS to operate without knowing r even under the more difficult heavy-tailed setting with $\tilde{O}(dr^{\frac{3}{2}} T^{\frac{1+\delta}{1+2\delta}} + d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}})$ regret bound, which is better than the trivial one in high-dimensional case, i.e. when $d \gtrsim T^{\frac{\delta^2}{(1+2\delta)(1+\delta)}}$. Intuitively, it obtains a useful rank $\hat{r}$ by truncating the estimated singular values at each batch. (4) The practical superiority of our LOTUS is then firmly validated in our simulations.

[1] $\tilde{O}$ ignores polylogarithmic factors. $d \coloneqq d_1 \vee d_2$ and $\tilde{D}_{rr} \coloneqq (D_{rr} - 1)\mathbb{1}_{\delta=1} + 1$ where D_{rr} is the r-th singular value of Θ^*.

5.2. Related Work

Besides the line of literature on stochastic low-rank matrix bandit with sub-Gaussian noise that is summarized in Chapter 4, we would introduce some other work that is related with our topic.

Bandit under Heavy-tailedness. Research on bandits with heavy-tailed rewards assumes the noise has finite $(1+\delta)$ moment, $\delta \in (0,1)$, and most existing algorithms follow two key strategies: truncation and median of means. Start with Bubeck et al. (2013), a UCB-based algorithm was proposed for MAB with heavy-tailed rewards, enjoying a logarithmic regret bound. To extend their study to the SLB setting, Medina & Yang (2016) developed two algorithms based on the truncation and median of means ideas, but both methods could only attain the regret bound of order $\tilde{O}(T^{\frac{3}{4}})$ when $\epsilon = 1$, which fails to fulfill our expectations. Shao et al. (2018) then refined their results on SLB and introduced two algorithms with improved regret bound. They also constructed a matching lower bound with T. Xue et al. (2020) investigated on the finite arm case and provided two SubLinUCB-based (Chu et al., 2011) algorithms. Recently, Kang & Kim (2023) borrowed the ideas from Huber regression and proposed an improved Huber bandit under finite arm sets. However, their work is confined to the low-dimensional bandit without sparsity, and their parameter vectors are presumed to be arm-dependent under the finite arm set. Another contemporary work (Xue et al., 2023) developed a nearly optimal algorithm for arbitrary arm sets with reduced computation in practice. Yet, none of these studies tackle the heavy-tailedness under the more challenging contextual high-dimensional bandits problem with sparsity, a useful niche our work aims to fill.

Matrix Recovery under Heavy-tailedness. All studies on low-rank matrix estimation revolve around two ideas: Convex approaches tend to replace the classic square loss with some more robust ones, like the renowned Huber loss (Huber, 1965; Sun et al., 2020). Tan et al. (2022) considered the sparse multitask regression under heavy-tailed noise, contrasting our focus on the trace regression problem. The two works most closely related to ours are Fan et al. (2021); Yu et al. (2023). Fan et al. (2021) established a two-step method for the robust trace regression, but they assumed the noise possesses finite $2k$ moment for $k > 1$ and their approximation error is not even proportional to the noise size. Yu et al. (2023) further employed the Huber loss to develop an enhanced regressor with error aligned with the noise scale as long as the noise has bounded variance. In our work, we further complement their result and revisit the Huber-type estimator robust to noise with only finite $(1+\delta)$

moment for any $\delta \in (0,1]$, and we deduce the error rate of order $\tilde{O}((d/n)^{\frac{\delta}{1+\delta}}\mathbb{E}(|\eta_t|^{1+\delta})^{\frac{1}{1+\delta}})$ scaling with the noise scale decently. On the other hand, nonconvex methods aim to seek local optima of the matrix recovery problem via gradient descent. The notable work Shen et al. (2022) developed a Riemannian sub-gradient method and attained the optimal statistical rate under heavy-tailed noises with bounded $(1+\delta)$ moment, but their work relies on some additional assumptions like the noise is symmetric or zero-median. In summary, our work stands as the first solution to address the trace regression problem under arbitrary heavy-tailed noise with only bounded $(1+\delta)$ moment ($\delta \in (0,1)$), which is significant on its own strengths.

5.3. Preliminaries

We will present the setting of LowHTR and introduce the common assumptions for theoretical analysis in this section. Denote T as the total horizon, which may be unknown to the agent. At each round $t \in [T]$, the agent is given an arm set $\mathcal{X}_t \subseteq \mathbb{R}^{d_1 \times d_2}$ ($d_1 \asymp d_2$) that can be fixed or varying over time. Then the agent chooses an arm $X_t \in \mathcal{X}_t$ and observes the associated stochastic reward y_t such that,

$$y_t = \langle X_t, \Theta^* \rangle + \eta_t, \tag{5.1}$$

where $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is an unknown parameter matrix with rank $r \ll d_1 \wedge d_2$ and η_t is the heavy-tailed noise. Specifically, we assume $\mathbb{E}(\eta_t|\mathcal{F}_t) = 0$ and $\mathbb{E}(|\eta_t|^{1+\delta}|\mathcal{F}_t) \leq c$ for some $\delta \in (0,1], c > 0$ conditional on the history filtration $\mathcal{F}_t = \{X_t, X_{t-1}, \eta_{t-1}, \ldots, X_1, \eta_1\}$, which indicates that $\mathbb{E}(y_t|\mathcal{F}_t) = \langle X_t, \Theta^* \rangle$. The compact SVD of Θ^* can be written as $\Theta^* = UDV^\top$ for some $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$, and we denote D_{ii} as its i-th largest singular value. Furthermore, we define $X_t^* := \arg\max_{X \in \mathcal{X}_t} \langle X, \Theta^* \rangle$ as the feature matrix of the optimal arm at round t, and the goal is to minimize the cumulative regret in total T rounds formulated as $R_T = \sum_{t=1}^{T} \langle X_t^*, \Theta^* \rangle - \langle X_t, \Theta^* \rangle$. Next, we present two mild and regular assumptions.

Assumption 5.3.1. We can find a sampling distribution $\mathcal{D}$ over $\mathcal{X}_t$ with the covariance matrix Σ, such that $\mathcal{D}$ is sub-Gaussian with parameter $\sigma^2 \asymp c_l := \lambda_{\min}(\Sigma) \asymp 1/(d_1 d_2)$.

Assumption 5.3.1 is commonly used in the modern low-rank matrix bandits (Kang et al., 2022; Lu et al., 2021), and can be easily satisfied in many cases. For instance, when $\mathcal{X}_t$ is a region in $\mathbb{R}^{d_1 \times d_2}$

(e.g., Euclidean unit ball), we can find such a sampling distribution if the convex hull of this region contains a ball with some constant radius. And when $\mathcal{X}_t$ is a finite set, it suffices if the arms are IID drawn from some sub-Gaussian distribution at each time. Note a random matrix $X \in \mathbb{R}^{d_1 \times d_2}$ follows sub-Gaussian distribution with parameter σ^2 if for any $t \in \mathbb{R}$ s.t.,

$$P(\langle A, X \rangle \geq \sqrt{2}\, \|A\|_{\mathrm{F}}\, t) \leq 2 \exp\left(-t^2/\sigma^2\right), \forall A \in \mathbb{R}^{d_1 \times d_2}.$$

Assumption 5.3.2. We have $\|\Theta^*\|_{\mathrm{F}} \leq S$, and for any $t \in [T], X \in \mathcal{X}_t$, it holds that $\|X\|_{\mathrm{F}} \leq S$.

Assumption 5.3.2 is very standard in contextual bandit literature. As a consequence, we can deduce that $\mathbb{E}(|y_t|^{1+\delta}|\mathcal{F}_t) \leq 2^\delta S^2 + 2^\delta c := b$. Based on the conditions on the sub-Gaussian parameter σ in Assumption 5.3.1, we can prove that $\|X\|_{\mathrm{F}}$ is bounded in a constant scale with high probability with its proof in Appendix D.1. But for simplicity and consistency with previous literature, we still impose this common assumption to bound $\|X\|_{\mathrm{F}}$ here. Note our work can be naturally extended to the generalized low-rank matrix bandit problem by further assuming the derivative of the inverse link function is bounded in the interval $[-S^2, S^2]$. Such an adaptation would result in the final regret bound being affected only by a constant factor, and we will leave it as our future work.

5.4. Methods

In this section, we present our novel LowTO With Estimated Subspaces (LOTUS) algorithm for the LowHTR problem. Our algorithm runs in a batched format adapted from the doubling trick (Besson & Kaufmann, 2018). And inspired by the success of the two-stage framework in ESTR (Jun et al., 2019), in each batch our algorithm also first recovers the subspaces spanned by Θ^*, and then invokes a new approach called LowTO that heavily penalizes on columns and rows complementary to our estimated subspaces. Contrasting prior works, our algorithm could dynamically switch between the exploration and exploitation stages so as to be agnostic to the horizon T, which is significantly more useful. We further improve LOTUS to operate without knowing the sparsity r, which further enhances its practicality.

Initially, we will introduce the nuclear penalized Huber-type low-rank matrix estimator under heavy-tailed noise as follows. Contracting the results in Yu et al. (2023), we further prove that our Huber-type estimator is robust to arbitrary heavy-tailed noise with the finite $(1+\delta)$ moment for $\delta \in (0, 1)$ on the trace regression problem.

Algorithm 8 LowTO With Estimated Subspaces (LOTUS)

Input: Arm set $\mathcal{X}_t$, sampling distribution $\mathcal{D}_t$, $\delta, T_0, \eta, \lambda, \{\lambda_{i,\perp}\}_{i=1}^{+\infty}$.
Initialization: The history buffer index set $\mathcal{H}_1 = \{\}$, the exploration buffer index set $\mathcal{H}_2 = \{\}$.

1: Pull arm $X_t \in \mathcal{X}_t$ according to $\mathcal{D}_t$ and observe payoff y_t. Then add (X_t, y_t) into $\mathcal{H}_1$ and $\mathcal{H}_2$ for $t \leq T_0$.
2: **for** $i = 1, 2, \ldots$ until the end of iterations **do**
3: Set the exploration length $T_1 = \min\left\{\left[\frac{d^{2+4\delta} r^{1+\delta}}{D_{rr}^{2+2\delta}} 2^{i(1+\delta)}\right]^{\frac{1}{1+3\delta}}, 2^i\right\}$.
4: For iteration t from $|\mathcal{H}_1|+1$ to $|\mathcal{H}_1|+T_1$, pull arm $X_t \in \mathcal{X}_t$ according to $\mathcal{D}_t$ and observe payoff y_t. Then add (X_t, y_t) into $\mathcal{H}_1$ and $\mathcal{H}_2$
5: Obtain the estimate $\widehat{\Theta}$ based on Eqn. (5.3) with $\mathcal{H}_2$, where we set $\tau_i \asymp \left(|\mathcal{H}_2|/(d + \ln(2^{i+1}/\epsilon))\right)^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}}, \lambda_i \asymp \sigma\left((d + \ln(2^{i+1}/\epsilon))/|\mathcal{H}_2|\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}$.
6: Calculate the full SVD of $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\, \widehat{D}\, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U} \in \mathbb{R}^{d_1 \times r}, \widehat{V} \in \mathbb{R}^{d_2 \times r}$.
7: For $T_2 = 2^i - T_1$ rounds, invoke LowTO with $\delta, [\widehat{U}, \widehat{U}_\perp], [\widehat{V}, \widehat{V}_\perp], \lambda, \lambda_{i,\perp}, \mathcal{H}_1$ and obtain the updated $\mathcal{H}_1$.

5.4.1. Low-rank Matrix Estimation. Suppose we collect n pairs of data $\{(X_i, y_i)\}$ according to some distribution satisfying Assumption 5.3.1 for X_i and the model of Eqn. (5.1) for the associated y_i after time n. Define the Huber loss (Huber, 1965) $l_\tau(\cdot)$ parameterized by the robustification $\tau > 0$ (Sun et al., 2020) as:

$$l_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \leq \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases}$$

To obtain a low-rank matrix estimate, we use the nuclear norm penalization as a convex surrogate for the rank and implement the following nuclear norm regularized Huber regressor to recover the subspaces under heavy-tailedness:

$$\widehat{\Theta} = \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \hat{L}_{\tau,[n]}(\Theta) + \lambda \|\Theta\|_{\text{nuc}}, \ \hat{L}_{\tau,[n]}(\Theta) = \frac{1}{n} \sum_{i \in [n]} l_\tau\left(y_i - \langle X_i, \Theta \rangle\right), \tag{5.2}$$

where τ and λ stand for the Huber loss robustification and the nuclear norm penalization parameters, respectively.

We then establish the following statistical properties of the estimator defined in Eqn. (5.2):

THEOREM 5.4.1. *By extending Assumption 5.3.1 with any order of σ and c_l, With probability at least $1-\epsilon$, the low-rank estimator $\widehat{\Theta}$ in Eqn. (5.2) with $\tau \asymp (n/(d + \ln(1/\epsilon)))^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}}$ and*

$\lambda \asymp \sigma\left((d+\ln(1/\epsilon))/n\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}$ *satisfies*

$$\left\|\widehat{\Theta}-\Theta^*\right\|_F \leq C_1 \frac{\sigma}{c_l}\left(\frac{d+\ln(1/\epsilon)}{n}\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}\sqrt{r},$$

for some constant C_1 *as long as we have* $n \gtrsim dr\nu^3, d, \nu^2$, *and* $(d-\ln(\epsilon))\sqrt{r\nu^3}$ *with* $\nu = \sigma^2/c_l$.

The proof of Theorem 5.4.1 involves a construction of the restricted strong convexity for the empirical Huber loss function $\hat{L}_\tau(\cdot)$ and a deduction of an upper bound for $\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}}$, and the details are presented in Appendix D.4. Note Theorem 5.4.1 generally holds without any restriction on the scale of σ and c_l. Provided the noise has a finite variance, i.e., $\delta = 1$, the deduced l_2-error rate aligns with the minimax value (Fan et al., 2019) under the standard penalized low-rank estimator with sub-Gaussian noise. Based on our knowledge, this is the first error bound in the trace regression problem under noise with finite $(1+\delta)$ moment $(\delta < 1)$ assuming nothing further.

To solve the convex optimization problem in Eqn. (5.2), we adopt the local adaptive majorize-minimization (LAMM) method (Fan et al., 2018; Sun et al., 2020; Yu et al., 2023) that is fast to use and scalable to large datasets. This method constructs an isotropic quadratic function to upper bound the Huber loss and utilizes a majorize-minimization algorithm for finding the optimal solution. One noteworthy advantage of this procedure is that the minimizer often yields a closed-form solution. Due to the space limit, we defer more details and the pseudocode to Appendix D.3.

5.4.2. LOTUS: The Rank r is Known. We will present our LOTUS algorithm in this subsection. To improve the two-stage framework introduced in Jun et al. (2019) which requires the knowledge of T and to further yield robust performance against heavy-tailedness, our LOTUS adaptively switches between exploration and exploitation in a batch manner without knowing T, and is equipped with a new LowTO algorithm designed for heavy-tailed rewards. The LOTUS algorithm is presented in Algorithm 8, with three core steps introduced in detail as follows:

Adaptive Exploration and Exploitation: Drawing inspiration from the doubling trick (Besson & Kaufmann, 2018), after some warm-up iterations of size T_0, our LOTUS operates with batches until termination where the batch sizes increase exponentially as $\{2^i\}_{i=1}^{+\infty}$. We define $\mathcal{H}_1$ and $\mathcal{H}_2$ as the history and exploration buffer index sets, where after time t all the indexes $[t]$ of past observations are included in $\mathcal{H}_1$ while $\mathcal{H}_2$ only contains sample indexes particularly used for subspace estimation of Θ^*. At the i-th batch of length 2^i, we first set $T_1^i = \min\{(d^{2+4\delta}r^{1+\delta}2^{i+i\delta}/D_{rr}^{2+2\delta})^{\frac{1}{1+3\delta}}, 2^i\}$ as

the exploration length, and we randomly sample T_1 arms according to the sampling distribution in Assumption 5.3.1 and put their indexes into both $\mathcal{H}_1$ and $\mathcal{H}_2$. Subsequently, we obtain an estimate $\widehat{\Theta}$ based on Eqn.(5.2) with samples indexed by $\mathcal{H}_2$, and then leverage the recovered subspaces in the remaining $T_2^i = 2^i - T_1^i$ rounds as the exploitation phase, where we invoke a new algorithm named LowTO. The details of this exploitation phase will be elaborated in the following two points. As shown in Algorithm 8 line 8, indexes of observations under LowTO are only added to $\mathcal{H}_1$ but not $\mathcal{H}_2$ and hence will not be used for matrix estimation. Unlike the traditional doubling trick that restarts the algorithm at each batch, our algorithm facilitates interaction across different batches. Specifically, at the i-th batch, it utilizes all the samples in $\mathcal{H}_1$ and $\mathcal{H}_2$ accumulated from the previous batches for more informed decision-making. Another point to highlight is that our LOTUS algorithm can also be run in a more randomized manner with the same regret bound: at the i-th batch, there is an option to explore with a probability of $T_1^i/2^i$ and to exploit with the remaining probability. We defer its pseudocode to Appendix D.2. For simplicity, we consider our original approach in this work, which involves an initial exploration phase of deterministic length followed by the use of LowTO.

Subspace Transformation: At the i-th batch, after we randomly sample arms for a carefully designed duration and add their observations into $\mathcal{H}_2$, we first acquire the estimated $\widehat{\Theta}$ based on the current $\mathcal{H}_2$ as shown in Eqn. (5.3). With the knowledge of r, then we can obtain its corresponding full SVD as $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\widehat{D}[\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U} \in \mathbb{R}^{d_1 \times r}, \widehat{U}_\perp \in \mathbb{R}^{d_1 \times (d_1 - r)}, \widehat{V} \in \mathbb{R}^{d_2 \times r}$ and $\widehat{V}_\perp \in \mathbb{R}^{d_2 \times (d_2 - r)}$.

$$\widehat{\Theta} = \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \hat{L}_{\tau_i, \mathcal{H}_2}(\Theta) + \lambda_i \|\Theta\|_{\text{nuc}} \tag{5.3}$$

Intuitively, Theorem 5.4.1 implies that our estimated column and row subspaces should align with the ground truth U, V. Borrowing the ideas from ESTR (Jun et al., 2019), we aim to transform the original LowHTR into the linear bandit problem under heavy-tailed rewards with some sparsity feature. Specifically, we first orthogonally rotate the actions set $\mathcal{X}_j$ in the exploitation phase as

$$\mathcal{X}_j^- = \left\{ [\widehat{U}, \widehat{U}_\perp]^\top X [\widehat{V}, \widehat{V}_\perp] : X \in \mathcal{X}_j \right\}, \tag{5.4}$$

$$\Theta^{*,\prime} = [\widehat{U}, \widehat{U}_\perp]^\top \Theta^* [\widehat{V}, \widehat{V}_\perp]. \tag{5.5}$$

Define the total dimension $p \coloneqq d_1 d_2$ and the effective dimension $k \coloneqq p - (d_1 - r)(d_2 - r)$. We perform a tailored vectorization of the arm set $\mathcal{X}_j^-$ as in Algorithm 9 line 4 to obtain a new arm set $\mathcal{X}_t' \subseteq \mathbb{R}^p$, and denote θ^* to be the corresponding rearranged version of $\mathrm{vec}(\Theta^{*,\prime})$ such that $\theta^*_{k+1:p} = \mathrm{vec}(\Theta^{*,\prime}_{r+1:d_1, r+1:d_2})$. Then it holds that $\theta^*_{k+1:p}$ is nearly zero based on the results in Stewart (1990) and Theorem 5.4.1. The formal result is shown as follows for the i-th batch with probability at least $1 - \epsilon$:

$$\left\|\theta^*_{k+1:p}\right\|_2 \lesssim S_\perp \coloneqq \frac{r\sigma^2 c^{\frac{2}{1+\delta}}}{c_l^2 D_{rr}^2} \left(\frac{d + \ln(1/\epsilon)}{|\mathcal{H}_2|}\right)^{\frac{2\delta}{1+\delta}}, \tag{5.6}$$

with the parameter setting that

$$\tau_i \asymp \left(|\mathcal{H}_2|/(d + \ln(1/\epsilon))\right)^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}},\ \lambda_i \asymp \sigma \left((d + \ln(1/\epsilon))/|\mathcal{H}_2|\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}},$$

Its complete proof is presented in Appendix D.6. Consequently, we can simplify the LowHTR problem to an equivalent p-dimensional linear bandits under heavy-tailedness with a unique sparse pattern, i.e., the final $(p - k)$ entries of θ^* are almost zero based on Eqn. (5.6).

Following the recovery of row and column subspaces of Θ^* and the particular arm set transformation after T_1^i rounds in the i-th batch, we will leverage the resulting almost-low-dimensional structure by using the following LowTO algorithm for the rest of the batch's duration.

LowTO Algorithm: To begin with, we reformulate the resulting p-dimensional linear bandit problem under heavy-tailed rewards in the following way: at round t, the agent chooses an arm $x_t \in \mathcal{X}_t'$ of dimension p where $\mathcal{X}_t'$ is a rearranged vectorization of $\mathcal{X}_t^-$ as defined in Algorithm 9 line 4, and observes a noisy payoff $y_t = x_t^\top \theta^* + \eta_t$ mixed with some heavy-tailed noise η_t.

Our LowTO algorithm is presented in Algorithm 9. Inspired by LowOFUL in the ESTR method (Jun et al., 2019), to exploit the additional pattern of θ^* shown in Eqn. (5.6), we propose the almost-low-dimensional truncation under OFU (LowTO) algorithm As shown in Algorithm 9 line 2, our LowTO also truncates each entry of $M^{-1/2} x_i y_i$ for $i = 1, \ldots, t-1$ at time t by some increasing threshold b_t, Different from linear bandits under heavy-tailedness, when calculating the estimator $\hat{\theta}$ in Algorithm 9 line 3, we put a weighted regularizer as the diagonal matrix $\Lambda = \mathrm{diag}(\lambda, \ldots, \lambda, \lambda_\perp, \ldots, \lambda_\perp)$ with λ only applied to the first k coordinates. By amplifying $\lambda_\perp$, we ensure greater penalization is applied to the final $p - k$ elements of $\hat{\theta}$ leading to their diminished values, and this phenomenon is well intended under the almost-low-dimensional structure. Subsequently, we utilize a UCB-based

Algorithm 9 LowTO

Input: $T, \delta, [\widehat{U}, \widehat{U}_\perp], [\widehat{V}, \widehat{V}_\perp], \lambda_0, \lambda_\perp, \mathcal{H}_1$.

Stage $M = \sum_{(x,y)\in\mathcal{H}_1'} xx^\top + \Lambda = \sum_{t=1}^{|\mathcal{H}_1'|} x_{s,t}x_{s,t}^\top + \Lambda$, $X^\top = [x_{s,1}, \ldots, x_{s,|\mathcal{H}_1'|}]$, $[u_1, \ldots, u_p]^\top = M^{-\frac{1}{2}}X^\top$

$$\text{with } \mathcal{H}_1' = \Big\{\Big(x_{s,t}^\top = [\mathrm{vec}(\widehat{U}^\top X\widehat{V})^\top, \mathrm{vec}(\widehat{U}^\top X\widehat{V}_\perp)^\top,$$
$$\mathrm{vec}(\widehat{U}_\perp^\top X\widehat{V})^\top, \mathrm{vec}(\widehat{U}_\perp^\top X\widehat{V}_\perp)^\top], y_{s,t} = y\Big) : (X, y) \in \mathcal{H}_1\Big\}.$$
$$\Lambda = \mathrm{diag}([\underbrace{\lambda_0, \ldots, \lambda_0}_{k}, \underbrace{\lambda_\perp, \ldots, \lambda_\perp}_{p-k}])$$

1: **for** $t = 1$ **to** T **do**

2: Get $\hat{y}_i = [y_{s,1}\mathbb{1}_{u_{i,1}y_{s,1}\le b_{t-1}}, \ldots, y_{t-1}\mathbb{1}_{u_{i,|\mathcal{H}_1|+t-1}y_{t-1} \le b_{t-1}}]^\top$ for $i \in [p]$, where $\hat{y}_i \in \mathbb{R}^{|\mathcal{H}_1|+t-1}$.

3: Calculate $\hat{\theta}_{t-1} = M^{-1/2}[u_1^\top\hat{y}_1, \ldots, u_p^\top\hat{y}_p]^\top$.

4: Transform the arm set $\mathcal{X}_t$ as

$$\mathcal{X}_t' = \Big\{[\mathrm{vec}(\widehat{U}^\top X\widehat{V})^\top, \mathrm{vec}(\widehat{U}^\top X\widehat{V}_\perp)^\top, \mathrm{vec}(\widehat{U}_\perp^\top X\widehat{V})^\top,$$
$$\mathrm{vec}(\widehat{U}_\perp^\top X\widehat{V}_\perp)^\top]^\top \in \mathbb{R}^p : X \in \mathcal{X}_t\Big\}.$$

5: Pull $x_t = \arg\max_{x\in\mathcal{X}_t'} x^\top\hat{\theta}_{t-1} + \beta_{t-1}\|x\|_{M^{-1}}$ and observe the reward y_t.

6: Restore x_t into its original matrix form X_t and then add (X_t, y_t) into $\mathcal{H}_1$.

7: Update $M = M + x_tx_t^\top$, $X^\top = [X^\top, x_t]$ and $[u_1, \ldots, u_p]^\top = M^{-1/2}X^\top$.

8: **return** The history buffer $\mathcal{H}_1$.

criterion to choose the pulled arm according to Algorithm 9 line 5, where we also decrease the variation of the last $p-k$ elements with M^{-1} to further reduce their impact on the decision-making. It is also noteworthy that we always reuse all the past observations stored in $\mathcal{H}_1$ at each batch when initializing the matrix M, which can facilitate a consistent and accurate estimator $\hat{\theta}$ in the early stage of the exploitation phase. And the randomly drawn samples in $\mathcal{H}_1$ contain more stochasticity and thus are more preferable for the parameter estimation.

We then state the regret bound of LowTO in Theorem 5.4.2:

THEOREM 5.4.2. *Suppose the input $\mathcal{H}_1$ is of size $H \lesssim T$ and we run our LowTO algorithm for T rounds. By setting $b_t = (b/\log(2p/\epsilon))^{\frac{1}{1+\delta}}(t+H)^{\frac{1-\delta}{2+2\delta}}$, $\beta_t = 4\sqrt{p}b^{\frac{1}{1+\delta}}\log(2p/\epsilon)^{\frac{\delta}{1+\delta}}(t+H)^{\frac{1-\delta}{2+2\delta}} + \sqrt{\lambda_0}S + \sqrt{\lambda_\perp}S_\perp$ with $\lambda_\perp = S^2T_2/(k\log(1+\frac{S^2T}{k\lambda_0}))$, with probability at least $1-\epsilon$, the regret of LowTO can be bounded by:*

$$\tilde{O}\left(\sqrt{kp}\,(T+H)^{\frac{1}{1+\delta}} + \sqrt{kT} + S_\perp T\right),$$

where $S_\perp$ is the upper bound of $\|\theta_{k+1:p}\|_2$ as shown in Eqn. (5.6) depending on $|\mathcal{H}_2|$.

In standard linear bandit under heavy-tailed noise case, we can recover the same regret bound of TOFU in the order of $\tilde{O}(p \cdot T^{\frac{1}{1+\delta}})$ by setting $S_\perp = S$ and $\lambda_\perp = \lambda$.

Overall regret: Now we are ready to present the overall regret bound for LOTUS in the following Theorem 5.4.3.

THEOREM 5.4.3. *By using the configuration of LowTO described in Theorem 5.4.2 and the parameter values of LOTUS shown in Algorithm 8 for each batch, and set ϵ as $\epsilon/2^{i+1}$ in β_t (formulated in Theorem 5.4.2) for the i-th batch. Then with probability at least $1-\epsilon$, it holds that*

$$R(T) \leq \tilde{O}\left(d^{\frac{2+4\delta}{1+3\delta}} r^{\frac{1+\delta}{1+3\delta}} T^{\frac{1+\delta}{1+3\delta}} / D_{rr}^{\frac{2+2\delta}{1+3\delta}} + d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}}\right),$$

under the condition that $T_1 \geq 5d^{\frac{1+2\delta}{\delta}} r^{\frac{1+\delta}{2\delta}} / D_{rr}^{\frac{1+\delta}{\delta}}$. Furthermore, we can simplify the above result as

$$R(T) \leq \begin{cases} \tilde{O}\left(d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{2}} / D_{rr}\right), \delta = 1; \\ \tilde{O}\left(d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}}\right), \delta < 1, T \gtrsim (dr)^{\frac{1+\delta}{2\delta}} / D_{rr}^{\frac{2(1+\delta)^2}{\delta(1-\delta)}}. \end{cases}$$

Note the regret bound in Theorem 5.4.3 improves upon the one attained for a simple linear bandit reduction, which contains the order of d^2. When the rewards have bounded variance, i.e., $\delta = 1$, our regret bound matches the modern one for low-rank matrix bandit under sub-Gaussian noise up to logarithmic terms (Kang et al., 2022; Lu et al., 2021).

5.4.3. LOTUS: The Rank r is Unknown. While all existing algorithms for low-rank matrix bandits require prior knowledge of the rank r, this information is never revealed to agents in real-world applications, and hence misspecification of r will not only undermine the theoretical foundations but also severely compromise the performance of these methods. To solve this crucial challenge, in this section we aim to enhance our LOTUS algorithm to be agnostic to r even under the more complex heavy-tailed scenario. For the Lasso bandit, which is another popular and easier high-dimensional bandit with sparsity, some algorithms (Ariu et al., 2022; Oh et al., 2021) free of the sparsity index have been recently introduced. However, when compared with our work, all of them necessitate some additional assumptions on the structure of the underlying parameter as well as the sampling distribution. For example, Oh et al. (2021) further assumes that the active entries of the parameter vector are relatively independent and the skewness of the sampling distribution is bounded. This fact substantiates the huge difficulty of devising an efficient algorithm for LowHRT

without additional conditions. Note our work also opens up a potential avenue for exploring low-rank matrix bandits without the need for knowledge about r, and we believe that completely addressing this intriguing problem must require more specific assumptions and investigations.

To improve our batched-explore-then-exploit-based LOTUS algorithm, an intuitive idea is to estimate the effective rank of $\widehat{\Theta}$ right after the matrix recovery in each batch. By trimming the estimated singular values $\{D_{ii}\}_{i=1}^{d}$ with some craftily designed increasing sequence that is deduced from Theorem 5.4.1, we could obtain a useful rank $\hat{r}$ with $\hat{r} \leq r$ and then only focus on the top-$\hat{r}$ row and column subspaces. We can demonstrate that all the ground truth singular values $\{D_{ii}\}_{i=\hat{r}+1}^{d}$ omitted are nearly null and hence negligible. Therefore, by penalizing the subspaces parallel to those omitted directions with a similar idea used in our original LOTUS, we could enjoy the low-rank benefit of LowHTR. Specifically, to modify line 6 and line 7 in Algorithm 8, we abuse the notation here and denote $\widehat{D}$ as the singular value matrix of $\widehat{\Theta}$ that is deduced in line 5. Subsequently, we estimate the useful rank $\hat{r}$ as

$$\hat{r} = \min\left\{ i \in [d+1] : \widehat{D}_{ii} \leq C_1 \frac{\sigma\sqrt{i}}{c_l} \left(\frac{d + \ln\left(2^{i+1}/\epsilon\right)}{|\mathcal{H}_2|} \right)^{\frac{\delta}{1+\delta}} \cdot c^{\frac{1}{1+\delta}} \right\} - 1 \wedge 1,$$

where C_1 is some specific constant in Theorem 5.4.1 and $\widehat{D}_{(d+1)(d+1)}$ is set to be 0 to avoid the empty set case. Afterward, we rewrite the full SVD of $\widehat{\Theta}$ as $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\, \widehat{D}\, [\widehat{V}, \widehat{V}_\perp]^\top$ with $\widehat{U} \in \mathbb{R}^{d_1 \times \hat{r}}, \widehat{V} \in \mathbb{R}^{d_2 \times \hat{r}}$ for each batch in line 6. In new line 7 of our improved LOTUS, we then input the new $[\widehat{U}, \widehat{U}_\perp]$ and $[\widehat{V}, \widehat{V}_\perp]$ with the estimated rank $\hat{r}$ as described above, and the effective dimension k in the following subspace estimation and LowTO implementation will become $k = p - (d_1 - \hat{r})(d_2 - \hat{r})$. Note $\hat{r}$ might differ across different batches, but $\hat{r} \leq r$ consistently holds. Conclusively, we can obtain the following regret bound of our improved LOTUS algorithm agnostic to r:

THEOREM 5.4.4. *By using the same setting and conditions of LOTUS as described in Theorem 5.4.3 and Algorithm 8 with* $T_1 = \min\left\{ d \cdot 2^{\frac{i(1+\delta)}{1+2\delta}}, 2^i \right\}$ *in line 3 of Algorithm 8, and utilizing the estimated useful rank* $\hat{r}$ *to set the corresponding value of* k *at each batch, the cumulative regret of our LOTUS agnostic to* r *can be bounded as*

$$R(T) \leq \tilde{O}\left(d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}} + d r^{\frac{3}{2}} T^{\frac{1+\delta}{1+2\delta}} \right),$$

with probability at least $1 - \epsilon$.

The above regret bound is efficient under the high-dimensional scenario, i.e., $d \gtrsim T^{\frac{\delta^2}{(1+2\delta)(1+\delta)}}$. While generally there exists a disparity between our derived regret bound in cases where r remains undisclosed and the optimal one, as previously discussed in this section, it would prove exceptionally difficult to devise an algorithm for LowHTR that remains agnostic to r while achieving an enhanced regret bound. Solving this issue would necessitate the formulation of more specific assumptions on the underlying structure of the arm matrices and Θ^*.

Moreover, we will showcase the superior efficiency of our LOTUS algorithm in both scenarios, whether the agent possesses knowledge of r or not, in the following experimental results in Section 5.6.

5.5. Regret Lower Bound

In this section, we provide a lower bound for the expected cumulative regret in LowHTR particularly regarding the order of T. The result is given as follows:

THEOREM 5.5.1. *Under the LowHTR problem with d, r, T and $S = 1$ in Assumption 5.3.2, there exists an instance with a fixed $\mathcal{X}_t$ containing $(d-1)r$ arms for which any algorithm must suffer an expected regret of order $\Omega(d^{\frac{\delta}{1+\delta}} r^{\frac{\delta}{1+\delta}} T^{\frac{1}{1+\delta}})$, i.e., $\mathbb{E}(R_T) \gtrsim d^{\frac{\delta}{1+\delta}} r^{\frac{\delta}{1+\delta}} T^{\frac{1}{1+\delta}} \gtrsim T^{\frac{1}{1+\delta}}$.*

Theorem 5.5.1 demonstrates that our LOTUS could attain the lower bound for LowHTR regarding the order of T when r is given. And this lower bound is tight with $r = d$ and finite arm sets since it matches the minimax rate for standard linear bandits under heavy-tailed noise (Xue et al., 2020). Further exploring the regret lower bound for d and r under LowHTR is notably challenging, given the fact that even the simpler low-rank matrix bandits under sub-Gaussian noise this problem is not thoroughly studied (Kang et al., 2022). And the regret lower bound may differ in the order of d when the arm set is infinitely large and arbitrary (Shao et al., 2018). We will leave them as future directions.

5.6. Experimental Results

We demonstrate that our proposed LOTUS yields superior performance over the existing LowESTR algorithm (Lu et al., 2021) in the presence of heavy-tailed noise under a suite of simulations. Since our work is the first one to study the LowHTR problem and currently there is no existing method for comparison, we utilize the LowESTR algorithm specifically designed for the sub-Gaussian noise

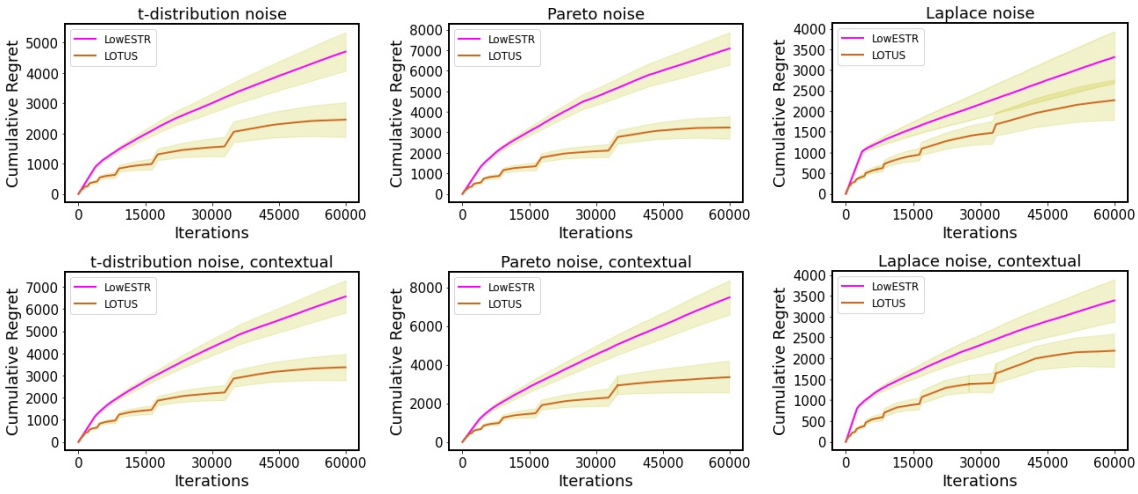


FIGURE 5.1. Plots of cumulative regrets of LowESTR and our proposed LOTUS with fixed or changing contextual arm set under t-distribution, Pareto, and Laplace heavy-tailed noise. We use the LOTUS algorithm agnostic to r in the first three experiments displayed in the first row, and we utilize the value of r in LOTUS in experiments shown in the second row.

to validate the robustness of our proposed LOTUS. LowESTR also borrows the idea of the two-stage framework from ESTR, and it improves upon ESTR on the computational efficiency of the matrix recovery step. It requires both the knowledge of the horizon T and the rank r as inputs. In the following experiments, we showcase that it becomes vulnerable and achieves suboptimal performance under heavy-tailed noise in practice as expected.

We consider two different settings of the parameter matrices Θ^* with $d_1 = d_2 = 10$ and $r = 2$. For the first scenario, we set the parameter matrix as a diagonal matrix $\Theta^* = \text{diag}([7, 4, 0, \ldots, 0])$. The arm set is fixed where we draw 500 random matrices from $\{X \in \mathbb{R}^{10\times 10} : \|X\|_{\text{F}} \leq 1\}$ in the beginning. And we implement the improved LOTUS algorithm introduced in Subsection 5.4.3 that is unaware of the rank r in this scenario. For the second case, we consider a more challenging parameter matrix Θ^* such that its first row represents a random vector of norm 7 and its second row is a perpendicular vector of norm 4 with other entries set to 0. Contrasting the first scenario, we consider a contextual arm set with 10 feature matrices drawn from $\{X \in \mathbb{R}^{10\times 10} : \|X\|_{\text{F}} \leq 1\}$ at each round. And we use the original LOTUS algorithm introduced in Subsection 5.4.2 requiring the knowledge of $r = 2$. For the heavy-tailed noise η_t, we consider the following three types of distribution for both scenarios introduced above:

- **Student's t-distribution:** The density function is given as $f(x) \asymp (1 + x^2/\nu)^{-\frac{\nu+1}{2}}$ with degree of freedom parameter $\nu > 0$ and $x \in \mathbb{R}$. By setting $\nu = 1.7$, it has infinite variance but finite 1.5 moment bounded by 6. The heavy-tail index is equal to 1.60.[2]
- **Pareto distribution:** The density function is given as $f(x) \asymp \alpha/(x+1)^{\alpha+1}$ for some shape parameter $\alpha > 0$ and $x > 0$. By setting $\alpha = 1.9$, it also has infinite variance but finite 1.5 moment bounded by 5. And the heavy-tail index is equal to 2.20.
- **Laplace distribution:** The density distribution is formulated as $f(x) \asymp \exp(-|x|/b)$ with some scale parameter b for $x \in \mathbb{R}$. By setting $b = 1$, the distribution possesses a finite variance bounded by 2. The heavy-tail index of this distribution is 1.36.

According to Figure 5.1, we observe that our LOTUS algorithm consistently exhibits superior and more resilient performance across all six scenarios compared to LowESTR. This advantage is particularly evident when dealing with distributions with a higher heavy-tail index, which is aligned with our expectations. On the contrary, LowESTR performs fairly in the presence of Laplace noise with a finite variance but struggles when faced with Pareto noise possessing stronger heavy-tailedness. Furthermore, it is noteworthy that the cumulative regret of the LOTUS algorithm exhibits a batch-wise increase, with a progressively clearer sub-linear pattern emerging in subsequent batches. This fact firmly validates the practical superiority of our LOTUS algorithm under both cases when the rank r is presented or not.

[2] A greater heavy-tail index exceeding 1 indicates that the distribution possesses stronger fluctuation and heavy-tailedness. (Hoaglin et al., 2000)

CHAPTER 6

Conclusion and Future Work

In this dissertation, we aim to solve some important and unexplored challenges of the bandit problems from both theoretical and practical perspectives. In Chapter 2, we introduce a new problem of Lipschitz bandits in the presence of adversarial corruptions, and we originally provide efficient algorithms against both weak adversaries and strong adversaries when agnostic to the total corruption budget C. The robustness and efficiency of our proposed algorithms is then validated under comprehensive experiments. In Chapter 3, we propose the first online continuous hyperparameter optimization method for contextual bandit algorithms named CDT given the continuous hyperparameter search space. Our framework can attain sublinear regret bound in theory, and is general enough to handle the hyperparameter tuning task for most contextual bandit algorithms. Multiple synthetic and real experiments with multiple GLB algorithms validate the remarkable efficiency of our framework compared with existing methods in practice. In the meanwhile, we propose the Zooming TS algorithm with Restarts, which is the first work on Lipschitz bandits under the *switching* environment. In Chapter 4, we discussed the generalized linear low-rank matrix bandit problem. We proposed two novel and efficient frameworks called G-ESTT and G-ESTS, and these two methods could achieve decent bounds of regret under some mild conditions. The practical superiority of our proposed frameworks is also validated under comprehensive experiments. And finally in Chapter 5, we introduce and examine the new problem of LowHTR, and we propose a robust algorithm named LOTUS that can be agnostic to T and even the rank r with a slightly milder regret bound. We also develop a matching lower bound to demonstrate our LOTUS is nearly optimal in the order of T. Meanwhile, we prove that our Huber-type estimator could solve the trace regression problem under arbitrary heavy-tailed noise with finite $(1+\delta)$ moment ($\delta \in (0, 1]$) and its Frobenious norm error is of scale $\tilde{O}((d/n)^{\frac{\delta}{1+\delta}}\mathbb{E}(|\eta|^{1+\delta})^{\frac{1}{1+\delta}})$ (η is the random noise). The practical superiority of our proposed method is validated under simulations.

There are several directions for our future work. For the Lipschitz bandits, how to propose an algorithm attaining the regret lower bound without knowing the rank r under adversarial corruptions is unknown, and conducting a comprehensive study on the non-stationary Lipschitz bandit problem also remains intriguing and unexplored. For the model selection of bandits, considering a continuous candidate space has hardly been explored and remains an interesting future direction. For the high-dimensional bandits without knowing the sparsity value (e.g. low rank r), it remains compelling to close the regret gap of the lower bound under some mild assumptions.

APPENDIX A

Appendix for Chapter 2

A.1. Analysis of Theorem 2.4.1

We modify the proof in Kleinberg et al. (2019) by dividing the cumulative regret into two parts, where the first part controls the error coming from the stochastic rewards and the second part deals with the extra error from adversarial corruptions in the following Appendix A.1.2. In the beginning we will present some auxiliary lemmas for preparation.

A.1.1. Useful Lemmas.

DEFINITION A.1.1. *We call it a clean process for Algorithm 1, if for each time $t \in [T]$ and each active arm $v \in \mathcal{X}$ at any time t, we have $|f(v) - \mu(v)| \leq r(v)$.*

Here we expand some notations from Algorithm 1: we denote $n_t(v)$ as the number of times the arm v has been pulled until the round t, and $f_t(x), r_t(x)$ as the corresponding average stochastic rewards and confidence radius respectively at time t such that,

$$r_t(x) = \sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n_t(x)}} + \frac{C}{n_t(x)}.$$

Note in our Algorithm 1 we do not write this subscript t for these components since there is no ambiguity in the description. And W.l.o.g we assume the optimal arm $x_* = \arg\max_{x\in\mathcal{X}} \mu(x)$ is unique in $\mathcal{X}$.

LEMMA A.1.1.1. *Given the adversarial corruptions are at most C, for Algorithm 1, the probability of a clean process is at least $1 - \delta$.*

PROOF. For each time $t \in [T]$, consider an arm $x \in \mathcal{X}$ that is active by the end of time t. Recall that when Algorithm 1 pulls the arm x, the reward is sampled IID from some unknown distribution $\mathbb{P}_x$ with expectation $\mu(x)$. And in the meanwhile, the stochastic reward may be corrupted by the adversary. Define random variables $U_{x,s}$ and values $C_{x,s}$ for $1 \leq s \leq n_t(x)$ as follows: for $s \leq n_t(x)$,

$U_{x,s}$ is the stochastic reward from the s-th time arm x is played and $C_{x,s}$ is the corruption injected on $U_{x,s}$ before the agent observing it. By applying Bernstein's Inequality, it naturally holds that

$$
\begin{aligned}
&P\left(|f_t(x)-\mu(x)| \geq r_t(x)\right) = P\left(|f_t(x)-\mu(x)| \geq \sqrt{\frac{4\ln T + 2\ln(2/\delta)}{n_t(x)}} + \frac{C}{n_t(x)}\right) \\
=&P\left(\left|\sum_{s=1}^{n_t(x)} \frac{U_{x,s}}{n_t(x)} + \sum_{s=1}^{n_t(x)} \frac{C_{x,s}}{n_t(x)} - \mu(x)\right| \geq \sqrt{\frac{4\ln T + 2\ln(2/\delta)}{n_t(x)}} + \frac{C}{n_t(x)}\right) \\
\leq&P\left(\left|\sum_{s=1}^{n_t(x)} \frac{U_{x,s}}{n_t(x)} - \mu(x)\right| + \sum_{s=1}^{n_t(x)} \frac{|C_{x,s}|}{n_t(x)} \geq \sqrt{\frac{4\ln T + 2\ln(2/\delta)}{n_t(x)}} + \frac{C}{n_t(x)}\right) \\
\overset{\text{(i)}}{\leq}&P\left(\left|\sum_{s=1}^{n_t(x)} \frac{U_{x,s}}{n_t(x)} - \mu(x)\right| \geq \sqrt{\frac{4\ln T + 2\ln(2/\delta)}{n_t(x)}}\right) \leq 2\cdot\exp\left(-\frac{n_t(x)}{2} \times \frac{4\ln T + 2\ln(2/\delta)}{n_t(x)}\right) \\
&= \delta T^{-2},
\end{aligned}
$$

where the inequality (i) comes from the fact that the total corruption budget is at most C. Since there are at most t active arms by time t, by taking the union bound over all active arms it holds that,

$$
P\left(\forall \text{ active arm } x \text{ at round } t, |f_t(x)-\mu(x)| \leq r_t(x)\right) \geq 1-\delta T^{-1}, \quad \forall t \in [T].
$$

Finally, we take the union bound over all round $t \leq T$, and it holds that,

$$
P\left(\forall t \leq T, \forall \text{ active arm } x \text{ at round } t, |f_t(x)-\mu(x)| \leq r_t(x)\right) \geq 1-\delta T^{-1},
$$

which implies that the probability of a clean process is at least $1-\delta$. □

LEMMA A.1.1.2. *If it is a clean process and the optimal arm $x_* \in B(v, r_t(v))$, then $\mathcal{B}(v, r_t(v))$ could never be eliminated from Algorithm 1 for any $t \in [T]$ and active arm v at round t.*

PROOF. Recall that from Algorithm 1, at round t the ball $\mathcal{B}(u, r_t(u))$ would be discarded if we have for some active arm v s.t.

$$
f_t(v) - r_t(v) > f_t(u) + 2r_t(u).
$$

If $x_* \in \mathcal{B}(u, r_t(u))$, then it holds that

$$
f_t(u) + 2r_t(u) \overset{\text{(i)}}{\geq} \mu(u) + r_t(u) \geq \mu(u) + D(u, x_*) \overset{\text{(ii)}}{\geq} \mu(x_*),
$$

where inequality (i) is due to the clean process and inequality (ii) comes from the fact that $\mu(\cdot)$ is a Lipschitz function. On the other hand, we have that for any active arm v,

$$\mu(v) \geq f_t(v) - r_t(v), \quad \mu(x_*) \geq \mu(v).$$

Therefore, it naturally holds that

$$f_t(v) - r_t(v) \leq f_t(u) + 2r_t(u).$$

□

LEMMA A.1.1.3. *If it is a clean process, then for any time t and any (previously) active arm v we have $\Delta(v) \leq 3r_t(v)$. Furthermore, we could deduce that $D(u, v) \geq \min\{\Delta(u), \Delta(v)\}/3$ for any pair of (previously) active arms (u, v) by the time horizon T.*

PROOF. Let S_t be the set of all arms that are active or were once active at round t. Suppose an arm x_t is played at time t. If x_t is just played for one time, i.e. x_t is just activated at time t, then we naturally have that,

$$\Delta(x_t) \leq 1 \leq 3r_t(x_t),$$

since the diameter of $\mathcal{X}$ is at most 1. Otherwise, if x_t was played before, i.e. x_t is chosen based on the selection rule instead of the activation rule, we will claim that

$$\mu(x_*) \leq f_t(x_t) + 2r_t(x_t) \leq \mu(x_t) + 3r_t(x_t),$$

under a clean process. First we will show that $f_t(x_t) + 2r_t(x_t) \geq \mu(x_*)$. Recall that the optimal arm x_* is never eliminated according to A.1.1.2 under a clean process and hence is covered by some confidence ball, i.e. $x_* \in \mathcal{B}(x', r_t(x')), \exists x' \in S_t$. Then based on the selection rule, it holds that

$$f_t(x_t) + 2r_t(x_t) \geq f_t(x') + 2r_t(x') \geq \mu(x') + r_t(x') \geq \mu(x_*) + r_t(x') - D(x_*, x') \geq \mu(x_*).$$

On the other hand, it holds that,

$$f_t(x_t) + 2r_t(x_t) \leq \mu(x_t) + 3r_t(x_t)$$

since it is a clean process. And these two results directly imply that

$$\mu(x_*) - \mu(x_t) = \Delta(x_t) \leq 3r_t(x_t). \tag{A.1}$$

For the other active arms $v \in S_t$ that was played before time t, let $s < t$ be the last time arm v was played, where we have $f_t(v) = f_s(v)$ and $r_t(v) = r_s(v)$, and then based on Eqn. (A.1) it holds that $\Delta(v) \leq 3r_s(v) = 3r_t(v)$.

Furthermore, we will show that $D(u, v) \geq \min\{\Delta(u), \Delta(v)\}/3$ for any pair of active arms (u, v) by the time horizon T. W.l.o.g we assume that v was activated before u, and u was first activated at some time s'. Then if v was active at the time s' it naturally holds that $D(u, v) > r_{s'}(v) \geq \Delta(v)/3$ according to the activation rule. If v was removed at the time s' then we also have $D(u, v) > r_{s'}(v)$ since u was not among the discarded region, and hence $D(u, v) \geq \Delta(v)/3$ holds as well. And this concludes our proof. □

A.1.2. Proof of Theorem 2.4.1. We modify the original argument for Zooming algorithm (Kleinberg et al., 2019) to decently resolve the presence of adversarial corruptions. In summary, we could bound the cumulative regret of order $\tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right)$: the first term is the regret caused by the stochastic rewards, which is identical to the regret we have without any corruptions; the second quantity bounds the additional regret caused by the corruptions.

Denote S_T as the active (or previously active) arm set across the time horizon T. Then based on Lemma A.1.1.3, for any $x \in S_T$ it holds that,

$$\Delta(x) \leq 3r_T(x) = 3\sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n_T(x)}} + \frac{3C}{n_T(x)}.$$

And this indicates that

$$\Delta(x) n_T(x) \leq 3\sqrt{\left(4\ln(T) + 2\ln\left(\frac{2}{\delta}\right)\right) n_T(x)} + 3C. \tag{A.2}$$

Then we denote

$$B_{i,T} = \left\{ v \in S_T \,:\, 2^i \leq \frac{1}{\Delta(v)} < 2^{i+1} \right\}, \quad \text{where } S_T = \bigcup_{i=0}^{+\infty} B_{i,T},$$

and write $r_i = 2^{-i}$. Then for arbitrary $u, v \in B_{i,T}, i \geq 0$, we have

$$\frac{r_i}{2} < \Delta(u) \leq r_i, \quad \frac{r_i}{2} < \Delta(v) \leq r_i,$$

which implies that $D(x, y) > r_i/6$ under a clean process based on Lemma A.1.1.3. Based on the definition of the zooming dimension d_z, it follows that $|B_{i,T}| \leq O(r_i^{d_z})$. Subsequently, for any $0 < \rho < 1$ it holds that

$$\sum_{\substack{v \in S_T, \\ \Delta(v) > \rho}} 1 \leq \sum_{i < -\log_2(\rho)} O(r_i^{-d_z}) = O\left(\frac{1}{\rho^{d_z}}\right). \tag{A.3}$$

Now we define the set I as:

$$I \coloneqq \left\{ v \in S_T : C \leq \sqrt{\left(4\ln(T) + 2\ln\left(\frac{2}{\delta}\right)\right) n_T(v)} \right\}.$$

When an arm v is in the set I, the cumulative regret in terms of it would be more related to the stochastic errors other than the adversarial attacks. Subsequently, we could divide the cumulative regret into two quantities:

$$\begin{aligned}
Regret_T &= \sum_{v \in S_T} \Delta(v) n_T(v) = \sum_{v \in S_T \cap I} \Delta(v) n_T(v) + \sum_{v \in S_T \cap I^c} \Delta(v) n_T(v) \\
&= \sum_{\substack{v \in S_T \cap I, \\ \Delta(v) \leq \rho_1}} \Delta(v) n_T(v) + \sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} \Delta(v) n_T(v) + \sum_{\substack{v \in S_T \cap I^c, \\ \Delta(v) \leq \rho_2}} \Delta(v) n_T(v) + \sum_{\substack{v \in S_T \cap I^c, \\ \Delta(v) > \rho_2}} \Delta(v) n_T(v) \\
&\overset{\text{(i)}}{\leq} \rho_1 T + 2 \sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} 3\sqrt{\left(4\ln(T) + 2\ln\left(\frac{2}{\delta}\right)\right) n_T(v)} + \rho_2 T + 2 \sum_{\substack{v \in S_T \cap I^c, \\ \Delta(v) > \rho_2}} 3C \\
&\overset{\text{(ii)}}{\lesssim} \rho_1 T + \sqrt{\ln\left(\frac{T}{\delta}\right)} \sqrt{\left(\sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} n_T(v)\right)\left(\sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} 1\right)} + \rho_2 T + C \sum_{\substack{v \in S_T \cap I^c, \\ \Delta(v) > \rho_2}} 1 \\
&\lesssim \rho_1 T + \sqrt{\ln\left(\frac{T}{\delta}\right)} \sqrt{\left(\sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} n_T(v)\right)\left(\sum_{\substack{v \in S_T \cap I, \\ \Delta(v) > \rho_1}} 1\right)} + \rho_2 T + C \sum_{\substack{v \in S_T \cap I^c, \\ \Delta(v) > \rho_2}} 1 \\
&\overset{\text{(iii)}}{\lesssim} \rho_1 T + \sqrt{\ln\left(\frac{T}{\delta}\right)} \sqrt{T} \left(\frac{1}{\rho_1}\right)^{\frac{d_z}{2}} + \rho_2 T + C \left(\frac{1}{\rho_2}\right)^{d_z}.
\end{aligned} \tag{A.4}$$

The inequality (i) comes from the definition of set I and Eqn. (A.2), and inequality (ii) is due to the Cauchy-Schwarz inequality where $\lesssim$ denotes "less in order". Furthermore, we get inequality (iii) based on Eqn. (A.3). Note Eqn. (A.4) holds for arbitrary $\rho_1, \rho_2 \in (0,1)$, and hence by taking

$$\rho_1 = T^{-\frac{1}{d_z+2}} \ln(T)^{\frac{1}{d_z+2}}, \quad \rho_2 = T^{-\frac{1}{d_z+1}} C^{\frac{1}{d_z+1}},$$

we have

$$Regret_T = O\left(\ln(T)^{\frac{1}{d_z+2}} T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right) = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right).$$

And this concludes our proof. □

REMARK A.1.1. *Note we could replace the second term of $r(x)$ with $\min\{1, C/n(x)\}$, i.e.*

$$r_t(x) = \sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n_t(x)}} + \min\left\{1, \frac{C}{n_t(x)}\right\},$$

since we know each instance of attack is assumed to be upper bounded by 1. And all our analyses and Lemmas introduced above could be easily verified. Specifically, the core Lemma A.1.1.1 still holds as

$$\sum_{s=1}^{n_t(x)} \frac{C_{x,s}}{n_t(x)} \leq \sum_{s=1}^{n_t(x)} \frac{1}{n_t(x)} = 1.$$

A.2. Analysis of Theorem 2.4.4

A.2.1. Useful Lemmas. We first present some supportive Lemmas.

LEMMA A.2.0.1. *For a sequence of IID Bernoulli trials with a fix success probability p, then with probability $1-\delta$, we could at most observe $[(1-p)\ln(1/\delta)/p]$ failures until the first success.*

PROOF. This is based on the property of negative binomial distribution: after we complete the first N trials, the probability of no success is $(1-p)^N$. To ensure this value is less than δ, we get

$$N = \log_{1-p}(\delta) = \frac{\ln(1/\delta)}{\ln(1/(1-p))} = \frac{\ln(1/\delta)}{\ln(1+p/(1-p))}.$$

By using the inequality $\ln(x+1) \leq x, \forall x > -1$, we could take $N = [(1-p)\ln(1/\delta)/p]$. □

LEMMA A.2.0.2. (Adapted from Lemma 3.3 (Lykouris et al., 2018)) *In Algorithm 2, for any layer whose tolerance level exceeds the unknown C, i.e. any layer with index $i \in [l^*]$ s.t. $v_i \geq C$, with probability at least $1-\delta$, this layer suffers from at most corruptions of amount* $(\ln(1/\delta) + 2e - 1)$.

PROOF. The proof of this Lemma is an adaptation from the proof of Lemma 3.3 in Lykouris et al. (2018), and we present the detailed proof here for completeness:
In the beginning, we introduce an important result (Lemma 1 in Beygelzimer et al. (2011)): Let $X_1, \ldots, X_T$ be a real-valued martingale difference sequence, i.e. $\forall t \in [T], \mathbb{E}(X_t | X_{t-1}, \ldots, X_1) = 0$. And $X_t \leq R$. Denote $V = \sum_{t=1}^{T} \mathbb{E}(X_t^2 | X_{t-1}, \ldots, X_1)$. Then for any $\delta > 0$, it holds that,

$$P\left(\sum_{t=1}^{T} X_t > R\ln\left(\frac{1}{\delta}\right) + \frac{e-2}{R} \cdot V\right) \leq \delta.$$

Assume a layer whose tolerance level $\tilde{C}$ is no less than C, and hence the probability of pulling this layer would be $1/\tilde{C} \leq 1/C$. For this layer, let $\tilde{C}_x^t$ be the corruption that is observed at round t when arm x is pulled, $x \in X$. Then at any time t, if the adversary selects corruption $c_t(a)$ then we know $\tilde{C}_x^t$ is equal to $c_t(a)$ with probability $1/\tilde{C}$ and 0 otherwise. Denote the filtration $\tilde{\mathcal{F}}_t$ containing all the realizations of random variables before time t. And hence at time t the adversary could contaminate the stochastic rewards of $\mathcal{X}$ according to $\tilde{\mathcal{F}}_t$. Let $\tilde{a}_t$ be the arm that would be selected if this layer is chosen at the time t. Since our Algorithm 2 is deterministic in terms of the active region conditioned on selecting each layer, and the pulled arm is randomly selected from the active region. Therefore, the selection of $\tilde{a}_t$ is also independent with $\tilde{C}_x^t$ given $\tilde{\mathcal{F}}_t$. We construct the martingale as:

$$X_t = \left|\tilde{C}_x^t\right| - \mathbb{E}\left(\left|\tilde{C}_x^t\right| \mid \tilde{\mathcal{F}}_t\right).$$

Therefore, it holds that

$$\mathbb{E}(X_t^2 | X_{t-1}, \ldots, X_1) = \frac{1}{\tilde{C}}\left(|c_t(a)| - \frac{|c_t(a)|}{\tilde{C}}\right)^2 + \frac{\tilde{C}-1}{\tilde{C}}\left(\frac{|c_t(a)|}{\tilde{C}}\right)^2 \leq 2\frac{|c_t(a)|}{C},$$

since we have that $C \leq \tilde{C}$ and $|c_t(a)| \leq 1$. And conclusively it holds that

$$V = \sum_{t=1}^{T} \mathbb{E}(X_t^2 | X_{t-1}, \ldots, X_1) \leq \sum_{t=1}^{T} 2\frac{|c_t(a)|}{C} \leq 2.$$

Furthermore, it naturally holds that $X_t \leq 1$ due to the fact that $|c_t(a)| \leq 1$. Based on Lemma 1 in Beygelzimer et al. (2011) we introduced above, with probability at least $1-\delta$, it holds that

$$\sum_{t=1}^{T} X_t \leq \ln\left(\frac{1}{\delta}\right) + 2(e-2).$$

On the other hand, we can trivially deduce that the expected corruption injected in this layer is at most 1 since we have total amount of corruptions C and the probability of choosing this layer at each time is fixed as $1/\tilde{C} \leq 1/C$. Conclusively, we have with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\left|\tilde{C}_x^t\right| = \sum_{t=1}^{T} X_t + \mathbb{E}\left(\sum_{t=1}^{T}\left|\tilde{C}_x^t\right| \mid \tilde{\mathcal{F}}_t\right) \leq \ln\left(\frac{1}{\delta}\right) + 2(e-2) + 1 = \ln\left(\frac{1}{\delta}\right) + 2e - 1.$$

And this completes the proof. □

Definition A.2.1. *We call it a clean process for Algorithm 2, if for any time $t \in [T]$, any layer $l \in [l^*]$ whose tolerance level $v_l \geq C$, any active region $A \in \mathcal{A}_l$ and any $x \in A$ at time t, we have*

$$|f_{l,A} - \mu(x)| \leq \frac{1}{2^{m_l}} + \sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A}}$$

hold for some $0 < \delta < 1$.

To facilitate our analysis in the rest of this section, we expand notations here for Algorithm 2. Similar as in Appendix A.1, we would add the subscript time t to some notations used in Algorithm 2.

- $m_{l,t}$: epoch index of layer l at time t;
- $n_{l,t}$: number of selecting the layer l at time t since the last refresh (line 10 of Algorithm 2) on the layer l;
- $\mathcal{A}_{l,t}$: active arm set of layer l at time t;
- $n_{l,A,t}$: number of selecting the layer l and active region $A \in \mathcal{A}_{l,t}$ by time t since the last refresh on the layer l;
- $f_{l,A,t}$: average stochastic rewards of selecting the layer l and active region $A \in \mathcal{A}_{l,t}$ by time t since the last refresh on the layer l.

We also denote l_0 as the minimum index of layer whose tolerance level just surpasses C, i.e. $l_0 = \arg\min\{l \in [l^*] : v_l \geq C\}$. Therefore, we get a clean process defined in Definition A.2.1 *iff.*

the following set Φ holds:

$$\Phi = \Bigg\{ |f_{l,A,t} - \mu(x)| \leq \frac{1}{2^{m_l}} + \sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}} : \tag{A.5}$$
$$\forall x \in A,\, \forall A \in \mathcal{A}_{l,t},\, \forall l \in \{l_0, l_0+1, \ldots, l^*\},\, \forall t \in [T] \Bigg\}.$$

Note we only need to prove the set Φ holds at the end of each epoch for the analysis of Algorithm 2. W.l.o.g. we will just prove the regret bound in Theorem 2.4.4 of Algorithm 2.

COROLLARY A.2.1. *With probability at least* $1 - \frac{\delta}{4}$, *we select one time of layer* l_0 *at most every* $BC\log(4T/\delta)$ *times of other layers simultaneously.*

PROOF. The proof is straight forward based on Lemma A.2.0.1. According to the construction of $\{v_l\}_{l=1}^{l^*}$, it holds that $C \leq v_{l_0} < BC$. This implies that the probability of sampling layer l_0 at each round is at least $\frac{1}{BC}$. Therefore, after sampling layer l_0 in line 2 of Algorithm 2, with probability at least $1 - \frac{\delta}{4T}$, we would sample all the other layers for at most

$$BC\frac{\log(4T/\delta)}{1 - \frac{1}{BC}} \leq BC\log(4T/\delta)$$

times. Since we know the number of time sampling layer l_0 is naturally at most T, by taking the union bound, we conclude the proof of Corollary A.2.1. □

LEMMA A.2.1.1. *For algorithm 2, the probability of a clean process is at least* $1 - \frac{3}{4}\delta$, *i.e.* $P(\Phi) \geq 1 - \frac{3}{4}\delta$.

PROOF. For each layer l whose tolerance level surpasses C, i.e. $l \geq l_0$, we know the probability of sampling this layer in line 2 of Algorithm 2 is at most $1/C$, and this indicates that with probability at least $1 - \delta_1$, this layer suffers from at most $(-\ln(\delta_1) + 2e - 1)$ levels of corruptions based on Lemma A.2.0.2. Note the number of layers is less than $\log_B(T)$. This indicates that by taking the union bound on all layers whose tolerance levels surpass C, we have with probability at least $1 - \delta_1$, all these layers suffer from at most $\left(\ln\left(\frac{\log_B(T)}{\delta_1}\right) + 2e - 1\right)$ levels of corruptions across the time horizon T. And note

$$\ln\left(\frac{\log_B(T)}{\delta_1}\right) + 2e - 1 \leq \ln\left(\frac{T}{\delta_1}\right)$$

since it is natural to have $T/\log_B(T) \geq e^3$. Then for any time t, any layer $l \geq l_0$ and any active region $A \in \mathcal{A}_{l,t}$, define $x_{A,s}$, $C_{A,s}$ and random variables $U_{A,s}$ as the s-th time arm pulled, the stochastic reward from pulling $x_{A,s}$ and the corruption injected on $U_{A,s}$ for $1 \leq s \leq n_{l,A,t}$. Also denote

$$r_{l,A,t} = \frac{1}{2^{m_{l,t}}} + \sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}}.$$

With probability at least $1 - \delta/4$, from the above argument, we know that all layers with the index at least l_0 suffer from at most $\ln\left(\frac{4T}{\delta}\right)$ levels of corruptions across the time horizon T. Denote this event as Ψ, i.e. $P(\Psi) \geq 1 - \delta/4$, then under Ψ it holds that

$$\begin{aligned}
&P\left(|f_{l,A,t} - \mu(x)| \leq r_{l,A,t},\ \forall x \in A\right) \\
&= P\left(\left|\sum_{s=1}^{n_{l,A,t}} \frac{U_{x,s}}{n_{l,A,t}} + \sum_{s=1}^{n_{l,A,t}} \frac{C_{x,s}}{n_{l,A,t}} - \mu(x)\right| \leq r_{l,A,t},\ \forall x \in A\right) \\
&\geq P\left(\left|\sum_{s=1}^{n_{l,A,t}} \frac{U_{x,s}}{n_{l,A,t}} - \sum_{s=1}^{n_{l,A,t}} \frac{\mu(x_{A,s})}{n_{l,A,t}}\right| + \left|\sum_{s=1}^{n_{l,A,t}} \frac{\mu(x_{A,s})}{n_{l,A,t}} - \mu(x)\right| + \left|\sum_{s=1}^{n_{l,A,t}} \frac{C_{x,s}}{n_{l,A,t}}\right| \leq r_{l,A,t}, \forall x \in A\right) \\
&\overset{\text{(i)}}{\geq} P\left(\left|\sum_{s=1}^{n_{l,A,t}} \frac{U_{x,s}}{n_{l,A,t}} - \sum_{s=1}^{n_{l,A,t}} \frac{\mu(x_{A,s})}{n_{l,A,t}}\right| + \left|\sum_{s=1}^{n_{l,A,t}} \frac{\mu(x_{A,s})}{n_{l,A,t}} - \mu(x)\right| \leq \frac{1}{2^{m_{l,t}}} + \sqrt{\frac{2\ln(4T^2/\delta)}{n_{l,A,t}}}, \forall x \in A\right) \\
&\overset{\text{(ii)}}{\geq} P\left(\left|\sum_{s=1}^{n_{l,A,t}} \frac{U_{x,s}}{n_{l,A,t}} - \sum_{s=1}^{n_{l,A,t}} \frac{\mu(x_{A,s})}{n_{l,A,t}}\right| \leq \sqrt{\frac{2\ln(4T^2/\delta)}{n_{l,A,t}}}\right) \\
&\geq 1 - \frac{\delta}{2} \cdot T^{-2}.
\end{aligned}$$

Inequality (i) is due to the definition of event Ψ and inequality (ii) comes from the fact that the diameter of A is at most $1/2^{m_{l,t}}$ and $\mu(\cdot)$ is a Lipschitz function. We know that at most T active regions would be played across time T. By taking the union bound on all rounds $t \in [T]$ and all active regions that have been played, it holds that

$$P\left(|f_{l,A,t} - \mu(x)| \leq r_{l,A,t},\ \forall x \in A,\ \forall A \in \mathcal{A}_{l,t},\ \forall l \in \{l_0, l_0+1, \ldots, l^*\},\ \forall t \in [T]\right) \geq 1 - \frac{\delta}{2}$$

under the event Ψ. Since $P(\Psi) \geq 1 - \delta/4$, overall it holds that

$$P\left(|f_{l,A,t} - \mu(x)| \leq r_{l,A,t},\ \forall x \in A,\ \forall A \in \mathcal{A}_{l,t},\ \forall l \in \{l_0, l_0+1, \ldots, l^*\},\ \forall t \in [T]\right) \geq 1 - \frac{3\delta}{4},$$

i.e. $P(\Phi) \geq 1 - 3\delta/4$. And this concludes our proof. □

LEMMA A.2.1.2. *We have* $r_{l,A,t} \leq 2/2^{m_{l,t}}$ *if* $n_{l,A,t} = 6\ln(4T/\delta) \cdot 4^{m_{l,t}}$.

PROOF. Based on the formulation of $r_{l,A,t}$

$$r_{l,A,t} = \frac{1}{2^{m_{l,t}}} + \sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}}.$$

It suffices to show that

$$\sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}} \leq \frac{1}{2^{m_{l,t}}} \tag{A.6}$$

by taking $n_{l,A,t} = 6\ln(4T/\delta) \cdot 4^{m_{l,t}}$. Firstly, we have that

$$\begin{aligned}\sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} \leq 2\sqrt{\frac{\ln(T) + \ln(4/\delta)}{6\ln(4T/\delta) \cdot 4^{m_{l,t}}}} &\leq 2\sqrt{\frac{\ln(T) + \ln(4/\delta)}{(3+2\sqrt{2})\ln(4T/\delta) \cdot 4^{m_{l,t}}}} \\ &\leq (2\sqrt{2}-2)\frac{1}{2^{m_{l,t}}}\end{aligned}$$

Secondly, it holds that

$$\frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}} \leq \frac{1}{3+2\sqrt{2}}\frac{1}{4^{m_{l,t}}} \leq (3-2\sqrt{2})\frac{1}{2^{m_{l,t}}}.$$

Combining the above two results, we have Eqn. (A.6) holds, which concludes our proof. □

LEMMA A.2.1.3. *Under a clean process, for any layer* l *whose tolerance level* v_l *is no less than* C*, i.e.* $l \geq l_0$*, it holds that*

$$\Delta(x) \leq 16/2^{m_{l,t}}, \quad \forall x \in A, \forall A \in \mathcal{A}_{l,t}, \forall t \in [T].$$

PROOF. We will show that under a clean process Φ, the optimal arm x_* would never be eliminated from layers whose tolerance levels are no less than C. Obviously, the optimal arm x_* is in the covering when $m_{l,t} = 1$, where the whole arm space $\mathcal{X}$ is covered. Assume the layer l_t reaches the end of epoch $m_{l_t,t}$ at time t (i.e. $m_{l_t,t+1} = m_{l_t,t} + 1$), and the optimal arm x_* is contained in some active region $A_* \in A_{l_t,t}$. Then under a clean process, for any active region $A_0 \in A_{l_t,t}$ it holds that,

$$f_{l_t,A_*,t} \geq \mu(x_*) - r_{l_t,A_*,t} \geq \mu(x_*) - 2/2^{m_{l_t,t}} \tag{A.7}$$

$$f_{l_t,A_0,t} \leq \mu(x) + r_{l_t,A_0,t} \leq \mu(x) + 2/2^{m_{l_t,t}}, \forall x \in A_0 \tag{A.8}$$

based on Lemma A.2.1.2 since we have $n_{l_t,A,t} = 6\ln(4T/\delta)\cdot 4^{m_{l_t,t}}, \forall A \in A_{l_t,t}$ in the end of the epoch. And since $\mu(x_*) \geq \mu(x), \forall x \in A_0$, it holds that

$$f_{l_t,A_0,t} - f_{l_t,A_*,t} \leq 4/2^{m_{l_t,t}}. \tag{A.9}$$

This implies that A_* will not be removed. Note the above argument holds for any epoch index and any layer whose corruption level surpasses C, and hence the optimal arm x_* would never be eliminated from layers whose tolerance levels are no less than C.

To prove Lemma A.2.1.3. When $m_{l,t} = 1$, it naturally holds since $\Delta(x) \leq 1 \leq 16/2^1$. Otherwise, let A_* be the covering that contains the optimal arm x_* for layer l in the previous epoch $m_{l,t} - 1$, and according to the above argument it is well defined. And we know x is also alive in the previous epoch, where we denote A_x as the covering that contains x in the previous epoch $m_{l,t} - 1$. Denote t_0 as the time the last epoch reaches the end of layer l ($m_{l,t} - 1 = m_{l,t_0}$), and then it holds that

$$\Delta(x) \leq f_{l,A_*,t_0} - f_{l,A_x,t_0} + 2r_{l,A_*,t_0} = f_{l,A_*,t_0} - f_{l,A_x,t_0} + \frac{4}{2^{m_{l,t_0}}} = f_{l,A_*,t_0} - f_{l,A_x,t_0} + \frac{8}{2^{m_{l,t}}}$$

since $r_{l,A_*,t_0} = r_{l,A_x,t_0} = 4/2^{m_{l,t_0}}$ at the end of the epoch m_{l,t_0}. On the other hand, since A_x was not eliminated at the end of the epoch m_{l,t_0}, based on the same argument used with Eqn. (A.7), (A.8), (A.9), we have that

$$f_{l,A_*,t_0} - f_{l,A_x,t_0} \leq \frac{4}{2^{m_{l,t_0}}} = \frac{8}{2^{m_{l,t}}},$$

and this fact indicates that

$$\Delta(x) \leq \frac{16}{2^{m_{l,t}}}.$$

Note this result holds for any layer whose tolerance level surpasses C and any $t \in [T]$. This implies Lemma A.2.1.3 holds conclusively. □

A.2.2. Proof of Theorem 2.4.4.

Proof. If the corruption budget $C \leq \ln(4T/\delta)$, then all the layers' tolerance levels exceed the unknown C, in which case based on Lemma A.2.1.1, with probability at least $1 - 3\delta/4$, it holds that $\forall x \in A, \forall A \in \mathcal{A}_{l,t}$, $\forall l \in [l^*]$, $\forall t \in [T]$

$$|f_{l,A,t} - \mu(x)| \leq \frac{1}{2^{m_l}} + \sqrt{\frac{4\ln(T) + 2\ln(4/\delta)}{n_{l,A,t}}} + \frac{\ln(T) + \ln(4/\delta)}{n_{l,A,t}}.$$

We denote $Regret_T(l)$ as the cumulative regret encountered from the layer l across time T, which implies that

$$Regret_T = \sum_{l=1}^{l^*} Regret_T(l).$$

For any fixed layer $l \in [l^*]$, we will then show that $Regret_T(l) = \tilde{O}(T^{\frac{d_z+1}{d_z+2}})$. Based on Lemma A.2.1.3, we know that for any layer l, any arm played after the epoch m would at most incurs a regret of volume $16/2^m$. Note at epoch m, the active arm set consists of $1/2^m$-coverings for some region in $\{x \in \mathcal{X} : \Delta(x) \leq 16/2^m\}$. Therefore, the number of active regions at this epoch m could be upper bounded by $\alpha 2^{d_z m}$ for some constant $\alpha > 0$. And for each active region, we will pull it for exactly $6\ln(4T/\delta) \cdot 4^m$ times in epoch m. Therefore, the total regret incurred in the epoch m for any layer would at most be

$$\alpha 2^{d_z m} \times 6\ln(4T/\delta) \cdot 4^m \times 16/2^m = 192\alpha \ln(4T/\delta) 2^{(d_z+1)m}.$$

Therefore, the total cumulative regret we experience for any layer l could be upper bounded as:

$$\begin{aligned} Regret_T(l) &\leq \sum_{m=1}^{M} 192\alpha \ln\left(\frac{4T}{\delta}\right) 2^{(d_z+1)m} + \frac{8}{2^M}T \\ &\leq 192\alpha \ln\left(\frac{4T}{\delta}\right) \frac{2^{(d_z+1)(M+1)} - 2^{d_z+1}}{2^{d_z+1}-1} + \frac{8}{2^M}T \\ &\leq 384\alpha \ln\left(\frac{4T}{\delta}\right) 2^{(d_z+1)M} + \frac{8}{2^M}T, \end{aligned}$$

where the second term bound the total regret after finishing the epoch M. Note we could take M as any integer here, even if the epoch M doesn't exist, our bound still works. By taking M as the closest integer to the value $\left(\ln\left(\frac{T}{48\alpha\ln(4T/\delta)}\right)/\left[(d_z+2)\ln(2)\right]\right)$. It holds that

$$Regret_T(l) \lesssim T^{\frac{d_z+1}{d_z+2}} \ln(T/\delta)^{\frac{1}{d_z+2}}, \qquad \forall l \in [l^*].$$

Therefore, it holds that with probability at least $1 - 3\delta/4 \geq 1 - \delta$,

$$Regret_T = \sum_{l=1}^{l^*} Regret_T(l) \lesssim T^{\frac{d_z+1}{d_z+2}} \ln(T/\delta)^{\frac{1}{d_z+2}} \cdot \log_2(T) = \tilde{O}(T^{\frac{d_z+1}{d_z+2}}).$$

On the other hand, if the corruption budget $C > \ln(4T/\delta)$, then not all the layers could tolerate the unknown total budget level C. We denote l_0 as the minimum index of the layer that is resilient

to C as defined in Eqn. A.5. Therefore, we could use the above argument to similarly deduce that:

$$\sum_{l=l_0}^{l^*} Regret_T(l) \lesssim T^{\frac{d_z+1}{d_z+2}} \ln(T/\delta)^{\frac{1}{d_z+2}} \cdot \log_2(T) = \tilde{O}(T^{\frac{d_z+1}{d_z+2}}). \tag{A.10}$$

For the first $(l_0 - 1)$ layers that are vulnerable to attacks, we could control their regret by using the cross-layer region elimination idea. Specifically, it holds that $v_{l_0} \leq BC$, then based on Corollary A.2.1, we know that with probability at least $1 - \delta/4$, we select one time of layer l_0 at most every $BC\log(4T/\delta)$ times of the first $(l_0 - 1)$ non-robust layers. Since the active regions in a lower-index layer are always a subset of the active regions for the layer with a higher index according to our cross-layer elimination rule in Algorithm 2. We know when the layer l_0 stays at the epoch m, any arm played in the layer $1, 2, \ldots, l_0$ would at most incurs a regret $16/2^m$. Therefore, when the layer l_0 stays in epoch m, we have probability at least $1 - 3\delta/4 - \delta/4 = 1 - \delta$, the total regret incurred from the first l_0 layers altogether could be bounded as

$$BC\log(4T/\delta) \times \alpha 2^{d_z m} \times 6\ln(4T/\delta) \cdot 4^m \times 16/2^m = 192BC\alpha\ln(4T/\delta)^2 2^{(d_z+1)m}.$$

Conclusively, it holds that

$$\begin{aligned}
\sum_{l=1}^{l_0} Regret_T(l) &\leq \sum_{m=1}^{M} 192BC\alpha\ln\left(\frac{4T}{\delta}\right)^2 2^{(d_z+1)m} + \frac{8}{2^M}T \\
&\leq 192BC\alpha\ln\left(\frac{4T}{\delta}\right)^2 \frac{2^{(d_z+1)(M+1)} - 2^{d_z+1}}{2^{d_z+1} - 1} + \frac{8}{2^M}T \\
&\leq 384BC\alpha\ln\left(\frac{4T}{\delta}\right)^2 2^{(d_z+1)M} + \frac{8}{2^M}T,
\end{aligned}$$

for arbitrary M. Similarly, then we can simply take M as the closest positive integer to the value

$$\left(\ln\left(\frac{T}{48\alpha BC\ln(4T/\delta)}\right)/\left[(d_z+2)\ln(2)\right]\right),$$

and we have that

$$\sum_{l=1}^{l_0} Regret_T(l) \lesssim T^{\frac{d_z+1}{d_z+2}} \left(BC\ln(T/\delta)^2\right)^{\frac{1}{d_z+2}}. \tag{A.11}$$

Combine the results from Eqn. A.10 and Eqn. A.11, with probability at least $1 - \delta$, it holds that

$$Regret_T = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}}\left(B^{\frac{1}{d_z+2}}C^{\frac{1}{d_z+2}} + 1\right)\right) = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}}\left(C^{\frac{1}{d_z+2}} + 1\right)\right).$$

And this completes our proof. □

A.3. Analysis of Theorem 2.4.6

A.3.1. Useful Lemmas.

LEMMA A.3.0.1. (Part of Theorem 3.2 and 5.3 in Pacchiano et al. (2020)) *If the regret of the optimal base algorithm could be bounded by* $U_*(T,\delta) = O(c(\delta)T^\alpha)$ *for some function* $c : \mathbb{R} \to \mathbb{R}$ *and constant* $\alpha \in [1/2, 1)$*, the regret of EXP.P and CORRAL with smoothing transformation as the master algorithms are shown in Table A.1:*

TABLE A.1. Table for Lemma A.3.0.1

	Known α*, Unknown* $c(\delta)$
EXP3.P	$\tilde{O}\left(T^{\frac{1}{2-\alpha}}c(\delta)\right)$
CORRAL	$\tilde{O}\left(T^{\alpha}c(\delta)^{\frac{1}{\alpha}}\right)$

The proof of this Lemma involves lots of technical details and is presented in Pacchiano et al. (2020) elaborately. And hence we would omit the proof here.

A.3.2. Proof of Theorem 2.4.6.

PROOF. The proof of our Theorem 2.4.6 is based on the above Lemma A.3.0.1. According to Theorem 2.4.1, with probability at least $1/\delta$, the regret bound of our Algorithm 1 could be bounded as

$$Regret_T = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}}\right) = \tilde{O}\left(T^{\frac{d_z+1}{d_z+2}} + C^{\frac{1}{d_z+2}}T^{\frac{d_z+1}{d_z+2}}\right).$$

Due to the fact that d_z is upper bounded by d and $C = O(T)$, it further holds that

$$Regret_T = \tilde{O}\left(\left(C^{\frac{1}{d_z+2}}+1\right)T^{\frac{d_z+1}{d_z+2}}\right) = \tilde{O}\left(\left(C^{\frac{1}{d+2}}+1\right)T^{\frac{d+1}{d+2}}\right).$$

Therefore, by taking $\alpha = \frac{d+1}{d+2}$ (known) and $c(\delta) = \left(C^{\frac{1}{d+2}}+1\right)$ (unknown) and plugging them into Lemma A.3.0.1, we have that

$$\mathbb{E}(Regret_T) = \begin{cases} \tilde{O}\left((C^{\frac{1}{d+2}}+1)T^{\frac{d+2}{d+3}}\right) & \text{EXP3.P,} \\ \tilde{O}\left((C^{\frac{1}{d+1}}+1)T^{\frac{d+1}{d+2}}\right) & \text{CORRAL.} \end{cases}$$

And this concludes our proof. □

A.4. Analysis of Theorem 2.4.7

Under the assumption that the diameter of $\mathcal{X}$ is at most 1, we could also assume that $\mu(x) \in [0,1], \forall x \in X$ due to the Lipschitzness of $\mu(\cdot)$ w.l.o.g. in this section.

A.4.1. Useful Lemmas.

LEMMA A.4.0.1. *In Algorithm 10, for any batch $i \in \left[\left\lceil \frac{T}{H} \right\rceil\right]$, the sum of stochastic rewards could be bounded as*

$$\left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} y_t \right| \leq 2H + \sqrt{2H \log(\frac{12T}{H\delta})}$$

simultaneously with probability at least $1 - \delta/3$.

PROOF. For arbitrary batch index $i \in \left[\left\lceil \frac{T}{H} \right\rceil\right]$, it holds that

$$\begin{aligned}
&P\left(\left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} y_t \right| \geq 2H + \sqrt{2H \log(\frac{12T}{H\delta})} \right) \\
=&P\left(\left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(x_t) + c_t(x_t) + \eta_t \right| \geq 2H + \sqrt{2H \log(\frac{12T}{H\delta})} \right) \\
\leq&P\left(\sum_{t=(i-1)H+1}^{\min\{iH,T\}} |\mu(x_t)| + \sum_{t=(i-1)H+1}^{\min\{iH,T\}} |c_t(x_t)| + \left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \eta_t \right| \geq 2H + \sqrt{2H \log(\frac{12T}{H\delta})} \right) \\
\leq&P\left(\left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \eta_t \right| \geq \sqrt{2H \log(\frac{12T}{H\delta})} \right) \overset{\text{(i)}}{\leq} \frac{H}{6T}\delta.
\end{aligned}$$

The inequality (i) comes from the fact that $\sum_{t=(i-1)H+1}^{iH} \eta_t$ is sub-Gaussian with parameter H. Therefore, by taking a union bound on all $\left\lceil \frac{T}{H} \right\rceil$ batches, it holds that

$$P\left(\forall i \in \left\lceil \frac{T}{H} \right\rceil : \left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} y_t \right| \leq 2H + \sqrt{2H \log(\frac{12T}{H\delta})} \right) \geq 1 - \frac{\delta}{3}.$$

And this concludes the proof of Lemma A.4.0.1. □

A.4.2. Proof of Theorem 2.4.7.

PROOF. We are ready to prove Theorem 2.4.7 now. Since we have $\lceil \log_2(T) \rceil$ base algorithms where the i-th base algorithm is our Robust Zooming Algorithm (Algorithm 1) with tolerance level

2^i, we can denote the base algorithm set as $W = \{2^i\}_{i=1}^{\lceil \log_2(T) \rceil}$ in terms of their tolerance levels. For any round $t \in [T]$, let w_t denote the base algorithm chosen from W. And denote $x_t(w)$, $w \in W$ as the arm pulled if the base algorithm w is chosen in the beginning of its batch. In other words, we have $x_t = x_t(w_t)$. Denote C_i as the total budget of corruptions in the i-th batch and hence $C = \sum_{i=1}^{\lceil T/H \rceil} C_i$, where recall that C is the unknown total budget of corruptions. And we also write $C_* = \max_i C_i$ as the maximum budget in a single batch. Let w_* be the element in W such that $C_* \leq w_* < 2C_*$. Therefore, we could decompose the cumulative regret into the following two quantities:

$$Regret_T = \underbrace{\sum_{t=1}^{T} (\mu(x_*) - \mu(x_t(w_*)))}_{\text{Quantity (I)}} + \underbrace{\sum_{t=1}^{T} (\mu(x_t(w_*)) - \mu(x_t(w_t)))}_{\text{Quantity (II)}} \tag{A.12}$$

And it suffices to bound these two quantities respectively. We know the Quantity (I) could be further represented as

$$\sum_{t=1}^{T} (\mu(x_*) - \mu(x_t(w_*))) = \sum_{i=1}^{\lceil \frac{T}{H} \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} (\mu(x_*) - \mu(x_t(w_*))) .$$

Here we will use the results from Theorem 2.4.1. Note by setting the probability rate as $\delta/3$ in Algorithm 1, we can prove that we have a clean process with probability at least $1 - \delta/3$ (line 5 in Algorithm 10). Although we run the Algorithm 1 here in a batch fashion and the total rounds is T, we can still easily show that with probability at least $1 - \delta/3$ we have a clean process for all batches. This is because the proof of Lemma A.1.1.1 only relies on taking a union bound over all rounds T where whether a restart is proceeded doesn't matter at all. According to Theorem 2.4.1 and the choice of w_*, the cumulative regret of each batch could be upper bounded by the order of

$$\tilde{O}\left(H^{\frac{d_z+1}{d_z+2}} + C_*^{\frac{1}{d_z+1}} H^{\frac{d_z}{d_z+1}}\right) = \tilde{O}\left(H^{\frac{d_z+1}{d_z+2}} + C_*^{\frac{1}{d+1}} H^{\frac{d}{d+1}}\right),$$

since $C_* \leq H$ naturally holds by definition. Therefore, it holds that

$$\text{Quantity (I)} = \tilde{O}\left(\left\lceil \frac{T}{H} \right\rceil \left(H^{\frac{d_z+1}{d_z+2}} + C_*^{\frac{1}{d+1}} H^{\frac{d}{d+1}}\right)\right) = \tilde{O}\left(TH^{\frac{-1}{d_z+2}} + TC_*^{\frac{1}{d+1}} H^{\frac{-1}{d+1}}\right), \tag{A.13}$$

with probability at least $1-\delta/3$. For Quantity (II), according to Lemma A.4.0.1, for any batch $i \in \left[\left\lceil \frac{T}{H} \right\rceil\right]$ the sum of stochastic rewards could be bounded by

$$\left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} y_t \right| \leq 2H + \sqrt{2H \log\left(\frac{12T}{H\delta}\right)}$$

simultaneously with probability at least $1-\delta/3$. We denote the event Ω as

$$\Omega = \left\{ \forall i \in \left\lceil \frac{T}{H} \right\rceil : \left| \sum_{t=(i-1)H+1}^{\min\{iH,T\}} y_t \right| \leq 2H + \sqrt{2H \log\left(\frac{12T}{H\delta}\right)} \right\},$$

and it holds that $P(\Omega) \geq 1-\delta/3$. And under the event Ω, from Theorem 6.3 in Auer et al. (2002b), we know with probability at least $1-\delta/3$, it holds that

$$\text{Quantity (II)} = \tilde{O}\left(\sqrt{H^2 \frac{T}{H}}\right) = \tilde{O}\left(\sqrt{TH}\right). \tag{A.14}$$

Specifically, in the statement of Theorem 6.3 (Auer et al., 2002b), we have $K = \lceil \log_2(T) \rceil, \delta = \delta/3, T = \lceil \frac{T}{H} \rceil$ here. And we multiply the regret bound in Theorem 6.3 (Auer et al., 2002b) by $\left(2H + \sqrt{2H \log(\frac{12T}{H\delta})}\right)$ as well since the original EXP3.P algorithm requires the magnitude of rewards not exceeding 1. Conclusively, by combining the results from Eqn. A.13 and Eqn. A.14 and taking a union bound on the probability rates, with probability at least $1-\delta/3-\delta/3-\delta/3 = 1-\delta$, we have that

$$Regret_T = \tilde{O}\left(TH^{\frac{-1}{d_z+2}} + TC_*^{\frac{1}{d+1}} H^{\frac{-1}{d+1}} + \sqrt{TH}\right).$$

By taking $H = T^{\frac{d+2}{d+4}}$ and using the fact that $d_z \leq d$, it holds that

$$\begin{aligned} Regret_T = \tilde{O}\left(T^{\frac{d+3}{d+4}} + C_*^{\frac{1}{d+1}} T^{\frac{d^2+4d+2}{(d+1)(d+4)}}\right) &= \tilde{O}\left(T^{\frac{d+3}{d+4}} + C_*^{\frac{1}{d+1}} T^{\frac{d+2}{d+3}}\right) \\ &= \tilde{O}\left(T^{\frac{d+3}{d+4}} + C^{\frac{1}{d+1}} T^{\frac{d+2}{d+3}}\right), \end{aligned}$$

with probability at least $1-\delta$. □

A.5. Additional Algorithms

A.5.1. BoB Robust Zooming Algorithm. Due to the space limit, we defer the pseudocode of BoB Robust Zooming algorithm here in Algorithm 10. We can observe that the top layer is an EXP3.P algorithm, which chooses the corruption level used for Robust Zooming algorithm in each

Algorithm 10 BoB Robust Zooming Algorithm

Input: Arm metric space $(\mathcal{X}, D)$, time horizon T, probability rate δ, batch size H.

1: Budget set for base algorithms $I = \{2^i\}_{i=1}^N$, $N = \lceil \log_2(T) \rceil$, $\alpha = 2\sqrt{\ln(\frac{3NT}{\delta})}$, $\gamma = \min\left\{\frac{3}{5}, 2\sqrt{\frac{3N\ln(N)}{5T}}\right\}$, weight $w_i = 1, i \in [N]$, cumulative sum $s = 0$.

2: **for** $t = 1$ **to** T **do**

3: **if** $t \in \{kH + 1 : k \in \mathbb{N}\}$ **then**

4: For $i = 1, \ldots, N$ set

$$p_i = (1-\gamma)\frac{w_i}{\sum_{j=1}^N w_j} + \frac{\gamma}{N}.$$

5: Choose the base algorithm index i' randomly with probability $\{p_i\}_{i=1}^N$.

6: Refresh the chosen Robust Zooming algorithm (Algorithm 1 with $C = 2^{i'}$) with active arm set $J = \{\}$, active space $\mathcal{X}_{act} = \mathcal{X}$ and probability rate $\delta/3$.

7: Run the chosen Robust Zooming algorithm and receive the reward y_t.

8: Update the chosen Robust Zooming algorithm according to Algorithm 1 and set $s = s + y_t$.

9: **if** $t \in \{kH : k \in \mathbb{N}^+\}$ **then**

10: Let $s = s / \left[p_{i'}\left(2H + \sqrt{2H\log(\frac{12T}{H\delta})}\right)\right]$.

11: Update EXP3.P component for index i': $w_{i'} = w_{i'} \exp\left(\frac{\gamma}{3N}\left(s + \frac{\alpha}{p_{i'}\sqrt{NT}}\right)\right)$.

batch adaptively. For each batch, we run our Robust Zooming algorithm with the chosen corruption level from the top layer, and use the accumulative rewards collected in each batch to update the components of EXP3.P (i.e. line 10 of Algorithm 10). Note we normalize the cumulative reward by dividing it with $(2H+\sqrt{2H\log(\frac{12T}{H\delta})})$, and this is because that we could prove that the magnitude of the cumulative reward at each batch would be at most $(2H + \sqrt{2H\log(\frac{12T}{H\delta})})$ with high probability as shown in Lemma A.4.0.1. And the EXP3.P algorithm (Auer et al., 2002b) requires the magnitude of reward should at most be 1 with our chosen values of α and γ. The regret bound of Algorithm 10 is given in Theorem 2.4.7 of our main paper.

A.6. Discussion on Lower Bounds

We now propose Theorem 2.4.2 and Theorem 2.4.5 with their detailed proof in Section A.6.1 and Section A.6.2 respectively, where we provide a pair of lower bounds for the strong adversary and the weak adversary.

A.6.1. Lower Bound for Strong Adversaries. We repeat our Theorem 2.4.2 for reference here and then provide a detailed proof as follows:

Theorem 2.4.2 *Under the strong adversary with corruption budget C, for any zooming dimension $d_z \in \mathbb{Z}^+$, there exists an instance such that any algorithm (even is aware of C) must suffer from the regret of order $\Omega\left(C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}}\right)$ with probability at least 0.5.*

PROOF. Here we consider the metric space $([0,1)^d, l_\infty)$. For arbitrary $\epsilon \in (0, \frac{1}{2})$, we can equally divide the space $[0,1]^d$ into $1/\epsilon^d$ small l_∞ balls whose diameters are equal to ϵ by discretizing each axis. (W.l.o.g we assume 1 is divisible by ϵ for simplicity since otherwise we could take $\lfloor 1/\epsilon^d \rfloor$ instead.) For example, if $d = 2$ and $\epsilon = \frac{1}{2}$, then we can divide the space into $2^2 = 4$ l_∞ balls: $[0, 0.5)^2, [0, 0.5) \times [0.5, 1), [0.5, 1) \times [0, 0.5), [0.5, 1)^2$. We denote these balls as $\{A_i\}_{i=1}^{1/\epsilon^d}, [0,1)^d = \cup_{i=1}^{1/\epsilon^d} A_i$ and their centers as $\{c_i\}_{i=1}^{1/\epsilon^d}$. (e.g. the center of $[0, 0.5)^2$ is $(0.25, 0.25)$.) Subsequently, we could define a set of functions $\{f_i(\cdot)\}_{i=1}^{1/\epsilon^d}$ as

$$f_i(x) = \begin{cases} \frac{\epsilon}{2} - \|x - c_i\|_\infty, & x \in A_i; \\ 0, & x \notin A_i. \end{cases}$$

We can easily verify that $f_i(\cdot)$ is a 1-Lipschitz function. For the zooming dimension, if ϵ is of constant scale, then the zooming dimension will become 0. However, in our analysis here, we would let ϵ rely on T and be sufficiently small so that the zooming dimension is d. If the underlying expected reward function is $f_k(\cdot)$ and there is no random noise, consider the strong adversary that shifts the reward of the arm down to whenever the pulled arm is in A_k and doesn't attack the reward otherwise. This attack could be done for roughly $\lfloor C/\epsilon \rfloor$ times. Intuitively, the learner can do no better than pull each arm in $[0,1]^d$ uniformly. This implies that roughly the learner should do $\lfloor C/\epsilon \rfloor \lfloor 1/\epsilon^d \rfloor$ rounds of uniform exploration before the attack budget C is used up, where the learner pulls arms outside A_k for approximately $\lfloor C/\epsilon \rfloor \cdot \lfloor (1-\epsilon^d)/\epsilon^d \rfloor$ times. Take $\epsilon = \left(\frac{C}{T}\right)^{\frac{1}{d+1}}$, we know that roughly the learner should do $\lfloor C/\epsilon \rfloor \lfloor 1/\epsilon^d \rfloor = T$ rounds of uniform exploration, and the cumulative regret is at least

$$\left\lfloor \frac{C}{\epsilon} \right\rfloor \cdot \left\lfloor \frac{(1-\epsilon^d)}{\epsilon^d} \right\rfloor \cdot \epsilon = \Theta\left(C^{\frac{1}{d+1}}T^{\frac{d}{d+1}}\right) = \Theta\left(C^{\frac{1}{d_z+1}}T^{\frac{d_z}{d_z+1}}\right).$$

For a more rigorous argument, note that for the k-th instance $f_k(\cdot)$, the adversary could maliciously replace the reward with 0 until the arm in A_k is pulled at least $\lfloor C/\epsilon \rfloor$ times. After $\lfloor C/\epsilon \rfloor \lfloor 1/2\epsilon^d \rfloor$ rounds, for any algorithm even with the information of value C, there must be at least $\lfloor 1/(2\epsilon^d) \rfloor$ balls among $\{A_i\}_{i=1}^{1/\epsilon^d}$ that have been pulled for at most $\lfloor C/\epsilon \rfloor$ times. As a consequence, when we

choose the problem instance k among these $\lfloor 1/(2\epsilon^d)\rfloor$ balls and set $\epsilon = \left(\frac{C}{T}\right)^{\frac{1}{d+1}}$, then we know that the regret of order

$$\epsilon \cdot \left\lfloor \frac{C}{\epsilon} \right\rfloor \cdot \left(\left\lfloor \frac{1}{2\epsilon^d} \right\rfloor - 1 \right) = \Theta\left(C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}} \right)$$

is unavoidable. This implies that the regret could be no worse than $\Omega(C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}})$ under the strong adversary with probability 0.5. □

For the stochastic Lipschitz bandit problem, based on Slivkins (2011) we know for any algorithm there exists one problem instance such that the expected regret is at least

$$\inf_{r_0 \in (0,1)} \left(r_0 T + C \log(T) \sum_{r=2^{-i}: i\in\mathbb{N}, r\geq r_0} \frac{N_z(r)}{r} \right),$$

where $N_z(r)$ is the zooming number. And hence the corruption-free lower bound $O\left(\ln(T)^{\frac{1}{d_z+2}} T^{\frac{d_z+1}{d_z+2}}\right)$ is optimal in terms of the zooming dimension d_z. Combining this result with our Theorem 2.4.2, we can conclude that for any algorithm, there exists a corrupted bandit instance where the algorithm must incur $\Omega\left(\max\left\{\ln(T)^{\frac{1}{d_z+2}} T^{\frac{d_z+1}{d_z+2}}, C^{\frac{1}{d_z+1}} T^{\frac{d_z}{d_z+1}}\right\}\right)$ cumulative regret, which coincides with the order of regret for our Robust Zooming algorithm. Conclusively, our algorithm obtains the optimal order of regret under the strong adversary.

We then restate our Theorem 2.4.3 for reference and then provide a detailed proof:

Theorem 2.4.3 *For any algorithm, when there is no corruption, we denote R_T^0 as the upper bound of cumulative regret in T rounds under our problem setting described in Section 2.3, i.e. $Regret_T \leq R_T^0$ with high probability, and it holds that $R_T^0 = o(T)$. Then under the strong adversary and unknown attacking budget C, there exists a problem instance on which this algorithm will incur linear regret $\Omega(T)$ with probability at least 0.5, if $C = \Omega(R_T^0/4^{d_z}) = \Omega(R_T^0)$.*

Proof. For the case that $d_z = 0$, we consider the metric space $([0,1], l_2)$ and define the Lipschitz function $f_1(\cdot)$ as □

$$f_1(x) = \begin{cases} 0.25 - |x - 0.25|, & x \in [0, 0.5]; \\ 0, & x \in (0.5, 1]. \end{cases},$$

and we assume there is no random noise and no adversarial corruption. (We call this instance I_0.) For any algorithm with $\mathbb{E}(Regret_T) \leq R_T^0$ when there is no adversarial corruption, we know that

$$\mathbb{E}(\#\text{ iterations playing arms in } (0.5,1]) \times 0.25 \leq \mathbb{E}(Regret_T) \leq R_T^0,$$

and hence $\mathbb{E}(\#\text{ iterations playing arms in } (0.5,1]) \leq 4R_T^0$. By Markov inequality, with probability at least 0.5, the number of iterations that play arms in $(0.5,1]$ is no more than $8R_T^0$.
Next, we define a new problem setting in the same metric space as:

$$f_2(x) = \begin{cases} 0.25 - |x - 0.25|, & x \in [0, 0.5]; \\ x - 0.5, & x \in (0.5, 1]. \end{cases}.$$

And under the setting of $f_2(\cdot)$ there is a malicious strong adversary with budget $C = 4R_T^0$ to attack using the following strategy: whenever the arm in $(0.5,1]$ is selected and the corruption budget has not been used up, the adversary moves the reward to 0. We call this instance I_1. Therefore, before the budget is used up, each selection of arm in $(0.5,1]$ returns a reward 0, and hence the agent can never tell the difference between I_0 and I_1 and would follow the same strategy under I_0 until the total corruption level reaches $C = 4R_T^0$ and then the adversary stops to contaminate the rewards. And this requires at least $C/0.5 = 2C = 8R_T^0$ rounds in which the agent chooses arms in $(0.5,1]$. Therefore, with probability of at least 0.5, the regret in T rounds is at least $(T - 8R_T^0)/4 = \Omega(T)$.
For $d_z > 0$, we use the metric space $([0,1]^d, \|\cdot\|_\infty)$ with $d = \lceil 2d_z \rceil$. We first partition the d-dimensional cube $[0,1]^d$ into 2^d sub-cubes with side length 0.5, i.e. equally divide the cube $[0,1]^d$ into 0.5-radius l_∞ balls whose diameters are equal to 0.5 by discretizing each axis. We denote these balls as $A_{i\,i=1}^{2^d}$ and the center of these balls as $c_{i\,i=1}^{2^d}$, e.g. $c_1 = [0.25]^d$. And we denote the vertex of each ball that matches the vertexes of $[0,1]^d$ as $v_{i\,i=1}^{2^d}$, e.g. $v_1 = [0]^d$. Subsequently, we could define the function $f_1(\cdot)$ as

$$f_1(x) = \begin{cases} 4^{\frac{-d}{d-d_z}} - \|x - c_1\|_\infty^{\frac{d}{d-d_z}}, & x \in A_1; \\ 0, & x \notin A_1. \end{cases}$$

and we assume there is no random noise and no adversarial corruption. (We call this instance I_0.) Since the regret of the algorithm under no corruption satisfies that $\mathbb{E}(Regret_T) \leq R_T^0$, and we know

that pulling any arm outside A_1 will incur a single regret of $4^{\frac{-d}{d-d_z}}$, and hence we have that

$$\mathbb{E}(\#\text{ iterations playing arms not in } A_1) \leq R_T^0 \cdot 4^{\frac{-d}{d-d_z}}.$$

Then by the pigeonhole principle, there exists a sub-ball $2 \leq i \leq 2^d$ such that the expected number of iterations to pull arms in A_i is no more than $R_T^0 \cdot 4^{\frac{-d}{d-d_z}}/(2^d-1)$. Without loss of generality, we assume $i = 2$, where $c_2 = [0.75, 0.25, \ldots, 0.25]$ and $v_2 = [1, 0, \ldots, 0]$. Similarly by using Markov Inequality, with probability at least 0.5, the number of iterations that play arms in A_2 is no more than $2R_T^0 \cdot 4^{\frac{-d}{d-d_z}}/(2^d-1)$.

Next, we define a new problem setting in the same metric space as:

$$f_2(x) = \begin{cases} 4^{\frac{-d}{d-d_z}} - \|x - c_1\|_\infty^{\frac{d}{d-d_z}}, & x \in A_1; \\ 2^{\frac{-d}{d-d_z}} - \|x - v_2\|_\infty^{\frac{d}{d-d_z}}, & x \in A_2; \\ 0, & x \notin A_1 \cup A_2. \end{cases}.$$

And under the setting of $f_2(\cdot)$ there is a malicious strong adversary with budget $C = 2R_T^0 \cdot 2^{\frac{-d}{d-d_z}}/(2^d-1) = \Theta(R_T^0/2^d)$ to attack the rewards. (Note $1 \leq d/(d-d_z) \leq 2$). Specifically, the adversary uses the following strategy: whenever the arm in A_2 is selected and the corruption budget has not been used up, the adversary moves the reward to 0. We call this instance I_1. Therefore, before the budget is used up, each selection of arm in A_2 returns a reward 0, and hence the agent can never tell the difference between I_0 and I_1 and would follow the same strategy under I_0 until the total corruption level reaches $C = 2R_T^0 \cdot 2^{\frac{-d}{d-d_z}}/(2^d-1)$, and then the adversary stops to contaminate the rewards. And this requires at least $C/2^{\frac{-d}{d-d_z}} = 2R_T^0 \cdot 4^{\frac{-d}{d-d_z}}/(2^d-1)$ rounds in which the agent chooses arms in A_2. Therefore, with probability of at least 0.5, the regret in T rounds is at least

$$\left(T - \frac{24^{\frac{-d}{d-d_z}} R_T^0}{2^d-1}\right) \times \left(2^{\frac{-d}{d-d_z}} - 4^{\frac{-d}{d-d_z}}\right) \geq \frac{3}{16}\left(T - \frac{32R_T^0}{2^d-1}\right) = \Omega(T).$$

□

A.6.2. Lower Bound for Weak Adversaries. Recall Theorem 2.4.5 in our main paper:

Theorem 2.4.5 *Under the weak adversary with corruption budget C, for any zooming dimension d_z, there exists an instance such that any algorithm (even is aware of C) must suffer from the regret of order $\Omega(C)$ with probability at least* 0.5.

PROOF. We can modify the argument of the previous subsection A.6.1 to validate Theorem 2.4.5. If $d_z = 0$, we could simply use the metric space $([0,1), l_2)$ and the reward function

$$\mu_1(x) = \begin{cases} \frac{1}{2} - |x - \frac{1}{4}|, & x \in [0, 0.5); \\ 0, & x \in [0.5, 1). \end{cases} \qquad \mu_2(x) = \begin{cases} 0, & x \in [0, 0.5); \\ \frac{1}{2} - |x - \frac{3}{4}|, & x \in [0.5, 1). \end{cases}$$

We can easily verify that the zooming dimension $d_z = 0$ holds. Assume there is no random noise, and at each iteration the weak adversary pushes the reward everywhere in $[0,1)$ to 0, which would use a 0.5 budget. Therefore, this attack could last for the first $\lfloor 2C \rfloor$ rounds, when the agent would just receive a 0 reward regardless of the pulled arm. For any algorithm, it would at least spend for $\lfloor C \rfloor$ rounds on either $[0, 0.5)$ or $[0.5, 1)$ with probability at least 0.5. By considering the above two reward functions, we know that it would incur $\Omega(C)$ regret with probability at least 0.5.

For $d_z > 0$, we set $d = \lceil 2d_z \rceil$ and consider the metric space $([0,1)^d, l_\infty)$. Similarly, we can equally divide the space $[0,1]^d$ into $1/2$ small l_∞ balls whose diameters are equal to $1/2$ by discretizing each axis. We denote these balls as $\{A_i\}_{i=1}^{2^d}$, $[0,1)^d = \cup_{i=1}^{2^d} A_i$ and their centers as $\{c_i\}_{i=1}^{2^d}$. (e.g. the center of $[0, 0.5)^2$ is $(0.25, 0.25)$.) Subsequently, we could define a set of functions $\{f_i(\cdot)\}_{i=1}^{1/2^d}$ as

$$\mu_i(x) = \begin{cases} 4^{\frac{-d}{d-d_z}} - \|x - c_i\|_\infty^{\frac{d}{d-d_z}}, & x \in A_i; \\ 0, & x \notin A_i. \end{cases}$$

We can easily verify that the zooming dimension of any instance is d_z. Assume there is no random noise, and at each iteration, the weak adversary pushes the reward everywhere in $[0,1)$ to 0, which would use a $4^{\frac{-d}{d-d_z}}$ budget. Therefore, this attack could last for the first $\lfloor 4^{\frac{d}{d-d_z}} C \rfloor$ rounds, when the agent would just receive a 0 reward regardless of the pulled arm. After $\lfloor 4^{\frac{d_z}{d-d_z}} C \rfloor$ rounds, for any algorithm even with the information of value C, there must be at least $\lfloor 2^{d-1} \rfloor$ balls among $\{A_i\}_{i=1}^{2^d}$ that have been pulled for at most $\lfloor 2^{(\frac{2d}{d-d_z} - d)} C \rfloor = \Theta(C)$ times. As a consequence, as for the problem instance k among these $\lfloor 2^{d-1} \rfloor$ balls, the regret incurred $\Omega(C)$. Similarly, this means that any algorithm must incur $\Omega(C)$ regret with probability 0.5. □

A.7. Additional Experimental Details

Note in our main paper we assume that $\sigma = 1$, and our pseudocodes of Algorithms are based on this assumption. When we know a better upper bound for σ, we could easily modify the components

in each algorithm based on σ. For example, we could modify the confidence radius of any active arm x in Algorithm 1 as

$$r(x) = \sigma\sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n(x)}} + \frac{C}{n(x)}.$$

Next, we exhibit the setup of algorithms involved in our experiments as follows:

- **Zooming algorithm (Kleinberg et al., 2019):** We use the same setting for stochastic Lipschitz bandit as in Kleinberg et al. (2019), and set the radius for each arm as:

$$r(x) = \sigma\sqrt{\frac{4\ln(T) + 2\ln(2/\delta)}{n(x)}}.$$

- **RMEL (ours):** We use the same parameter setting for RMEL as shown in our Algorithm 2. And based on the experimental results in Figure 2.1, this method apparently works best under different kinds of attacks and reward functions.
- **BoB Robust Zooming algorithm (ours):** We use the same parameter setting with σ for BoB Robust Zooming algorithm as shown in Algorithm 10 without restarting the algorithm after each batch since we found that restarting will sometimes abandon useful information empirically. This BoB-based approach also works well according to Figure 2.1.

The numerical results of final cumulative regrets in our simulations in Section 2.5 (Figure 2.1) are displayed in Table A.2.

Note our RMEL (Algorithm 2) is designed to defend against the weak adversary in the theoretical analysis, and hence to be consistent, we also consider the weak adversary for both types of attacks under the same experimental setting and three levels of corrupted budgets. Recall that in the previous experiments in Section 2.5, the adversary will contaminate the stochastic rewards only if the pulled action is in the specific region (Oracle: benign arm, Garcelon: targeted arm region), and otherwise the adversary will not spend its budget. And hence it is a strong adversary whose action relies on the current arm. To adapt these two attacks into a weak-adversary version, we could simply inject both sorts of attacks at each round based on their principles at each round: the Oracle will uniformly push the expected rewards of all "good arms" below the expected reward of the worst arm with an additional margin of 0.1 with probability 0.5 at the very beginning of each round. And the Garcelon will modify the expected rewards of all arms outside the targeted region into a random Gaussian noise $N(0, 0.01)$ with probability 0.5 ahead of the agent's action.

Consequently, adversaries may consume the corruption budget at each round regardless of the pulled arm, and we expect that they will run out of their total budget in fewer iterations than the strong adversary does. We use the same experimental settings as in Section 2.5, and the results are exhibited in Table A.3.

From Table A.3, we can see the experimental results under the weak adversary are consistent with those under the strong adversary. The state-of-the-art Zooming algorithm is evidently vulnerable to the corruptions, while our proposed algorithms, especially RMEL, could yield robust performance across multiple settings consistently. We can also observe that compared with the strong adversary, the weak adversary is less malicious than expected.

Another remark is that the adversarial settings used in our experiments may not be consistent with the assumption that $|c_t(x)| \leq 1$, while we find that (1). by modifying the original attacks and restricting the attack volume to be at most one with truncation, we can get a very similar result as shown in Table A.2 and Table A.3. (2). actually we can change the assumption to $|c_t(x)| \leq u$ where u is an arbitrary positive constant for the theoretical analysis.

	Algorithm	Budget (C)	Oracle	Garcelon
Triangle reward function	Zooming	0	366.58	366.58
		3000	10883.51	10660.17
		4500	11153.78	11487.59
	RMEL	0	512.46	512.46
		3000	921.95	504.78
		4500	928.27	1542.17
	BoB Robust Zooming	0	461.16	461.16
		3000	495.06	531.37
		4500	1323.97	736.85
	Algorithm	Budget (C)	Oracle	Garcelon
Sine reward function	Zooming	0	315.94	315.94
		3000	5289.65	3174.26
		4500	5720.30	3174.29
	RMEL	0	289.86	289.86
		3000	442.66	289.29
		4500	862.90	332.71
	BoB Robust Zooming	0	435.44	435.44
		3000	414.54	746.96
		4500	1887.35	1148.09
	Algorithm	Budget (C)	Oracle	Garcelon
Two dim reward function	Zooming	0	3248.54	3248.54
		3000	8730.73	8149.79
		4500	9496.83	13672.00
	RMEL	0	2589.32	2589.32
		3000	5660.10	2590.77
		4500	6265.09	2872.64
	BoB Robust Zooming	0	3831.94	3831.94
		3000	6310.29	4217.74
		4500	6932.09	4380.19

TABLE A.2. Numerical values of final cumulative regrets of different algorithms under the experimental settings used in Figure 2.1 in Section 2.5 (strong adversaries).

	Algorithm	Budget (C)	Oracle	Garcelon
Triangle reward function	Zooming	0	366.58	366.58
		3000	10861.72	10660.18
		4500	10862.75	10661.99
	RMEL	0	512.46	512.46
		3000	624.29	620.96
		4500	623.50	634.59
	BoB Robust Zooming	0	461.16	461.16
		3000	545.27	561.77
		4500	552.66	569.51
	Algorithm	Budget (C)	Oracle	Garcelon
Sine reward function	Zooming	0	315.94	315.94
		3000	5178.81	2636.73
		4500	5186.28	2799.22
	RMEL	0	289.86	289.86
		3000	280.62	277.22
		4500	284.94	288.06
	BoB Robust Zooming	0	435.44	435.44
		3000	450.21	439.08
		4500	461.13	456.36
	Algorithm	Budget (C)	Oracle	Garcelon
Two dim reward function	Zooming	0	3248.54	3248.54
		3000	6380.37	6517.29
		4500	6991.41	6854.05
	RMEL	0	2589.32	2589.32
		3000	3198.06	2940.93
		4500	4231.88	4067.16
	BoB Robust Zooming	0	3831.94	3831.94
		3000	4019.08	3335.67
		4500	4901.20	4054.05

TABLE A.3. Numerical values of final cumulative regrets of different algorithms under the experimental settings introduced in Appendix A.7 (weak adversaries).

APPENDIX B

Appendix for Chapter 3

B.1. Supportive Experimental Details

B.1.1. Simulations on the Optimal Hyperparameter Value in Grid Search. To further validate the necessity of dynamic hyperparameter tuning, we conduct a simulation for UCB algorithms LinUCB, UCB-GLM, GLOC and TS algorithms LinTS, GLM-TSL with a grid search of exploration parameter in $\{0.1, 0.5, 1, 1.5, 2, \ldots, 10\}$ and then report the best parameter value under different settings. Specifically, we set $d = 10, T = 8000, K = 60, 120$, and choose arm $x_{t,a}$ and θ^* randomly in $\{x : \|x\| \leq 1\}$. Rewards are simulated from $N(x_{t,a}^\top \theta^*, 0.5)$ for LinUCB, LinTS, and from Bernoulli$(1/(1 + \exp(-x_{t,a}^\top \theta^*)))$ for UCB-GLM, GLOC and GLM-TSL. The results are displayed in Table B.1, where we can see that the optimal hyperparameter values are distinct and far from the theoretical ones under different algorithms or settings. Moreover, the theoretical optimal exploration rate should be identical under different values of K for most algorithms shown here, but in practice the best hyperparameter to use depends on K, which also contradicts with the theoretical result.

Bandit type	Linear bandit		Generalized linear bandit		
Algorithm	LinUCB	LinTS	UCB-GLM	GLOC	GLM-TSL
$K = 60$	2.5	1	1.5	4.5	1.5
$K = 120$	3	1.5	2.5	5	2

TABLE B.1. The optimal exploration parameter value in grid search for LinUCB, LinTS, UCB-GLM, GLOC and GLM-TSL based on average cumulative regret of 5 repeated simulations.

B.1.2. Simulations to Validate the Lipschitzness of Hyperparameter Configuration. We also conduct another simulation to show it is reasonable and fair to assume the expected reward is an almost-stationary Lipschitz function w.r.t. hyperparameter values. Specifically, we set $d = 6, T = 3000, K = 60$, and for each time we run LinUCB and LinTS by using our CDT

framework, but also obtain the results by choosing the exploration hyperparameter in the set $\{0.3, 0.45, 0.6, \ldots, 8.85, 9\}$ respectively. For the first 200 rounds we use the random selection for sufficient exploration, and hence we omit the results for the first 200 rounds. After the warming-up period, we divide the rest of iterations into 140 groups uniformly, where each group contains 20 consecutive iterations. Then we calculate the mean of the obtained reward of each hyperparameter value in the adjacent 20 rounds, and centralize the mean reward across different hyperparameters in each group (we call it group mean reward). Afterward, we can calculate the mean and standard deviation of the group mean reward for different hyperparameter values across all groups. The results are shown in Figure B.1, where we can see the group mean reward can be decently represented by a stationary Lipschitz continuous function w.r.t hyperparameter values. Conclusively, we could formulate the hyperparameter optimization problem as a stationary Lipschitz bandit after sufficient exploration in the long run. And in the very beginning we can safely believe there is also only finite number of change points. This fact firmly authenticates our problem setting and assumptions.

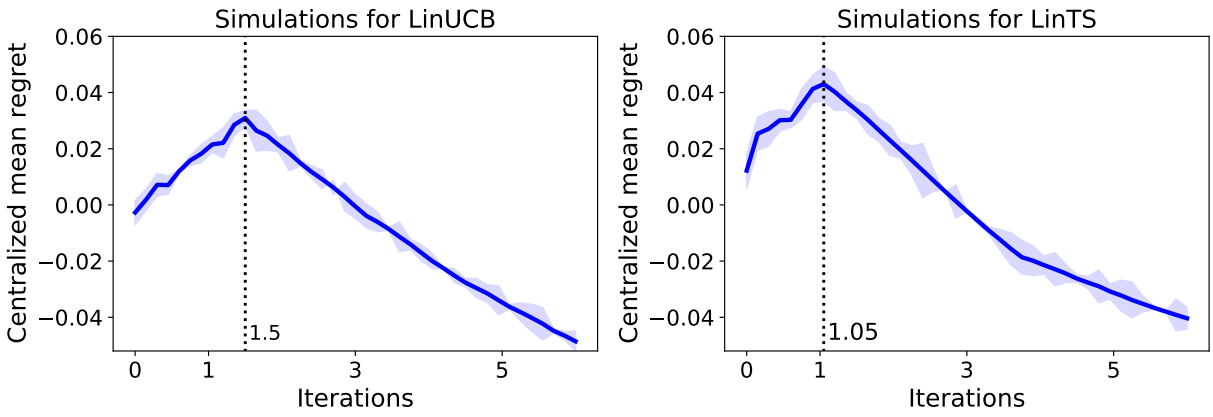


FIGURE B.1. Average cumulative regret and its standard deviation of group mean reward for different hyperparameter values across all groups.

B.1.3. Simulations for Algorithm 3. We also conduct empirical studies to evaluate our proposed Zooming TS algorithm with Restarts (Algorithm 3) in practice. Here we generate the dataset under the *switching* environment, and abruptly change the underlying mean function for several times within the time horizon T. The methods used for comparison as well as the simulation setting are elaborated as follows:

Methods. We compare our Algorithm 3 (we call it Zooming TS-R for abbreviation) with two contenders: (1) Zooming algorithm (Kleinberg et al., 2019): this algorithm is designed for the static Lipschitz bandit, and would fail in theory under the *switching* environment; (2) Oracle: we assume this algorithm knows the exact time for all switching points, and would renew the Zooming algorithm when reaching a new stationary environment. Although this algorithm could naturally perform well, but it is infeasible in reality. Therefore, we would just use Oracle as a skyline here, and a direct comparison between Oracle and our Algorithm 3 is inappropriate.

Settings. Assume the set of arm is $[0, 1]$. The unknown mean function $f_t(x)$ is chosen from two classes of reward functions with different smoothness around their maximum:

(1) $\{0.9 - 0.9|x - a|, x \in [0, 1] : a = 0.05, 0.25, 0.45, 0.70, 0.95\}$ (triangle function);

(2) $\left\{\frac{2}{3\pi} \sin\left(\frac{3\pi}{2}(x - a + \frac{1}{3})\right), x \in [0, 1] : a = 0.05, 0.25, 0.45, 0.70, 0.95\right\}$.

We set $T = 90,000$ and $c(T) = 3$, and choose the location of changing points at random in the very beginning. The random noise is generated according to $N(0, 0.1)$. The value of epoch size H is set as suggested by our theory $H = 10\lceil(T/c(T))^{3/4}\rceil$. For each class of reward functions, we run the simulations for 20 times and report the average cumulative regret as well as the standard deviation for each contender in Figure B.2. (The change points are fixed for each repetition to make the average value meaningful.)

Figure B.2 shows the performance comparisons of three different methods under the *switching* environment measured by the average cumulative regret. We can see that Oracle is undoubtedly the best since it knows the exact times for all change points and hence restart our Zooming TS algorithm accordingly. The traditional Zooming algorithm ranks the last w.r.t both mean and standard deviation since it doesn't take the non-stationarity issue into account at all, and would definitely fail when the environment changes. This fact also coincides with our expectations precisely. Our proposed algorithm has an obvious advantage over the traditional Zooming algorithm when the change points exist, and we can see that our algorithm could adapt to the environment change quickly and smoothly.

B.1.4. Additional Details and Results for Section 3.5.

B.1.4.1. *Baselines with A Large Candidate Set.* To further make a fair comparison and validate the high superiority of our proposed CDT framework over the existing OP, TL (or Syndicated) which relies on a user-defined hyperparameter candidate set, we explore whether CDT will consistently

outperform if baselines are running with a large tuning set. Here we replace the original tuning set $C_1 = \{0.1, 1, 2, 3, 4, 5\}$ with a finer set $C_2 = \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$. And the new results are shown in the following Table B.2 (original results in Section 3.5 are in gray).

Candidate Set		C1		C2	
Algorithm	Setting	TL/Syndicated	OP	TL/Syndicated	OP
LinUCB	Simulations	343.14	383.62	356.23	389.91
	Movielens	346.16	390.10	359.10	408.67
LinTS	Simulations	828.41	869.30	874.34	925.29
	Movielens	519.09	666.35	516.62	667.77
UCB-GLM	Simulations	271.45	350.85	298.68	367.97
	Movielens	381.00	397.58	406.29	412.62
GLM-TSL	Simulations	433.27	445.43	448.21	458.71
	Movielens	446.74	678.91	458.23	718.46
Laplace-TS	Simulations	510.03	568.81	530.29	567.10
	Movielens	949.51	1063.92	958.10	1009.23
GLOC	Simulations	406.28	417.30	414.82	427.05
	Movielens	571.36	513.90	568.91	520.72
SGD-TS	Simulations	448.29	551.63	458.09	557.04
	Movielens	1016.72	1084.13	1038.94	1073.91

TABLE B.2. Cumulative regrets of baselines under different hyperparameter tuning sets.

Therefore, we can observe that the performance overall becomes worse under C_2 compared with the original C_1. In other words, adding lots of elements to the tuning set will not help improve the performance of existing algorithms. We believe this is because the theoretical regret bound of TL (Syndicated) also depends on the number of candidates k in terms of $\sqrt{k}$ (Ding et al., 2022b). There is no theoretical guarantee for OP. After introducing so many redundant values in the candidate set, the TL (Syndicated) and OP algorithms would get disturbed and waste lots of concentration on those unnecessary candidates.

In conclusion, we believe the existing algorithms relying on user-tuned candidate sets would perform well if the size of the candidate set is reasonable and the candidate set contains some value very close to the optimal hyperparameter value. However, in practice, finding the unknown optimal hyperparameter value is a black-box problem, and it's impossible to construct a candidate set satisfying the above requirements at the beginning. If we discretize the interval finely, then the large size of the candidate set would hurt the performance as well. On the other hand, our proposed

CDT could adaptively "zoom in" on the regions containing this optimal hyperparameter value automatically, without the need of pre-specifying a "good" set of hyperparameters. And CDT could always yield robust results according to the extensive experiments we did in Section 3.5.

On the other hand, these results also imply an interesting fact. Note it is doable to first discretize the continuous space and then implement an algorithm with discrete candidate sets, such as Syndicated (Ding et al., 2022b). However, we observe that finely discretizing the hyperparameter space will significantly hurt the practical performance and hence is wasteful and inefficient. Intuitively, it is inefficient to place lots of "probes" in other regions that do not contain the optimal point, and we should place probes in more promising regions via adaptive discretization methodology. In theory, the uniform discretization idea will lead to regret bound of order $T^{\frac{d+1}{d+2}}$ with covering dimension d and the zooming idea will incur $T^{\frac{d_z+1}{d_z+2}}$ regret with zooming dimension d_z, and we know $d_z \leq d$ and d_z could be significantly smaller than d under various cases. Therefore, we believe the same phenomena will occur in the non-stationary Lipschitz bandits and also our hyperparameter tuning framework as well.

B.1.4.2. *Ablation Study on the Choice of T_1 and T_2.* For T_1, we set it to $T^{2/(p+3)}$ where p stands for the number of hyperparameters according to Theorem 3.4.2. Specifically, for LinUCB, LinTS, UCB-GLM, GLM-TSL and Laplace-TS, we choose it to be 118. For GLOC and SGD-TS, we set it as 45. Here we also rerun our experiments in Section 3.5 with $T_1 = 0$ (no warm-up) since we believe a long warm-up period will abandon lots of useful information, and then we report the results after this change:

We can observe that the results are almost identical from Table B.3. For T_2, Theorem 3.4.2 suggests that $T_2 = O\left(T^{(p+2)/(p+3)}\right)$. In our original experiments, we choose $T_2 = 3T^{(p+2)/(p+3)}$. To take an ablation study on T_2 we take $T_2 = kT^{(p+2)/(p+3)}$ for $k = 1, 2, 3$ in each experiment, and to see whether our CDT framework is robust to the choice of k.

According to Table B.4, we can observe that overall $k = 2$ and $k = 3$ perform better than $k = 1$. We believe it is because, in the long run, the optimal hyperparameter would tend to be stable, and hence some restarts are unnecessary and inefficient. Note by choosing $k = 1$ our proposed CDT still outperforms the existing TL and OP tuning algorithms overall. For $k = 2$ and $k = 3$, we can observe that their performances are comparable, which implies that the choice of k is quite robust in practice. We believe it is due to the fact that our proposed Zooming TS algorithm could always

Algorithm	Setting	$T_1 = 0$	$T_1 = T^{2/(p+3)}$
LinUCB	Simulation	298.28	303.14
	Movielens	313.29	307.19
LinTS	Simulation	677.03	669.45
	Movielens	343.18	340.85
UCB-GLM	Simulation	299.74	300.54
	Movielens	314.41	311.72
GLM-TSL	Simulation	339.49	333.07
	Movielens	428.82	432.47
Laplace-TS	Simulation	520.29	520.35
	Movielens	903.16	900.10
GLOC	Simulation	414.70	418.05
	Movielens	455.39	461.78
SGD-TS	Simulation	430.05	425.98
	Movielens	843.91	838.06

TABLE B.3. Ablation study on the role of T_1 in our CDT framework.

Algorithm	Setting	$k = 1$	$k = 2$	$k = 3$
LinUCB	Simulation	328.28	300.62	298.28
	Movielens	310.06	303.10	313.29
LinTS	Simulation	717.77	670.90	677.03
	Movielens	360.12	352.19	343.18
UCB-GLM	Simulation	314.01	316.95	299.74
	Movielens	347.92	325.58	314.41
GLM-TSL	Simulation	320.21	331.43	339.49
	Movielens	439.98	428.91	428.82
Laplace-TS	Simulation	565.15	540.61	520.29
	Movielens	948.10	891.91	903.16
GLOC	Simulation	417.05	414.70	415.05
	Movielens	441.85	455.39	462.24
SGD-TS	Simulation	450.14	430.05	414.57
	Movielens	852.98	843.91	830.35

TABLE B.4. Ablation study on the role of T_2 in our CDT framework.

adaptively approximate the optimal point. Although it is unknown which one is better in practice under different cases, our comprehensive simulations show that choosing either one in practice will work well and outperform all the existing methods. In conclusion, these results suggest that we have a universal way to set the values of T_1 and T_2 according to the theoretical bounds, and we do not need to tune them for each particular dataset. In other words, the performance of our CDT tuning framework is robust to the choice of T_1, T_2 under different scenarios.

B.2. Supportive Remarks

REMARK B.2.1. (Justifications on assumptions) We further explain the motivations of the Lipschitzness and piecewise stationarity assumptions of the expected reward function for hyperparameter tuning of bandit algorithms.

For Lipschitzness, we get the motivation of our formulation shown in Eqn. 3.3 and Eqn. 3.4 from the hyperparameter tuning work on the offline machine learning algorithms. Specifically, Bayesian optimization is widely considered as the state-of-the-art and most popular hyperparameter tuning method, which assumes that the underlying function is sampled from a Gaussian process in the given space. By selecting a value x in the space and obtaining the corresponding reward, Bayesian optimization could update its estimation of the underlying function, especially in the neighbor of x sequentially. And it also relies on a user-defined kernel function, whose selection is also purely empirical and lacks theoretical support. In our work, we use a similar idea as Bayesian optimization: close hyperparameters tend to yield similar values with other conditions fixed. And this natural extension motivates the Lipschitz assumption made in our paper. Therefore, it is fair to make a similar and analogous assumption (close hyperparameters yield similar results given other conditions fixed) for the hyperparameter tuning of bandit algorithms in our work. We validate this assumption using a suite of simulations in Appendix B.1.

For the piecewise stationarity, as we mention in Section 3.3, it is inappropriate to assume the strict stationarity of the bandit algorithm performance under the same hyperparameter value setting across time T. As an example, for most UCB and TS-based bandit algorithms (e.g. LinUCB, LinTS, UCB-GLM, GLM-UCB, GLM-TSL, etc.), the exploration degree of an arm is a multiplier of the exploration rate and the uncertainty of an arm. In the beginning, a moderate value of the exploration rate may lead to a large exploration degree for the arm since the uncertainty is large. On the contrary, in the long run, a moderate value of exploration rate will lead to a minor exploration degree for the arm since its value has been well estimated with small uncertainty. Therefore, a fixed hyperparameter setting may suggest different results across different stages of time, and hence it is unreasonable to expect the strong stationarity of the hyperparameter tuning for bandit algorithms at all time steps. On the other hand, it would be very inefficient to assume a completely non-stationary environment as in Ding et al. (2022b) which uses EXP3. In very close time steps, we could anticipate that the same hyperparameter setting would yield a very similar

result in expectation since the uncertainty of any arm would be close. And using a non-stationary environment will totally waste this information and hence is inefficient. Therefore, it is very well motivated to use a partial non-stationarity assumption that lies in the middle ground between the above two extremes. Note our proposed tuning method yields very promising results in extensive experiments under our formulations. And the stationary environment can be regarded as a special case of our switching environment setting where the functions in between all change points are the same.

Finally, we will explain why it is excessively difficult to present theoretical validation regarding these assumptions in our paper. As we mentioned, our formulation is motivated by Bayesian optimization, arguably the most popular method for hyperparameter tuning for offline machine learning algorithms. And we use a similar idea: similar hyperparameters tend to yield similar values while other conditions are fixed. However, people could hardly provide any theory backing for the analogous assumption of Bayesian optimization for any offline machine learning algorithms (e.g. regression, classification), and hyperparameter tuning is widely considered as a black-box problem for offline machine learning algorithms. Not to mention that the theoretical analysis of hyperparameter tuning for any bandit algorithm is much more challenging than that of offline machine learning algorithms since historical observations along with hyperparameter values will affect the online selection simultaneously for the bandit algorithms, and we can use different hyperparameters in different rounds for bandit algorithms. Conclusively, our formulation is natural and well-motivated.

B.3. Detailed Proof on the Zooming Dimension

In the beginning, we would reload some notations for simplicity. Here we could omit the time subscript (or superscript) t since the following result could be identically proved for each round t. Assume the Lipschitz function f is defined on $\mathbb{R}^{p_c}$, and $v^* := \arg\max_{v \in A} f(v)$ denotes the maximal point (w.l.o.g. assume it's unique), and $\Delta(v) = f(v^*) - f(v)$ is the "badness" of the arm v. We then naturally denote A_r as the r-optimal region at the scale $r \in (0, 1]$, i.e. $A_r = \{v \in A : r/2 < \Delta(v) \leq r\}$. The r-zooming number could be denoted as $N_z(r)$. And the zooming dimension could be naturally denoted as p_z. Note that by the Assouad's embedding theorem, any compact doubling metric space $(A, \mathrm{Dist}(\cdot, \cdot))$ can be embedded into the Euclidean space with some type of metric.

Therefore, for all compact doubling metric spaces with cover dimension p_c, it is sufficient to study on the metric space $([0,1]^{p_c}, \|\cdot\|^l)$ for some $l \in (0, +\infty]$ instead.

We will rigorously prove the following two facts regarding the r-zooming number $N_z(r)$ of (A, f) for arbitrary compact set $A \subseteq \mathbb{R}^{p_c}$ and Lipschitz function $f(\cdot)$ defined on A:

- $0 \leq p_z \leq p_c$.
- The zooming dimension could be much smaller than p_c under some mild conditions. For example, if the payoff function f is greater than $\|v^* - v\|^\beta$ in scale in a (non-trivial) neighborhood of v^* for some $\beta \geq 1$, i.e. $f(v^*) - f(v) \geq C(\|v^* - v\|^\beta)$ as $\|v^* - v\| \leq r$ for some $C > 0$ and $r = \Theta(1)$, then it holds that $p_z \leq (1 - 1/\beta)p_c$. Note $\beta = 2$ when we have $f(\cdot)$ is C^2-smooth and strongly concave in a neighborhood of v^*, which subsequently implies that $p_z \leq p_c/2$.

PROOF. Due to the compactness of A, it suffices to prove the results when $A = [0,1]^{p_c}$. By the definition of the zooming dimension p_z, it naturally holds that $p_z \geq 0$. On the other side, since the space A is a closed and bounded set in $\mathbb{R}^{p_c}$, we assume the radius of A is no more than S, which consequently implies that the $r/16$-covering number of A is at most the order of

$$\left(\frac{S}{\frac{r}{16}}\right)^{p_c} = (16S)^{p_c} \cdot r^{-p_c}.$$

Since we know $A_r \subseteq A$, it holds that $p_z \leq p$. Secondly, if the payoff function f is locally greater than $\|v^* - v\|^\beta$ in scale for some $\beta \geq 1$, i.e. $f(v^*) - f(v) \geq C(\|v^* - v\|^\beta)$, then there exists $C \in \mathbb{R}$ and $\delta > 0$ such that as long as $C\|v - v^*\|^\beta \leq \delta$ we have $f(v^*) - f(v) \geq C\|v - v^*\|^\beta$. Therefore, for $0 < r < \delta$, it holds that,

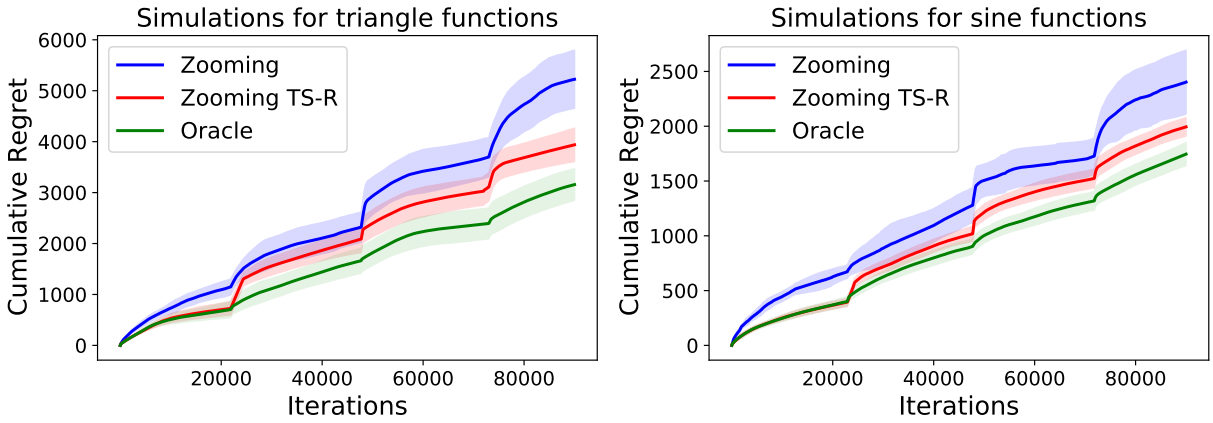
$$\{v : r \geq f(v^*) - f(v) > r/2\} \subseteq \{v : C\|v - v^*\|^\beta \leq r\} = \left\{v : \|v - v^*\| \leq \left(\frac{r}{C}\right)^{\frac{1}{\beta}}\right\}$$

It holds that the r-covering number of the Euclidean ball with center v^* and radius $(r/c)^{(1/\beta)}$ is of the order of

$$\left(\frac{\left(\frac{r}{C}\right)^{\frac{1}{\beta}}}{\frac{r}{16}}\right)^{p_c} = \left(\frac{16}{C^{\frac{1}{\beta}}}\right)^{p_c} \cdot r^{-(1-\frac{1}{\beta})p_c}$$

which explicitly implies that $p_z \leq (1 - 1/\beta)p_c$. □

FIGURE B.2. Cumulative regret plots of Zooming TS-R, Zooming and Oracle algorithms under the *switching* environment.



B.4. Intuition of our Thompson Sampling update

Intuitively, we consider a Gaussian likelihood function and Gaussian conjugate prior to design our Thompson Sampling version of zooming algorithm, and here we would ignore the clipping step for explanation. Suppose the likelihood of reward $\tilde{y}_t$ at time t, given the mean of reward $I(v_t)$ for our pulled arm v_t, follows a Gaussian distribution $N(I(v_t), s_0^2)$. Then, if the prior of $I(v_t)$ at time t is given by $N(\hat{f}_t(v_t), s_0^2/n_t(v_t))$, we could easily compute the posterior distribution at time $t+1$,

$$P(I(v_t)|\tilde{y}_t) \propto P(\tilde{y}_t|I(v_t))P(I(v_t)),$$

as $N(\hat{f}_{t+1}(v_t), s_0^2/n_{t+1}(v_t))$. We can see this result coincides with our design in Algorithm 3 and its proof is as follows:

PROOF.

$$\begin{aligned}
P(I(v_t)|\tilde{y}_t) &\propto P(\tilde{y}_t|I(v_t))P(I(v_t)) \\
&\propto \exp\left\{-\frac{1}{2s_0^2}[(I(v_t)-\tilde{y}_t)^2 + n_t(v_t)(I(v_t)-f_t(v_t))^2]\right\} \\
&\propto \exp\left\{-\frac{1}{2s_0^2}[(n_t(v_t)+1)I(v_t)^2 - 2(\tilde{y}_t + n_t(v_t)f_t(v_t))I(v_t)]\right\} \\
&\propto \exp\left\{-\frac{n_{t+1}(v_t)}{2s_0^2}\left[I(v_t)^2 - 2\frac{(\tilde{y}_t + n_t(v_t)f_t(v_t))}{n_{t+1}(v_t)}I(v_t)\right]\right\} \\
&\propto \exp\left\{-\frac{n_{t+1}(v_t)}{2s_0^2}(I(v_t)-f_{t+1}(v_t))^2\right\}
\end{aligned}$$

Therefore, the posterior distribution of $I(v_t)$ at time $t+1$ is $N(f_{t+1}(v_t), s_0^2 \frac{1}{n_{t+1}(v_t)})$. □

This gives us an intuitive explanation why our Zooming TS algorithm works well when we ignore the clipped distribution step. And we have stated that this clipping step is inevitable in Lipschitz bandit setting in our main paper since (1) we'd like to avoid underestimation of good active arms, i.e. avoid the case when their posterior samples are too small. (2) We could at most adaptively zoom in the regions which contains v^* instead of exactly detecting v^*, and this inevitable loss could be mitigated by setting a lower bound for TS posterior samples. Note that although the intuition of our Zooming TS algorithm comes from the case where contextual bandit rewards follow a Gaussian distribution, we also prove that our algorithm can achieve a decent regret bound under the *switching* environment and the optimal instance-dependent regret bound under the stationary Lipschitz bandit setting.

B.5. Proof of Theorem 3.4.1

B.5.1. Stationary Environment Case. To prove Theorem 3.4.1, we will first focus on the stationary case, where $f_t \coloneqq f,\ \forall t \in [T]$. When the environment is stationary, we could omit the subscript (or superscript) t in some notations as in Section B.3 for simplicity: Assume the Lipschitz function is f, and $v^* \coloneqq \arg\max_{v \in A} f(v)$ denotes the maximal point (w.l.o.g. assume it's unique), and $\Delta(v) = f(v^*) - f(v)$ is the "badness" of the arm v. We then naturally denote A_r as the r-optimal region at the scale $r \in (0, 1]$, i.e. $A_r = \{v \in A : r/2 < \Delta(v) \leq r\}$. The r-zooming number could be denoted as $N_z(r)$. And the zooming dimension could be naturally denoted as p_z. Note we could omit the subscript (or superscript) t for the notations just mentioned above since all these values would be fixed through all rounds under the stationary environment.

B.5.1.1. *Useful Lemmas and Corollaries.* Recall that $\hat{f}_t(v)$ is the average observed reward for arm $v \in A$ by time t. And we call all the observations (pulled arms and observed rewards) over T total rounds as a process.

Definition B.5.1. *We call it a clean process, if for each time $t \in [T]$ and each strategy $v \in A$ that has been played at least once at any time t, we have $|\hat{f}_t(v) - f(v)| \leq r_t(v)$.*

Lemma B.5.1.1. *The probability that, a process is clean, is at least $1 - 1/T$.*

PROOF. Fix some arm v. Recall that each time an algorithm plays arm v, the reward is sampled IID from some distribution $\mathbb{P}_v$. Define random variables $U_{v,s}$ for $1 \leq s \leq T$ as follows: for $s \leq n_T(v)$, $U_{v,s}$ is the reward from the s-th time arm v is played, and for $s > n_T(v)$ it is an independent sample from $\mathbb{P}_v$. For each $k \leq T$ we can apply Chernoff bounds to $\{U_{v,s} : 1 \leq s \leq k\}$ and obtain that:

$$P\left(\left|\frac{1}{k}\sum_{s=1}^{k} U_{v,s} - f(v)\right| \geq \sqrt{\frac{13\tau_0^2 \ln T}{2k}}\right) \leq 2 \cdot \exp\left(-\frac{k}{2\tau_0^2}\frac{13\tau_0^2 \ln T}{2k}\right)$$

$$= 2\exp\left(\frac{13}{4}\ln T\right) = 2T^{-3.25} \leq T^{-3}, \tag{B.1}$$

since we can trivially assume that $T \geq 16$. Let N be the number of arms activated all over rounds T; note that $N \leq T$. Define X-valued random variables $\{x_i\}_{i=1}^{T}$ as follows: x_j is the $\min(j, N)$-th arm activated by time T. For any $x \in A$ and $j \leq T$, the event $\{x = x_j\}$ is independent of the random variables $\{U_{x,s}\}$: the former event depends only on payoffs observed before x is activated, while the latter set of random variables has no dependence on payoffs of arms other than x. Therefore, Eqn. (B.1) is still valid if we replace the probability on the left side with conditional probability, conditioned on the event $\{x = x_j\}$. Taking the union bound over all $k \leq T$, it follows that:

$$P(\forall t \leq T, |f(v) - \hat{f}_t(v)| \leq r_t(v) \,|\, x_j = v) \geq 1 - T^{-2}, \quad \forall v \in A, j \in [T],$$

Integrating over all arms v we get

$$P(\forall t \leq T, |f(x_j) - \hat{f}_t(x_j)| \leq r_t(x_j)) \geq 1 - T^{-2}, \quad \forall j \in [T].$$

Finally, we take the union bound over all $j \leq T$, and it holds that,

$$P(\forall t \leq T, j \leq T, |f(x_j) - \hat{f}_t(x_j)| \leq r_t(x_j)) \geq 1 - T^{-1},$$

and this obviously implies the result. □

LEMMA B.5.1.2. *If it is a clean process, then $B(v, r_t(v))$ could never be eliminated from Algorithm 3 for any $t \in [T]$ and arm v that is active at round t, given that $v^* \in B(v, r_t(v))$.*

PROOF. Recall that from Algorithm 3, at round t the ball $B(u, r_t(u))$ would be permanently removed if we have for some active arm v s.t.

$$\hat{f}_t(v) - r_t(v) > \hat{f}_t(u) + 2r_t(u).$$

If we have that $v^* = \arg\max_{x \in A} f(x) \in B(u, r_t(u))$, then it holds that

$$\hat{f}_t(u) + 2r_t(u) \geq f(u) + r_t(u) \geq f(u) + \text{Dist}(u, v^*) \geq f(v^*),$$

where the first inequality is due to the clean process and the last one comes from the fact that f is a Lipschitz function. On the other hand, we have that for any active arm v,

$$f(v) \geq \hat{f}_t(v) - r_t(v), \quad f(v^*) \geq f(v).$$

Therefore, it holds that

$$\hat{f}_t(v) - r_t(v) \leq \hat{f}_t(u) + 2r_t(u).$$

And this inequality concludes our proof. □

LEMMA B.5.1.3. *If it is a clean process, then for any time t and any active strategy v that has been played at least once before time t we have $\Delta(v) \leq 5\mathbb{E}[r_t(v)]$. Furthermore, it holds that $\mathbb{E}(n_t(v)) \leq O(\ln(T)/\Delta(v)^2)$.*

PROOF. Let S_t be the set of all arms that are active at time t. Suppose an arm v_t is played at time t and was previously played at least twice before time t. Firstly, We would claim that

$$f(v^*) \leq I_t(v_t) \leq f(v_t) + 3r_t(v_t)$$

holds uniformly for all t with probability at least $1 - \delta$, which directly implies that $\Delta(v_t) \leq 3r_t(v_t)$ with high probability uniformly. First we show that $I_t(v_t) \geq f(v^*)$. Indeed, recall that all arms are covered at time t, so there exists an active arm v_t^* that covers v^*, meaning that v^* is contained in the confidence ball of v_t^*. And based on Lemma B.5.1.2 the confidence ball containing v^* could never be eliminated at round t when it's a clean process. Recall $Z_{t,v}$ is the i.i.d. standard normal random variable used for any arm v in round t (Eqn. (3.6)). Since arm v_t was chosen over v_t^*, we

have $I_t(v_t) \geq I_t(v_t^*)$. Since this is a clean process, it follows that

$$I_t(v_t^*) = \hat{f}_t(v_t^*) + s_0\sqrt{\frac{1}{n_t(v_t^*)}}Z_{t,v_t^*} \geq f(v_t^*) + s_0\sqrt{\frac{1}{n_t(v_t^*)}}Z_{t,v_t^*} - r_t(v_t^*) \tag{B.2}$$

Furthermore, according to the Lipschitz property we have

$$f(v_t^*) \geq f(v^*) - \text{Dist}(v_t^*, v^*) \geq f(v^*) - r_t(v_t^*). \tag{B.3}$$

Combine Eqn. (B.2) and (B.3), we have

$$\begin{aligned} I_t(v_t) \geq I_t(v_t^*) &\geq f(v^*) + s_0\sqrt{\frac{1}{n_t(v_t^*)}}Z_{t,v_t^*} - 2r_t(v_t^*) \\ &= f(v^*) + \sqrt{\frac{52\pi\tau_0^2\ln(T)}{n_t(v_t^*)}}\left(Z_{t,v_t^*} - \frac{1}{\sqrt{2\pi}}\right) \geq f(v^*), \end{aligned} \tag{B.4}$$

where we get the last inequality since we truncate the random variable Z_{t,v_t^*} by the lower bound $1/\sqrt{2\pi}$ according to the definition. On the other hand, we have

$$I_t(v_t) \leq f(v_t) + r_t(v_t) + s_0\sqrt{\frac{1}{n_t(v_t)}}Z_{t,v_t} = f(v_t) + \left(1 + 2\sqrt{2\pi}Z_{t,v_t}\right)r_t(v_t) \tag{B.5}$$

Therefore, by combining Eqn. (B.4) and (B.5) we have that

$$\Delta(v_t) \leq \left(1 + 2\sqrt{2\pi}Z_{t,v_t}\right)r_t(v_t). \tag{B.6}$$

And we know that $Z_{t,:}$ is defined as $Z_{t,:} = \max\{1/\sqrt{2\pi}, \tilde{Z}_{t,:}\}$ where $\tilde{Z}_{t,:}$ is IID drawn from standard normal distribution. In other words, Z_{t,v_t} follows a clipped normal distribution with the following PDF:

$$f(x) = \begin{cases} \phi(x) + (1 - \Phi(x))\delta\left(x - \frac{1}{\sqrt{2\pi}}\right), & x \geq \frac{1}{\sqrt{2\pi}}; \\ 0, & x < \frac{1}{\sqrt{2\pi}}; \end{cases}$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and CDF of standard normal distribution. And we have

$$\mathbb{E}(Z_{t,v_t}) \leq \frac{1}{\sqrt{2\pi}} + \int_{\frac{1}{\sqrt{2\pi}}}^{+\infty} x\phi(x)dx \leq \frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{4\pi}} \leq \sqrt{\frac{2}{\pi}}$$

By taking expectation on Eqn. (B.6), we have $\Delta(v_t) \leq 5\mathbb{E}(r_t(v_t))$. Next, we would show that $\mathbb{E}(n_t(v_t)) \leq O(\ln(T))/\Delta(v_t)^2$. Based on Eqn. (B.5) and the definition of $r_t(\cdot)$, we could deduce

that

$$\sqrt{n_t(v_t)} \leq \sqrt{\frac{13}{2}\tau_0^2 \ln(T)}(1+2\sqrt{2\pi}Z_{t,v_t})\frac{1}{\Delta(v_t)},$$

which thus implies that

$$n_t(v_t) \leq \frac{13}{2}\tau_0^2 \ln(T)(1+2\sqrt{2\pi}Z_{t,v_t})^2\frac{1}{\Delta(v_t)^2} = O(\ln(T))(1+2\sqrt{2\pi}Z_{t,v_t})^2\frac{1}{\Delta(v_t)^2}. \tag{B.7}$$

By simple calculation, we could show that

$$\mathbb{E}(Z_{t,v_t}^2) \leq \frac{1}{2\pi} + \int_{\frac{1}{\sqrt{2\pi}}}^{+\infty} x^2\phi(x)dx \leq \frac{1}{\pi} + \frac{1}{2} \leq 1$$

$$\Rightarrow \quad \mathbb{E}\left[(1+2\sqrt{2\pi}Z_{t,v_t})^2\right] \leq 1 + 4\sqrt{2\pi}\sqrt{\frac{2}{\pi}} + 8\pi < +\infty.$$

After revisiting Eqn. (B.7), we can show that $\mathbb{E}(n_t(v_t)) \leq O(\ln(T))/\Delta(v_t)^2$. Now suppose arm v is only played once at time t, then $r_t(v) > 1$ and thus the lemma naturally holds. Otherwise, let s be the last time arm v has been played according to the selection rule, where we have $r_t(v) = r_s(v)$, and then based on Eqn. (B.5) it holds that

$$I_t(v) \leq f(v) + \left(1+2\sqrt{2\pi}Z_{s,v}\right) r_t(v).$$

And then we could show that $\Delta(v) \leq 5\mathbb{E}(r_t(v))$. By using an identical argument as before, we could show that $\mathbb{E}(n_t(v)) \leq O(\ln(T))/\Delta(v)^2$. □

Lemma B.5.1.4. *Let $X_1, \ldots, X_n$ be independent σ^2-sub-Gaussian random variables. Then for every $t > 0$,*

$$P\left(\max_{1,\leq,n} X_i \geq \sqrt{2\sigma^2(\ln(T)+t)}\right) \leq e^{-t}.$$

Proof. Let $u = \sqrt{2\sigma^2(\ln(n)+t)}$, we have

$$P\left(\max_{1,\leq,n} X_i \geq u\right) = P(\exists i, X_i \geq u) \leq \sum_{i=1}^{n} P(X_i \geq u) \leq ne^{-\frac{u^2}{2\sigma^2}} = e^{-t}.$$

□

B.5.1.2. *Proof of Theorem 3.4.1 under stationary environment.*

Proof. By Lemma B.5.1.1 we know that it is a clean process with probability at least $1 - \frac{1}{T}$. In other words, denote the event $\Omega := \{\text{clean process}\}$, and then we have that $P(\Omega) \geq 1 - \frac{1}{T}$. And

according to Lemma B.5.1.2 we're aware that the active confidence balls containing the best arm can't be removed in a clean process. Remember that we use S_T as the set of all arms that are active in the end, and denote

$$B_{i,T} = \left\{ v \in S_T \,:\, 2^i \le \frac{1}{\Delta(v)} < 2^{i+1} \right\}, \quad \text{where } S_T = \bigcup_{i=0}^{+\infty} B_{i,T},$$

where $i \ge 0$. Then, under the event Ω, by using Corollary B.5.1.3 we have $\mathbb{E}(n_T(v)|\Omega) \le O(\ln T)/\Delta(v)^2$, and hence it holds that

$$\sum_{v \in B_{i,T}} \Delta(v)\mathbb{E}(n_T(v)|\Omega) \le O(\ln T) \sum_{v \in B_{i,t}} \frac{1}{\Delta(v)} \le O(\ln T) \cdot 2^i |B_{i,t}|$$

Denote $r_i = 2^{-i}$, we have

$$\sum_{v \in B_{i,T}} \Delta(v)\mathbb{E}(n_T(v)|\Omega) \le O(\ln T) \cdot \frac{1}{r_i} |B_{i,t}|$$

Next, we would show that for any active arms u, v we have

$$\text{Dist}(u,v) > \frac{1}{4\sqrt{2\pi \ln(T)}} \min\{\Delta(u), \Delta(v)\} \tag{B.8}$$

with probability at least $1 - \frac{1}{T}$. W.l.o.g assume u has been activated before v. Let s be the time when v has been activated. Then by the philosophy of our algorithm we have that $\text{Dist}(u,v) > r_s(v)$. Then according to Eqn. (B.6) in the proof Lemma B.5.1.3, it holds that $r_s(v) \ge \frac{1}{2\sqrt{2\pi}Z}\Delta(v)$ for some random variable Z following the clipped standard normal distribution. Define the event $\Upsilon = \{Z_{t,v_t} < 2\sqrt{\ln(T)} \text{ for all } t \in [T]\}$, then based on Lemma B.5.1.4 we have $P(\Upsilon) \ge 1 - \frac{1}{T}$. Then under the event Υ, we have $r_s(v) \ge \frac{1}{4\sqrt{2\pi \ln(T)}}\Delta(v)$, which then implies that Eqn. (B.8) holds under Υ. Since for arbitrary $x, y \in B_{i,T}$ we have

$$\frac{r_i}{2} < \Delta(x) \le r_i, \quad \frac{r_i}{2} < \Delta(y) \le r_i,$$

which implies that under the event Υ

$$\text{Dist}(x,y) > \frac{1}{4\sqrt{2\pi \ln(T)}} \min\{\Delta(x), \Delta(y)\} > \frac{r_i}{8\sqrt{2\pi \ln(T)}}.$$

Therefore, x and y should belong to different sets of $(r_i/8\sqrt{2\pi\ln(T)})$-diameter-covering. It follows that $|B_{i,T}| \leq N_z(r_i/8\sqrt{2\pi\ln(T)}) \leq O(\ln(T)^p)cr_i^{p_z} \leq \tilde{O}(cr_i^{p_z})$. Recall $N_z(r)$ is defined as the minimal number of balls of radius no more than r required to cover A_r. As a result, under the events Ω and Υ, it holds that

$$\sum_{v\in B_{i,T}} \Delta(v)\mathbb{E}(n_T(v)\,|\,\Omega\cap\Upsilon) \leq O(\ln T)\cdot\frac{1}{r_i}N_z(r_i) \tag{B.9}$$

Therefore, based on Eqn. (B.9), we have

$$\begin{aligned}
R_L(T) &= \sum_{v\in S_T}\Delta(v)\mathbb{E}(n_T(v)) \\
&= P(\Omega\cap\Upsilon)\sum_{v\in S_T}\Delta(v)\mathbb{E}(n_T(v)\,|\,\Omega\cap\Upsilon) + P(\Omega^c\cup\Upsilon^c)\sum_{v\in S_T}\Delta(v)\mathbb{E}(n_T(v)\,|\,\Omega^c\cup\Upsilon^c) \\
&\leq \sum_{v\in S_T:\Delta(v)\leq\rho}\Delta(v)\mathbb{E}(n_T(v)\,|\,\Omega\cap\Upsilon) + \sum_{v\in S_T:\Delta(v)>\rho}\Delta(v)\mathbb{E}(n_T(v)\,|\,\Omega\cap\Upsilon) + \frac{2}{T}\cdot T \\
&\leq \rho T + \sum_{i<\log_2(\frac{1}{\rho})}\frac{1}{r_i}\tilde{O}(cr_i^{-p_z}) + 2 \\
&\leq \rho T + \tilde{O}(1)\sum_{i<\log_2(\frac{1}{\rho})}\frac{1}{r_i}cr_i^{-p_z} + 2 \\
&\leq \rho T + \tilde{O}(1)\sum_{k=0}^{\lfloor\log_{1/2}2\rho\rfloor} c2^{k(p_z+1)} + 2 \\
&\leq \rho T + \tilde{O}(1)\cdot 2\cdot 2^{\lfloor\log_{1/2}2\rho\rfloor(p_Z+1)} + 2 \\
&\leq \rho T + \tilde{O}(1)\left(\frac{1}{2\rho}\right)^{p_z+1} + 2
\end{aligned}$$

By choosing ρ in the scale of

$$\rho = \tilde{O}\left(\frac{1}{T}\right)^{\frac{1}{p_z+2}},$$

it holds that

$$R_L(T) = \tilde{O}\left(T^{\frac{p_z+1}{p_z+2}}\right).$$

B.5.2. *Switching* (Non-stationary) Environment Case. Since there are $c(T)$ change points for the environment Lipschitz functions $f_t(\cdot)$, i.e.

$$\sum_{t=1}^{T-1} \mathbf{1}[\exists m \in A \,:\, f_t(m) \neq f_{t+1}(m)] = c(T).$$

Given the length of epochs as H, we would have $\lceil T/H \rceil$ epochs overall. And we know that among these $\lceil T/H \rceil$ different epochs, at most $c(T)$ of them contain the change points. For the rest of epochs that are free of change points, the cumulative regret could be bounded by the result we just deduced for the stationary case above. And the cumulative regret in any epoch with stationary environment could be bounded as $H^{(p_{z,*}+1)/(p_{z,*}+2)}$. Specifically, we could partition the T rounds into $m = \lceil T/H \rceil$ epochs:

$$[T_1 + 1, T] = [\omega_0 = T_1 + 1, \omega_1) \cup [\omega_1, \omega_2) \cup \cdots \cup [\omega_{m-1}, \omega_m = T + 1),$$

where $\omega_{i+1} = \omega_i + H$ for $i = 0, \ldots, m-2$. Denote all the change points as $T_1 \leq \rho_1 < \cdots < \rho_{c(T)} \leq T$, and then define

$$\Omega = \{ \cup[\omega_i, \omega_{i+1}) : \rho_j \in [\omega_i, \omega_{i+1}), \exists j = 1, \ldots c;\, i = 0, \ldots, m-1\}.$$

Then it holds that $|\Omega| \leq Hc(T)$. Therefore, it holds that

$$R_L(T) \leq \tilde{O}\left(Hc(T) + \left(\frac{T}{H} + 1\right) H^{\frac{p_{z,*}+1}{p_{z,*}+2}} \right) \leq \tilde{O}\left(Hc(T) + \frac{T}{H} \cdot H^{\frac{p_{z,*}+1}{p_{z,*}+2}} \right),$$

where the first part bound the regret of non-stationary epochs and the second part bound that of stationary ones. By taking $H = (T/c(T))^{(p_{z,*}+2)/(p_{z,*}+3)}$, it holds that

$$R_L(T) \leq \tilde{O}\left(T^{\frac{p_{z,*}+2}{p_{z,*}+3}} c(T)^{\frac{1}{p_{z,*}+3}} \right).$$

And this concludes our proof for Theorem 3.4.1. □

B.6. Algorithm 3 with unknown $c(T)$ and $p_{z,*}$

B.6.1. Introduction of Algorithm 11. When both the number of change points $c(T)$ over the total time horizon T and the zooming dimension $p_{z,*}$ are unknown, we could adapt the BOB idea used in Cheung et al. (2019); Zhao et al. (2020) to choose the optimal epoch size H based on

the EXP3 meta algorithm. In the following, we first describe how to use the EXP3 algorithm to choose the epoch size dynamically even if $c(T)$ and $p_{z,*}$ are unknown. Then we present the regret analysis in Theorem B.6.1 and its proof.

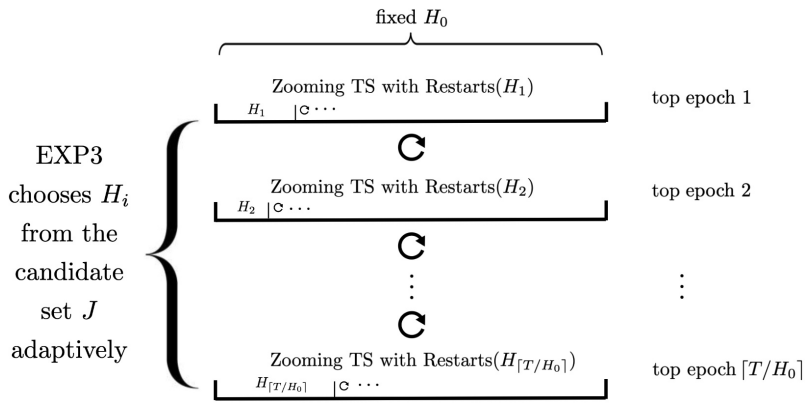


FIGURE B.3. An illustration of Zooming TS algorithm with double restarts when $c(T)$ is agnostic.

Although the zooming dimension $p_{z,*}$ is unknown, it holds that $p_{z,*} \leq p_{\mathrm{c}}$, and hence we could simply use the upper bound of $p_{z,*}$ (denoted as p_{u}) as p_{c} instead (recall p_{c} is the covering dimension). Note that the upper bound $p_{z,*}$ could be more specific when we have some prior knowledge of the reward Lipschitz function $f(\cdot)$: for example, as we mentioned in Appendix B.3, if the function $f(\cdot)$ is known to be C^2-smooth and strongly concave in a neighborhood of its maximum defined in $\mathbb{R}^{p_{\mathrm{c}}}$, it holds that $p_{z,*} \leq p_{\mathrm{c}}/2$ and then we could use $p_{\mathrm{u}} = p_{\mathrm{c}}/2$ as the upper bound. Note that we also use the BOB mechanism in the CDT framework for hyperparameter tuning in Algorithm 4, where we treat the zooming TS algorithm with Restarts as the meta algorithm to select the hyperparameter setting in the upper layer, and then use the selected configuration for the bandit algorithm in the lower layer. However, here we would use BOB mechanism differently: we firstly divide the total horizon T into several epochs of the same length H_0 (named top epoch), where in each top epoch we would restart the Algorithm 3. And in the $i-$th top epoch the restarting length H_i (named bottom epoch) of Algorithm 3 could be chosen from the set $J = \{J_i := \lceil k \rceil \,:\, k \geq 1, k = H_0/2^{i-1}, i = 1, 2, \dots\}$, where the chosen bottom epoch size could be adaptively tuned by using EXP3 as the meta algorithm. Here we restart the zooming TS algorithm from two perspectives, where we first restart the zooming TS algorithm with Restarts (Algorithm 3) in each top epoch of some fixed length H_0, and then for each top epoch the restarting length H_i for Algorithm 3 would be tuned on the fly based on the

previous observations (Cheung et al., 2019). Therefore, we would name this method Zooming TS algorithm with Double Restarts.

As for how to choose the bottom epoch size H_i in each top epoch of length H_0, we implement a two-layer framework: In the upper layer, we use the adversarial MAB algorithm EXP3 to pull the candidate from $J = \{J_i\}$. And then in the lower layer we use it as the bottom epoch size for Algorithm 3. When a top epoch ends, we would update the components in EXP3 based on the rewards witnessed in this top epoch. The illustration of this double restarted strategy is depicted in Figure B.3. And the detailed procedure is shown in Algorithm 11.

THEOREM B.6.1. *By using the (top) epoch size as* $H_0 = \lceil T^{(p_u+2)/(p_u+4)} \rceil$, *the expected total regret of our Zooming TS algorithm with Double Restarts (Algorithm 11) under the switching environment over time* T *could be bounded as*

$$R_L(T) \leq \tilde{O}\left(T^{\frac{p_u+2}{p_u+3}} \cdot \max\left\{c(T)^{\frac{1}{p_u+3}}, T^{\frac{1}{(p_u+3)(p_u+4)}}\right\}\right).$$

Specifically, it holds that

$$R_L(T) \leq \begin{cases} T^{\frac{p_u+2}{p_u+3}} c(T)^{\frac{1}{p_u+3}}, & c(T) \geq T^{\frac{1}{p_u+4}}, \\ T^{\frac{p_u+3}{p_u+4}}, & c(T) < T^{\frac{1}{p_u+4}}, \end{cases}$$

where $p_u \leq p_c$ *is the upper bound of* $p_{z,*}$.

Therefore, we observe that if $c(T)$ is large enough, we could obtain the same regret bound as in Theorem 3.4.1 given $p_{z,*}$.

B.6.2. Proof of Theorem B.6.1.

PROOF. The proof of Theorem B.6.1 relies on the recent usage of the BOB framework that was firstly introduced in Cheung et al. (2019) and then widely used in various bandit-based model selection work (Ding et al., 2022a; Zhao et al., 2020). To be consistent we would use the notations in Algorithm 11 in this proof, and we would also recall these notations here for readers' convenience: for the i-th bottom epoch, we assume the candidate H_{j_i} is pulled from the set J in the beginning, where j_i is the index of the pulled candidate. At round t, given the current bottom epoch length H_{j_i} for some i, we pull the arm $v_t(H_{j_i}) \in A$ and then collect the stochastic reward Y_t. We also define $c_i(T)$ as the number of change points during each top epoch, and hence it naturally holds that

$\sum_{i=1}^{\lceil T/H_0 \rceil} c_i(T) = c(T)$. Given these notations, the expected cumulative regret could be decomposed into the following two parts:

$$
\begin{aligned}
R_L(T) = \mathbb{E}\left[\sum_{t=1}^{T} f_t(v_t^*) - f_t(v_t)\right] &= \mathbb{E}\left[\sum_{i=1}^{\lceil T/H_0 \rceil} \sum_{t=(i-1)H_0+1}^{\min\{T, iH_0\}} f_t(v_t^*) - f_t(v_t(H_{j_i}))\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{i=1}^{\lceil T/H_0 \rceil} \sum_{t=(i-1)H_0+1}^{\min\{T, iH_0\}} f_t(v_t^*) - f_t(v_t(H^*))\right]}_{\text{Quantity (I)}} \\
&+ \underbrace{\mathbb{E}\left[\sum_{i=1}^{\lceil T/H_0 \rceil} \sum_{t=(i-1)H_0+1}^{\min\{T, iH_0\}} f_t(v_t(H^*)) - f_t(v_t(H_{j_i}))\right]}_{\text{Quantity (II)}},
\end{aligned} \tag{B.10}
$$

where H^* could be any restarting period in J, and we expect it could approximate the optimal choice $H^{\text{opt}} = (T/c(T))^{(p_u+2)/(p_u+3)}$ in Theorem 3.4.1. (Here we replace $p_{z,*}$ by p_u in Theorem 3.4.1 since the underlying $p_{z,*}$ is mostly unspecified in reality.) According to the proof of Theorem 3.4.1 in Appendix B.7, the Quantity (I) could be bounded as:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{\lceil T/H_0 \rceil} \sum_{t=(i-1)H_0+1}^{\min\{T, iH_0\}} f_t(v_t^*) - f_t(v_t(H^*))\right] &\leq \sum_{i=1}^{\lceil T/H_0 \rceil} H^* c_i(T) + \frac{H_0}{H^*}(H^*)^{\frac{p_u+2}{p_u+4}} \\
&= H^* c(T) + T(H^*)^{-\frac{1}{p_u+2}}
\end{aligned}
$$

However, it is clear that each candidate in J could at most be the length of top epoch size H_0, which we set to be $\lceil T^{(p_u+2)/(p_u+4)} \rceil$, and hence it would be more challenging if the optimal choice $H^{\text{opt}} = (T/c(T))^{(p_u+2)/(p_u+3)}$ is larger than H_0. To deal with this issue, we bound the expected cumulative regret in two different cases separately:

(1) If $H^{\text{opt}} = (T/c(T))^{(p_u+2)/(p_u+3)} \leq H_0$, which is equivalent to

$$
\left(\frac{T}{c(T)}\right)^{\frac{p_u+2}{p_u+3}} \leq H_0 \Leftrightarrow \left(\frac{T}{c(T)}\right)^{\frac{p_u+2}{p_u+3}} \leq T^{\frac{p_u+2}{p_u+4}} \Leftrightarrow c(T) \geq T^{\frac{1}{p_u+4}},
$$

Algorithm 11 Zooming TS algorithm with Double Restarts

Input: Time horizon T, space A, upper bound $p_{\mathrm{u}} \leq p_{\mathrm{c}}$.

the (top) epoch size $H_0 = \lceil T^{(p_{\mathrm{u}}+2)/(p_{\mathrm{u}}+4)} \rceil$, $N = \lceil \log_2(H_0) \rceil + 1$, $J = \{H_i = \lceil H_0/2^{i-1} \rceil\}_{i=1}^{N}$.

Initialize the exponential weights $w_j(1) = 1$ for $j = 1, \ldots, |J|$.

Initialize the exploration parameter for the EXP3 algorithm as $\alpha = \min\left\{1, \sqrt{\frac{|J|\log(|J|)}{(e-1)\lceil T/H_0 \rceil}}\right\}$.

for $i = 1$ **to** $\lceil T/H_0 \rceil$ **do**

Update probability distribution for selecting candidates in J based on EXP3 as:

$$p_j(i) = \frac{\alpha}{|J|} + (1-\alpha)\frac{w_j(i)}{\sum_{k=1}^{|J|} w_k(i)},\ j = 1, \ldots, |J|.$$

Pull j_i from $\{1, 2, \ldots, |J|\}$ according to the probability distribution $\{p_j(i)\}_{j=1}^{|J|}$.

Run Zooming TS algorithm with Restarts using the (bottom) epoch size H_{j_i} for $t = (i-1)H_0 + 1$ **to** $\min\{T, iH_0\}$, and collect the pulled arm $v_t(H_{j_i})$ and reward Y_t at each iteration.

Update components in EXP3: $r_j(i) = 0$ for all $j \neq j_i$; $r_j(i) = \sum_{k=(i-1)H_0+1}^{\min\{T,iH_0\}} Y_k/p_j(i)$ if $j = j_i$, and then

$$w_j(i+1) = w_j(i)\exp\left(\frac{\alpha}{|J|} r_j(i)\right),\ j = 1, \ldots, |J|.$$

then we know that there exists some $H^+ \in J$ such that $H^+ \leq (T/c(T))^{(p_u+2)/(p_u+3)} \leq 2H^+$. By setting $H^* = H^+$, the Quantity (I) could be bounded as:

$$\begin{aligned}\text{Quantity (I)} &= \tilde{O}\left(H^+ c(T) + T(H^*)^{-\frac{1}{p_u+2}}\right) \\ &= \tilde{O}\left(H^{\mathrm{opt}} c(T) + T(H^{\mathrm{opt}})^{-\frac{1}{p_u+2}}\right) = \tilde{O}\left(T^{\frac{p_u+2}{p_u+3}} c(T)^{\frac{1}{p_u+3}}\right).\end{aligned}$$

For the Quantity (II), we could bound it based on the results in Auer et al. (2002b). Specifically, from Corollary 3.2 in Auer et al. (2002b), the expected cumulative regret of EXP3 could be upper bounded by $2Q\sqrt{(e-1)LK\ln(K)}$, where Q is the maximum absolute sum of rewards in any epoch, L is the number of rounds and K is the number of arms. Under our setting, we can set $Q = H_0, L = \lceil T/H_0 \rceil$ and $K = |J| = O(\ln(H_0))$. So we could bound Quantity (II) as:

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{\lceil T/H_0 \rceil} \sum_{t=(i-1)H_0+1}^{\min\{T,iH_0\}} f_t(v_t(H^*)) - f_t(v_t(H_{j_i}))\right] \leq 2\sqrt{e-1}H_0\sqrt{\frac{T}{H_0}|J|\ln(|J|)} = \tilde{O}(\sqrt{TH_0}) \\ &= \tilde{O}\left(T^{\frac{p_u+3}{p_u+4}}\right) = \tilde{O}\left(T^{\frac{p_u+2}{p_u+3}} T^{\frac{1}{(p_u+3)(p_u+4)}}\right) = \tilde{O}\left(T^{\frac{p_u+2}{p_u+3}} c(T)^{\frac{1}{p_u+3}}\right), \qquad \text{(B.11)}\end{aligned}$$

where we have the last equality since we assume that $c(T) \geq T^{1/(p_u+4)}$. Therefore, we have finished the proof for this case. (2) If $H^{\mathrm{opt}} = (T/c(T))^{(p_u+2)/(p_u+3)} > H_0$, which is equivalent to

$$\left(\frac{T}{c(T)}\right)^{\frac{p_u+2}{p_u+3}} > H_0 \Leftrightarrow \left(\frac{T}{c(T)}\right)^{\frac{p_u+2}{p_u+3}} > T^{\frac{p_u+2}{p_u+4}} \Leftrightarrow c(T) < T^{\frac{1}{p_u+4}},$$

then we know that H^{opt} is greater than all candidates in J, which means that we could not bound the Quantity (I) based on the previous argument. By simply using $H^* = H_0$, it holds that

$$\text{Quantity (I)} = \tilde{O}\left(H_0 c(T) + T \cdot H_0^{-\frac{1}{p_u+2}}\right) = \tilde{O}\left(T^{\frac{p_u+3}{p_u+4}}\right).$$

For Quantity (II), based on Eqn. (B.11), we have

$$\text{Quantity (II)} = \tilde{O}\left(T^{\frac{p_u+3}{p_u+4}}\right).$$

Combining the case (1) and (2), it holds that

$$R_L(T) \leq \begin{cases} T^{\frac{p_u+2}{p_u+3}} c(T)^{\frac{1}{p_u+3}}, & c(T) \geq T^{\frac{1}{p_u+4}}, \\ T^{\frac{p_u+3}{p_u+4}}, & c(T) < T^{\frac{1}{p_u+4}}. \end{cases}$$

And this concludes our proof. □

B.7. Analysis of Theorem 3.4.2

B.7.1. Additional Lemma.

LEMMA B.7.1. *(Proposition 1 in Li et al. (2017)) Define $V_{n+1} = \sum_{t=1}^{n} X_t X_t^T$, where X_t is drawn IID from some distribution in unit ball $\mathbb{B}^d$. Furthermore, let $\Sigma := E[X_t X_t^T]$ be the second moment matrix, let $B, \delta_2 > 0$ be two positive constants. Then there exists positive, universal constants C_1 and C_2 such that $\lambda_{\min}(V_{n+1}) \geq B$ with probability at least $1 - \delta_2$, as long as*

$$n \geq \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta_2)}}{\lambda_{\min}(\Sigma)}\right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}.$$

LEMMA B.7.2. *(Theorem 2 in Abbasi-Yadkori et al. (2011)) For any $\delta < 1$, under our problem setting in Section 3.3, it holds that for all $t > 0$,*

$$\left\|\hat{\theta}_t - \theta^*\right\|_{V_t} \leq \beta_t(\delta),$$
$$\forall x \in \mathbb{R}^d, |x^\top(\hat{\theta}_t - \theta^*)| \leq \|x\|_{v_t^{-1}} \beta_t(\delta),$$

with probability at least $1-\delta$*, where*

$$\beta_t(\delta) = \sigma\sqrt{\log\left(\frac{(\lambda+t)^d}{\delta^2\lambda^d}\right)} + \sqrt{\lambda}S.$$

In this subsection we denote $\alpha^*(\delta) \coloneqq \beta_T(\delta)$.

LEMMA B.7.3. *(Filippi et al. (2010)) Let* $\lambda > 0$*, and* $\{x_i\}_{i=1}^t$ *be a sequence in* $\mathbb{R}^d$ *with* $\|x_i\| \leq 1$*, then we have*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \leq 2\log\left(\frac{\det(V_{t+1})}{\det(\lambda I)}\right) \leq 2d\log\left(1+\frac{t}{\lambda}\right),$$

$$\sum_{s=1}^{t} \|x_s\|_{V_s^{-1}} \leq \sqrt{T\left(\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}}\right)} \leq \sqrt{2dt\log\left(1+\frac{t}{\lambda}\right)}.$$

LEMMA B.7.4. *(Agrawal & Goyal (2013)) For a Gaussian random variable* Z *with mean* m *and variance* σ^2*, for any* $z \geq 1$*,*

$$P(|Z-m| \geq z\sigma) \leq \frac{1}{\sqrt{\pi}z}e^{-z^2/2}.$$

LEMMA B.7.5 (Adapted from Lemma B.7.2). *For any* $\delta < 1$*, under our problem setting in Section 3.3 with the regularization hyper-parameter* $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ $(\lambda_{\min} > 0)$*, it holds that for all* $t > 0$*,*

$$\left\|\hat{\theta}_t - \theta^*\right\|_{V_t} \leq \beta_t(\delta),$$

$$\forall x \in \mathbb{R}^d, |x^\top(\hat{\theta}_t - \theta^*)| \leq \|x\|_{V_t^{-1}}\beta_t(\delta),$$

with probability at least $1-\delta$*, where*

$$\beta_t(\delta) = \sigma\sqrt{\log\left(\frac{(\lambda_{\min}+t)^d}{\delta^2\lambda_{\min}^d}\right)} + \sqrt{\lambda_{\max}}S.$$

PROOF. The proof of this Lemma is trivial given Lemma B.7.2. For any $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, according to Lemma B.7.2 it holds that, for all $t > 0$,

$$\left\|\hat{\theta}_t - \theta^*\right\|_{V_t} \leq \beta_t(\delta),$$

$$\forall x \in \mathbb{R}^d, |x^\top(\hat{\theta}_t - \theta^*)| \leq \|x\|_{V_t^{-1}}\beta_t(\delta),$$

with probability at least $1-\delta$, where

$$\beta_t(\delta) = \sigma\sqrt{\log\left(\frac{(\lambda+t)^d}{\delta^2\lambda^d}\right)} + \sqrt{\lambda}S \leq \sigma\sqrt{\log\left(\frac{(\lambda_{\min}+t)^d}{\delta^2\lambda_{\min}^d}\right)} + \sqrt{\lambda_{\max}}S.$$

□

B.7.2. Proof of Theorem 3.4.2. Recall the partition of the cumulative regret as:

$$R(T) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T_1}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t{}^\top\theta^*)\right)\right]}_{\text{Quantity (A)}} + \underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*)\right)\right]}_{\text{Quantity (B)}}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*\right) - \mu(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*)\right)\right]}_{\text{Quantity (C)}}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*\right) - \mu(x_t(\alpha(i_t)|\mathcal{F}_t)^\top\theta^*)\right)\right]}_{\text{Quantity (D)}}.$$

For Quantity (A), it could be easily bounded by the length of warming up period as:

$$\mathbb{E}\left[\sum_{t=1}^{T_1}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t{}^\top\theta^*)\right)\right] \leq T_1 = O\left(T^{\frac{2}{p+3}}\right) \leq O\left(T^{\frac{p+2}{p+3}}\right). \tag{B.12}$$

For Quantity (B), it depicts the cumulative regret of the contextual bandit algorithm that runs with the theoretical optimal hyperparameter $\alpha^*(t)$ all the time. Therefore, if we implement any state-of-the-arm contextual generalized linear bandit algorithms (e.g. Filippi et al. (2010); Li et al. (2010; 2017)), it holds that

$$\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu(x_{t,*}^\top\theta^*) - \mu(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*)\right)\right] \leq \tilde{O}(\sqrt{T-T_1}) = \tilde{O}(\sqrt{T}). \tag{B.13}$$

For Quantity (C), it represents the cumulative difference of regret under the theoretical optimal hyperparameter combination $\alpha^*(t)$ with two lines of history $\mathcal{F}_t$ and $\mathcal{F}_t^*$. Note for most GLB algorithms, the most significant hyperparameter is the exploration rate, which directly affect the

decision-making process. Regarding the regularization hyperparameter λ, it is used to make V_t invertible and hence would be set to 1 in practice. And in the long run it would not be influential. Moreover, there is commonly no theoretical optimal value for λ, and it could be set to an arbitrary constant in order to obtain the $\tilde{O}(\sqrt{T})$ bound of regret. For theoretical proof, this hyperparameter (λ) is also not significant: for example, if the search interval for λ is $[\lambda_{\min}, \lambda_{\max}]$, then we can easily modify the Lemma B.7.3 as:

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \leq 2\log\left(\frac{\det(V_{t+1})}{\det(\lambda I)}\right) \leq 2d\log\left(1+\frac{t}{\lambda_{\min}}\right),$$

$$\sum_{s=1}^{t} \|x_s\|_{V_s^{-1}} \leq \sqrt{T\left(\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}}\right)} \leq \sqrt{2dt\log\left(1+\frac{t}{\lambda_{\min}}\right)}.$$

We will offer a more detailed explanation to this fact in the following proof of bounding Quantity (C). Furthermore, other parameters such as the stepsize in a loop of gradient descent will not be crucial either since the final result would be similar after the convergence criterion is met. Therefore, w.l.o.g we would only assume there is only one exploration rate hyperparameter here to bound Quantity (C). Recall that $\alpha(t)$ is the combination of all hyperparameters, and hence we could denote this exploration rate hyperparameter as $\alpha(t)$ in this part since there is no more other hyperparameter. Here we would use LinUCB and LinTS for the detailed proof, and note that regret bound of all other UCB and TS algorithms could be similarly deduced. We first reload some notations: recall we denote $V_t = \lambda I + \sum_{i=1}^{t-1} x_i x_i^\top, \hat{\theta}_t = V_t^{-1} \sum_{i=1}^{t-1} x_i y_i$ where x_t is the arm we pulled at round t by using our tuned hyperparameter $\alpha(i_t)$ and the history based on our framework all the time. And we denote

$$X_t = \arg\max_{x \in \mathcal{X}_t} x^\top \hat{\theta}_t + \alpha^*(t) \|x\|_{V_t^{-1}}$$

Similarly, we denote $\tilde{V}_t = \lambda I + \sum_{i=1}^{t-1} \tilde{X}_i \tilde{X}_i^\top, \tilde{\theta}_t = \tilde{V}_t^{-1} \sum_{i=1}^{t-1} \tilde{X}_i \tilde{y}_i$, where $\tilde{X}_t$ is the arm we pulled by using the theoretical optimal hyperparameter $\alpha^*(t)$ under the history of always using $\{\alpha^*(s)\}_{s=1}^{t-1}$, and $\tilde{y}_t$ is the corresponding payoff we observe at round t. Therefore, it holds that,

$$\tilde{X}_t = \arg\max_{x \in \mathcal{X}_t} x^\top \tilde{\theta}_t + \alpha^*(t) \|x\|_{\tilde{V}_t^{-1}}.$$

By using these new definitions, the Quantity (C) could be formulated as:

$$\underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t^*)^\top\theta^*\right)-\mu(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*)\right)\right]}_{\text{Quantity (C)}}=\mathbb{E}\left[\sum_{t=T_1+1}^{T}\mu(\tilde{X}_t^\top\theta^*)-\mu(X_t^\top\theta^*)\right]$$

For LinUCB, since the Lemma B.7.2 holds for any sequence $(x_1,\ldots,x_t)$, and hence we have that with probability at least $1-\delta$,

$$\left\|\hat{\theta}-\theta\right\|_{V_t}\leq\beta_t(\delta)\leq\alpha^*(T,\delta), \tag{B.14}$$

where

$$\beta_t(\delta)=\sigma\sqrt{\log\left(\frac{(\lambda+t)^d}{\delta^2\lambda^d}\right)}+\sqrt{\lambda}S=\alpha^*(t).$$

And we will omit δ for simplicity. For LinUCB, we have that

$$\begin{aligned}X_t^\top\hat{\theta}_t+\alpha^*(t)\left\|X_t\right\|_{V_t^{-1}}&\geq\tilde{X}_t^\top\hat{\theta}_t+\alpha^*(t)\left\|\tilde{X}_t\right\|_{V_t^{-1}}\\&\geq\tilde{X}_t^\top\theta^*+\alpha^*(t)\left\|\tilde{X}_t\right\|_{V_t^{-1}}+\tilde{X}_t^\top(\hat{\theta}_t-\theta^*)\geq\tilde{X}_t^\top\theta^*.\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}X_t^\top\theta^*+\alpha^*(t)\left\|X_t\right\|_{V_t^{-1}}+X_t^\top(\hat{\theta}_t-\theta^*)&\geq\tilde{X}_t^\top\theta^*\\X_t^\top\theta^*+2\alpha^*(t)\left\|X_t\right\|_{V_t^{-1}}&\geq\tilde{X}_t^\top\theta^*,\end{aligned}$$

which implies that

$$(\tilde{X}_t-X_t)^\top\theta^*\leq2\alpha^*(T)\left\|X_t\right\|_{V_t^{-1}}.$$

By Lemma B.7.3 and choosing $T_1=T^{2/(p+3)}$, it holds that,

$$\sum_{t=T_1+1}^{T}\left\|X_t\right\|_{V_t^{-1}}\leq\sum_{t=T_1+1}^{T}\left\|X_t\right\|\sqrt{\lambda_{\min}(V_t)}=O(T\times T^{-1/(p+3)})=O(T^{(p+2)/(p+3)}).$$

And then it holds that,

$$\sum_{t=T_1+1}^{T}\left(\tilde{X}_t^T\theta-X_t\theta\right)=\tilde{O}\left(\alpha^*(T)\sum_{t=T_1+1}^{T}\left\|\tilde{X}_t\right\|_{V_t^{-1}}\right)=\tilde{O}(T^{(p+2)/(p+3)}). \tag{B.15}$$

Note $\beta_t(\delta)$ contain the regularizer parameter λ, and it's often set to some constant (e.g. 1) in practice. If we tune λ in the search interval $[\lambda_{\min}, \lambda_{\max}]$, then we can still have the identical bound as in Eqn. (B.14) by using the fact that

$$\beta_t(\delta) = \sigma\sqrt{\log\left(\frac{(\lambda+t)^d}{\delta^2\lambda^d}\right)} + \sqrt{\lambda}S \leq \sigma\sqrt{\log\left(\frac{(\lambda_{\min}+t)^d}{\delta^2\lambda_{\min}^d}\right)} + \sqrt{\lambda_{\max}}S.$$

This result is deduced in our Lemma B.7.5, which implies that tuning the regularizer hyperparameter would not affect the order of final regret bound in Eqn. (B.15). Therefore, as we mentioned earlier, we could only consider the exploration rate as the unique hyperparameter for theoretical analysis.

For LinTS, we have that

$$\begin{aligned}
X_t^\top\hat{\theta}_t + \alpha^*(T)\left\|X_t\right\|_{V_t^{-1}} Z_t &\geq \tilde{X}_t^\top\hat{\theta}_t + \alpha^*(T)\left\|\tilde{X}_t\right\|_{V_t^{-1}}\tilde{Z}_t \\
&\geq \tilde{X}_t^\top\theta^* + \alpha^*(T)\left\|\tilde{X}_t\right\|_{V_t^{-1}}\tilde{Z}_t + \tilde{X}_t^\top(\hat{\theta}_t - \theta^*) \\
&\geq \tilde{X}_t^\top\theta^* + \alpha^*(T)\left\|\tilde{X}_t\right\|_{V_t^{-1}}\tilde{Z}_t + \left\|\tilde{X}_t\right\|_{V_t^{-1}}\left\|\hat{\theta}_t - \theta^*\right\|_{V_t} \\
&\geq \tilde{X}_t^\top\theta + (\alpha^*(T)\tilde{Z}_t - \alpha^*(T))\left\|\tilde{X}_t\right\|_{V_t^{-1}},
\end{aligned}$$

where Z_t and $Z_{t,*}$ are IID normal random variables, $\forall t$. And then we could deduce that

$$\begin{aligned}
X_t^\top\theta^* + \alpha^*(T)\left\|X_t\right\|_{V_t^{-1}} Z_t + X_t^\top(\hat{\theta}_t - \theta^*) &\geq \tilde{X}_t^\top\theta + (\alpha^*(T)\tilde{Z}_t - \alpha^*(T))\left\|\tilde{X}_t\right\|_{V_t^{-1}} \\
X_t^\top\theta^* + \alpha^*(T)\left\|X_t\right\|_{V_t^{-1}} Z_t + \alpha^*(T)\left\|X_t\right\|_{V_t^{-1}} &\geq \tilde{X}_t^\top\theta + (\alpha^*(T)\tilde{Z}_t - \alpha^*(T))\left\|\tilde{X}_t\right\|_{V_t^{-1}} \\
(\tilde{X}_t - X_t)^\top\theta^* &\leq (\alpha^*(T) - \alpha^*(T)\tilde{Z}_t)\left\|\tilde{X}_t\right\|_{V_t^{-1}} + (\alpha^*(T) + \alpha^*(T)Z_t)\left\|X_t\right\|_{V_t^{-1}} := K_t
\end{aligned}$$

where K_t is normal random variable with

$$\mathbb{E}(K_t) \leq 2\alpha(T)T^{-1/(p+3)},\ \mathrm{SD}(K_t) \leq \sqrt{2}\alpha^*T^{-1/(p+3)}.$$

Consequently, we have

$$\sum_{t=T_1+1}^{T}\left(\tilde{X}_t^T\theta - X_t^T\theta\right) \leq \sum_{t=T_1+1}^{T} K_t := K$$

$$\mathbb{E}(K) = 2\alpha^*(T)T^{(p+2)/(p+3)} = \tilde{O}(T^{\frac{p+2}{p+3}}),\ \mathrm{SD}(K) \leq \sqrt{2}\alpha^*T^{\frac{p+1}{2p+6}} = O(T^{\frac{p+1}{2p+6}}).$$

Based on Lemma B.7.4, we have

$$P\left(\sum_{t=T_1+1}^{T}\left(\tilde{X}_t^T\theta - X_t^T\theta\right) \geq K > (2\alpha^* + \sqrt{2})T^{\frac{p+2}{p+3}}\right) \leq \frac{1}{c\sqrt{\pi}\sqrt{T}}e^{-c^2T/2}. \tag{B.16}$$

This probability upper bound is minimal and negligible, which means the bound on its expected value (Quantity (C)) could be easily deduced. Note we could use this procedure to bound the regret for other UCB and TS bandit algorithms, since most of the proof for GLB algorithms are closely related to the rate of $\sum_{t=T_1+1}^{T}\|X_t\|_{V_t^{-1}}$ and the consistency of $\hat{\theta}_t$. In conclusion, we have that Quantity (C) could be upper bounded by the order $\tilde{O}(T^{\frac{p+2}{p+3}})$.

For Quantity (D), which is the extra regret we paid for hyperparameter tuning in theory. Recall we assume $\mu(x_t(\alpha|\mathcal{F}_t)^\top\theta^*) = g_t(\alpha) + \eta_{\mathcal{F}_t,\alpha}$ for some time-dependent Lipschitz function g_t. And $(\eta_{\mathcal{F}_t,\alpha} - \mathbb{E}[\eta_{\mathcal{F}_t,\alpha}])$ is IID sub-Gaussian with parameter τ^2 where $\mathbb{E}[\eta_{\mathcal{F}_t,\alpha}]$ depends on the history $\mathcal{F}_t$. Denote $\nu_{\mathcal{F}_t,\alpha} = \eta_{\mathcal{F}_t,\alpha} - \mathbb{E}[\eta_{\mathcal{F}_t,\alpha}]$ is the IID sub-Gaussian random variable with parameter τ^2, then we have that

$$y_t = g_t(\alpha(i_t)) + \nu_{\mathcal{F}_t,\alpha(i_t)} + E[\eta_{\mathcal{F}_t,\alpha(i_t)}] + \epsilon_t$$

Since $\nu_{\mathcal{F}_t,\alpha(i_t)}, \epsilon_t$ is IID sub-Gaussian random variable independent with $\mathcal{F}_t$, we denote $\tilde{\epsilon}_{\mathcal{F}_t,\alpha(i_t)} = \nu_{\mathcal{F}_t,\alpha(i_t)} + \epsilon_t$ as the IID sub-Gaussian noise with parameter $\tau^2 + \sigma^2$. And then we have

$$y_t = g_t(\alpha(i_t)) + E[\eta_{\mathcal{F}_t,\alpha(i_t)}] + \tilde{\epsilon}_{\mathcal{F}_t,\alpha(i_t)}, \quad \mathbb{E}(y_t) = g_t(\alpha(i_t)) + E[\eta_{\mathcal{F}_t,\alpha(i_t)}]$$
$$\mu(x_t(\alpha|\mathcal{F}_t)^\top\theta^*) = g_t(\alpha) + E[\eta_{\mathcal{F}_t,\alpha}].$$

For Quantity (D), recall it could be formulated as:

$$\underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*\right) - \mu(x_t(\alpha(i_t)|\mathcal{F}_t)^\top\theta^*\right)\right)\right]}_{\text{Quantity (D)}}.$$

Since both terms in Quantity (D) are based on the same line of history $\mathcal{F}_t$ at iteration t, and the value of $E[\eta_{\mathcal{F}_t,\alpha}]$ only depends on the history filtration $\mathcal{F}_t$ but not the value of α. Therefore, it

holds that

$$\underbrace{\mathbb{E}\left[\sum_{t=T_1+1}^{T}\left(\mu\left(x_t(\alpha^*(t)|\mathcal{F}_t)^\top\theta^*\right)-\mu(x_t(\alpha(i_t)|\mathcal{F}_t)^\top\theta^*\right)\right)\right]}_{\text{Quantity (D)}}=\sum_{t=T_1+1}^{T}g_t(\alpha^*(t))-\mathbb{E}[g_t(\alpha(i_t))]$$

$$\leq\sum_{t=T_1+1}^{T}\sup_{\alpha\in A}g_t(\alpha)-\mathbb{E}[g_t(\alpha(i_t))].$$

Therefore, Quantity (D) could be regarded as the cumulative regret of a non-stationary Lipschitz bandit and the noise is IID sub-Gaussian with parameter $\tau_0^2=(\tau^2+\sigma^2)$. We assume that, under the *switching* environment, the Lipschitz function $g_t(\cdot)$ would be piecewise stationary and the number of change points is of scale $\tilde{O}(1)$. Therefore, Quantity (D) can be upper bounded the cumulative regret of our Zooming TS algorithm with restarted strategy given $c(T)=\tilde{O}(1)$. By choosing $T_2=(T-T_1)^{(p+2)/(p+3)}=\Theta(T^{(p+2)/(p+3)})$, and according to Theorem 3.4.1, it holds that,

$$\sum_{t=T_1+1}^{T}\sup_{\alpha\in A}g_t(\alpha)-\mathbb{E}[g_t(\alpha(i_t))]\leq\tilde{O}\left(T^{\frac{p+2}{p+3}}\right). \tag{B.17}$$

By combining the results deduced in Eqn. (B.12), Eqn. (B.13), Eqn. (B.15) (or Eqn. (B.16)) and Eqn. (B.17), we finish the proof of Theorem 3.4.2 for linear bandits. For generalized linear bandits, under the default and standard assumption in the generalized linear bandit literature that the derivative of $\mu(\cdot)$ could be upper bounded by some constant given $|x|\leq S$, the regret could be bounded by further multiplying a constant in the same order.

□

APPENDIX C

Appendix for Chapter 4

C.1. Clarification about σ_0^2

It is a common assumption that the random noise η_t in Eqn. (4.2) is a sub-Gaussian random variable in GLM, and here we would briefly explain this assumption.

LEMMA C.1.0.1. (Sub-Gaussian property for GLM residuals) *For any generalized linear model with a probability density function or probability mass function of the canonical form*

$$f(Y = y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right),$$

where the function $b(\cdot)$ is Lipschitz with parameter k_μ. Then we can conclude that the random variable $(Y - b'(\theta)) = (Y - \mu(\theta))$ satisfies sub-Gaussian property with parameter at most $\sqrt{\phi\, k_\mu}$.

Proof. We prove the Lemma C.1.0.1 based on its definition directly. For any $t \in \mathbb{R}$, we have:

$$\begin{aligned}
\mathbb{E}\left[\exp\{t\left(Y - b'(\theta)\right)\}\right] &= \int_{-\infty}^{+\infty} \exp\left\{t(y - b'(\theta)) + \frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\} dy \\
&= \int_{-\infty}^{+\infty} \exp\left\{\frac{(\theta + \phi t)y - b(\theta + \phi t)}{\phi} + c(y, \phi)\right\} \\
&\qquad \times \exp\left\{\frac{b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right\} dy \\
&= \exp\left\{\frac{b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right\} \\
&\overset{(i)}{=} \exp\left\{\frac{t^2 \phi\, b''(\theta + \delta\phi t)}{2}\right\} \leq \exp\left\{\frac{t^2\, \phi\, k_\mu}{2}\right\} := \exp\left\{\frac{t^2\, \sigma_0^2}{2}\right\},
\end{aligned}$$

where the equality (i) is based on the remainder of Taylor expansion. □

This theorem tells us that it is a standard assumption that the noise η_t in Eqn. (4.2) is a sub-Gaussian random variable. For instance, if we assume the inverse link function $\mu(\cdot)$ is globally Lipschitz with parameter k_μ, we can simply take $\sigma_0^2 = k_\mu \phi$. And this assumption also widely holds under a class of GLMs such as the most popular Logistic model.

C.2. Proof of Theorem 4.4.1

C.2.1. Useful Lemmas.

LEMMA C.2.0.1. (Sub-guassian moment bound) *For sub-Gaussian random variable X with parameter σ^2, i.e.*

$$\mathbb{E}(\exp(sX)) \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}.$$

Then we have $Var(X) = \mathbb{E}(X^2) \leq 4\sigma^2$.

Proof. It holds that,

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_0^{+\infty} P(X^2 > t)\, dt \\
&= \int_0^{+\infty} P(|X| > \sqrt{t})\, dt \\
&\leq 2\int_0^{+\infty} \exp\left(\frac{-t^2}{4\sigma^2}\right) dt \\
&= 4\sigma^2 \int_0^{+\infty} e^{-u}\, du, \quad u = t/(2\sigma^2) \\
&= 4\sigma^2
\end{aligned}$$

□

LEMMA C.2.0.2. *(*Generalized Stein's Lemma, *(Diaconis et al., 2004)) For a random variable X with continously differentiable density function $p : \mathbb{R}^d \to \mathbb{R}$, and any continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. If the expected values of both $\nabla f(X)$ and $f(X) \cdot S(X)$ regarding the density p exist, then they are identical, i.e.*

$$\mathbb{E}[f(X) \cdot S(X)] = \mathbb{E}[\nabla f(X)].$$

This is a very famous result in the area of Stein's method, and we would omit its proof.

LEMMA C.2.0.3. *(Minsker (2018)) Let $Y_1, \ldots, Y_n \in \mathbb{R}^{d_1 \times d_2}$ be a sequence of independent real random matrices, and assume that*

$$\sigma_n^2 \geq \max\left(\left\|\sum_{j=1}^n \mathbb{E}(Y_j Y_j^\top)\right\|_{op}, \left\|\sum_{j=1}^n \mathbb{E}(Y_j^\top Y_j)\right\|_{op}\right).$$

Then for any $t \in \mathbb{R}^+$ and $\nu \in \mathbb{R}^+$, it holds that,

$$P\left(\left\|\sum_{j=1}^{n} \tilde{\psi}_\nu(Y_j) - \sum_{j=1}^{n} \mathbb{E}(Y_j)\right\|_{op} \geq t\sqrt{n}\right) \leq 2(d_1+d_2)\exp\left(\nu t\sqrt{n} + \frac{\nu^2\sigma_n^2}{2}\right)$$

The detailed proof of this lemma is based on a series of work proposed in Minsker (2018). And we would omit it here as well. Based on Lemma C.2.0.2 and C.2.0.3, we would propose the following Lemma C.2.0.4 adapted from the work in Yang et al. (2017). And this Lemma serves as a crux for the proof of Theorem 4.4.1.

LEMMA C.2.0.4. *$L : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ is the loss function defined in Eqn. (4.6). Then by setting*

$$t = \sqrt{2(d_1+d_2)M(4\sigma_0^2 + S_f^2)\log\left(\frac{2(d_1+d_2)}{\delta}\right)},$$

$$\nu = \frac{t}{(4\sigma_0^2 + S_f)M(d_1+d_2)\sqrt{T_1}} = \sqrt{\frac{2\log\left(\frac{2(d_1+d_2)}{\delta}\right)}{T_1(d_1+d_2)M(4\sigma_0^2+S_f^2)}},$$

we have with probability at least $1-\delta$, it holds that

$$P\left(\|\nabla L(\mu^*\Theta^*)\|_{op} \geq \frac{2t}{\sqrt{T_1}}\right) \leq \delta,$$

where $\mu^* = \mathbb{E}[\mu'(\langle X, \Theta^*\rangle)] \geq c_\mu > 0$.

Proof. Based on the definition of our loss function $L(\cdot)$ in Eqn. (4.6), we have that

$$\begin{aligned}
\nabla_x L(\mu^*\Theta^*) &= 2\mu^*\Theta^* - \frac{2}{T_1}\sum_{i=1}^{T_1}\tilde{\psi}_\nu(y\cdot S(x)) \\
&= 2\mathbb{E}[\mu'(\langle X_1, \Theta^*\rangle)]\Theta^* - \frac{2}{T_1}\sum_{i=1}^{T_1}\tilde{\psi}_\nu(y_i\cdot S(X_i)) \\
&\overset{\text{(i)}}{=} 2\mathbb{E}[\mu(\langle X_1, \Theta^*\rangle)S(X_1)] - \frac{2}{T_1}\sum_{i=1}^{T_1}\tilde{\psi}_\nu(y_i\cdot S(X_i)) \\
&\overset{\text{(ii)}}{=} 2\left[\mathbb{E}(Y_1\cdot S(X_1)) - \frac{1}{T_1}\sum_{i=1}^{T_1}\tilde{\psi}_\nu(y_i\cdot S(X_i))\right]
\end{aligned}$$

where we have (i) due to the generalized Stein's Lemma (Lemma C.2.0.2), and (ii) comes from the fact that the random noise $\eta_1 = y_1 - \mu(\langle X_1, \Theta^*\rangle)$ is zero-mean and independent with X_1. Therefore,

in order to implement the Lemma C.2.0.3, we can see that it suffices to get σ^2 defined as:

$$\sigma^2 = \max\left(\left\|\sum_{j=1}^{n}\mathbb{E}[y_j^2 S(X_j)S(X_j)^\top]\right\|_{\text{op}}, \left\|\sum_{j=1}^{n}\mathbb{E}[y_j^2 S(X_j)^\top S(X_j)]\right\|_{\text{op}}\right).$$

It holds that,

$$\begin{aligned}
\left\|\sum_{j=1}^{T_1}\mathbb{E}[y_j^2 S(X_j)S(X_j)^\top]\right\|_{\text{op}} &\leq T_1 \times \left\|\mathbb{E}[y_1^2 S(X_1)S(X_1)^\top]\right\|_{\text{op}} \\
&= T_1 \times \left\|\mathbb{E}[(\eta_1 + \mu(\langle X_1, \Theta^*\rangle))^2 S(X_1)S(X_1)^\top]\right\|_{\text{op}} \\
&= T_1 \times \left\|\mathbb{E}[\eta_1^2 S(X_1)S(X_1)^\top] + \mathbb{E}[\mu(\langle X_1, \Theta^*\rangle))^2 S(X_1)S(X_1)^\top]\right\|_{\text{op}} \\
&= T_1 \times \left\|\mathbb{E}(\eta_1^2)\mathbb{E}[S(X_1)S(X_1)^\top] + \mathbb{E}[\mu(\langle X_1, \Theta^*\rangle))^2 S(X_1)S(X_1)^\top]\right\|_{\text{op}} \\
&\overset{\text{(i)}}{\leq} T_1 \times \left\|4\sigma_0^2\,\mathbb{E}[S(X_1)S(X_1)^\top] + S_f^2\,\mathbb{E}[S(X_1)S(X_1)^\top]\right\|_{\text{op}} \\
&= (4\sigma_0^2 + S_f^2)T_1 \times \left\|\mathbb{E}[S(X_1)S(X_1)^\top]\right\|_{\text{op}}
\end{aligned}$$

where the inequality (i) comes from the fact that $|\mu(\langle X_1, \Theta^*\rangle)| \leq S_f$, and $S(X_1)S(X_1)^\top$ is always positive semidefinite. Next, without loss of generality we assume X_i are independent across rows, since if X_i are independent across columns we can study the value of $\left\|\mathbb{E}[S(X_1)^\top S(X_1)]\right\|_{\text{op}}$ given the fact that the largest singular values of $S(X_1)^\top S(X_1)$ and $S(X_1)S(X_1)^\top$ are identical for arbitrary X_1. We know that $\mathbb{E}[S(X_1)S(X_1)^\top]$ is always symmetric and positive semidefinite, and hence we have for any $u \in \mathbb{R}^{d_1}$ with $\|u\| = 1$

$$\begin{aligned}
u^\top \mathbb{E}[S(X_1)S(X_1)^\top]u = \mathbb{E}[u^\top S(X_1)S(X_1)^\top u] &= \mathbb{E}\left[\left\|S(X_1)^\top u\right\|^2\right] \\
&= \sum_{j=1}^{d_2}\mathbb{E}\left[\left(\sum_{i=1}^{d_1} S_{i,j}(X_1)u_i\right)^2\right] \\
&= \sum_{j=1}^{d_2}\mathbb{E}\left[\sum_{i=1}^{d_1} S_{i,j}(X_1)^2 u_i^2\right] \leq d_2 M,
\end{aligned}$$

and this result implies that $\left\|\mathbb{E}[S(X_1)S(X_1)^\top]\right\|_{\text{op}} \leq d_2 M \leq (d_1 + d_2)M$ Therefore, we have that

$$\left\|\sum_{j=1}^{T_1}\mathbb{E}[y_j^2 S(X_j)S(X_j)^\top]\right\|_{\text{op}} \leq (4\sigma_0^2 + S_f^2)(d_1 + d_2)T_1 M.$$

And similarly, we can prove that

$$\left\| \sum_{j=1}^{T_1} \mathbb{E}[y_j^2 S(X_j)^\top S(X_j)] \right\|_{\text{op}} \leq (4\sigma_0^2 + S_f^2)(d_1 + d_2) T_1 M.$$

Therefore, we can take $\sigma^2 = (4\sigma_0^2 + S_f^2)(d_1 + d_2)T_1 M$ consequently. By using Lemma C.2.0.3, we have

$$P\left(\|\nabla L(\mu^* \Theta^*)\|_{\text{op}} \geq \frac{2t}{\sqrt{T_1}} \right) \leq 2(d_1 + d_2) \exp\left(-\nu t \sqrt{T_1} + \frac{\nu^2 (4\sigma_0^2 + S_f^2) M (d_1 + d_2) T_1}{2} \right)$$

By plugging the values of t and ν in Lemma C.2.0.4, we finish the proof. □

C.2.2. Proof of Theorem 4.4.1. Since the estimator $\widehat{\Theta}$ minimizes the regularized loss function defined in Eqn. (4.6), we have

$$L(\widehat{\Theta}) + \lambda_{T_1} \left\| \widehat{\Theta} \right\|_{\text{nuc}} \leq L(\mu^* \Theta^*) + \lambda_{T_1} \|\mu^* \Theta^*\|_{\text{nuc}}.$$

And due to the fact that $L(\cdot)$ is a quadratic function, we have the following expression based on multivariate Taylor's expansion:

$$L(\widehat{\Theta}) - L(\mu^* \Theta^*) = \langle \nabla L(\mu^* \Theta^*), \Theta \rangle + 2 \|\Theta\|_F^2, \quad \text{where } \Theta = \widehat{\Theta} - \mu^* \Theta^*.$$

By rearranging the above two results, we can deduce that

$$\begin{aligned} 2\|\Theta\|_F^2 &\leq -\langle \nabla L(\mu^* \Theta^*), \Theta \rangle + \lambda_{T_1} \|\mu^* \Theta^*\|_{\text{nuc}} - \lambda_{T_1} \left\| \widehat{\Theta} \right\|_{\text{nuc}} \\ &\overset{\text{(i)}}{\leq} \|\nabla L(\mu^* \Theta^*)\|_{\text{op}} \|\Theta\|_{\text{nuc}} + \lambda_{T_1} \|\mu^* \Theta^*\|_{\text{nuc}} - \lambda_{T_1} \left\| \widehat{\Theta} \right\|_{\text{nuc}}, \end{aligned} \tag{C.1}$$

where (i) comes from the duality between matrix operator norm and nuclear norm. Next, we represent the saturated SVD of Θ^* in the main paper as $\Theta^* = UDV^\top$ where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$, and here we would work on its full version, i.e.

$$\Theta^* = (U, U_\perp) \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} (V, V_\perp)^\top = (U, U_\perp) D^* (V, V_\perp)^\top,$$

where we have $U_\perp \in \mathbb{R}^{d_1 \times (d_1 - r)}$, $D^* \in \mathbb{R}^{d_1 \times d_2}$ and $V_\perp \in \mathbb{R}^{d_2 \times (d_2 - r)}$. Furthermore, we define

$$\Lambda = (U, U_\perp)^\top \Theta (V, V_\perp) = \begin{pmatrix} U^\top \Theta V & U^\top \Theta V_\perp \\ U_\perp^\top \Theta V & U_\perp^\top \Theta V_\perp \end{pmatrix} = \Lambda_1 + \Lambda_2$$

where we write

$$\Lambda_1 = \begin{pmatrix} 0 & 0 \\ 0 & U_\perp^\top \Theta V_\perp \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} U^\top \Theta V & U^\top \Theta V_\perp \\ U_\perp^\top \Theta V & 0 \end{pmatrix}$$

Afterwards, it holds that

$$\begin{aligned} \left\| \widehat{\Theta} \right\|_{\text{nuc}} &= \|\mu^* \Theta^* + \Theta\|_{\text{nuc}} = \left\| (U, U_\perp)(\mu^* D^* + \Lambda)(V, V_\perp)^\top \right\|_{\text{nuc}} \\ &= \|\mu^* D^* + \Lambda\|_{\text{nuc}} + \|\mu^* D^* + \Lambda_1 + \Lambda_2\|_{\text{nuc}} \\ &\geq \|\mu^* D^* + \Lambda_1\|_{\text{nuc}} - \|\Lambda_2\|_{\text{nuc}} \\ &= \|\mu^* D\|_{\text{nuc}} + \|\Lambda_1\|_{\text{nuc}} - \|\Lambda_2\|_{\text{nuc}} \\ &= \|\mu^* \Theta^*\|_{\text{nuc}} + \|\Lambda_1\|_{\text{nuc}} - \|\Lambda_2\|_{\text{nuc}}, \end{aligned}$$

which implies that

$$\|\mu^* \Theta^*\|_{\text{nuc}} - \left\| \widehat{\Theta} \right\|_{\text{nuc}} \leq \|\Lambda_2\|_{\text{nuc}} - \|\Lambda_1\|_{\text{nuc}} \tag{C.2}$$

Combine Eqn. (C.1) and (C.2), we have that

$$2 \|\Theta\|_F^2 \leq \left(\|\nabla L(\mu^* \Theta^*)\|_{\text{op}} + \lambda_{T_1} \right) \|\Lambda_2\|_{\text{nuc}} + \left(\|\nabla L(\mu^* \Theta^*)\|_{\text{op}} - \lambda_{T_1} \right) \|\Lambda_1\|_{\text{nuc}}$$

Then, we refer to the setting in our Lemma C.2.0.4, and we choose $\lambda = 4t/\sqrt{T_1}$ where the value of t is determined in Lemma C.2.0.4, i.e.

$$\lambda_{T_1} = 4 \sqrt{\frac{2(4\sigma_0^2 + S_f^2) M (d_1 + d_2) \log(2(d_1 + d_2)/\delta)}{T_1}},$$

we know that $\lambda_{T-1} \geq 2 \|\nabla L(\mu^* \Theta^*)\|_{\text{op}}$ with probability at least $1 - \delta$ for any $\delta \in (0, 1)$. Therefore, with probability at least $1 - \delta$, we have

$$2 \|\Theta\|_F^2 \leq \frac{3}{2} \lambda_{T_1} \|\Lambda_2\|_{\text{nuc}} - \frac{1}{2} \lambda_{T_1} \|\Lambda_1\|_{\text{nuc}} \leq \frac{3}{2} \lambda_{T_1} \|\Lambda_2\|_{\text{nuc}}$$

. Since we can easily verify that the rank of Λ_2 is at most $2r$, and by using Cauchy-Schwarz Inequality we have that

$$2\left\|\Theta\right\|_F^2 \leq \frac{3}{2}\lambda_{T_1}\sqrt{2r}\left\|\Lambda_2\right\|_F \leq \frac{3}{2}\lambda_{T_1}\sqrt{2r}\left\|\Lambda\right\|_F = \frac{3}{2}\lambda_{T_1}\sqrt{2r}\left\|\Theta\right\|_F,$$

which implies that

$$\left\|\Theta\right\|_F \leq \frac{3}{4}\sqrt{2r}\lambda_{T_1} = 6\sqrt{\frac{(4\sigma_0^2 + S_f^2)M(d_1+d_2)r\log(\frac{2(d_1+d_2)}{\delta})}{T_1}},$$

and it concludes our proof. □

C.3. Theorem C.3.1 and its analysis

C.3.1. Theorem C.3.1.

THEOREM C.3.1. (Regret of LowGLM-UCB) *Under Assumption 4.3.4 and 4.3.5, for any fixed failure rate $\delta \in (0,1)$, if we run the LowGLM-UCB algorithm with $\rho_t(\delta) = \alpha_{t+T_1}(\delta/2)$ and*

$$\lambda_\perp \asymp \frac{c_\mu S_0^2 T}{k\log(1+\frac{c_\mu S_0^2 T}{k\lambda_0})},$$

then the bound of regret for LowGLM-UCB ($Regret_{T_2}$) achieves $\widetilde{O}(k\sqrt{T} + TS_\perp)$, with probability at least $1-\delta$.

C.3.2. Proposition C.3.1 with its proof. We firstly present the following important Proposition C.3.1 for obtaining the upper confidence bound.

PROPOSITION C.3.1. *For any δ, t such that $\delta \in (0,1)$, $t \geq 2$, and for $\beta_t^x(\delta)$ defined in Eqn. (4.11) and (4.12), with probablity $1-\delta$, it holds that*

$$|\mu(x^\top\theta^*) - \mu(x^\top\hat{\theta}_t)| \leq \beta_{t+T_1}^x(\delta), \tag{C.3}$$

simultaneously for all $x \in \mathbb{R}$ and all $t \geq 2$.

C.3.2.1. *Technical Lemmas.*

LEMMA C.3.1.1. (Adapted from Abbasi-Yadkori et al., 2011, Theorem 1) *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration and $\{x_t\}_{t=0}^\infty$ be an $\mathbb{R}^d$-valued stochastic process adapted to $\mathcal{F}_t$. Let $\{\eta_t\}_{t=0}^\infty$ be a real-valued stochastic*

process such that η_t is adapted to $\mathcal{F}_t$ and is conditionally σ_0-sub-Gaussian for some $\sigma_0 > 0$, i.e.

$$\mathbb{E}[\exp(\lambda\eta_t)|\mathcal{F}_t] \leq \exp\left(\frac{\lambda^2\sigma_0^2}{2}\right), \qquad \forall\lambda \in \mathbb{R}.$$

Consider the martingale $S_t = \sum_{k=1}^{t} \eta_k x_k$ and the process $V_t = \sum_{k=1}^{t} x_k x_k^\top + \Lambda$ when $t \geq 2$. And Λ is fixed and independent with sample random variables after time m. For any $\delta > 0$, with probability at least $1-\delta$, we have the following result simultaneously for all $t \geq m+1$:

$$\|S_t\|_{V_t^{-1}} \leq \sigma_0\sqrt{\log(\det(V_t)) - \log(\delta^2\det(\Lambda))}.$$

We defer the proof for this lemma to Section C.3.2.3 since a lot of technical details are involved.

LEMMA C.3.1.2. *For any two symmetric positive definite matrix $A, B \in R^{p\times p}$ such that $A \preceq B$, we have $AB^{-1}A \preceq A$.*

PROOF. Since $A \preceq B$ and both of them are invertible matrices, we have $B^{-1} \preceq A^{-1}$ directly based on positive definiteness property. Conjugate with A on both sides we can directly obtain $AB^{-1}A \preceq A$. □

LEMMA C.3.1.3. (Valko et al., 2014, Lemma 5) *For any $T \geq 1$, let $V_{T+1} = \sum_{i=1}^{T} x_i x_i^\top + \Lambda \in R^p$ where $\Lambda = diag\{\lambda_1, \ldots, \lambda_p\}$. And we assume that $\|x_i\|_2 \leq S$. Then:*

$$\log\frac{|V_{T+1}|}{|\Lambda|} \leq \max_{\{t_i\}_{i=1}^p} \sum_{i=1}^{p} \log\left(1 + \frac{S^2 t_i}{\lambda_i}\right),$$

where the maximum is taken over all possible positive real numbers $\{t_i\}_{i=1}^p$ such that $\sum_{i=1}^p t_i = T$

PROOF. We aim to bound the determinant $|V_{T+1}|$ under the coordinate constrains $\|x_i\|_2 \leq S$. Let's denote

$$U(x_1, \ldots, x_T) = |\Sigma + \sum_{t=1}^{T} x_t x_t^\top|.$$

Based on the property of the sum of rank-1 matrices (e.g. Valko et al., 2014, Lemma 4), we know that the maximum of $U(x_1, \ldots, x_T)$ is reached when all x_t are aligned with the axes:

$$U(x_1, \ldots, x_T) = \max_{\substack{x_1,\ldots,x_T; \\ x_t \in S\cdot\{e_1,\ldots,e_N\}}} |\Sigma + \sum_{t=1}^{T} x_t x_t^\top| = \max_{\substack{t_1,\ldots,t_N \text{positive integers}; \\ \sum_{i=1}^N t_i = T}} |\mathrm{diag}(\lambda_i + t_i)|$$
$$\leq \max_{\substack{t_1,\ldots,t_N \text{positive integers}; \\ \sum_{i=1}^N t_i = S^2 T}} \prod_{i=1}^{N} (\lambda_i + S^2 t_i).$$

□

C.3.2.2. *Proof of Proposition C.3.1.*

Proof. Recall our definition of $g_t(\theta)$ and its gradient accordingly as

$$g_t(\theta) = \sum_{i=1}^{T_1} \mu({x_{s_1,i}}^\top \theta) x_{s_1,i} + \sum_{k=1}^{t-1} \mu(x_k^\top \theta) x_k + \Lambda\theta,$$

$$\nabla_\theta g_t(\theta) = \sum_{i=1}^{T_1} \mu'({x_{s_1,i}}^\top \theta) x_{s_1,i} x'_{s_1,i} + \sum_{k=1}^{t-1} \mu'(x_k^\top \theta) x_k x_k^\top + \Lambda \overset{\text{(i)}}{\succeq} c_\mu M_t(c_\mu), \tag{C.4}$$

where the relation (i) holds if $\theta \in \Theta_0$. Based on Assumptions, we know the gradient $\nabla_\theta g_t(\theta)$ is continuous. Then the Fundamental Theorem of Calculus will imply that

$$g_t(\theta^*) - g_t(\hat{\theta}_t) = G_t(\theta^* - \hat{\theta}_t),$$

where

$$G_t = \int_0^1 \nabla_\theta g_t(s\theta^* + (1-s)\hat{\theta}_t)\, ds.$$

Since we assume that the inverse link function $\mu(\cdot)$ is $k_\mu-$Lipshitz, and the matrix G_t is always invertible due to the fact that at least we have $G_t \succeq \Lambda$, we can obtain the following result. Notice the inequality (i) comes from the fact that $G_t \succeq c_\mu M_t(c_\mu)$ and hence $M_t(c_\mu)^{-1}/c_\mu \succeq G_t^{-1}$.

$$|\mu(x^\top\theta^*) - \mu(x^\top\hat{\theta}_t)| \leq k_\mu |x^\top(\theta^* - \hat{\theta}_t)| = k_\mu |x^\top G_t^{-1}(g_t(\theta^*) - g_t(\hat{\theta}_t))|$$
$$\leq k_\mu \|x\|_{G_t^{-1}} \left\|g_t(\theta^*) - g_t(\hat{\theta}_t)\right\|_{G_t^{-1}} \overset{\text{(i)}}{\leq} \frac{k_\mu}{c_\mu} \|x\|_{M_t(c_\mu)^{-1}} \left\|g_t(\theta^*) - g_t(\hat{\theta}_t)\right\|_{M_t(c_\mu)^{-1}}.$$

In addition, based on the definition of $\hat{\theta}_t$ in Equation (4.10), we have $g_t(\hat{\theta}_t) - g_t(\theta^*) = \sum_{k=1}^{T_1}(y_{s_1,k} - \mu(x_{s_1,k}^\top \theta^*))x_{s_1,k} + \sum_{k=1}^{t-1}(y_k - \mu(x_k^\top \theta^*))x_k - \Lambda\theta^* = \sum_{k=1}^{T_1}\eta_{s_1,k}x_{s_1,k} + \sum_{k=1}^{t-1}\eta_k x_k - \Lambda\theta^*$. Therefore,

$$
\begin{aligned}
|\mu(x^\top\theta^*) - \mu(x^\top\hat{\theta}_t)| &\leq \frac{k_\mu}{c_\mu}\|x\|_{M_t(c_\mu)^{-1}} \left\|g_t(\hat{\theta}_t) - g_t(\theta^*)\right\|_{M_t^{-1}(c_\mu)} \\
&\leq \frac{k_\mu}{c_\mu}\|x\|_{M_t(c_\mu)^{-1}} \left(\left\|\sum_{k=1}^{T_1}\eta_{s_1,k}x_{s_1,k} + \sum_{k=1}^{t-1}\eta_k x_k\right\|_{M_t^{-1}(c_\mu)} + \|\Lambda\theta^*\|_{M_t^{-1}(c_\mu)}\right). \qquad \text{(C.5)}
\end{aligned}
$$

Now, let's use Lemma C.3.1.1 to bound the term $\left\|\sum_{k=1}^{T_1}\eta_{s_1,k}x_{s_1,k} + \sum_{k=1}^{t-1}\eta_k x_k\right\|_{M_t^{-1}(c_\mu)}$. If we define the filtration $\mathcal{F}_t := \{\{x_t, x_{t-1}, \eta_{t-1}, \ldots, x_1, \eta_1\} \cup \{x_{s_1,k}, \eta_{s_1,k}\}_{k=1}^{T_1}\}$, then for any $\delta \in (0,1)$, with probability $1-\delta$, it holds that for all $t \geq 2$,

$$
\left\|\sum_{k=1}^{T_1}\eta_{s_1,k}x_{s_1,k} + \sum_{k=1}^{t-1}\eta_k x_k\right\|_{M_t^{-1}(c_\mu)} \leq \sigma_0\sqrt{\log\left(\frac{|M_t(c_\mu)|}{|\frac{\Lambda}{c_\mu}|}\right) - 2\log(\delta)},
$$

where based on Lemma C.3.1.3,

$$
\begin{aligned}
\log\left(\frac{|M_t(c_\mu)|}{|\frac{\Lambda}{c_\mu}|}\right) &\leq \max_{\substack{t_i\geq 0,\\ \sum_{i=1}^t t_i = t+T_1}} \sum_{i=1}^{p}\log\left(1 + \frac{c_\mu S_0^2 t_i}{\lambda_i}\right) \\
&\leq k\log\left(1 + \frac{c_\mu S_0^2}{k\lambda_0}(t+T_1)\right) + (d-k)\log\left(1 + \frac{c_\mu S_0^2}{(d-k)\lambda_\perp}(t+T_1)\right) \\
&\leq k\log\left(1 + \frac{c_\mu S_0^2}{k\lambda_0}(t+T_1)\right) + \frac{c_\mu S_0^2}{\lambda_\perp}(t+T_1). \qquad \text{(C.6)}
\end{aligned}
$$

And next by Lemma C.3.1.2, we have

$$
\|\Lambda\theta^*\|_{M_t^{-1}(c_\mu)} = c_\mu\left\|\frac{\Lambda}{c_\mu}\theta^*\right\|_{M_t^{-1}(c_\mu)} \leq \sqrt{c_\mu}\,\|\theta^*\|_\Lambda \leq \sqrt{c_\mu}(\sqrt{\lambda_0}S_0 + \sqrt{\lambda_\perp}S_\perp). \qquad \text{(C.7)}
$$

Combine Equation (C.6) and (C.7) into Equation (C.5), we finish our proof. □

Since Equation (C.3) in Proposition C.3.1 holds simultaneously for all $x \in \mathbb{R}$ and $t \geq 1$, the following conclusion holds.

COROLLARY C.3.1. *For any random variable z defined in $\mathbb{R}$, we have the following holds*

$$
|\mu(z^\top\theta^*) - \mu(z^\top\hat{\theta}_t)| \leq \beta_{t+T_1}^z(\delta),
$$

with probability at least $1-\delta$. *Furthermore, for any sequence of random variable* $\{z_t\}_{t=2}^{T}$, *with probability* $1-\delta$ *it holds that*

$$|\mu(z_t^\top \theta^*) - \mu(z_t^\top \hat{\theta}_t)| \leq \beta_{t+T_1}^{z_t}(\delta),$$

simultaneously for all $t \geq 1$.

C.3.2.3. *Proof of Lemma C.3.1.1.* For the proof of Lemma C.3.1.1 we will need the following two lemmas, and we will use the same notations as in Lemma C.3.1.1 in this section.

LEMMA C.3.1.4. *Let* $\lambda \in \mathbb{R}^d$ *be arbitrary and consider any* $t \geq 0$

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s(\lambda^\top x_s)}{\sigma_0} - \frac{1}{2}(\lambda^\top x_s)^2\right]\right).$$

Let τ *be a stopping time with respect to the filtration* $\{\mathcal{F}_t\}_{t=0}^{+\infty}$. *Then* M_t^λ *is a.s. well defined and* $\mathbb{E}(M_\tau^\lambda) \leq 1$.

PROOF. We claim that $\{M_t^\lambda\}$ is a supermartingale. Let

$$D_t^\lambda = \exp\left(\frac{\eta_s(\lambda^\top x_s)}{\sigma_0} - \frac{1}{2}(\lambda^\top x_s)^2\right)$$

Observe that by conditional σ_0-sub-Gaussianity of η_t we have $\mathbb{E}[D_t^\lambda|\mathcal{F}_{t-1}] \leq 1$. Clearly, D_t^λ and M_t^λ is $\mathcal{F}_t$-measurable. Moreover,

$$\mathbb{E}[M_t^\lambda|\mathcal{F}_{t-1}] = \mathbb{E}[M_1^\lambda \cdots D_{t-1}^\lambda D_t^\lambda|\mathcal{F}_{t-1}] = D_1^\lambda \ldots D_{t-1}^\lambda \mathbb{E}[D_t^\lambda|\mathcal{F}_{t-1}] \leq M_{t-1}^\lambda,$$

which implies that M_t^λ is a supermartingale with its expected value upped bounded by 1. To show that M_t^λ is well defined. By the convergence theorem for nonnegative supermartingales, $\lim_{t\to\infty} M_t^\lambda$ is a.s. well-defined, which indicates that M_τ^λ is also well-defined for all $\tau \in \mathbb{N}^+ \cup \{+\infty\}$. By Fatou's Lemma, it holds that

$$\mathbb{E}[M_\tau^\lambda] = \mathbb{E}[\liminf_{t\to\infty} M_{\min\{t,\tau\}}^\lambda] \leq \liminf_{t\to\infty} \mathbb{E}[M_{\min\{t,\tau\}}^\lambda] \leq 1.$$

□

LEMMA C.3.1.5. *For any positive semi-definite matrix $P \in \mathbb{R}^{d\times d}$ and positive definite matrix $Q \in \mathbb{R}^{d\times d}$, and any $x, a \in \mathbb{R}^d$, it holds that*

$$\|x-a\|_P^2 + \|x\|_Q^2 = \left\|x-(P+Q)^{-1}Pa\right\|_{P+Q}^2 + \|a\|_P^2 - \|Pa\|_{(P+Q)^{-1}}^2 .$$

This lemma could be easily proved based on elementary calculation and hence its proof would be omitted here.

LEMMA C.3.1.6. *Let τ be a stopping time with $\tau > m$ on the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then for $\delta > 0$, with probability $1-\delta$,*

$$||S_\tau||_{V_\tau^{-1}}^2 \le 2\sigma_0^2 \log\left(\frac{det(V_\tau)^{1/2} det(\Lambda)^{-1/2}}{\delta}\right).$$

PROOF. W.l.o.g., assume that $\sigma_0 = 1$. Denote

$$\tilde{V}_t = V_t - \Lambda = \sum_{s=1}^{t} x_s x_s^\top, \qquad M_t^\lambda = \exp\left((\lambda^\top S_t) - \frac{1}{2}||\lambda||_{\tilde{V}_t}^2\right).$$

Note by Lemma C.3.1.4, we naturally have that $\mathbb{E}[M_t^\lambda] \le 1$.

Since in round $m+1$, we get the diagonal positive definite matrix Λ with its elements independent with samples after round m. Let z be a Gaussian random variable that is independent with other random variables after round m with covariance Λ^{-1}. Define

$$M_t = \mathbb{E}[M_t^z | \mathcal{F}_\infty], \quad t > m,$$

where $\mathcal{F}_\infty$ is the tail σ-algebra of the filtration. Clearly, it holds that $\mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[M_\tau^z | z, \mathcal{F}_\infty]] \le \mathbb{E}[1] \le 1$. Let f be the density of z and for a positive definite matrix P let $c(P) = \sqrt{(2\pi)^d/\det(P)}$. Then for $t > m$ it holds that,

$$\begin{aligned} M_t &= \int_{\mathbb{R}^d} \exp\left((\lambda^\top S_t) - \frac{1}{2}||\lambda||_{\tilde{V}_t}^2\right) f(\lambda) d\lambda \\ &= \frac{1}{c(\Lambda)} \exp\left(\frac{1}{2}\|S_t\|_{\tilde{V}_t^{-1}}^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\{\left\|\lambda - \tilde{V}_t^{-1}S_t\right\|_{\tilde{V}_t}^2 + \|\lambda\|_\Lambda^2\right\}\right) d\lambda. \end{aligned}$$

Based on Lemma C.3.1.5, it holds that

$$\left\|\lambda - \tilde{V}_t^{-1}S_t\right\|_{\tilde{V}_t}^2 + \|\lambda\|_\Lambda^2 = \left\|\lambda - V_t^{-1}S_t\right\|_{V_t}^2 + \left\|\tilde{V}_t^{-1}S_t\right\|_{\tilde{V}_t}^2 - \|S_t\|_{V_t^{-1}}^2$$

, and this implies that

$$M_t = \frac{1}{c(\Lambda)} \exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|\lambda - V_t^{-1}S_t\right\|^2_{V_t}\right) d\lambda$$
$$= \left(\frac{\det(\Lambda)}{\det(V_t)}\right)^{1/2} \exp\left(-\frac{1}{2}\left\|\lambda - V_t^{-1}S_t\right\|^2_{V_t}\right).$$

Now, from $\mathbb{E}[M_\tau] \leq 1$, we have that for $\tau > m$

$$P\left(\|S_\tau\|^2_{V_\tau^{-1}} > \log\left(\frac{\det(V_\tau)}{\delta^2\det(\Lambda)}\right)\right) = P\left(\frac{\exp\left(\frac{1}{2}\|S_\tau\|^2_{V_\tau^{-1}}\right)}{\delta^{-1}(\det(V_\tau)/\det(\Lambda))^{1/2}} > 1\right)$$
$$\leq \mathbb{E}\left[\frac{\exp\left(\frac{1}{2}\|S_\tau\|^2_{V_\tau^{-1}}\right)}{\delta^{-1}(\det(V_\tau)/\det(\Lambda))^{1/2}}\right] \leq \mathbb{E}[M_\tau]\delta \leq \delta.$$

□

Combining Lemma C.3.1.4-C.3.1.6. We now contruct a stopping time and define the bad event:

$$B_t(\delta) \coloneqq \left\{w : \|S_t\|^2_{V_t^{-1}} > \sigma_0^2 \log\left(\frac{\det(V_t)}{\delta^2\det(\Lambda)}\right)\right\}.$$

And we are interested in bounding the probability that $\cup_{t>m} B_t(\delta)$ happens. Define $\tau(w) = \min\{t > m : w \in B_t(\delta)\}$. Then τ is a stopping time and it holds that,

$$\cup_{t>m} B_t(\delta) = \{w : \tau(w) < \infty\}.$$

Then we have that

$$P\left[\cup_{t>m} B_t(\delta)\right] = P[m < \tau < \infty] = P\left[\|S_\tau\|^2_{V_\tau^{-1}} > \sigma_0^2 \log\left(\frac{\det(V_\tau)}{\delta^2\det(\Lambda)}\right), \tau > m\right] \leq \delta.$$

This concludes our proof of Lemma C.3.1.1.

C.3.3. Proposition C.3.2 with its proof. We denote the optimal action $x^* = \arg\max_{x\in\mathcal{X}_0} \mu(x^\top\theta^*)$.

PROPOSITION C.3.2. *For all $\delta \in (0,1)$, with probability $1-\delta$, it holds that*

$$\mu(x^{*\top}\theta^*) - \mu(x_t^\top\theta^*) \leq 2\beta^{x_t}_{t+T_1}\left(\frac{\delta}{2}\right),$$

simultaneously for all $t \in \{2, 3, \ldots, T_2\}$.

PROOF. According to Corollary C.3.1, outside of the event of measure can be bounded by $\delta/2$:

$$\mu(x_t^\top \hat{\theta}_t) - \mu(x_t^\top \theta^*) \leq \beta_{t+T_1}^{x_t}\left(\frac{\delta}{2}\right) \quad \text{for all } t \in \{2, 3, \ldots, T_2\}.$$

Similarly, with probability at least $1 - \delta/2$ it holds that

$$\mu(x^{*\top} \theta^*) - \mu(x^{*\top} \hat{\theta}_t) \leq \beta_{t+T_1}^{x^*}\left(\frac{\delta}{2}\right) \quad \text{for all } t \in \{2, 3, \ldots, T_2\}.$$

Besides, by the choice of x_t in Algorithm 6

$$\begin{aligned}
\mu(x^{*\top} \hat{\theta}_t) - \mu(x_t^\top \hat{\theta}_t) &= \mu(x^{*\top} \hat{\theta}_t) + \beta_{t+T_1}^{x^*}\left(\frac{\delta}{2}\right) - \mu(x_t^\top \hat{\theta}_t) - \beta_{t+T_1}^{x^*}\left(\frac{\delta}{2}\right) \\
&\leq \mu(x_t^\top \hat{\theta}_t) + \beta_{t+T_1}^{x_t}\left(\frac{\delta}{2}\right) - \mu(x_t^\top \hat{\theta}_t) - \beta_{t+T_1}^{x^*}\left(\frac{\delta}{2}\right) \\
&= \beta_{t+T_1}^{x_t}\left(\frac{\delta}{2}\right) - \beta_{t+T_1}^{x^*}\left(\frac{\delta}{2}\right).
\end{aligned}$$

By combining the former inequalities we finish our proof. □

C.3.4. Proof of Theorem C.3.1.

PROOF. Based on Proposition C.3.2 we have

$$\mu(x^{*\top} \theta^*) - \mu(x_t^\top \theta^*) \leq 2\beta_{t+T_1}^{x_t}\left(\frac{\delta}{2}\right) = 2\alpha_{t+T_1}\left(\frac{\delta}{2}\right) \|x_t\|_{M_t^{-1}(c_\mu)} \leq 2\alpha_T\left(\frac{\delta}{2}\right) \|x_t\|_{M_t^{-1}(c_\mu)}.$$

Since we know that $\mu(x^{*\top}\theta^*) - \mu(x^\top \theta^*) \leq k_\mu(x^{*\top}\theta^* - x^\top \theta^*) \leq 2k_\mu S_0^2$ for all possible action x, and we can safely expect that $\alpha_{T_2}(\delta/2) > k_\mu S_0^2$ (at least by choosing $\sigma_0 = k_\mu \max\{S_0^2, 1\}$), then the regret of Algorithm 6 can be bounded as

$$\begin{aligned}
Regret_{T_2} &\leq 2k_\mu S_0^2 + \sum_{t=2}^{T_2} \min\{\mu(x^{*\top}\theta^*) - \mu(x_t^\top \theta^*), 2k_\mu S_0^2\} \\
&\leq 2k_\mu S_0^2 + 2\alpha_T\left(\frac{\delta}{2}\right) \sum_{t=2}^{T_2} \min\{\|x_t\|_{M_t^{-1}(c_\mu)}, 1\} \\
&\overset{\text{(i)}}{\leq} 2k_\mu S_0^2 + 2\alpha_T\left(\frac{\delta}{2}\right) \sqrt{T_2} \sqrt{\sum_{t=2}^{T_2} \min\{\|x_t\|^2_{M_t^{-1}(c_\mu)}, 1\}}.
\end{aligned}$$

where the ineuqlity (i) comes from Cauchy-Schwarz inequality. And a commonly-used fact (e.g. Abbasi-Yadkori et al. (2011), Lemma 11) yields that

$$\sum_{i=2}^{t} \min\{\|x_i\|^2_{M_i^{-1}(c_\mu)}, 1\} \le 2\log\left(\frac{|M_{t+1}(c_\mu)|}{|M_2(c_\mu)|}\right) \le 2\log\left(\frac{|M_{t+1}(c_\mu)|}{|\frac{\Lambda}{c_\mu}|}\right)$$
$$\le k\log\left(1+\frac{c_\mu S_0^2}{k\lambda_0}(t+T_1)\right) + \frac{c_\mu S_0^2}{\lambda_\perp}(t+T_1).$$

Finally, by using the argument in Eqn. (C.6) and then plugging in the chosen value for $\lambda_\perp = \frac{c_\mu S_0^2 T}{k\log(1+\frac{c_\mu S_0^2 T}{k\lambda_0})}$, we have

$$Regret_{T_2} \le 2k_\mu S_0^2 +$$
$$\frac{2k_\mu}{c_\mu}\left(\sigma_0\sqrt{2k\log\left(1+\frac{c_\mu S_0^2}{k\lambda_0}T\right) - 2\log\left(\frac{\delta}{2}\right)} + \sqrt{c_\mu}\left(\sqrt{\lambda_0}S_0 + \sqrt{\frac{c_\mu S_0^2 T}{k\log\left(1+\frac{c_\mu S_0^2}{k\lambda_0}T\right)}}S_\perp\right)\right)$$
$$\times\sqrt{T_2}\sqrt{4k\log\left(1+\frac{c_\mu S_0^2}{k\lambda_0}T\right)},$$

which gives us the final bound in Theorem C.3.1. □

C.4. Consistency of $\hat{\theta}_t^{\text{new}}$ in Algorithm 6

W.l.o.g. we assume that $\{\theta : \|\theta-\theta^*\|_2 \le 1\} \subseteq \Theta^*$, or otherwise we can modify the contraint of c_μ in Assumption 4.3.5 as $c_\mu := \inf_{\{x\in\mathcal{X}_0, \|\theta-\theta^*\|_2\le 1\}} \mu'(x^\top\theta) > 0$. And we also assume that $\|x\|_2 \le 1$ for $x \in \mathcal{X}_0$.

Adapted from the proof of Theorem 1 in Li et al. (2017), define $G(\theta) = g(\theta) - g(\theta^*) = \sum_{i=1}^{T_1}(\mu(x_{s_1,i}^\top\theta) - \mu(s_1, i^\top\theta^*))x_{s_1,i} + \sum_{i=1}^{n}(\mu(x_i^\top\theta) - \mu(x_i^\top\theta^*))x_i + \Lambda(\theta-\theta^*)$. W.l.o.g we suppose $c_\mu \le 1$ based on argument in Appendix C.7. Then it holds that for any $\theta_1, \theta_2 \in \mathbb{R}^p$

$$G(\theta_1) - G(\theta_2) =$$
$$\left[\sum_{i=1}^{T_1}(\mu'(x_{s_1,i}^\top\theta) - \mu(s_1, i^\top\theta^*))x_{s_1,i}x_{s_1,i}^\top + \sum_{i=1}^{n}(\mu'(x_i^\top\theta) - \mu(x_i^\top\theta^*))x_i x_i^\top + \Lambda\right](\theta_1-\theta_2).$$

By denoting $V = \sum_{i=1}^{T_1} x_{s_1,i}x_{s_1,i}^\top + \sum_{i=1}^{n} x_i x_i^\top + \Lambda$. We have

$$(\theta_1-\theta_2)^\top(G(\theta_1) - G(\theta_2)) \ge (\theta_1-\theta_2)^\top(c_\mu V)(\theta_1-\theta_2) > 0$$

Therefore, the rest of proof would be identical to that of Step 1 in the proof of Theorem 1 in Li et al. (2017). Based on the step 1 in the proof of Theorem 1 in Li et al. (2017), we have

$$\|G(\theta)\|_{V^{-1}}^2 \geq c_\mu^2 \lambda_{\min}(V) \|\theta - \theta^*\|_2^2 .$$

as long as $\|\theta - \theta^*\|_2 \leq 1$. Then Lemma A of Chen et al. (1999) and Lemma 7 of Li et al. (2017) suggest that we have

$$\left\|\hat{\theta} - \theta^*\right\| \leq \frac{4\sigma}{c_\mu}\sqrt{\frac{p + \log(1/\delta)}{\sigma^2}} \leq 1,$$

when $\lambda_{\min}(V) \geq 16\sigma^2[p + \log(1/\delta)]/c_\mu^2$ for any $\delta > 0$. Therefore, it suffices to show that the condition $\lambda_1 \geq 16\sigma^2[p + \log(1/\delta)]/c_\mu^2$ for any $\delta > 0$ holds with high probability (e.g. $1 - \delta$), and we utilize the Proposition 1 of Li et al. (2017), which is given as follows:

Proposition (Proposition 1 of Li et al. (2017)): *Define* $V_n = \sum_{t=1}^n x_t x_t^\top (+\Lambda)$ *where* x_i *is drawn iid from some distribution* ν *with suppost in the unit ball,* $\mathbb{B}^d$. *Furthermore, let* $\Sigma = \mathbb{E}(x_t x_t^\top)$ *be the second moment matrix, and* B *and* δ *be two positive constants. Then, there exists positive universal constants* C_1 *and* C_2 *such that* $\lambda_{\min}(V_n) \geq B$ *with probability at least* $1 - \delta$, *as long as*

$$n \geq \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)}\right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}$$

Therefore, we can dedeuce that $\left\|\hat{\theta}_t - \theta^*\right\|_2 \leq 1$ holds with probability at least $1 - \delta$ as long as $T_1 \geq ((\hat{C}_1\sqrt{p} + \hat{C}_2\sqrt{\log(1/\delta)})/\lambda_1)^2 + 2B/\lambda_1$ holds for some absolute constants $\hat{C}_1, \hat{C}_2$ with the definition $B \coloneqq 16\sigma^2(p + \log(1/\delta))/c_\mu^2$. Notice that this condition could easily hold if $\lambda_1 \asymp \sigma^2$ is not diminutive in magnitude. Otherwise, we believe a tighter bound exists in that case, and we will leave it as a future work.

We also present an intuitive explanation for this consistency result: Li et al. (2017) proved the consistency of the MLE $\hat{\theta}_t$ without the regularizer. Regarding the penalty $\theta^\top \Lambda \theta$, for the first k entries of $\hat{\theta}_t$ the penalized parameter λ_0 is small, and hence it will have mild effect after sufficient warm-up rounds T_1. For the remaining $(p - k)$ elements suffering large penalty, the estimated $\hat{\theta}_{t,k+1:p}$ would be ultra small in magnitude as desired since we argue that after the transformation $\theta^*_{k+1:p}$ will also be insignificant. This implies that $\left\|\hat{\theta}_{t,k+1:p} - \theta^*_{k+1:p}\right\|_2$ is well contronlled. As a result, the estimated $\hat{\theta}_t$ tends to be consistent.

C.5. Analysis of Theorem 4.4.2

C.5.1. Proof of Theorem 4.4.2.

PROOF. Let us define $r_t = \max_{X \in \mathcal{X}} \mu(\langle X, \Theta^* \rangle) - \mu(\langle X_t, \Theta^* \rangle)$, the instantaneous regret at time t. We can easily bound the regret for stage 1 as $\sum_{t=1}^{T_1} r_t \leq 2S_f T_1$. For the second stage, we have a bound according to Theorem C.3.1 (Theorem C.8.1):

$$\sum_{t=T_1+1}^{T} r_t \leq \widetilde{O}(k\sqrt{T} + \sqrt{\lambda_0 kT} + TS_\perp) \leq \widetilde{O}\left(k\sqrt{T} + \sqrt{\lambda_0 kT} + T\frac{(d_1+d_2)Mr}{T_1 D_{rr}^2} \log\left(\frac{d_1+d_2}{\delta}\right)\right).$$

Therefore, the overall regret is:

$$\sum_{t=1}^{T} r_t \leq \widetilde{O}\left(2S_f T_1 + k\sqrt{T} + \sqrt{\lambda_0 kT} + T\frac{(d_1+d_2)Mr}{T_1 D_{rr}^2} \log\left(\frac{d_1+d_2}{\delta}\right)\right).$$

After plugging the choice of T_1 given in Theorem 4.4.2, it holds that

$$\begin{aligned}\sum_{t=1}^{T} r_t &\leq \widetilde{O}\left((\frac{\sqrt{r(d_1+d_2)M}}{D_{rr}} + \sqrt{\lambda_0 k} + k)\sqrt{T}\right) \lesssim \widetilde{O}\left((\frac{\sqrt{r(d_1+d_2)M}}{D_{rr}} + k)\sqrt{T}\right) \\ &= \widetilde{O}\left(\frac{\sqrt{(d_1+d_2)MrT}}{D_{rr}}\right).\end{aligned}$$

□

C.6. Details of Theorem 4.4.3

C.6.1. Proof of Theorem 4.4.3.

PROOF. Here we will overload the notation a little bit. Under the new arm feature set and parameter set after rotation, let X^* be the best arm and X_t be the arm we pull at round t for stage 2. And we denote $x_{t,sub}$ be the vectorization of X_t after removing the last $p-k$ covariates, and similarly define x^*_{sub} and θ^*_{sub} as the subtracted version of $\text{vec}(X^*)$ and $\text{vec}(\Theta^*)$ respectively. We use $r_t = \mu(\langle X^*, \Theta^* \rangle) - \mu(\langle X_t, \Theta^* \rangle)$ as the instantaneous regret at round t for stage 2. Then it

holds that, for $t \in [T_2]$

$$\begin{aligned}
r_t &= \mu(\langle X^*, \Theta^* \rangle) - \mu({x^*_{sub}}^\top \theta^*_{sub}) + \mu({x^*_{sub}}^\top \theta^*_{sub}) - \mu(x_{t,sub}^\top \theta^*_{sub}) + \mu(x_{t,sub}^\top \theta^*_{sub}) - \mu(\langle X_t, \Theta^* \rangle) \\
&\leq k_\mu |\langle X^*, \Theta^* \rangle - {x^*_{sub}}^\top \theta^*_{sub}| + k_\mu |\langle X_t, \Theta^* \rangle - x_{t,sub}^\top \theta^*_{sub}| + \mu({x^*_{sub}}^\top \theta^*_{sub}) - \mu(x_{t,sub}^\top \theta^*_{sub}) \\
&\leq k_\mu (\left\| \widehat{U}_\perp^\top X^* \widehat{V}_\perp \right\|_F + \left\| \widehat{U}_\perp^\top X_t \widehat{V}_\perp \right\|_F) \left\| \widehat{U}_\perp^\top U D V^\top \widehat{V}_\perp \right\|_F + \mu({x^*_{sub}}^\top \theta^*_{sub}) - \mu(x_{t,sub}^\top \theta^*_{sub}) \\
&\leq 2k_\mu S_0 \frac{d_1 d_2 r}{T_1 D_{rr}^2} \log\left(\frac{d_1+d_2}{\delta}\right) + \mu({x^*_{sub}}^\top \theta^*_{sub}) - \mu(x_{t,sub}^\top \theta^*_{sub}).
\end{aligned}$$

Therefore, the overall regret can be bounded as

$$2S_f T_1 + \sum_{t=1}^{T_2} r_t \leq 2S_f T_1 + 2k_\mu S_0 \frac{d_1 d_2 r}{D_{rr}^2 T_1} T_2 + \sum_{t=1}^{T_2} \mu({x^*_{sub}}^\top \theta^*_{sub}) - \mu(x_{t,sub}^\top \theta^*_{sub}).$$

Since efficient low dimensional generalized linear bandit algorithm can achieve regret $\widetilde{O}(\epsilon\sqrt{d}T)$ where ϵ is the misspecified rate, d is the dimension of parameter and T is the time horizon when no sparsity (low-rank structure) presents in the model. After plugging our carefully chosen T_1, the regret is

$$\begin{aligned}
&2S_f T_1 + 2k_\mu S_0 \frac{(d_1+d_2)Mr}{T_1 D_{rr}^2} \log\left(\frac{d_1+d_2}{\delta}\right) T_2 + \widetilde{O}\left(\frac{(d_1+d_2)Mr}{T_1 D_{rr}^2} \sqrt{(d_1+d_2)r T_2}\right) \\
&= \widetilde{O}\left((\frac{\sqrt{r^{3/2}(d_1+d_2)^{3/2}M}}{D_{rr}} + k)\sqrt{T} \right) = \widetilde{O}\left(\frac{\sqrt{r^{3/2}(d_1+d_2)^{3/2}M}}{D_{rr}} \sqrt{T} \right). \qquad \text{(C.8)}
\end{aligned}$$

□

C.7. Explanation of V_t replacing $M_t(c_\mu)$

Technically we can always assume $c_\mu \in (0, 1]$ since we can always choose $c_\mu = 1$ when it can take values greater than 1. And when $c_\mu \leq 1$ it holds that,

$$M_t(c_\mu) = \sum_{i=1}^{t-1} x_i x_i^\top + \frac{\Lambda}{c_\mu} \succeq \sum_{i=1}^{t-1} x_i x_i^\top + \Lambda = V_t.$$

Therefore, we can easily keep the exactly identical outline of our proof of the bound of regret for Algorithm 6 after replacing $M_t(c_\mu)$ by V_t everywhere, and the result only change by a constant factor of $1/\sqrt{c_\mu}$, which would not be too large in most cases. However, in our algorithm and proof we still use $M_t(c_\mu)$ for a better theoretical bound.

C.8. Additional Algorithms

C.8.1. PLowGLM-UCB. We could modify Algorithm 6 by only recomputing $\hat{\theta}_t$ and whenever $|M_t(c_\mu)|$ increases by a constant factor $C > 1$ in scale, and consequently we only need to solve the Eqn. (4.10) for $O(\log(T_2))$ times up to the horizon T_2, which significantly alleviate the computational complexity. The pseudo-code of PLowGLM-UCB is given in Algorithm 12.

Algorithm 12 PLowGLM-UCB

Input: $T_2, k, \mathcal{X}_0$, the probability rate δ, penalization parameters $(\lambda_0, \lambda_\perp)$, the constant C.

Initialize $M_1(c_\mu) = \sum_{i=1}^{T_1} x_{s_1,i}\, x_{s_1,i}^\top + \Lambda/c_\mu$.
for $t \geq 1$ **do**
  if $|M_t(c_\mu)| > C|M_\tau(c_\mu)|$ **then**
    Estimate $\hat{\theta}_t$ according to (4.10).
    $\tau = t$
  Choose arm $x_t = \arg\max_{x \in \mathcal{X}_0} \{\mu(x^\top \hat{\theta}_\tau) + \rho_\tau(\delta) \|x\|_{M_t^{-1}(c_\mu)}\}$, receive y_t.
  Update $M_{t+1}(c_\mu) \longleftarrow M_t(c_\mu) + x_t x_t^\top$.

Theorem C.8.1 shows the regret bound of PLowGLM-UCB under Assumption 4.3.4 and 4.3.5.

THEOREM C.8.1. (Regret of PLowGLM-UCB) *For any fixed failure rate $\delta \in (0,1)$, if we run the PLowGLM-UCB algorithm with $\rho_t(\delta) = \alpha_{t+T_1}(\delta/2)$ and*

$$\lambda_\perp \asymp \frac{c_\mu S_0^2 T}{k \log(1 + \frac{c_\mu S_0^2 T}{k\lambda_0})}.$$

Then the regret of PLowGLM-UCB ($Regret_{T_2}$) satisfies, with probability at least $1-\delta$

$$\widetilde{O}(k\sqrt{T_2} + \sqrt{\lambda_0 k T} + T S_\perp) \cdot \sqrt{C} = \widetilde{O}(k\sqrt{T} + T S_\perp) \cdot \sqrt{C}.$$

Similarly, for PLowUCB-GLM we can also prove that the regret bound increase at most by a constant multiplier $\sqrt{C}$ by using the same lemma and argument we show in the following Section C.8.2. And we can get the bound of regret for PLowGLM-UCB in problem dependence case, and the bound will be exactly the same as that we have shown in Theorem C.3.1 except a constant multiplier $\sqrt{C}$.

C.8.2. Proof of Theorem C.8.1. We use similar sketch of proof for Theorem 5 in Abbasi-Yadkori et al. (2011). First, we show the following lemma:

LEMMA C.8.1.1. (Abbasi-Yadkori et al. (2011), Lemma 12) *Let A and B be two positive semi-definite matrices such that $A \preceq B$. Then, we have that*

$$\sup_{x \neq 0} \frac{x^\top A x}{x^\top B x} \leq \frac{|A|}{|B|}.$$

Then we can outline the proof of Theorem C.8.1 as follows.

PROOF. Let τ_t be the value of τ at step t in Algorithm C.8.1. By an argument similar to the one used in proof of Theorem C.3.1, we deduce that for any $x \in \mathbb{R}$ and all $t \geq 2$ simultaneously,

$$\begin{aligned}
|\mu(x^\top \theta^*) - \mu(x^\top \hat{\theta}_{\tau_t})| &\leq \frac{k_\mu}{c_\mu} \left\| g_{\tau_t}(\theta^*) - g_{\tau_t}(\hat{\theta}_{\tau_t}) \right\|_{M_{\tau_t}^{-1}(c_\mu)} \|x\|_{M_{\tau_t}^{-1}(c_\mu)} \\
&= \frac{k_\mu}{c_\mu} \left\| g_{\tau_t}(\theta^*) - g_{\tau_t}(\hat{\theta}_{\tau_t}) \right\|_{M_{\tau_t}^{-1}(c_\mu)} \left\| M_{\tau_t}^{-\frac{1}{2}}(c_\mu)\, x \right\|_2 \\
&\leq \frac{k_\mu}{c_\mu} \left\| g_{\tau_t}(\theta^*) - g_{\tau_t}(\hat{\theta}_{\tau_t}) \right\|_{M_{\tau_t}^{-1}(c_\mu)} \left\| M_t^{-\frac{1}{2}}(c_\mu)\, x \right\|_2 \sqrt{\frac{|M_{\tau_t}^{-1}(c_\mu)|}{|M_t^{-1}(c_\mu)|}} \\
&\leq \frac{k_\mu}{c_\mu} \sqrt{C} \left\| g_{\tau_t}(\theta^*) - g_{\tau_t}(\hat{\theta}_{\tau_t}) \right\|_{M_{\tau_t}^{-1}(c_\mu)} \|x\|_{M_t^{-1}(c_\mu)} \leq \sqrt{C} \beta^x_{t+T_1}(\delta).
\end{aligned}$$

where the last inequality comes from the proof of Proposition C.3.1 similarly. The rest of the proof will be mostly identical to that of Theorem C.3.1 and hence we would copy it here for completeness: Based on Proposition C.3.2 we have

$$\begin{aligned}
\mu(x^{*\top} \theta^*) - \mu(x_t^\top \theta^*) \leq 2\sqrt{C} \beta^{x_t}_{t+T_1}\left(\frac{\delta}{2}\right) &= 2\sqrt{C} \alpha_{t+T_1}\left(\frac{\delta}{2}\right) \|x_t\|_{M_t^{-1}(c_\mu)} \\
&\leq 2\sqrt{C} \alpha_T \left(\frac{\delta}{2}\right) \|x_t\|_{M_t^{-1}(c_\mu)}.
\end{aligned}$$

Since we have that $\alpha_{T_2}(\delta/2) > k_\mu S_0^2$, the Regret of Algorithm 12 can be bounded as

$$\begin{aligned}
Regret_{T_2} &\leq 2k_\mu S_0^2 + \sum_{t=2}^{T_2} \min\{\mu(x^{*\top} \theta^*) - \mu(x_t^\top \theta^*), 2k_\mu S_0^2\} \\
&\leq 2k_\mu S_0^2 + 2\sqrt{C} \alpha_T \left(\frac{\delta}{2}\right) \sum_{t=2}^{T_2} \min\{\|x_t\|_{M_t^{-1}(c_\mu)}, 1\} \\
&\leq 2k_\mu S_0^2 + 2\sqrt{C} \alpha_T \left(\frac{\delta}{2}\right) \sqrt{T_2} \sqrt{\sum_{t=2}^{T_2} \min\{\|x_t\|^2_{M_t^{-1}(c_\mu)}, 1\}}.
\end{aligned}$$

Algorithm 13 Generalized Explore Subspace Then Transform (G-ESTT)

Input: Action set $\{\mathcal{X}_t\}$, $T, T_1, \mathcal{D}$, the probability rate δ, parameters for stage 2: $\lambda, \lambda_\perp$.

Stage 1: Subspace Estimation

1: Randomly choose $X_t \in \mathcal{X}$ according to $\mathcal{D}$ and record X_t, Y_t for $t = 1, \dots T_1$.

2: Obtain $\widehat{\Theta}$ by solving the following equation:

$$\widehat{\Theta} = \arg \min_{\Theta \in R^{d_1 \times d_2}} \frac{1}{T_1} \sum_{i=1}^{T_1} \{b(\langle X_i, \Theta \rangle) - y_i \langle X_i, \Theta \rangle\} + \lambda_{T_1} \|\Theta\|_{\text{nuc}}.$$

3: Obtain the full SVD of $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp] \, \widehat{D} \, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U}$ and $\widehat{V}$ contains the first r left-singular vectors and the first r right-singular vectors respectively.

Stage 2: Almost Low Rank Generalized Linear Bandit

4: Rotate the admissible parameter space: $\Theta' := [\widehat{U}, \widehat{U}_\perp]^\top \Theta [\widehat{V}, \widehat{V}_\perp]$, and transform the parameter set as:

$$\Theta_0 := \{\text{vec}(\Theta'_{1:r,1:r}), \text{vec}(\Theta'_{r+1:d_1,1:r}), \text{vec}(\Theta'_{1:r,r+1:d_2}), \text{vec}(\Theta'_{r+1:d_1,r+1:d_2})\}.$$

5: **for** $t \geq T - T_1$ **do**

6: Rotate the arm feature set: $\mathcal{X}'_t := [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X}_t [\widehat{V}, \widehat{V}_\perp]$.

7: Define the vectorized arm set so that the last $(d_1 - r) \cdot (d_2 - r)$ components are almost negligible:

$$\mathcal{X}_{0,t} := \{\text{vec}(\mathcal{X}'_{\{1:r,1:r\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,1:r\},t}), \text{vec}(\mathcal{X}'_{\{1:r,r+1:d_2\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,r+1:d_2\},t})\}.$$

8: Invoke LowGLM-UCB (PLowGLM-UCB or LowUCB-GLM) with the arm set $\mathcal{X}_{0,t}$, the parameter space Θ_0, the low dimension $k = (d_1 + d_2)r - r^2$ and penalization parameter $(\lambda_0, \lambda_\perp)$ for one round. Update the matrix $M_t(c_\mu)$ or V_t accordingly.

where the last ineuqlity comes from Cauchy-Schwarz inequality. Finally, by a self-normalized martingale inequality (Abbasi-Yadkori et al. (2011), Lemma 11) and and then plugging in the chosen value for $\lambda_\perp = \dfrac{c_\mu S_0^2 T}{k \log(1 + \frac{c_\mu S_0^2 T}{k\lambda_0})}$, we have

$$\begin{aligned} Regret_{T_2} &\leq 2k_\mu S_0^2 + \frac{2k_\mu}{c_\mu}\sqrt{C} \\ &\times \left(\sigma_0 \sqrt{2k \log\left(1 + \frac{c_\mu S_0^2}{k\lambda_0} T\right) - 2\log\left(\frac{\delta}{2}\right)} + \sqrt{c_\mu} \left(\sqrt{\lambda_0} S_0 + \sqrt{\frac{c_\mu S_0^2 T}{k \log\left(1 + \frac{c_\mu S_0^2}{k\lambda_0} T\right)}} S_\perp \right) \right) \\ &\times \sqrt{T_2} \sqrt{4k \log\left(1 + \frac{c_\mu S_0^2}{k\lambda_0} T\right)}, \end{aligned}$$

which gives us the final bound in Theorem C.8.1. □

C.8.3. Algorithms for the Contextual Setting. To show algorithm G-ESTT and G-ESTS for the contextual setting, where the arm set $\mathcal{X}_t = \{X_{i,t}\}$ may vary over time $t = [T]$, we would firstly update some notations besides the ones we have defined in Section 4.4.2. We denote the

Algorithm 14 Generalized Explore Subspace Then Subtract (G-ESTS)

Input: Action set $\{\mathcal{X}_t\}$, $T, T_1, \mathcal{D}$, the probability rate δ, parameters for stage 2: $\lambda, \lambda_\perp$.

Stage 1: Subspace Estimation

1: **for** $t = 1$ **to** T_1 **do**

2: Pull arm $X_t \in \mathcal{X}$ according to the distribution $\mathcal{D}$, observe payoff y_t.

3: Obtain $\widehat{\Theta}$ by solving the following equation:

$$\widehat{\Theta} = \arg \min_{\Theta \in R^{d_1 \times d_2}} \frac{1}{T_1} \sum_{i=1}^{T_1} \{b(\langle X_i, \Theta \rangle) - y_i \langle X_i, \Theta \rangle\} + \lambda_{T_1} \|\Theta\|_{\text{nuc}}.$$

4: Obtain the full SVD of $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp] \, \widehat{D} \, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U}$ and $\widehat{V}$ contains the first r left-singular vectors and the first r right-singular vectors respectively.

Stage 2: Low Rank Generalized Linear Bandit

5: Rotate the admissible parameter space: $\Theta' \coloneqq [\widehat{U}, \widehat{U}_\perp]^\top \Theta [\widehat{V}, \widehat{V}_\perp]$, and transform the parameter set as:

$$\Theta_0 \coloneqq \{\text{vec}(\Theta'_{1:r,1:r}), \text{vec}(\Theta'_{r+1:d_1,1:r}), \text{vec}(\Theta'_{1:r,r+1:d_2}), \text{vec}(\Theta'_{r+1:d_1,r+1:d_2})\}.$$

6: **for** $t \geq T - T_1$ **do**

7: Rotate the arm feature set: $\mathcal{X}'_t \coloneqq [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X}_t [\widehat{V}, \widehat{V}_\perp]$.

8: Define the vectorized arm set so that the last $(d_1 - r) \cdot (d_2 - r)$ components are almost negligible, and then drop the last $(d_1 - r) \cdot (d_2 - r)$ components:

$$\mathcal{X}_{0,sub,t} \coloneqq \{\text{vec}(\mathcal{X}'_{\{1:r,1:r\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,1:r\},t}), \text{vec}(\mathcal{X}'_{\{1:r,r+1:d_2\},t})\}.$$

9: Invoke any modern generalized linear (contextual) bandit algorithm with the arm set $\mathcal{X}_{0,sub,t}$, the parameter space $\Theta_{0,sub}$, and the low dimension $k = (d_1 + d_2)r - r^2$ for one round.

time-dependent action set $\mathcal{X}_t$ after rotation as:

$$\mathcal{X}'_t = [\widehat{U}, \widehat{U}_\perp]^\top \mathcal{X} [\widehat{V}, \widehat{V}_\perp],$$

And we modify the notations of the vectorized arm set for G-ESTT and G-ESTS defined in Eqn. (4.3), (4.13) accordingly for each iteration:

$$\mathcal{X}_{0,t} \coloneqq \{\text{vec}(\mathcal{X}'_{\{1:r,1:r\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,1:r\},t}), \text{vec}(\mathcal{X}'_{\{1:r,r+1:d_2\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,r+1:d_2\},t})\},$$

$$\mathcal{X}_{0,sub,t} \coloneqq \{\text{vec}(\mathcal{X}'_{\{1:r,1:r\},t}), \text{vec}(\mathcal{X}'_{\{r+1:d_1,1:r\},t}), \text{vec}(\mathcal{X}'_{\{1:r,r+1:d_2\},t})\}.$$

Details can be found in Algorithm 13 and 14.

C.9. Additional Experimental Details

C.9.1. Parameter Setup for Simulations. Here we present our parameter setting for algorithms involved in our experiment in Section 4.5.

Basic setup: horizon $T = 45000$. For the case where $d_1 = d_2 = 12$ and $r = 2$ we extend the

horizon until 75000 in figures to display the superiority of our proposed algorithms more clearly. The 480 (1000) random matrices are sampled uniformly from $d_1 d_2$-dimensional unit sphere.

LowESTR: (same setup as in Lu et al. (2021))

- failure rate: $\delta = 0.01$, the standard deviation: $\sigma = 0.01$ and the steps of stage 1: $T_1 = 1800$.
- penalization parameter in stage 1: $\lambda_{T_1} = 0.01\sqrt{\frac{1}{T_1}}$, and the gradient decent step size: 0.01.
- $B = 1, B_\perp = \frac{\sigma^2(d_1+d_2)^3 r}{T_1 D_{r,r}^2}, \lambda = 1, \lambda_\perp = \frac{T_2}{k\log(1+T_2/\lambda)}$, grid search for $\sqrt{\beta_t}$ with multiplier in $\{0.2, 1, 5\}$.

SGD-TS: (details in Ding et al. (2021))

- grid search for exploration rates in $\{0.1, 1, 10\}$.
- grid search for C in $\{1, 3, 5, 7\}$.
- grid search for initial step sizes in $\{0.01, 0.1, 1, 5, 10\}$.

G-ESTT: (LowGLM-UCB in Stage 2)

- failure rate: $\delta = 0.01$, and the steps of stage 1: $T_1 = 1800$.
- $S_0 = 1, \Theta = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \leq 1\}$ for the case $r = 1$, and $S_0 = 5, \Theta = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \leq 5\}$ for the case $r = 2$.
- penalization in solving Eqn. (4.6) with λ_{T_1} suggested in Theorem 4.4.1. (We believe that a simple grid search near this value would be better.)
- p_{ij} set to be centered normal distribution with standard deviation $1/d$ in Stage 1. Specifically, at each round we randomly select a matrix $X_{rand,t}$ based on this $\{p_{ij}\}$ elementwisely, and then pull the arm that is closest to $X_{rand,t}$ w.r.t. $\|\cdot\|_F$ among all candidates in the arm set.
- proximal gradient descent with backtracking line search solving Eqn. (4.6), step size set to 0.1.
- $\lambda_0 = 1, \lambda_\perp = \frac{c_\mu^2 S_0^2 T_2}{k\log\left(1+\frac{c_\mu S_0^2 T_2}{k\lambda_0}\right)}, S_\perp = \frac{d_1 d_2 r}{T_1 D_{rr}^2}\log\left(\frac{d_1+d_2}{\delta}\right)$, grid search for exploration bonus with multiplier in $\{0.2, 1, 5\}$.

G-ESTS: (SGD-TS in Stage 2)

- The steps of stage 1: $T_1 = 1800$.
- penalization in solving Eqn. (4.6) with λ_{T_1} suggested in Theorem 4.4.1. (We believe that a simple grid search near this value would be better.)

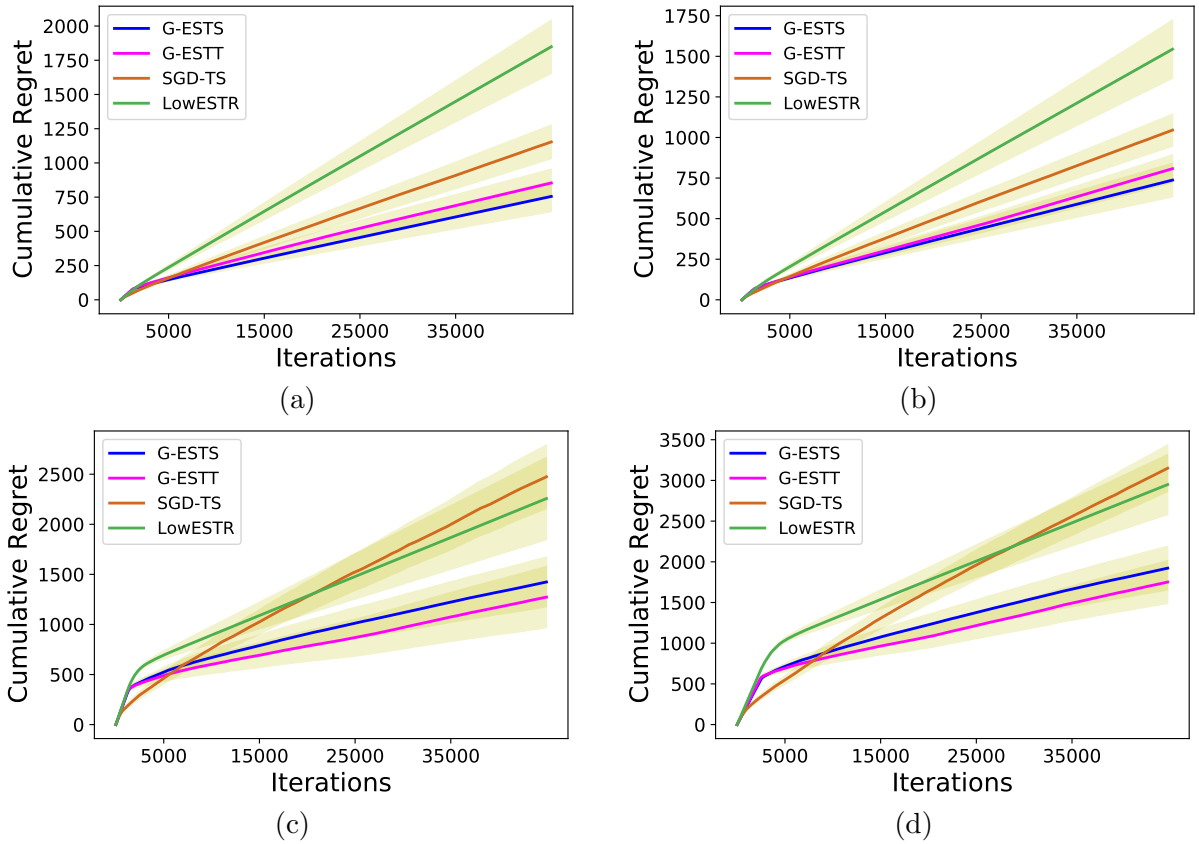


FIGURE C.1. Plots of regret curves of algorithm G-ESTT, G-ESTS, SGD-TS and LowESTR under four settings (1000 arms). (a): diagonal Θ^* $d_1 = d_2 = 10, r = 1$; (b): diagonal Θ^* $d_1 = d_2 = 12, r = 1$; (c): non-diagonal Θ^* $d_1 = d_2 = 10, r = 2$; (d): non-diagonal Θ^* $d_1 = d_2 = 12, r = 2$.

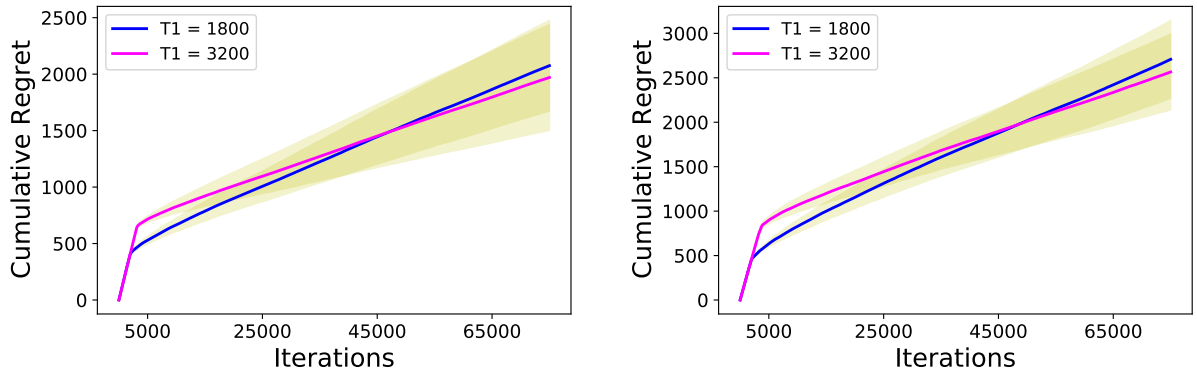
- p_{ij} set to be centered normal distribution with standard deviation $1/d$ in Stage 1. Specifically, at each round we randomly select a matrix $X_{rand,t}$ based on this $\{p_{ij}\}$ elementwisely, and then pull the arm that is closest to $X_{rand,t}$ w.r.t. $\|\cdot\|_F$ among all candidates in the arm set.
- proximal gradient descent with backtracking line search solving Eqn. (4.6), step size set to 0.1.
- use the same setup for SGD-TS as we have listed.

C.9.2. Additonal experimental results. Here we display the regret curves of algorithms under four settings with 1000 arms in Figure C.1, where our proposed G-ESTS and G-ESTT also dominate other methods regarding both accuracy and computation.

C.9.3. Comparison between G-ESTT and G-ESTS. In this section we compare the performance of our two frameworks G-ESTT and G-ESTS, and it is obvious that both these two proposed methods work better than the existing LowESTR and state-of-the-art generalized linear

bandit algorithms under our problem setting based on Figure 4.1 and C.1. Notice that G-ESTT and G-ESTS perform similarly well under the scenario $r = 1$ (G-ESTS is slightly better). However, for the case $r = 2$, we find that G-ESTT achieve less cumulative regret than G-ESTS does. We believe it is because that, on the one hand, G-ESTS depends more on the precision of estimate $\widehat{\Theta}$, which becomes more challenging for the case $r = 2$. On the other hand, for G-ESTS how to reuse the random-selected actions in stage 1 is also tricky, and we will leave it as a future work. Therefore, G-ESTT (with LowUCB-GLM) quickly takes the lead in the very beginning of stage 2 since LowUCB-GLM can yield a consistent estimator early in stage 2 by reclaiming the randomly-chosen actions.

However, we find that G-ESTS is incredibly faster than other methods (including G-ESTT) as it only spends about one tenth of the running time of LowESTR until convergence as shown in Table 4.1. Notice that G-ESTT with LowUCB-GLM is a little bit slower since it utilizes more samples for estimation in each iteration for better performance. Moreover, we conduct another simulation for the case $r = 2, d_1 = d_2 = 12$ where we additionally choose $T_1 = 3200$, and the results are displayed in Figure C.2 after 100 times repeated simulations. We observe that by appropriately enlarging the length of stage 1 (T_1), G-ESTS would perform better in the long run as we expect, since a more accurate estimation of Θ^* could be obtained. Therefore, we can conclude our proposed G-ESTS could perform prominently with parsimonious computation by mildly tuning the length of stage 1 (T_1).



(a) (b)

FIGURE C.2. Plots of regret curves of algorithm G-ESES the scenario $d_1 = d_2 = 12, r = 2$ under $T_1 = 1800$ and $T_1 = 3200$ (a): fixed 480 arms; (b): fixed 1000 arms.

TABLE C.1. Comparison between our proposed Stein's lemma-based method and the log-likelihood maximization method for low-rank matrix subspace estimation.

Case	Low-rank detection method	Regret	Transformed error
Figure 4.1(a)	Stein's lemma-based method	G-ESTT:723.27, G-ESTS:510.80	0.086
	Log-likelihood maximization	G-ESTT:724.96, G-ESTS:515.25	0.089
Figure 4.1(c)	Stein's lemma-based method	G-ESTT:1088.26, G-ESTS:1106.71	0.542
	Log-likelihood maximization	G-ESTT:1136.54, G-ESTS:1198.39	0.583

C.9.4. Comparison with other matrix subspace detection methods. To pre-check the efficiency of our Stein's lemma-based method for subspace estimation, we also tried the nuclear-norm regularized log-likelihood maximization with its details introduced in the following Appendix C.9. Particularly, we could solve the regularized negative log-likelihood minimization problem with nuclear norm penalty as shown in Eqn. (C.9).

Specifically, we consider the two cases of our simulations: 480 arms, d =10, r =1 (Figure 4.1(a) case) and 480 arms, d =10, r =2 (Figure 4.1(c) case). We used the same setting as described in Appendix C.9 above (T_1=1800, T=45000), and implemented proximal gradient descent with the backtracking line search for optimization. The average regret cumulative regret along with the average transformed error $\left\|\theta^*_{(k+1):p}\right\|_2$ defined in Eqn. (4.8) are reported in Table C.1.

Therefore, we can see that our low-rank matrix detection method outperforms the regularized log-likelihood maximization method, especially when the underlying parameter matrix is complicated (Figure 4.1(c) case). This is also consistent with our theoretical analysis, as we will show in the following Appendix C.10 that the theoretical bound of loss $\left\|\hat{\Theta} - \Theta^*\right\|_{\mathrm{F}}^2$ is of order d^3r/T_1 using the regularized log-likelihood maximization method, which is worse than the convergence rate of our proposed method in Theorem 4.4.1.

C.10. Bonus: Matrix Estimation with Restricted Strong Convexity

C.10.1. Methodology. As we have mentioned in our main paper, we can achieve a decent matrix recovery rate regarding the Frobenius norm by using generalized first-order Stein's Lemma on Eqn. (4.6). For the completeness of our work, we also approach the matrix estimation problem by using the restricted strong convexity theory alternatively to see whether we could get the same covergence rate $O(\sqrt{d_1+d_2)^3r/T_1})$ in GLM as in the linear case under the stronger assumptions of sub-Gaussian property. Specifically, we use the regularized negative log-likelihood minimization

with nuclear norm penalty for the loss function in stage 1, and consequently we are able to get the same bound as in the linear case. Notice that this work is also non-trivial since constructing the restricted strong convexity for the generalized linear low-rank matrix estimation requires us to use a truncation argument and a peeling technique (Raskutti et al., 2010), which is completely different that used in simple linear case (Lu et al., 2021). Therefore, to facilitate further study in this area and for the completeness of our work, we would present the detailed proof here in the following as a bonus. Loss function: we consider the following well-defined regularized negative log-likelihood minimization problem with nuclear norm penalty in stage 1:

$$\widehat{\Theta} = \arg\min_{\Theta \in R^{d_1 \times d_2}} L_{T_1}(\Theta) + \lambda_{T_1} \|\Theta\|_{\text{nuc}}, \quad \text{where}$$
$$L_{T_1}(\Theta) = \frac{1}{T_1} \sum_{i=1}^{T_1} \{b(\langle X_i, \Theta \rangle) - y_i \langle X_i, \Theta \rangle\}, \tag{C.9}$$

Note the problem defined in Eqn. (C.9) is convex and hence can be easily solved by gradient-based algorithms. (Boyd et al., 2004; Kingma & Ba, 2014; Wang et al., 2023) Next, we first present different assumptions with notations reloaded:

Assumption C.10.1. There exists a sampling distribution $\mathcal{D}$ over $\mathcal{X}$ with covariance matrix of $\text{vec}(X)$ as $\Sigma \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$, such that $\lambda_{\min}(\Sigma) = \lambda_1$ and $\text{vec}(X)$ is sub-Gaussian with parameter $\sigma = \lambda_2$ such that λ_1 / λ_2^2 can be absolutely bounded.

Assumption C.10.2. The norm of true parameter Θ^* and feature matrices in $\mathcal{X}$ is bounded: there exists $S \in \mathbb{R}^+$ such that for all arms $X \in \mathcal{X}$, $\|X\|_F, \|\Theta^*\|_F \leq S$; $\|X\|_{\text{op}}, \|\Theta^*\|_{\text{op}} \leq S_2$ ($S_2 \leq S$).

Assumption C.10.3. The inverse link function $\mu(\cdot)$ is continuously differentiable, Lipschitz with constant k_μ. $c_\mu \geq \inf_{\Theta \in \Theta, X \in \mathcal{X}} \mu'(\langle X, \Theta \rangle) > 0$ and $c_\mu \geq \inf_{\{|x| < (S+2)\sigma c_2\}} \mu'(x) > 0$ for some constant c_2.

Here we could safely choose $\sigma = 1/\sqrt{d_1 d_2}$ (Lu et al., 2021) as default. Without loss of generality, we can assume that $c_- \sigma^2 \leq \{\lambda_1, \lambda_2^2\} \leq c_+ \sigma^2$ for some absolute constant c_-, c_+ for the simplicity of following theoretical analysis. Assumption C.10.1 implies that if X is sampled from the distribution

$\mathcal{D}$, then for any $\Delta \in \mathbb{R}^{d_1 \times d_2}$ satisfying $\|\Delta\|_F \leq 1$, we have:

$$\mathbb{E}[\langle X, \Delta \rangle^2] = \mathrm{vec}(\Delta)^\top \Sigma \mathrm{vec}(\Delta) \geq \lambda_1 \geq c_- \, \sigma^2 \coloneqq \alpha; \tag{C.10}$$

$$\mathbb{E}[\langle X, \Delta \rangle^4] \leq 16 \lambda_2^4 \leq 16 c_+^2 \, \sigma^4 \coloneqq \beta. \tag{C.11}$$

C.10.2. Theorem.

THEOREM C.10.4. (Bounds for GLM via another loss function in Eqn. (C.9)) *For any low-rank generalized linear model with samples $X_1 \ldots, X_{T_1}$ drawn from $\mathcal{X}$ according to $\mathcal{D}$ in Assumption C.10.1, and Assumption C.10.2, C.10.3 hold. Then the optimal solution to the nuclear norm regularization problem (C.9) with $\lambda_{T_1} = \Omega(\sigma \sqrt{(d - \log(\delta))/T_1})$ would satisfy:*

$$\left\| \widehat{\Theta} - \Theta^* \right\|_F^2 \asymp \frac{d}{T_1 \, \sigma^2} r \asymp \frac{d^3 r}{T_1}, \tag{C.12}$$

with probability at least $1 - \delta$ given the condition $d\, r \lesssim \sigma^2 T_1$ and $(1 + \sigma)^2 d\, r \lesssim T_1$ hold.

To prove this theorem, roughly speaking we firstly deduce the restricted strong convexity condition for our optimization problem with high probability, and then extend some previous results on the oracle inequality of estimation error.

C.10.3. Restricted Strong Convexity.

DEFINITION C.10.5. (Restricted strong convexity (RSC), (Negahban et al., 2012)). *Given the cost function $L_{T_1}(\Theta)$ defined in (4.6) and $X_1, \ldots, X_{T_1} \in \mathbb{R}^{d_1 \times d_2}$, the first-order Taylor-series error is defined as:*

$$\mathcal{E}_{T_1}(\Delta) \coloneqq L_{T_1}(\Theta^* + \Delta) - L_{T_1}(\Theta^*) - \langle \nabla L_{T_1}(\Theta^*), \Delta \rangle.$$

For a given norm $\|\cdot\|$ and regularizer $\Phi(\cdot)$, the cost function satisfies a restricted strong convexity (RSC) condition with radius $R > 0$, curvature $\kappa > 0$ and tolerance τ^2 if

$$\mathcal{E}_{T_1}(\Delta) \geq \frac{\kappa}{2} \|\Delta\|_F^2 - \tau_{T_1}^2 \Phi^2(\Delta), \quad \text{for all } \|\Delta\|_F \leq R.$$

THEOREM C.10.6. (RSC for GLM under distribution $\mathcal{D}$). *Consider any low-rank generalized linear model with samples $X_1 \ldots, X_{T_1}$ drawn from $\mathcal{X}$ according to $\mathcal{D}$ in Assumption C.10.1, and Assumption C.10.2 and C.10.3 hold. Then there exists constants c_3, c_4 such that with probability $1 - \delta$, we*

have the RSC *condition holds:*

$$\text{(C.13)}\quad \mathcal{E}_{T_1}(\Delta) \geq c_3\sigma^2 c_\mu \|\Delta\|_F^2 - (c_4\sigma^2 + 2\sigma)\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) c_\mu \|\Delta\|_{\text{nuc}}^2 \quad \text{for all } \|\Delta\|_F \leq 1$$

$$\text{with } \kappa = c_3\sigma^2 c_\mu,\ \tau_{T_1}^2 = (c_4\sigma^2 + 2\sigma)\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) c_\mu,\ R = 1,\ \|\cdot\| = \|\cdot\|_F \text{ and } \Phi(\cdot) = \|\cdot\|_{\text{nuc}}$$

for $T_1 = O(\log(\log_2(d)/\delta))$.

REMARK C.10.1. *The radius* R *in Theorem C.10.6 can be adapted to any finite positive constant keeping the same proof outline. And the required sample size* T_1 *only change in logarithmic power, which can be easily satisfied.*

C.10.3.1. *Proof of Theorem C.10.6.* To prove theorem C.10.6, we use a truncation argument and the peeling technique (Raskutti et al., 2010; Wainwright, 2019):
Using the property of the remainder in the Taylor series, we have

$$\mathcal{E}_{T_1}(\Delta) = \frac{1}{T_1}\sum_{i=1}^{T_1} \mu'\left(\langle X_i, \Theta^*\rangle + t\langle X_i, \Delta\rangle\right)\langle X_i, \Delta\rangle^2,$$

for some $t \in [0,1]$. Based on (C.10) and (C.11) we will set two truncation parameters $K_1^2 = 4\beta/\alpha$ and $K_2^2 = 4\beta S^2/\alpha$ for further use. For any $\|\Delta\|_F = \delta \in (0,1]$, we set $\tau = K_1\delta$ and a trunction function $\phi_\tau(v) = v^2 \cdot I_{\{|v|\leq 2\tau\}}$. Then we have:

$$\mathcal{E}_{T_1}(\Delta) \geq \frac{1}{T_1}\sum_{i=1}^{T_1} \mu'\left(\langle X_i, \Theta^*\rangle + t\langle X_i, \Delta\rangle\right)\phi_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i, \Theta^*\rangle|\leq K_2\}}.$$

The right hand side would always be 0 if $|\langle X_i, \Theta^*\rangle + t\langle X_i, \Delta\rangle| > 2K_1 + K_2$, which implies the following result based on Assumption C.10.3:

$$\mathcal{E}_{T_1}(\Delta) \geq c_\mu \frac{1}{T_1}\sum_{i=1}^{T_1} \phi_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i, \Theta^*\rangle|\leq K_2\}}.$$

Therefore, it suffices to how that for all $\delta \in (0,1]$ and for $\|\Delta\|_F = \delta$, we have:

$$\text{(C.14)}\qquad \frac{1}{T_1}\sum_{i=1}^{T_1} \phi_{\tau(\delta)}(\langle X_i, \Delta\rangle) I_{\{|\langle X_i, \Theta^*\rangle|\leq K_2\}} \geq a_1\delta^2 - a_2\|\Delta\|_{\text{nuc}}\,\delta,$$

for some parameters a_1 and a_2 since the inequality $\|\Delta\|_F \leq \|\Delta\|_{\text{nuc}}$ always holds. Note the fact that $\phi_{\tau(\delta)}(\langle X_i, \Delta\rangle) = \delta^2\phi_{\tau(1)}(\langle X_i, \Delta/\delta\rangle)$, then for any $\|\Delta\|_F = \delta$ such that $\delta \in (0,1]$, we can apply

bound (C.14) to the rescaled unit-norm matrix Δ/δ to obtain:

$$\frac{1}{T_1}\sum_{i=1}^{T_1}\phi_\tau(1)(\langle X_i, \Delta/\delta\rangle)I_{\{|\langle X_i,\Theta^*\rangle|\le K_2\}} \ge a_1 - a_2\left\|\Delta/\delta\right\|_{\text{nuc}},$$

which implies that it suffices to show (C.14) holds when $\delta = 1$, i.e.

$$\frac{1}{T_1}\sum_{i=1}^{T_1}\phi_\tau(\langle X_i, \Delta\rangle)I_{\{|\langle X_i,\Theta^*\rangle|\le K_2\}} \ge a_1 - a_2\left\|\Delta\right\|_{\text{nuc}}, \quad \text{for all } \left\|\Delta\right\|_F = 1.$$

Then we can construct another truncation function $\tilde{\phi}_\tau(v)$ with parameter at most $2\tau = 2K_1$ as

$$\tilde{\phi}_\tau(v) = v^2 I_{\{|v|\le\tau\}} + (v-2\tau)^2 I_{\{\tau<v\le 2\tau\}} + (v+2\tau)^2 I_{\{-2\tau\le v<-\tau\}}.$$

Then it suffices to show that

$$\frac{1}{T_1}\sum_{i=1}^{T_1}\tilde{\phi}_\tau(\langle X_i, \Delta\rangle)I_{\{|\langle X_i,\Theta^*\rangle|\le K_2\}} \ge a_1 - a_2\left\|\Delta\right\|_{\text{nuc}}, \quad \text{for all } \left\|\Delta\right\|_F = 1.$$

And for a given radius $r \ge 1$, define the random variable

$$Z_{T_1}(r) = \sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\le r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1}\tilde{\phi}_\tau(\langle X_i, \Delta\rangle)I_{\{|\langle X_i,\Theta^*\rangle|\le K_2\}} - \mathbb{E}\left(\tilde{\phi}_\tau(\langle X, \Delta\rangle)I_{\{|\langle X,\Theta^*\rangle|\le K_2\}}\right)\right|.$$

Firstly, we can prove that

$$\mathbb{E}[\tilde{\phi}_\tau(\langle X, \Delta\rangle)I_{\{|\langle X,\Theta^*\rangle|\le K_2\}}] \ge \frac{1}{2}\alpha, \tag{C.15}$$

by using the chosen values for K_1 and K_2 to show that

$$\mathbb{E}[\tilde{\phi}_\tau(\langle X, \Delta\rangle)] \ge \frac{3}{4}\alpha, \quad \mathbb{E}[\tilde{\phi}_\tau(\langle X, \Delta\rangle)I_{\{|\langle X,\Theta^*\rangle|> K_2\}}] \le \frac{1}{4}\alpha.$$

Specifically, since we have

$$\mathbb{E}[\tilde{\phi}_\tau(\langle X, \Delta\rangle)] \ge \mathbb{E}[\langle X, \Delta\rangle^2 I_{\{|\langle X,\Theta^*\rangle|\le\tau\}}] \ge \alpha - \mathbb{E}[\langle X, \Delta\rangle^2 I_{\{|\langle X,\Theta^*\rangle|>\tau\}}]$$

And we can show that the last term is at most $\alpha/4$ based on the Markov's inequality and Cauchy-Schwarz inequality:

$$\mathbb{E}[\langle X, \Delta\rangle^2 I_{\{|\langle X,\Theta^*\rangle|>\tau\}}] \le \sqrt{\mathbb{E}[\langle X, \Delta\rangle^4]}\sqrt{P([|\langle X, \Theta^*\rangle| > \tau])} \le \sqrt{\beta}\sqrt{\frac{\beta}{\tau^4}} \le \frac{\alpha}{4}.$$

And similarly we can prove that $\mathbb{E}[\tilde{\phi}_\tau(\langle X, \Delta\rangle) I_{\{|\langle X,\Theta^*\rangle|>K_2\}}] \leq \alpha/4$. On the other hand, by our choice $\tau = K_1$, the empirical process defining $Z_{T_1}(r)$ is based on functions bounded in absolute value by K_1^2. Thus, the functional Hoeffding inequality (Theorem 3.26 in Wainwright (2019)) implies that

$$P\left(Z_{T_1}(r) \geq \mathbb{E}(Z_{T_1}(r)) + \sigma r\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) + \frac{\alpha}{4}\right) \leq$$

$$\exp\left(-\frac{n\left(\sigma r\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) + \frac{\alpha}{4}\right)^2}{4K_1^4}\right). \tag{C.16}$$

To bound the expected value term $\mathbb{E}(Z_{T_1}(r))$, we introduce an i.i.d sequance of Rademacher variables $\{\varepsilon_i\}_{i=1}^{T_1}$ and then use the symmetrization argument:

$$\begin{aligned}
\mathbb{E}(Z_{T_1}(r)) &= \mathbb{E}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \tilde{\phi}_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i,\Theta^*\rangle|\leq K_2\}} - \mathbb{E}\left(\tilde{\phi}_\tau(\langle X, \Delta\rangle) I_{\{|\langle X,\Theta^*\rangle|\leq K_2\}}\right)\right|\right] \\
&= \mathbb{E}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \tilde{\phi}_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i,\Theta^*\rangle|\leq K_2\}} - \mathbb{E}\left(\frac{1}{T_1}\sum_{i=1}^{T_1} \tilde{\phi}_\tau(\langle Y_i, \Delta\rangle) I_{\{|\langle Y_i,\Theta^*\rangle|\leq K_2\}}\right)\right|\right] \\
&\leq \mathbb{E}_{X_i,Y_i}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \tilde{\phi}_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i,\Theta^*\rangle|\leq K_2\}} - \frac{1}{T_1}\sum_{i=1}^{T_1} \tilde{\phi}_\tau(\langle Y_i, \Delta\rangle) I_{\{|\langle Y_i,\Theta^*\rangle|\leq K_2\}}\right|\right] \\
&= \mathbb{E}_{X_i,Y_i,\varepsilon_i}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \varepsilon_i\left(\tilde{\phi}_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i,\Theta^*\rangle|\leq K_2\}} - \tilde{\phi}_\tau(\langle Y_i, \Delta\rangle) I_{\{|\langle Y_i,\Theta^*\rangle|\leq K_2\}}\right)\right|\right] \\
&\leq 2\mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \varepsilon_i \tilde{\phi}_\tau(\langle X_i, \Delta\rangle) I_{\{|\langle X_i,\Theta^*\rangle|\leq K_2\}}\right|\right]
\end{aligned} \tag{C.17}$$

$$\overset{\text{(i)}}{\leq} 8K_1\mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \left|\frac{1}{T_1}\sum_{i=1}^{T_1} \varepsilon_i\langle \Delta, X_i\rangle\right|\right] \overset{\text{(ii)}}{\leq} 8K_1\, r\cdot \mathbb{E}_{X_i,\varepsilon_i}\left[\left\|\frac{1}{T_1}\sum_{i=1}^{T_1} \varepsilon_i X_i\right\|_{\text{op}}\right],$$

where the inequality (i) comes from Rademacher contraction property and (ii) is by the duality between matrix $\|\cdot\|_2$ and $\|\cdot\|_{\text{nuc}}$ norms. Using the previous conclusion (Exercise 9.8 in Wainwright (2019)), we have

$$\mathbb{E}_{X_i,\varepsilon_i}\left[\left\|\frac{1}{T_1}\sum_{i=1}^{T_1}\varepsilon_i X_i\right\|_{\text{op}}\right] \leq \sigma c_5 c_+ \left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right), \tag{C.18}$$

where c_5 is an independent absolute constant. Combine (C.16), (C.17) and (C.18), we have

(C.19)

$$P\left(Z_{T_1}(r) \geq (8K_1c_5c_+ + 1)\sigma r\left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right)+\frac{\alpha}{4}\right) \leq \exp\left(-\frac{T_1\left(\sigma r\left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right)+\frac{\alpha}{4}\right)^2}{4K_1^4}\right).$$

According to (C.15) and (C.19), we prove the following conclusion for any fixed value of radium r:

$$P\left(\sup_{\substack{\|\Delta\|_F=1,\\ \|\Delta\|_{\text{nuc}}\leq r}} \mathcal{E}_{T_1}(\Delta) < \frac{1}{4}\alpha c_\mu - (8K_1c_5c_+ + 1)\left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right)\sigma c_\mu r\right) \leq$$

$$\exp\left(-\frac{T_1\left(\sigma r\left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right)+\frac{\alpha}{4}\right)^2}{4K_1^4}\right). \tag{C.20}$$

Since we have $\|\Delta\|_F = 1$, based on Cauchy-Schwarz inequality we have $1 \leq \|\Delta\|_{\text{nuc}} \leq \sqrt{d}$. To prove the RSC we use a peeling argument to extend r to all possible values. Define the event:

$$E := \Bigg\{\text{There exists } \Delta \text{ s.t. } \|\Delta\|_F = 1,\ \mathcal{E}_{T_1}(\Delta) < \frac{1}{4}\alpha c_\mu - (16K_1c_5c_+ + 2) \times \left(\sqrt{\frac{d_1}{T_1}}+\sqrt{\frac{d_2}{T_1}}\right)\sigma c_\mu \|\Delta\|_{\text{nuc}}\Bigg\} \tag{C.21}$$

$$V_i := \{2^{i-1} \leq \|\Delta\|_{\text{nuc}} < 2^i\}, \quad i = 1, \ldots, \left\lceil \frac{1}{2}\log_2(d)\right\rceil + 1.$$

Then we can conclude that $E \subseteq \bigcup_{i=1}^{\lceil \frac{1}{2}\log_2(d)\rceil+1}(E \cap V_i)$. And we can show the probability of each partition event $(E \cap V_i)$ can be upper bounded by (C.20):

$$\begin{aligned}
P(E \cap V_i) =& P\left(\sup_{\substack{\|\Delta\|_F=1,\\ 2^{i-1}\le\|\Delta\|_{\text{nuc}}<2^i}} \mathcal{E}_{T_1}(\Delta) < \frac{1}{4}\alpha c_\mu - (16K_1c_5c_+ + 2)\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right)\sigma c_\mu \|\Delta\|_{\text{nuc}} \right) \\
&\le P\left(\sup_{\substack{\|\Delta\|_F=1,\\ 2^{i-1}\le\|\Delta\|_{\text{nuc}}<2^i}} \mathcal{E}_{T_1}(\Delta) < \frac{1}{4}\alpha c_\mu - (8K_1c_5c_+ + 1)\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right)\sigma c_\mu 2^i \right) \\
&\le \exp\left(-\frac{T_1\left(2^i\sigma\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) + \frac{\alpha}{4}\right)^2}{4K_1^4} \right),
\end{aligned}$$

which implies that

$$P(E) \le \log_2(d)\exp\left(-\frac{T_1\left(2\sigma\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) + \frac{\alpha}{4}\right)^2}{4K_1^4} \right).$$

We complete our proof of Theorem C.10.6 by noticing that the constants c_3, c_4 in (C.13) only depend on the absolute constants c_5, c_+ and c_- through our proof. □

C.10.4. Technical Lemmas.

LEMMA C.10.6.1. (Bound for GLM with nuclear regualarization, (Negahban et al., 2012; Wainwright, 2019)) *Consider the negative log-likelihood cost function $L_{T_1}(\cdot)$ defined in 4.6 and observations $X_1, \ldots, X_{T_1}$ satisfy a specific* RSC *condtion in Definition 1, such that*

$$\mathcal{E}_{T_1}(\Delta) \ge \frac{\kappa}{2}\|\Delta\|_F^2 - \tau_{T_1}^2\|\Delta\|_{\text{nuc}}^2, \quad \text{for all } \|\Delta\| \le 1.$$

Then under the "good" event: $\mathcal{G}(\lambda_{T_1}) \coloneqq \{\|\nabla L_{T_1}(\Theta^*)\|_{\text{op}} \le \lambda_{T_1}/2\}$, and the following two conditions hold:

$$\tau_{T_1}^2 r \le \frac{\kappa}{128}, \quad 4.5\frac{\lambda_{T_1}^2}{\kappa^2}r \le 1.$$

Then any optimal solution to Eqn. C.9 satisfies the bound

$$\left\|\widehat{\Theta} - \Theta^*\right\|_F^2 \le 4.5\frac{\lambda_{T_1}^2}{\kappa^2}r. \tag{C.22}$$

□

C.10.5. Proof of Theorem C.10.4. According to Theorem C.10.6, there exists two absolute constants c_3, c_4 such that with probability at least $1-\delta$, we have the RSC condition holds:

$$\mathcal{E}_{T_1}(\Delta) \geq c_3\sigma^2 c_\mu \|\Delta\|_F^2 - (c_4\sigma^2 + 2\sigma)\left(\sqrt{\frac{d_1}{T_1}} + \sqrt{\frac{d_2}{T_1}}\right) c_\mu \|\Delta\|_{\text{nuc}}^2 \quad \text{for all } \|\Delta\|_F \leq 1.$$

To implement Lemma 1, we would like to to figure out the value for regularization parameter λ_{T_1} such that the event $\mathcal{G}(\lambda_{T_1})$ can hold with high probability and simultaneously the bound in (C.22) can be well controlled. The proof is by using the covering argument and Bernstein's inequality to bound the operator norm.

Let $\xi_i = \langle X_i, \Theta^* \rangle$, we have $\|\nabla L_{T_1}(\Theta^*)\|_{\text{op}} = \left\|\frac{1}{n}\sum_{i=1}^{T_1}(b'(\xi_i) - y_i)X_i\right\|_{\text{op}}$, and for all $i \in [T_1]$

$$\mathbb{E}[(b'(\xi_i) - y_i)X_i] = \mathbb{E}\left[X_i\, \mathbb{E}[b'(\xi_i) - y_i \,|\, X_i]\right] = 0.$$

Let $\mathcal{S}^{d_1}$ ($\mathcal{S}^{d_2}$) be the d_1 (d_2) dimensional Euclidean-norm unit sphere, and $\mathcal{N}^{d_1}$ ($\mathcal{N}^{d_2}$) be the 1/4 covering on $\mathcal{S}^{d_1}$ ($\mathcal{S}^{d_2}$) and $\Xi(A) = \sup_{\substack{u\in\mathcal{N}^{d_1},\\ v\in\mathcal{N}^{d_2}}} u^\top A v$ for all $A \in \mathbb{R}^{d_1\times d_2}$. By the proof of Lemma 1 in Fan et al. (2019), we know that

$$\|A\|_{\text{op}} \leq \frac{16}{7}\Xi(A). \tag{C.23}$$

Besides, based on the properties of Orlicz-1 norm and Orlicz-2 norm, we have:

$$\left\|(b'(\xi_i) - y_i)u^\top X_i v\right\|_{\psi_1} \leq \left\|(b'(\xi_i) - y_i)\right\|_{\psi_2}\left\|u^\top X_i v\right\|_{\psi_2} \leq c_6\sqrt{k_\mu}\lambda_2, \quad \text{for all } u \in S^{d_1},\, v \in S^{d_2}.$$

For some absolute constant c_6 (e.g. $c_6 = 6$). Then for any fixed $u \in S^{d_1}$, $v \in S^{d_2}$, by Berstein's inequality we have

$$P\left(\left|\frac{1}{T_1}\sum_{i=1}^{T_1}(b'(\xi_i) - y_i)u^\top X_i v\right| > t\right) \leq 2\exp\left[-c_7\min\left(\frac{T_1 t^2}{c_6^2 k_\mu \lambda_2^2}, \frac{T_1 t}{c_6\sqrt{k_\mu}\lambda_2}\right)\right].$$

Then by the combination over all the union bounds and relation (C.23) we can claim that

$$P\left(\left\|\frac{1}{T_1}\sum_{i=1}^{T_1}(b'(\xi_i) - y_i)X_i\right\|_{\text{op}} > \frac{16}{7}t\right) \leq 2\,7^{d_1+d_2}\exp\left[-c_7\min\left(\frac{T_1 t^2}{c_6^2 k_\mu \lambda_2^2}, \frac{T_1 t}{c_6\sqrt{k_\mu}\lambda_2}\right)\right].$$

Then the event $\{\|\nabla L_{T_1}(\Theta^*)\|_2 \geq \frac{16}{7}t\}$ holds with probability $1-\delta$ if

$$t = \sqrt{k_\mu}\lambda_2 \max\left(\sqrt{\frac{c_6(d_1+d_2)\log(7)+c_6\log(2/\delta)}{T_1}}, \frac{c_6(d_1+d_2)\log(7)+c_6\log(2/\delta)}{T_1}\right)$$
$$= \Omega\left(\sqrt{\frac{d_1+d_2-\log(\delta)}{T_1}}\,\sigma\right).$$

Since we assume $(d_1+d_2) \lesssim T_1$. By taking $\lambda_{T_1} = \frac{32}{7}t \asymp \sqrt{\frac{d_1+d_2-\log(\delta)}{T_1}}\,\sigma$. We complete the proof of Theorem C.10.4 and obtain the scale of the bound in (C.12) after plugging the chosen values of κ and λ_{T_1} into (C.22). □

Notice that the loss function here shown in Eqn. (C.9) is also convex and hence could be solved by a wide class of optimization methods (e.g. subgradient descent algorithm), and we have the convergence rate of matrix estimation as

$$\left\|\widehat{\Theta}-\Theta^*\right\|_F = \widetilde{O}\left(\sqrt{\frac{d^3 r}{T_1}}\right).$$

APPENDIX D

Appendix for Chapter 5

D.1. Remarks of Assumption 5.3.2

We will show that when a series of iid random matrices $X_{i}{}_{i=1}^{m}$ follows a sub-Gaussian distribution with parameter $\sigma \asymp \frac{1}{\sqrt{d_1 d_2}}$, then the scale of $\max_{i\in[m]} \|X_i\|_{\mathrm{F}}$ can be bounded by some constant up to some very small logarithmic terms. The results can be directly deduced from the following Lemma:

LEMMA D.1.0.1. *If iid random matrices $X_{i}{}_{i=1}^{m} \in \mathbb{R}^{d_1\times d_2}$ follows a sub-Gaussian distribution with parameter σ, then with probability at least $1-\delta$ it holds that:*

$$\|X_i\|_F \leq 4\sigma\sqrt{d_1 d_2} + 2\sqrt{2}\sigma\sqrt{\ln(\frac{m}{\delta})}, \quad \forall i \in [m].$$

PROOF. Denote $\mathcal{N}_{\frac{1}{2}}$ as the $\frac{1}{2}$-covering of the matrix space $\{X : \|X\|_{\mathrm{F}} \leq 1\}$, then it holds that $|\mathcal{N}_{\frac{1}{2}}| \leq (1+1/0.5)^{d_1 d_2} = 5^{d_1 d_2}$. And for $\|V\|_{\mathrm{F}} \leq 1$ we define $S(V)$ as the closest point in $\mathcal{N}_{\frac{1}{2}}$ such that $\|V - S(V)\|_{\mathrm{F}} \leq \frac{1}{2}$. Next, we can have that

$$\begin{aligned}\|X_i\|_{\mathrm{F}} = \max_{\|V\|_{\mathrm{F}}=1} \langle V, X_i\rangle &= \max_{\|V\|_{\mathrm{F}}=1} \langle V - S(V) + S(V), X_i\rangle \leq \max_{Z\in\mathcal{N}_{\frac{1}{2}}} \langle Z, X_i\rangle + \max_{\|W\|_{\mathrm{F}}=\frac{1}{2}} \langle W, X_i\rangle \\ &\leq \max_{Z\in\mathcal{N}_{\frac{1}{2}}} \langle Z, X_i\rangle + \frac{1}{2}\max_{\|W\|_{\mathrm{F}}=1} \langle W, X_i\rangle,\end{aligned}$$

which indicates that $\|X_i\|_{\mathrm{F}} \leq 2\max_{Z\in\mathcal{N}_{\frac{1}{2}}} \langle Z, X_i\rangle$. Therefore, it holds that for any $t > 0$

$$P\left(\|X_i\|_{\mathrm{F}} \geq t\right) \leq P\left(\max_{Z\in\mathcal{N}_{\frac{1}{2}}} \langle Z, X_i\rangle \geq \frac{1}{2}\right) \leq |\mathcal{N}_{\frac{1}{2}}| \cdot \exp\left(-\frac{t^2}{8\sigma^2}\right) \leq 5^{d_1 d_2} \cdot \exp\left(-\frac{t^2}{8\sigma^2}\right).$$

This fact indicates that

$$P\left(\|X_i\|_{\mathrm{F}} \geq 2\sqrt{2}\sigma\sqrt{\ln\left(\frac{1}{\delta}\right)} + 4\sigma\sqrt{d_1 d_2}\right) \leq \delta.$$

Algorithm 15 Randomized LOTUS

Input: Arm set $\mathcal{X}_t$, sampling distribution $\mathcal{D}_t$, $\delta, T_0, \eta, \lambda, \{\lambda_{i,\perp}\}_{i=1}^{+\infty}$.
Stage The history buffer index set $\mathcal{H}_1 = \{\}$, the exploration buffer index set $\mathcal{H}_2 = \{\}$.

Pull arm $X_t \in \mathcal{X}_t$ according to $\mathcal{D}_t$ and observe payoff y_t. Then add (X_t, y_t) into $\mathcal{H}_1$ and $\mathcal{H}_2$ for $t \leq T_0$.
for $i = 1, 2, \ldots$ until the end of iterations **do**
    Set the expected exploration length $T_1 = \min\left\{\left[\frac{d^{2+4\delta} r^{1+\delta}}{D_{rr}^{2+2\delta}} 2^{i(1+\delta)}\right]^{\frac{1}{1+3\delta}}, 2^i\right\}$.
    for $t = |\mathcal{H}_1| + 1 + |\mathcal{H}_1| + 2^i$ **do**
        if Randomly sample from Bernoulli$(T_1/2^i)$ and get 1 **then**
            Pull arm $X_t \in \mathcal{X}_t$ according to $\mathcal{D}_t$ and observe payoff y_t. Then add (X_t, y_t) into $\mathcal{H}_1$ and $\mathcal{H}_2$
        else
            Obtain the estimate $\widehat{\Theta}$ based on Eqn. (5.3) with $\mathcal{H}_2$, where we set $\tau_i \asymp \left(|\mathcal{H}_2|/(d + \ln(2^{i+1}/\epsilon))\right)^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}}, \lambda_i \asymp \sigma\left((d + \ln(2^{i+1}/\epsilon))/|\mathcal{H}_2|\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}$.
            Calculate the full SVD of $\widehat{\Theta} = [\widehat{U}, \widehat{U}_\perp]\, \widehat{D}\, [\widehat{V}, \widehat{V}_\perp]^\top$ where $\widehat{U} \in \mathbb{R}^{d_1 \times r}, \widehat{V} \in \mathbb{R}^{d_2 \times r}$.
            For the next round, invoke LowTO with $\delta, [\widehat{U}, \widehat{U}_\perp], [\widehat{V}, \widehat{V}_\perp], \lambda, \lambda_{i,\perp}, \mathcal{H}_1$ and obtain the updated $\mathcal{H}_1$.

Therefore, we have that

$$P\left(\max_{i\in[m]} \|X_i\|_{\mathrm{F}} < 2\sqrt{2}\sigma\sqrt{\ln\left(\frac{1}{\alpha}\right)} + 4\sigma\sqrt{d_1 d_2}\right) \geq (1-\alpha)^m = 1 - \delta, \quad \text{where } \alpha = 1 - (1-\delta)^{\frac{1}{m}}.$$

For any $m > 1$ and $x \in [0, 1]$, based on the taylor series of the function $f(x) = (1-x)^{\frac{1}{m}} = 1 - \frac{x}{m} - O(x^2)$, it holds that $1 - \frac{x}{m} > (1-x)^{\frac{1}{m}}$. And this fact leads to the final result:

$$P\left(\max_{i\in[T]} \|X_i\|_{\mathrm{F}} < 2\sqrt{2}\sigma\sqrt{\ln\left(\frac{T}{\delta}\right)} + 4\sigma\sqrt{d_1 d_2}\right) > 1 - \delta,$$

which indicates that $\max_{i\in[T]} \|X_i\|_{\mathrm{F}}$ can be uniformly bounded by a constant scale up to some minimal error. □

In our case with $\sigma \asymp \frac{1}{\sqrt{d_1 d_2}}$, with probability at least $1 - \delta$ it holds that

$$\max_{i\in[m]} \|X_i\|_{\mathrm{F}} \lesssim \frac{2\sqrt{2}}{\sqrt{d_1 d_2}}\sqrt{\ln\left(\frac{m}{\delta}\right)} + 4.$$

D.2. Alternative Version of LOTUS

As we mention in Subsection 5.4.2, we also have an alternative version of our LOTUS algorithm in a more randomized manner. Specifically, at each batch, our original version illustrated in Algorithm 8 uses the static explore-then-exploit framework, where it first randomly samples some arms from

Algorithm 16 LAMM Algorithm for the Solution to Eqn.(5.2)

Input: Initial $\widehat{\Theta}_0$, stopping threshold ϵ, α_0, ψ, λ.

1: **for** $i = 1, 2, \ldots$ until $\left\|\widehat{\Theta}_i - \widehat{\Theta}_{i-1}\right\|_{\mathrm{F}} \leq \epsilon$ **do**

2: Initialize $\widehat{\Theta}_i = \widehat{\Theta}_{i-1}, \alpha_i = \max(\alpha_0, \alpha_{i-1}/\psi)$ and $s_i = 0$.

3: **while** $F(\widehat{\Theta}_i; \widehat{\Theta}_{i-1}, \alpha_i) < \hat{L}_\tau(\widehat{\Theta}_i)$ **or** $s_i = 0$ **do**

4: $\widehat{\Theta}_i = S(\widehat{\Theta}_{i-1} - \alpha_i^{-1}\nabla \hat{L}_\tau(\widehat{\Theta}_{i-1}), \alpha_i^{-1}\lambda)$.

5: $s_i = s_i + 1, \alpha_i = \psi \cdot \alpha_i$.

the distribution $\mathcal{D}_t$ in Assumption 5.3.1 and then exploits the recovered low-rank subspaces with our LowTO method. However, we can mix these two exploration and exploitation steps in each batch. Specifically, we can explore by the sampling distribution D_t with the probability of $T_1^i/2^i$ at each time t, otherwise we will conduct the subspace transformation and LowTO algorithm based on the current $\mathcal{H}_t$. The full pseudocode is presented in Algorithm 15. We can expect the same order of regret as in Theorem 5.4.3 based on the fact that if we do a series of iid Bernoulli trials with probability p for n times, then with a high probability the sum of success will be close to np for large n up to some logarithmic terms.

D.3. Details of the LAMM Algorithm

We implement the LAMM algorithm that was first proposed in Fan et al. (2018) and recently extended to the matrix estimation setting (Yu et al., 2023) for the Huber-type estimator formulated in Eqn. (5.2). Here we use the unified framework proposed in Yu et al. (2023), and for the sake of completeness we will still present its details as follows:

LAMM is presented in Algorithm 16. The LAMM method is a very efficient and scalable algorithm under high-dimensional datasets, and its first crux is establishing an isotropic quadratic function that locally upper bounds the objective function $\hat{L}_\tau(\Theta)$ at each iteration until convergence. Based on the second-order Taylor expansion, given the previous estimate $\widehat{\Theta}_{t-1}$ at iteration $t-1$, we can define the quadratic function at iteration t as:

$$F(\Theta; \widehat{\Theta}_{t-1}, \alpha_k) = \hat{L}_\tau(\widehat{\Theta}_{t-1}) + \langle \nabla \hat{L}_\tau(\widehat{\Theta}_{t-1}), \Theta - \widehat{\Theta}_{t-1} \rangle + \frac{\alpha_t}{2} \left\|\Theta - \widehat{\Theta}_{t-1}\right\|_{\mathrm{F}}^2,$$

with some quadratic parameter $\alpha_t > 0$. This parameter needs to be sufficiently large as we illustrated above such that $\hat{L}_\tau(\widehat{\Theta}_t) \leq F(\widehat{\Theta}_t; \widehat{\Theta}_{t-1}, \alpha_t)$ holds where

$$\widehat{\Theta}_t = \arg\min_{\Theta \in R^{d_1 \times d_2}} F(\Theta; \widehat{\Theta}_{t-1}, \alpha_t) + \lambda \|\Theta\|_{\text{nuc}}.$$

We will use an iterative increment approach on α_t with some multiplier $\psi > 1$ to guarantee the quadratic function F majorizes the objective function $\hat{L}$ at each descent. This fact ensures the descent of the objective function at each iteration with a closed-formed solution. Specifically, to minimize the penalized isotropic quadratic function, we can deduce the solution in the following ways: for $k > 0$, define the soft-thresholding operator on a diagonal matrix $\Sigma = \text{diag}(\{\sigma_i\})$ as $S(\Sigma, k) = \text{diag}(\{\max(\sigma_i - k, 0)\})$. For any general matrix Θ with its SVD decomposition as $\Theta = U\Sigma V^\top$, we write $S(\Theta, k) = US(\Sigma, k)V^\top$. Then the solution of $\widehat{\Theta}_t$ can be represented as:

$$\widehat{\Theta}_t = S(\widehat{\Theta}_{t-1} - \alpha_t^{-1} \nabla \hat{L}_\tau(\widehat{\Theta}_{t-1}), \alpha_t^{-1}\lambda).$$

D.4. Analysis of Theorem 5.4.1

D.4.1. Preliminaries.

LEMMA D.4.0.1. *(Bernstein Inequality) Let X be a random variable with mean μ and variance σ^2. Assume we can find some $b > 0$ such that*

$$\mathbb{E}|X - \mu|^k \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \ k = 3, 4, 5, \ldots$$

Then it holds that

$$P\left(|X - \mu| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right), \ \forall t > 0.$$

COROLLARY D.4.1. *(Adapted from Bernstein Inequality) Let X be a random variable with mean μ and variance σ^2. Assume we can find some $b > 0$ such that*

$$\mathbb{E}|X - \mu|^k \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \ k = 3, 4, 5, \ldots$$

Then it holds that

$$P\left(X - \mu \geq \sqrt{2t}\sigma + 2bt\right) \leq \exp\left(-t\right), \ \forall t > 0.$$

PROOF. Based on Lemma D.4.0.1, we have that for any $t > 0$

$$P\left(X - \mu \geq \sqrt{2t}\sigma + 2bt\right) \leq \exp\left(-\frac{(\sqrt{2t}\sigma + 2bt)^2}{2\sigma^2 + 2b(\sqrt{2t}\sigma + 2bt)}\right) \leq \exp\left(-\frac{2\sigma^2 t + 4b^2t^2 + 4\sqrt{2}b\sigma t^{\frac{3}{2}}}{2\sigma^2 + 4\sigma^2 t + 2\sqrt{2}b\sigma\sqrt{t}}\right)$$
$$\leq \exp(-t).$$

□

DEFINITION D.4.1. *(Local Restricted Strong Convexity) For the empirical loss function $\hat{L}_\tau(\cdot)$, we can define the event of local restricted strong convexity $\mathcal{E}(s, l, \kappa)$ in terms of the radius parameter s, l and the curvature parameter κ as*

$$\mathcal{E}(s, l, \kappa) = \left\{ \inf_{\Theta \in \mathcal{M}(\Theta^*, s, l)} \frac{\langle \nabla \hat{L}_\tau(\Theta) - \nabla \hat{L}_\tau(\Theta^*), \Theta - \Theta^* \rangle}{\|\Theta - \Theta^*\|_F^2} \geq \kappa \right\},$$

where $\mathcal{M}(\Theta^*, s, l) = \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} : \|\Theta - \Theta^*\|_F \leq s, \|\Theta - \Theta^*\|_{nuc} \leq l \|\Theta - \Theta^*\|_F\right\}$.

We assume $d_1 \geq d_2$ without loss of generality, and denote $\widehat{\Delta} := \widehat{\Theta} - \Theta^*$ in the following argument. To start with, we will show that our target $\left\|\widehat{\Delta}\right\|_{\mathrm{F}}$ can be bounded conditioned on the event $\mathcal{E}(s, l, \kappa)$ and $\lambda \geq 2\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{\mathrm{op}}$.

THEOREM D.4.2. *Conditioned on the event $\lambda \geq 2\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{op}$ and the event $\mathcal{E}(s, l, \kappa)$ with $s \geq 9\sqrt{r}\frac{\lambda}{\kappa}$ and $l \geq 4\sqrt{2r}$, then we can deduce that*

$$\left\|\widehat{\Delta}\right\|_F = \left\|\widehat{\Theta} - \Theta^*\right\|_F \leq 9\sqrt{r} \cdot \frac{\lambda}{\kappa}.$$

PROOF. We will prove Theorem D.4.2 by contradiction. Assume we have that $\lambda \geq 2\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{\mathrm{op}}$ and $\mathcal{E}(s, l, \kappa)$ holds with $s \geq 9\sqrt{r}\frac{\lambda}{\kappa}$ and $l \geq 4\sqrt{2r}$, and we assume $\left\|\widehat{\Delta}\right\|_{\mathrm{F}} > 9\sqrt{r} \cdot \frac{\lambda}{\kappa}$ holds. Define $\widetilde{\Theta}_x = \Theta^* + x(\widehat{\Theta} - \Theta^*)$ as a function of $x \in [0, 1]$, then there exists some $\zeta \in (0, 1)$ such that $\widetilde{\Theta}_\zeta = \Theta^* + \zeta(\widehat{\Theta} - \Theta^*)$ satisfying $\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\mathrm{F}} = 9\sqrt{r} \cdot \frac{\lambda}{\kappa}$ since $\left\|\widetilde{\Theta}_x - \Theta^*\right\|_{\mathrm{F}}$ is a continuous function in terms of $x \in [0, 1]$. Furthermore, we define $Q(x) = \hat{L}_\tau(\widetilde{\Theta}_x) - \hat{L}_\tau(\Theta^*) - \langle \nabla \hat{L}_\tau(\Theta^*), \widetilde{\Theta}_x - \Theta^* \rangle$. Note $x \in [0, 1] \to Q(x)$ can be easily shown as a convex function: first, we observe that $\widetilde{\Theta}_x$ is a linear function of x, and the Huber loss function defined in Section 5.4.1 is convex (Huber, 1965), which implies that $\hat{L}_\tau(\widetilde{\Theta}_x)$ is convex. On the other hand, the inner product $\langle \nabla \hat{L}_\tau(\Theta^*), \widetilde{\Theta}_x - \Theta^* \rangle$ is bi-linear and hence naturally convex as well. Therefore, we know that $Q'(x) = \langle \nabla \hat{L}_\tau(\widetilde{\Theta}_x) - \nabla \hat{L}_\tau(\Theta^*), \widehat{\Theta} - \Theta^* \rangle$

is monotonically increasing. And it holds that

$$\text{(D.1)}\quad \zeta Q'(\zeta) \le \zeta Q'(1) \implies \langle \nabla \hat{L}_\tau(\widetilde{\Theta}_\zeta) - \nabla \hat{L}_\tau(\Theta^*), \widetilde{\Theta}_\zeta - \Theta^* \rangle \le \zeta \langle \nabla \hat{L}_\tau(\widehat{\Theta}) - \nabla \hat{L}_\tau(\Theta^*), \widehat{\Theta} - \Theta^* \rangle$$

To bound the right-hand side of Eqn. (D.1), since $\widehat{\Theta}$ is the solution to the convex optimization problem in Eqn. (5.2), then we have the sub-gradient condition as:

$$\langle \nabla \hat{L}_\tau(\widehat{\Theta}) + \lambda \widehat{Z}, \widehat{\Theta} - \Theta^* \rangle \le 0, \quad \text{where } \widehat{Z} \in \partial \left\| \widehat{\Theta} \right\|_{\text{nuc}}.$$

Due to the definition of the sub-gradient, it holds that $\|\Theta^*\|_{\text{nuc}} \ge \|\widehat{\Theta}\|_{\text{nuc}} + \langle \widehat{Z}, \Theta^* - \widehat{\Theta} \rangle$. By assuming $\lambda \ge 2 \left\| \nabla \hat{L}_\tau(\Theta^*) \right\|_{\text{op}}$, we can have that

$$\begin{aligned}\langle \nabla \hat{L}_\tau(\widehat{\Theta}) - \nabla \hat{L}_\tau(\Theta^*), \widehat{\Theta} - \Theta^* \rangle &\le \langle \lambda \widehat{Z}, \Theta^* - \widehat{\Theta} \rangle + \langle \nabla \hat{L}_\tau(\Theta^*), \Theta^* - \widehat{\Theta} \rangle \\ &\le \lambda \left(\|\Theta^*\|_{\text{nuc}} - \|\widehat{\Theta}\|_{\text{nuc}} \right) + \frac{\lambda}{2} \left\| \Theta^* - \widehat{\Theta} \right\|_{\text{nuc}} \le \frac{3\lambda}{2} \left\| \widehat{\Delta} \right\|_{\text{nuc}}\end{aligned}$$

To bound $\left\| \widehat{\Delta} \right\|_{\text{nuc}}$, we utilize the regular procedure (Negahban & Wainwright, 2011; Yu et al., 2023). We restate the notation and define the reduced SVD of Θ^* as $\Theta^* = U \Sigma V^\top$ with $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$. Then we denote two sets as:

$$\begin{aligned}\mathbb{M} &= \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} : \text{row}(\Theta) \subseteq \text{col}(V), \text{col}(\Theta) \subseteq \text{col}(U) \right\}, \\ \overline{\mathbb{M}}^\perp &= \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} : \text{row}(\Theta) \subseteq \text{col}(V)^\perp, \text{col}(\Theta) \subseteq \text{col}(U)^\perp \right\},\end{aligned}$$

and hence $\mathbb{M} \subseteq \overline{\mathbb{M}}$. Next we will show that $\|\widehat{\Delta}_{\overline{\mathbb{M}}^\perp}\|_{\text{nuc}} \le 3 \|\widehat{\Delta}_{\overline{\mathbb{M}}}\|_{\text{nuc}}$ in the following part. First, since $\widehat{\Theta}$ is the solution to the problem defined in Eqn.(5.2), we have that

$$\hat{L}_\tau(\widehat{\Theta}) + \lambda \left\| \widehat{\Theta} \right\|_{\text{nuc}} \le \hat{L}_\tau(\Theta^*) + \lambda \|\Theta^*\|_{\text{nuc}} \iff \hat{L}_\tau(\widehat{\Theta}) - \hat{L}_\tau(\Theta^*) \le \lambda \left(\|\Theta^*\|_{\text{nuc}} - \left\| \widehat{\Theta} \right\|_{\text{nuc}} \right).$$

For the left-hand side, it holds that

$$\begin{aligned}\hat{L}_\tau(\widehat{\Theta}) - \hat{L}_\tau(\Theta^*) &\ge \langle \nabla \hat{L}_\tau(\Theta^*), \widehat{\Theta} - \Theta^* \rangle \ge - \left\| \nabla \hat{L}_\tau(\Theta^*) \right\|_{\text{op}} \left\| \widehat{\Delta} \right\|_{\text{nuc}} \\ &\ge - \left\| \nabla \hat{L}_\tau(\Theta^*) \right\|_{\text{op}} \left(\left\| \widehat{\Delta}_{\mathbb{M}} \right\|_{\text{nuc}} + \left\| \widehat{\Delta}_{\overline{\mathbb{M}}^\perp} \right\|_{\text{nuc}} \right) \ge - \frac{\lambda}{2} \left(\left\| \widehat{\Delta}_{\overline{\mathbb{M}}} \right\|_{\text{nuc}} + \left\| \widehat{\Delta}_{\overline{\mathbb{M}}^\perp} \right\|_{\text{nuc}} \right). \qquad \text{(D.2)}\end{aligned}$$

And for the right-hand side, we have that

$$\left\|\widehat{\Theta}\right\|_{\text{nuc}} = \left\|\Theta^* + \widehat{\Delta}\right\|_{\text{nuc}} = \left\|\Theta^*_{\mathbb{M}} + \widehat{\Delta}_{\overline{\mathbb{M}}} + \widehat{\Delta}_{\overline{\mathbb{M}}^\perp}\right\|_{\text{nuc}} \geq \|\Theta^*_{\mathbb{M}}\|_{\text{nuc}} + \left\|\widehat{\Delta}_{\overline{\mathbb{M}}^\perp}\right\|_{\text{nuc}} - \left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{nuc}},$$

and hence we have that

$$\|\Theta^*\|_{\text{nuc}} - \left\|\widehat{\Theta}\right\|_{\text{nuc}} = \|\Theta^*_{\mathbb{M}}\|_{\text{nuc}} - \left\|\widehat{\Theta}\right\|_{\text{nuc}} \leq \left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{nuc}} - \left\|\widehat{\Delta}_{\overline{\mathbb{M}}^\perp}\right\|_{\text{nuc}}. \tag{D.3}$$

Combining the results from Eqn. (D.2) and Eqn. (D.3), we can deduce that $\|\widehat{\Delta}_{\overline{\mathbb{M}}^\perp}\|_{\text{nuc}} \leq 3\|\widehat{\Delta}_{\overline{\mathbb{M}}}\|_{\text{nuc}}$. Next, since we have that $\text{rank}(\widehat{\Delta}_{\overline{\mathbb{M}}}) \leq 2r$, then based on Cauchy-Schwarz inequality it holds that

$$\left\|\widehat{\Delta}\right\|_{\text{nuc}} \leq \left\|\widehat{\Delta}_{\mathbb{M}}\right\|_{\text{nuc}} + \left\|\widehat{\Delta}_{\overline{\mathbb{M}}_\perp}\right\|_{\text{nuc}} \leq \left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{nuc}} + \left\|\widehat{\Delta}_{\overline{\mathbb{M}}_\perp}\right\|_{\text{nuc}} \leq 4\left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{nuc}} \leq 4\sqrt{2r}\left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{F}} \leq 4\sqrt{2r}\left\|\widehat{\Delta}\right\|_{\text{F}}.$$

Therefore, we can show that $\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{nuc}} \leq 4\sqrt{2r}\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{F}}$. And remember that we assume $\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{F}} = 9\sqrt{r}\cdot\frac{\lambda}{\kappa}$. These facts indicate that $\widetilde{\Theta}_\zeta \in \mathcal{M}(\Theta^*, s, l)$ with $s \geq 9\sqrt{r}\frac{\lambda}{\kappa}$ and $l \geq 4\sqrt{2r}$. Therefore, based on the local restricted strong convexity, we have

$$\kappa\zeta\left\|\widehat{\Delta}\right\|_{\text{F}}\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{F}} = \kappa\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{F}}^2 \leq \langle\nabla\hat{L}_\tau(\widetilde{\Theta}_\zeta) - \nabla\hat{L}_\tau(\Theta^*), \widetilde{\Theta}_\zeta - \Theta^*\rangle.$$

For the left-hand side, it holds that

$$\kappa\zeta\left\|\widehat{\Delta}\right\|_{\text{F}}\left\|\widetilde{\Theta}_\zeta - \Theta^*\right\|_{\text{F}} = \kappa\zeta\left\|\widehat{\Delta}\right\|_{\text{F}} 9\sqrt{r}\frac{\lambda}{\kappa} = \zeta\lambda\left\|\widehat{\Delta}\right\|_{\text{F}} 9\sqrt{r},$$

and for the right-handed side, based on Eqn. (D.1) we have that

$$\begin{aligned}\langle\nabla\hat{L}_\tau(\widetilde{\Theta}_\zeta) - \nabla\hat{L}_\tau(\Theta^*), \widetilde{\Theta}_\zeta - \Theta^*\rangle &\leq \zeta\langle\nabla\hat{L}_\tau(\widehat{\Theta}) - \nabla\hat{L}_\tau(\Theta^*), \widehat{\Theta} - \Theta^*\rangle \\ &\leq \eta\frac{3\lambda}{2}\left\|\widehat{\Delta}\right\|_{\text{nuc}} \leq \zeta 6\sqrt{2}\lambda\sqrt{r}\left\|\widehat{\Delta}\right\|_{\text{F}}\end{aligned}$$

Consequently, we have $9 \leq 6\sqrt{2}$ that contradicts the fact, which means that

$$\left\|\widehat{\Delta}\right\|_{\text{F}} \leq 9\sqrt{r}\cdot\frac{\lambda}{\kappa}.$$

□

Next, we will show the event $\mathcal{E}(s, l, \kappa)$ and the event $\lambda \geq 2\left\|\nabla\hat{L}_\tau(\Theta^*)\right\|_{\text{op}}$ hold with high probability individually. Specifically, we will first give an upper bound of $\left\|\nabla\hat{L}_\tau(\Theta^*)\right\|_{\text{op}}$ in Theorem D.4.3 and then present the event $\mathcal{E}(s, l, \kappa)$ holds with high probability in Theorem D.4.4.

THEOREM D.4.3. *By taking* $\tau = (\frac{n}{5d-\ln(\epsilon)})^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}}$, *then with probability at least* $1-\epsilon$, *it holds that*

$$\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{op} \leq (10+11\sqrt{2})\sigma \left(\frac{n}{5d-\ln(\epsilon)}\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}.$$

PROOF. Define the zero-mean random matrix $\Gamma = \nabla \hat{L}_\tau(\Theta^*) - \mathbb{E}\nabla \hat{L}_\tau(\Theta^*)$, then we have that

$$\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}} = \left\|\nabla \hat{L}_\tau(\Theta^*) - \mathbb{E}\nabla \hat{L}_\tau(\Theta^*) + \mathbb{E}\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}} \leq \|\Gamma\|_{\text{op}} + \left\|\mathbb{E}\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}}.$$

Therefore, we could control these two terms separately. Denote $S^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$. For the second term, we have that

$$\nabla \hat{L}_\tau(\Theta^*) = -\frac{1}{n}\sum_{i=1}^n l'_\tau(y_i - \langle X_i, \Theta^*\rangle)X_i = -\frac{1}{n}\sum_{i=1}^n l'_\tau(\eta_i)X_i.$$

Therefore, we can deduce that

$$\begin{aligned}
\left\|\mathbb{E}\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}} &= \sup_{u\in S^{d_1-1}, v\in S^{d_2-1}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(l'_\tau(\eta_i)u^\top X_i v\right) \\
&= \sup_{u\in S^{d_1-1}, v\in S^{d_2-1}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\mathbb{E}\left(l'_\tau(\eta_i)u^\top X_i v|\mathcal{F}_i\right)\right) \\
&= \sup_{u\in S^{d_1-1}, v\in S^{d_2-1}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(u^\top X_i v \cdot \mathbb{E}\left(l'_\tau(\eta_i)|\mathcal{F}_i\right)\right)
\end{aligned}$$

By the expression of $l'_\tau(\cdot)$, we can deduce that

$$|\mathbb{E}\left(l'_\tau(\eta_i)|\mathcal{F}_i\right)| = |\mathbb{E}\left(l'_\tau(\eta_i) - \eta_i|\mathcal{F}_i\right)| \leq \mathbb{E}\left(\frac{|\eta_i|^{1+\delta}}{\tau^\delta}\Big|\mathcal{F}_i\right) \leq \frac{c}{\tau^\delta}$$

And since $u^\top X_i v$ is sub-Gaussian with the parameter σ^2, we have $\mathbb{E}(|u^\top X_i v|) \leq \sqrt{2\sigma^2}$. Conclusively, it holds that

$$\left\|\mathbb{E}\nabla \hat{L}_\tau(\Theta^*)\right\|_{\text{op}} \leq \frac{\sqrt{2}}{\tau^\delta} c \cdot \sigma. \tag{D.4}$$

To bound the operator norm of Γ, we use the regular covering technique: Let $\mathcal{N}_{\frac{1}{4}}^d$ be the $1/4$ covering of S^{d-1}, then we claim that

$$\|\Gamma\|_{\text{op}} \leq \frac{5}{2} \max_{u\in\mathcal{N}_{\frac{1}{4}}^{d_1}, v\in\mathcal{N}_{\frac{1}{4}}^{d_2}} u^\top \Gamma v. \tag{D.5}$$

To prove this result, for any $u \in S^{d_1-1}, v \in S^{d_2-1}$, we denote $S(u) \in \mathbb{R}^{d_1}$ ($S(v) \in \mathbb{R}^{d_2}$) as the nearest neighbor of u (v) in $N_{\frac{1}{4}}^{d_1}$ ($\mathcal{N}_{\frac{1}{4}}^{d_2}$) such that $\|u - S(u)\|_2, \|v - S(v)\|_2 \leq \frac{1}{4}$. We take u, v such that $u^\top \Gamma v = \|\Gamma\|_{\text{op}}$. Therefore, it holds that

$$\begin{aligned}\|\Gamma\|_{\text{op}} = u^\top \Gamma v &= S(u)^\top \Gamma S(v) + (u - S(u))^\top \Gamma v + u^\top \Gamma (v - S(v)) + (u - S(u))^\top \Gamma (v - S(v)) \\ &\leq \max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} u^\top \Gamma v + \frac{1}{4}\|\Gamma\|_{\text{op}} + \frac{1}{4}\|\Gamma\|_{\text{op}} + \frac{1}{16}\|\Gamma\|_{\text{op}} \leq \max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} u^\top \Gamma v + \frac{3}{5}\|\Gamma\|_{\text{op}},\end{aligned}$$

which leads to Eqn. (D.5). And then it holds that

$$\|\Gamma\|_{\text{op}} \leq \frac{5}{2} \max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}\left(l'_\tau(\eta_i) u^\top X_i v \right) - l'_\tau(\eta_i) u^\top X_i v \right].$$

To bound the right-hand side term, we aim to use a union bound of probability with Corollary D.4.1. Since $u^\top X_i v$ is sub-Gaussian with parameter σ for arbitrary $u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}$, then we have that for $k = 2, 3, \ldots$

$$\mathbb{E}|u^\top X_i v|^k = \int_0^\infty P\left(|u^\top X_i v|^k > t \right) dt \leq 2 \int_0^\infty \exp\left(-\frac{t^2}{2k\sigma^2} \right) dt \leq \frac{1}{2} \cdot k! \cdot (\sqrt{2}\sigma)^k.$$

The above results along with the fact that $|l'_\tau(\cdot)| \leq \tau$ can lead to the following inequality for $k = 2, 3, \ldots$:

$$\mathbb{E}\left| \sum_{i=1}^{n} l'_\tau(\eta_i) u^\top X_i v \right|^k \leq n \cdot \tau^{k-1-\delta} \mathbb{E}\left(|l'_\tau(\eta_i)|^{1+\delta} |u^\top X_i v|^k \right) \leq \frac{1}{2} \cdot k! \cdot \left(\sqrt{2}\sigma\tau \right)^{k-2} \cdot (2n\sigma^2 \tau^{1-\delta} c).$$

Based on Corollary D.4.1, it holds that

$$P\left(u^\top \Gamma v \geq 4\sqrt{x\sigma^2\tau^{1-\delta}c} \frac{1}{\sqrt{n}} + 4\sqrt{2}\sigma\tau \frac{x}{n} \right) \leq e^{-x}.$$

By taking the union bound on all $u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}$ and using the fact that $9^{d_1+d_2} \leq e^{5d}$, it holds that

$$P\left(\|\Gamma\|_{\text{op}} \geq \frac{5}{2} \max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} u^\top \Gamma v \geq 10\sigma\sqrt{c} \sqrt{\frac{5d - \ln(\epsilon)}{n}} \tau^{\frac{1-\delta}{2}} + 10\sqrt{2}\sigma\tau \frac{5d - \ln(\epsilon)}{n} \right) \leq \epsilon. \tag{D.6}$$

Combining the results in Eqn. (D.4) and Eqn. (D.6), we have that

$$P\left(\left\|\nabla\hat{L}_\tau(\Theta^*)\right\|_{\text{op}} \geq 10\sigma\sqrt{c}\sqrt{\frac{5d-\ln(\epsilon)}{n}}\tau^{\frac{1-\delta}{2}} + 10\sqrt{2}\sigma\tau\cdot\frac{5d-\ln(\epsilon)}{n} + \frac{\sqrt{2}}{\tau^\delta}c\cdot\sigma\right) \leq \epsilon.$$

By taking $\tau = \left(\frac{n}{5d-\ln(\epsilon)}\right)^{\frac{1}{1+\delta}}\cdot c^{\frac{1}{1+\delta}}$, we have that

$$P\left(\left\|\nabla\hat{L}_\tau(\Theta^*)\right\|_{\text{op}} \leq (10+11\sqrt{2})c^{\frac{1}{1+\delta}}\cdot\left(\frac{5d-\ln(\epsilon)}{n}\right)^{\frac{\delta}{1+\delta}}\right) \geq 1-\epsilon.$$

□

THEOREM D.4.4. *For any $s, l > 0$, if we take τ and n such that*

$$\tau \geq \max\left\{32\sigma^2 s\sqrt{\frac{1}{c_l}}, \left(\frac{64\sigma^2 c}{c_l}\right)^{\frac{1}{1+\delta}}\right\}$$

$$n \geq \max\left\{8\ln(9)(d_1+d_2), \left(225\sigma\sqrt{\ln(9)(d_1+d_2)}\frac{\tau l}{sc_l}\right)^2, \left(\frac{48\sigma^2}{c_l}\sqrt{-2\ln(\epsilon)}\right)^2, -\frac{\tau^2}{c_l s^2}\ln(\epsilon)\right\}.$$

Then with probability at least $1-\epsilon$, the local restricted strong convexity $\mathcal{E}(s,l,\kappa)$ holds with $\kappa = \frac{c_l}{4}$.

PROOF. Given the values of $s, l > 0$, for the sake of simplicity we denote the event Φ as $\Phi = \mathcal{M}(\Theta^*, s, l) = \left\{\Theta\in\mathbb{R}^{d_1\times d_2} : \|\Theta-\Theta^*\|_{\text{F}} \leq s, \|\Theta-\Theta^*\|_{\text{nuc}} \leq l\|\Theta-\Theta^*\|_{\text{F}}\right\}$. Since the Huber loss is convex and differentiable, we have

$$\begin{aligned}
D(\Theta) &\coloneqq \langle\nabla\hat{L}_\tau(\Theta) - \nabla\hat{L}_\tau(\Theta^*), \Theta-\Theta^*\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \left(l'_\tau(y_i - \langle X_i, \Theta^*\rangle) - l'_\tau(y_i - \langle X_i, \Theta\rangle)\right)\cdot\langle X_i, \Theta-\Theta^*\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^n \left(l'_\tau(y_i - \langle X_i, \Theta^*\rangle) - l'_\tau(y_i - \langle X_i, \Theta\rangle)\right)\cdot\langle X_i, \Theta-\Theta^*\rangle\cdot\mathbb{1}_{\Xi_i(\Theta)},
\end{aligned}$$

where the last inequality holds since Huber loss is convex, and $\Xi_i(\Theta)$ is defined as

$$\Xi_i(\Theta) = \left\{|\eta_i| \leq \frac{\tau}{2}\right\}\cap\left\{|\langle X_i, \Theta-\Theta^*\rangle| \leq \frac{\tau}{2s}\|\Theta-\Theta^*\|_{\text{F}}\right\}.$$

Note whenever $\Theta\in\Phi$ and $\Xi_i(\Theta)$ hold we have that

$$|y_i - \langle X_i, \Theta\rangle| \leq |y_i - \langle X_i, \Theta^*\rangle| + \frac{\tau}{2s}\cdot\|\Theta-\Theta^*\|_{\text{F}} \leq \tau.$$

Since we have $l''_\tau(u) = 1$ with $|u| \le \tau$, it holds that

$$D(\Theta) \ge \frac{1}{n}\sum_{i=1}^{n}\langle X_i, \Theta - \Theta^*\rangle^2 \cdot \mathbb{1}_{\Xi_i(\Theta)}.$$

Furthermore, we define the function $\phi_R(x)$ with some $R > 0$ as

$$\phi_R(x) = \begin{cases} x^2, & \text{if } |x| \le \frac{R}{2}; \\ (x-R)^2, & \text{if } \frac{R}{2} < x \le R; \\ (x+R)^2, & \text{if } -R \le x < -\frac{R}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

And we know $\phi_r(\cdot)$ is R-Lipschitz continuous with the properties that

$$\phi_{\alpha R}(\alpha x) = \alpha^2\phi_R(x)\ \forall \alpha > 0, \text{ and } x^2 \cdot \mathbb{1}_{|x|\le R/2} \le \phi_R(x) \le x^2 \cdot \mathbb{1}_{|x|\le R}.$$

Then we can deduce that

$$\begin{aligned}
\frac{D(\Theta)}{\|\Theta - \Theta^*\|_{\mathrm{F}}^2} &\ge \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\mathrm{F}}}\right)^2 \cdot \mathbb{1}_{\Xi_i(\Theta)} \ge \frac{1}{n}\sum_{i=1}^{n}\phi_{\frac{\tau}{2s}}\left(\frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\mathrm{F}}}\right) \cdot \mathbb{1}_{\{|\eta_i| \le \frac{\tau}{2}\}} \\
&\coloneqq \frac{1}{n}\sum_{i=1}^{n}\beta_{\tau,s}(X_i, \Theta, \eta_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\beta_{\tau,s}(X_i, \Theta, \eta_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right) \\
&\ge \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right) - \sup_{\Theta\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\beta_{\tau,s}(X_i, \Theta, \eta_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right)\right| \\
&\coloneqq A_1 - A_2.
\end{aligned}$$

For simplicity we write $\Delta = \Theta - \Theta^*$ as a function of Θ. To lower bound the first term A_1, we have that for any $i \in [n]$,

$$\begin{aligned}
\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right) &\ge \mathbb{E}\left[\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^2 \cdot \mathbb{1}_{\{|\langle X_i, \Delta\rangle| \le \frac{\tau}{4s}\|\Delta\|_{\mathrm{F}}\}} \cdot \mathbb{1}_{\{|\eta_i| \le \frac{\tau}{2}\}}\right] \\
&\ge \mathbb{E}\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^2 - \mathbb{E}\left[\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^2 \cdot \mathbb{1}_{\{|\langle X_i, \Delta\rangle| > \frac{\tau}{4s}\|\Delta\|_{\mathrm{F}}\}}\right] - \mathbb{E}\left[\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^2 \cdot \mathbb{1}_{\{|\eta_i| > \frac{\tau}{2}\}}\right] \\
&\coloneqq A_{11} - A_{12} - A_{13}.
\end{aligned}$$

Based on Assumption 5.3.1, we have $A_{11} \geq c_l$. Furthermore, it holds that

$$A_{12} \leq \sqrt{\mathbb{E}\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^4} \cdot \sqrt{\mathbb{E}\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^4 \Big/ \left(\frac{\tau}{4s}\right)^4} \leq 256\sigma^4 \cdot \frac{s^2}{\tau^2}$$
$$A_{13} \leq \left(\frac{2}{\tau}\right)^{1+\delta} \mathbb{E}\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^2 \cdot \mathbb{E}|\eta_i|^{1+\delta} \leq \frac{16}{\tau^{1+\delta}}\sigma^2 \cdot c.$$

By choosing that $\tau \geq \max\left\{32\sigma^2 s\sqrt{\frac{1}{c_l}}, \left(\frac{64\sigma^2 c}{c_l}\right)^{\frac{1}{1+\delta}}\right\}$, it holds that $A_{12} \leq \frac{c_l}{4}$ and $A_{13} \leq \frac{c_l}{4}$, which indicates that

$$\mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right) \geq \frac{c_l}{2}, \ \ \forall i \in [n],$$

which implies that

$$A_1 \geq \frac{c_l}{2} \tag{D.7}$$

Afterward, we'd like to upper-bound the term A_{12}. Since we have that $\forall i \in [n]$

$$0 \leq \beta_{\tau,s}(X_i, \Theta, \eta_i) \leq \frac{\tau^2}{16s^2}, \quad \mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right)^2 \leq \mathbb{E}\left(\frac{\langle X_i, \Delta\rangle}{\|\Delta\|_{\mathrm{F}}}\right)^4 \leq 16\sigma^4.$$

Then based on the Bousquet's inequality (Bousquet, 2002), with probability at least $1-\epsilon$ it holds that

$$\begin{aligned} A_2 &\leq \mathbb{E}A_2 + \sqrt{\mathbb{E}A_2} \cdot \frac{\tau}{2s}\sqrt{\frac{-\ln(\epsilon)}{n}} + 4\sigma^2\sqrt{\frac{-2\ln(\epsilon)}{n}} + \frac{\tau^2}{16s^2}\frac{-\ln(\epsilon)}{3n} \\ &\leq 2\mathbb{E}A_2 + 4\sigma^2\sqrt{\frac{-2\ln(\epsilon)}{n}} + + \frac{\tau^2}{16s^2}\frac{-4\ln(\epsilon)}{3n}. \end{aligned}$$

To bound the first term $\mathbb{E}A_2$, we use the regular Rademacher symmetrization argument by defining a series of iid Rademacher random variables $\{e_i\}$ with $\tilde{X}_i, \tilde{\eta}_i$ that are iid with X_i, η_i:

$$\begin{aligned} \mathbb{E}A_2 &= \mathbb{E}\left[\sup_{\Theta\in\Phi}\left|\frac{1}{n}\sum_{i=1}^n \beta_{\tau,s}(X_i, \Theta, \eta_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\beta_{\tau,s}(X_i, \Theta, \eta_i)\right)\right|\right] \\ &\leq \mathbb{E}\left[\sup_{\Theta\in\Phi}\left|\left(\frac{1}{n}\sum_{i=1}^n \beta_{\tau,s}(X_i, \Theta, \eta_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\beta_{\tau,s}(\tilde{X}_i, \Theta, \tilde{\eta}_i)\right)\right) e_i\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{\Theta\in\Phi}\frac{1}{n}\sum_{i=1}^n \beta_{\tau,s}(X_i, \Theta, \eta_i) e_i\right]. \end{aligned}$$

Denote the event $c(l) \coloneqq \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} : \|\Theta - \Theta^*\|_{\text{nuc}} \leq l \|\Theta - \Theta^*\|_{\text{F}}\right\}$. Recall that we define as:

$$\beta_{\tau,s}(X_i, \Theta, \eta_i) = \phi_{\frac{\tau}{2s}}\left(\frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\text{F}}}\right) \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}} = \phi_{\frac{\tau}{2s}}\left(\frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\text{F}}} \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right).$$

Define $c(t) = \frac{2s}{\tau}\phi_{\frac{\tau}{2s}}(t)$ and it is easy to show that $c(\cdot)$ is a 1-Lipschitz function. By using the Talagrand's concentration inequality (Wainwright, 2019), it holds that

$$\begin{aligned}
\mathbb{E}A_2 &\leq \frac{\tau}{s} \cdot \mathbb{E}\left[\sup_{\Theta \in c(l)} \frac{1}{n}\sum_{i=1}^n e_i \cdot \frac{2s}{\tau} \cdot \phi_{\frac{\tau}{2s}}\left(\frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\text{F}}} \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right] \\
&\frac{\tau}{s} \cdot \mathbb{E}\left[\sup_{\Theta \in c(l)} \frac{1}{n}\sum_{i=1}^n e_i \cdot \frac{2s}{\tau} \cdot \frac{\langle X_i, \Theta - \Theta^*\rangle}{\|\Theta - \Theta^*\|_{\text{F}}} \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right] \\
&\leq \frac{\tau}{s} \cdot \mathbb{E}\left[\sup_{\Theta \in c(l)} \frac{1}{n}\left\|\sum_{i=1}^n e_i X_i \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right\|_{\text{op}} \cdot \left\|\frac{\Theta - \Theta^*}{\|\Theta - \Theta^*\|_{\text{F}}}\right\|_{\text{nuc}}\right] \\
&\leq \frac{\tau l}{sn} \cdot \mathbb{E}\left\|\sum_{i=1}^n e_i X_i \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right\|_{\text{op}}.
\end{aligned}$$

By using the same technique in the proof of Theorem D.4.3, we can bound the operator norm by using the covering argument. Denote $\mathcal{N}_{\frac{1}{4}}^d$ be the $1/4$ covering of S^{d-1}, then it holds that

$$\mathbb{E}\left\|\sum_{i=1}^n e_i X_i \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right\|_{\text{op}} \leq \frac{5}{2} \cdot \mathbb{E}\left[\max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} \sum_{i=1}^n e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right].$$

Note for any pair of $u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}$, we have that

$$\mathbb{E}\left(\sum_{i=1}^n e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right) = 0$$

$$\mathbb{E}\left(\sum_{i=1}^n |e_i|^k |u^\top X_i v|^k \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right) \leq \mathbb{E}|u^\top X_i v|^k \leq \frac{1}{2} \cdot k! \cdot (\sqrt{2}\sigma)^{k-2} \cdot 2\sigma^2, \quad k = 2, 3, \ldots.$$

We can write the moment generating function $M(\lambda)$ of the random variable $\sum_{i=1}^{n} e_i \cdot u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}$ as:

$$\begin{aligned}
M(\lambda) &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} e_i \cdot u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda e_i \cdot u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right] \\
&\leq \prod_{i=1}^{n}\left[1 + \frac{\lambda^2 \cdot 2\sigma^2}{2} + \frac{\lambda^2 \cdot 2\sigma^2}{2}\left(\sum_{k=3}^{\infty}(|\lambda|\sqrt{2}\sigma)^{k-2}\right)\right] \\
&= \prod_{i=1}^{n}\left[1 + \frac{2\lambda^2\sigma^2}{2} \cdot \frac{1}{1-\sqrt{2}\sigma|\lambda|}\right] \\
&\leq \exp\left(n\lambda^2\sigma^2 \frac{1}{1-\sqrt{2}\sigma|\lambda|}\right), \qquad |\lambda| \leq \frac{1}{\sqrt{2}\sigma}.
\end{aligned}$$

Therefore, it holds that for any $s_0 > 0$

$$\begin{aligned}
\mathbb{E}\left[\max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} \sum_{i=1}^{n} e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right] &= \frac{1}{s_0}\mathbb{E}\left[\ln\left(\exp\left(s_0 \cdot \sum_{i=1}^{n} e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right)\right] \\
&\leq \frac{1}{s_0}\ln\left(\mathbb{E}\left[\max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} \exp\left(s_0 \cdot \sum_{i=1}^{n} e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right]\right) \\
&\leq \frac{1}{s_0}\ln\left(9^{d_1+d_2}\mathbb{E}\left[\exp\left(s_0 \cdot \sum_{i=1}^{n} e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right)\right]\right) \\
&= \frac{(d_1+d_2)\ln(9) + ns_0^2\sigma^2 \cdot \frac{1}{1-\sqrt{2}\sigma|s_0|}}{s_0}, \quad \forall |s_0| \leq \frac{1}{\sqrt{2}\sigma}.
\end{aligned}$$

By taking $s_0 = \frac{\sqrt{(d_1+d_2)\ln(9)}}{\sigma\cdot\sqrt{n}}$, and conditioned on $n \geq 8\ln(9)(d_1+d_2)$, we have that

$$\mathbb{E}\left[\max_{u \in \mathcal{N}_{\frac{1}{4}}^{d_1}, v \in \mathcal{N}_{\frac{1}{4}}^{d_2}} \sum_{i=1}^{n} e_i u^\top X_i v \cdot \mathbb{1}_{\{|\eta_i| \leq \frac{\tau}{2}\}}\right] \leq 3\sqrt{\ln(9)} \cdot \sqrt{n(d_1+d_2)} \cdot \sigma.$$

And this fact implies that

$$\mathbb{E}A_2 \leq \frac{15\tau\sigma l}{2s}\sqrt{\ln(9)}\sqrt{\frac{d_1+d_2}{n}}.$$

Conclusively, with probability at least $1-\epsilon$ we have that

$$A_2 \leq \frac{15\tau\sigma l}{s}\sqrt{\ln(9)}\sqrt{\frac{d_1+d_2}{n}} + 4\sigma^2\sqrt{\frac{-2\ln(\epsilon)}{n}} + + \frac{\tau^2}{16s^2}\frac{-4\ln(\epsilon)}{3n}.$$

Therefore, by ensuring that

$$n \geq \max\left\{8\ln(9)(d_1+d_2), \left(225\sigma\sqrt{\ln(9)(d_1+d_2)}\frac{\tau l}{sc_l}\right)^2, \left(\frac{48\sigma^2}{c_l}\sqrt{-2\ln(\epsilon)}\right)^2, -\frac{\tau^2}{c_l s^2}\ln(\epsilon)\right\},$$

we have

$$P\left(A_2 \leq \frac{c_l}{4}\right) \geq 1-\epsilon. \tag{D.8}$$

Given the results shown in Eqn. (D.7) and Eqn. (D.8), we have that with probability at least $1-\epsilon$, it holds that

$$\frac{\langle \nabla \hat{L}_\tau(\Theta) - \nabla \hat{L}_\tau(\Theta^*), \Theta - \Theta^* \rangle}{\|\Theta - \Theta^*\|_{\mathrm{F}}^2} \geq \frac{c_l}{4}, \quad \forall \Theta \in \Phi.$$

□

D.4.2. Proof of Theorem 5.4.1. Theorem 5.4.1 can be naturally proved based on the above Theorem D.4.2, Theorem D.4.3 and Theorem D.4.4. Here we assume c_l and σ are in constant scale in general, and for the LowHTR problem with $\sigma^2 \asymp c_l \asymp \frac{1}{d_1 d_2}$, our proof can be slightly modified as we discuss later.

By taking $\lambda \asymp \sigma\left(\frac{d-\ln(\epsilon)}{n}\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}$, and $\tau \asymp \left(\frac{n}{d-\ln(\epsilon)}\right)^{\frac{1}{1+\delta}} c^{\frac{1}{1+\delta}}$, we can guarantee that $\lambda \geq 2\left\|\nabla \hat{L}_\tau(\Theta^*)\right\|_{\mathrm{op}}$ with probability at least $1-\epsilon$ from Theorem D.4.3. By choosing $l \asymp 4\sqrt{2r}$ and $s = \frac{\tau}{32\sigma^2}\sqrt{c_l}$, then the conditions in Theorem D.4.2 can be satisfied as long as $n \gtrsim (d-\ln(\epsilon))\sqrt{r\nu^3}$ where we denote $\nu = \frac{\sigma^2}{c_l}$. Furthermore, under the above setting, we know the local restricted strong convexity $\mathcal{E}(s, l, c_l/4)$ holds with probability at least $1-\epsilon$ as long as the conditions in Theorem D.4.4 hold. By reviewing the conditions of Theorem D.4.4, we know it suffices to have $n \gtrsim d, \nu^2, dr\nu^3$. Therefore, with probability at least $1-2\epsilon$, the final error bound in Theorem D.4.2 indicates that

$$\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{c_l}\left(\frac{d+\ln(1/\epsilon)}{n}\right)^{\frac{\delta}{1+\delta}} c^{\frac{1}{1+\delta}}\sqrt{r}.$$

□

D.5. Proof of Theorem 5.4.2

We now prove the regret bound given in Theorem 5.4.2: We have $\|\theta^*\| \leq S$ based on Section 5.3 and $\|\theta^*_{k+1:p}\| \leq S_\perp$ for some small $S_\perp$. In the beginning, we have the transformed buffer set $\mathcal{H}_1'$ of size $H \coloneqq |\mathcal{H}_1'|$, and we write the pair information (X, y) in $\mathcal{H}_1'$ as $\{(x_{s,1}, y_{s,1}), \ldots, (x_{s,H}, y_{s,H})\}$.

And we denote $(x_{e,t}, y_{e,t})$ as the pair of pulled arm and corresponding stochastic payoff at round t. To abuse the notation, at round $t+1$ we denote $\{(x_i, y_i)\}_{i=1}^{t+H}$ as the pairs of observations in the initial buffer set and obtained by the end of round t in order.

At the beginning of the round t+1, the current $M := M_t$ can be written as $M_t = \sum_{i=1}^{H} x_{s,i} x_{s-i}^\top \sum_{j=1}^{t} x_{e,j} x_{e,j}^\top +$ Λ, where Λ is a positive diagonal matrix with λ occupying the first k diagonal entries and $\lambda_\perp$ the next $p-k$ entries. According to Algorithm 9, we denote $X_t \in \mathbb{R}^{(t+H)\times p}$ where each row of X_t is the feature vector of the pulled arm (in the history buffer set or not). Assume $t + H > p$, we denote its full SVD as $X_t = U_x \Sigma_x V_x^\top$ with $U_x \in \mathbb{R}^{(t+H)\times p}$ and $V_x \in \mathbb{R}^{p\times p}$. We also write $M_t = V_x(\Sigma^2 + \Lambda)V_x^\top \in \mathbb{R}^{p\times p}$. And we further denote

$$\begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_p^\top \end{pmatrix} = M_t^{-\frac{1}{2}} X_t^\top = V_x(\Sigma_x^2 + \Lambda)^{-\frac{1}{2}} \cdot \Sigma_x U_x^\top \preceq V_x U_x^\top = \begin{pmatrix} V_{x,11} & \cdots & V_{x,1p} \\ \vdots & \ddots & \vdots \\ V_{x,p1} & \cdots & V_{x,pp} \end{pmatrix} \cdot \begin{pmatrix} U_{x,1}^\top \\ \vdots \\ U_{x,p}^\top \end{pmatrix} \in \mathbb{R}^{p\times(t+H)}.$$

We first show that for all $i \in [p]$,

$$\|u_i\|_2 \le \|\sum_{j=1}^{p} V_{ij} U_j\|_2 = \sqrt{\sum_{j=1}^{p} V_{ij}^2 \|U_j\|_2^2} = 1$$

$$\|u_i\|_{1+\delta} \le (t+H)^{\frac{1}{1+\delta}-\frac{1}{2}} \cdot \|u_i\|_2 \le (t+H)^{\frac{1-\delta}{2(1+\delta)}},$$

where the last inequality is deduced from the Cauchy-Schwarz inequality. With the formulation of $\hat{\theta}_t$ in Algorithm 9 line 3, we have that

$$\begin{aligned}
\|\hat{\theta}_t - \theta^*\|_{M_t} &= \left\| M_t^{-\frac{1}{2}} \begin{pmatrix} u_1^\top \hat{y}_1 \\ \vdots \\ u_p^\top \hat{y}_p \end{pmatrix} - M_t^{-1} X_t^\top X_t \theta^* - M_t^{-1} \Lambda \theta^* \right\|_{M_t} \\
&\leq \left\| M_t^{-\frac{1}{2}} \begin{pmatrix} u_1^\top \hat{y}_1 \\ \vdots \\ u_p^\top \hat{y}_p \end{pmatrix} - M_t^{-\frac{1}{2}} \begin{pmatrix} u_1^\top \\ \vdots \\ u_p^\top \end{pmatrix} X_t \theta^* \right\|_{M_t} + \|\Lambda \theta^*\|_{M_t^{-1}} \\
&\leq \left\| \begin{pmatrix} u_1^\top (\hat{y}_1 - X_t \theta^*) \\ \vdots \\ u_p^\top (\hat{y}_p - X_t \theta^*) \end{pmatrix} \right\|_2 + \|\theta^*\|_\Lambda \\
&\leq \sqrt{\sum_{i=1}^{p} \left(u_i^\top (\hat{y}_i - X_t \theta^*) \right)^2} + \sqrt{\lambda_0} S + \sqrt{\lambda_\perp} S_\perp .
\end{aligned}$$

To present a bound on the first term, we divide it into two separate parts.

$$\begin{aligned}
u_i^\top (\hat{y}_i - X_t \theta^*) &= \sum_{j=1}^{t+H} u_{i,j} (\hat{y}_{i,j} - \mathbb{E}(y_j | \mathcal{F}_{j-1})) \\
&= \sum_{j=1}^{t+H} u_{i,j} \left[(\hat{y}_{i,j} - \mathbb{E}(\hat{y}_{i,j} | \mathcal{F}_{j-1})) - \mathbb{E}(y_j \mathbb{1}_{\{|u_{i,j} y_j| > b_t\}} | \mathcal{F}_{j-1}) \right] \\
&\leq \left| \sum_{j=1}^{t+H} u_{i,j} (\hat{y}_{i,j} - \mathbb{E}(\hat{y}_{i,j} | \mathcal{F}_{j-1})) \right| + \left| \sum_{j=1}^{t+H} u_{i,j} \mathbb{E}(y_j \mathbb{1}_{\{|u_{i,j} y_j| > b_t\}} | \mathcal{F}_{j-1}) \right| \coloneqq A_1 + A_2
\end{aligned}$$

For the first term A_1, based on Bernstein' inequality for martingales (Seldin et al., 2012), for any $i \in [p]$ it holds that with probability at least $1 - \frac{\epsilon}{p}$:

$$\begin{aligned}
A_1 &\leq 2 b_t \ln \left(\frac{2p}{\epsilon} \right) + \left| \frac{1}{2 b_t} \sum_{j=1}^{t+H} \mathbb{E} \left[u_{i,j}^2 \left(\hat{y}_{i,j} - \mathbb{E}(\hat{y}_{i,j} | \mathcal{F}_{j-1}) \right)^2 | \mathcal{F}_{j-1} \right] \right| \\
&\leq 2 b_t \ln \left(\frac{2p}{\epsilon} \right) + \frac{b_t}{2} \left| \sum_{j=1}^{t+H} \mathbb{E} \left[\left(\frac{u_{i,j} \left(\hat{y}_{i,j} - \mathbb{E}(\hat{y}_{i,j} | \mathcal{F}_{j-1}) \right)}{b_t} \right)^2 | \mathcal{F}_{j-1} \right] \right| \coloneqq 2 b_t \ln \left(\frac{2p}{\epsilon} \right) + \frac{b_t}{2} \left| \sum_{j=1}^{t+H} \mathbb{E} \left[T | \mathcal{F}_{j-1} \right] \right| .
\end{aligned}$$

Since we know that $|T| \leq 1$ and hence $\mathbb{E}(T^2) \leq \mathbb{E}(|T|^{1+\delta})$, and we can then deduce that

$$A_1 \leq 2b_t \ln\left(\frac{2p}{\epsilon}\right) + \frac{b_t}{2} \cdot \frac{\sum_{j=1}^{t+H} |u_{i,j}|^{1+\delta} \cdot b}{b_t^{1+\delta}} \leq 2b_t \ln\left(\frac{2p}{\epsilon}\right) + \frac{b}{2b_t^{\delta}}(t+H)^{\frac{1-\delta}{2}}.$$

Therefore, we know that with probability at least $1-\epsilon$ the following result holds for all $i \in [p]$ simultaneously:

$$\left|\sum_{j=1}^{t+H} u_{i,j}(\hat{y}_{i,j} - \mathbb{E}(\hat{y}_{i,j}|\mathcal{F}_{j-1}))\right| \leq 2b_t \ln\left(\frac{2p}{\epsilon}\right) + \frac{b}{2b_t^{\delta}}(t+H)^{\frac{1-\delta}{2}}.$$

For the term A_2, with the help of Holder's inequality, we have for all $i \in [p]$:

$$\begin{aligned}
A_2 &\leq \sum_{j=1}^{t+H} \mathbb{E}\left(|u_{i,j}y_j|^{1+\delta}\right)^{\frac{1}{1+\delta}} \cdot \mathbb{E}\left(\mathbb{1}_{|u_{i,j}y_j|>b_t}\right)^{\frac{\delta}{1+\delta}} \\
&\leq \sum_{j=1}^{t+H} |u_{i,j}| \cdot b^{\frac{1}{1+\delta}} \cdot P\left(|u_{i,j}y_j| > b_t\right)^{\frac{\delta}{1+\delta}} \\
&\leq \sum_{j=1}^{t+H} |u_{i,j}| \cdot b^{\frac{1}{1+\delta}} \cdot \left(\frac{|u_{i,j}|^{1+\delta} b}{b_t^{1+\delta}}\right) \leq \frac{b}{b_t^{\delta}} \cdot (t+H)^{\frac{1-\delta}{2}}.
\end{aligned}$$

Therefore, by taking

$$b_t = \left(\frac{b}{\ln\left(\frac{2p}{\epsilon}\right)}\right)^{\frac{1}{1+\delta}} \cdot (t+H)^{\frac{1-\delta}{2+2\delta}},$$

we can deduce that with probability at least $1-\epsilon$ the following result holds for all $i \in [p]$ simultaneously:

$$u_i^{\top}(\hat{y}_i - X_t\theta^*) \leq 4b^{\frac{1}{1+\delta}} \left(\ln\left(\frac{2p}{\delta}\right)\right)^{\frac{\delta}{1+\delta}} \cdot (t+H)^{\frac{1-\delta}{2+2\delta}}.$$

Therefore, with probability at least $1-\epsilon$ it holds that

$$\|\hat{\theta}_t - \theta^*\|_{M_t} \leq 2\sqrt{p} \cdot b^{\frac{1}{1+\delta}} \left(\ln\left(\frac{2p}{\delta}\right)\right)^{\frac{\delta}{1+\delta}} \cdot (t+H)^{\frac{1-\delta}{2+2\delta}} \coloneqq \beta_t(\epsilon).$$

Denote the optimal arm at time $t+1$ as $x_{e,t+1}^*$. Therefore, the instance regret at time $t+1$ can be bounded by

$$\begin{aligned}
{x_{e,t+1}^*}^\top \theta^* - x_{e,t+1}^\top \theta^* &= {x_{e,t+1}^*}^\top \theta^* - {x_{e,t+1}^*}^\top \hat{\theta}_t + {x_{e,t+1}^*}^\top \hat{\theta}_t - x_{e,t+1}^\top \hat{\theta}_t + x_{e,t+1}^\top \hat{\theta}_t - x_{e,t+1}^\top \theta^* \\
&\leq \beta_t(\epsilon)\|x_{e,t+1}^*\|_{M_t^{-1}} + x_{e,t+1}^\top \hat{\theta}_t + \beta_t(\epsilon)\|x_{e,t+1}\|_{M_t^{-1}} - x_{e,t+1}^\top \hat{\theta}_t - \|x_{e,t+1}^*\|_{M_t^{-1}} + \beta_t(\epsilon)\|x_{e,t+1}\|_{M_t^{-1}} \\
&\leq \min\{S^2, 2\beta_t(\epsilon)\|x_{e,t+1}\|_{M_t^{-1}}\}.
\end{aligned}$$

Therefore, with probability at least $1-\epsilon$, it holds that

$$\begin{aligned}
\sum_{t=1}^T r_t &= \sum_{t=1}^T \min\{S^2, 2\beta_t\left(\frac{\epsilon}{T}\right)\|x_{e,t+1}\|_{M_t^{-1}}\} \\
&\leq 2\beta_T\left(\frac{\epsilon}{T}\right)\sum_{t=1}^T \min\{\frac{S^2}{\beta_T\left(\frac{\epsilon}{T}\right)}, \|x_{e,t+1}\|_{M_t^{-1}}\} \leq 2\beta_T\left(\frac{\epsilon}{T}\right)\cdot\sqrt{T}\cdot\sqrt{\sum_{t=1}^T \min\{\|x_{e,t+1}\|_{M_t^{-1}}^2, 1\}}
\end{aligned}$$

We denote $\tilde{M}_{T+1} = \sum_{t=1}^T x_{e,t}x_{e,t}^\top + \Lambda$, and by Lemma 9 of Dani et al. (2008), it holds that

$$\begin{aligned}
\sqrt{\sum_{t=1}^T \min\{\|x_{e,t+1}\|_{M_t^{-1}}^2, 1\}} &\leq 2\ln\left(\frac{\det(\tilde{M}_{T+1})}{\det(\Lambda)}\right) \\
&\leq 2k\cdot\ln\left(1+\frac{S^2}{k\lambda_0}T\right) + 2(p-k)\ln\left(1+\frac{S^2}{(p-k)\lambda_\perp}T\right) \\
&\leq 2k\cdot\ln\left(1+\frac{S^2}{k\lambda_0}T\right) + \frac{2S^2}{\lambda_\perp}T \leq 4k\cdot\ln\left(1+\frac{S^2}{k\lambda_0}T\right),
\end{aligned}$$

by taking that $\lambda_\perp = \frac{S^2T}{k\ln\left(1+\frac{S^2}{k\lambda_0}T\right)}$. Therefore, with probability at least $1-\epsilon$, it holds that

$$\begin{aligned}
R(T) &\leq 2\sqrt{T}\cdot\sqrt{4k\cdot\ln\left(1+\frac{S^2}{k\lambda_0}T\right)}\cdot\left[2\sqrt{p}\cdot b^{\frac{1}{1+\delta}}\left(\ln\left(\frac{2p}{\delta}\right)\right)^{\frac{\delta}{1+\delta}}\cdot(T+H)^{\frac{1-\delta}{2+2\delta}} + \sqrt{\lambda_0}S + \sqrt{\lambda_\perp}S_\perp\right] \\
&= \widetilde{O}\left(\sqrt{kp}\cdot T^{\frac{1}{1+\delta}} + \sqrt{kT} + S_\perp T\right).
\end{aligned}$$

□

D.6. Proof of Eqn. (5.6)

Our argument is adapted from the proof of Theorem 3 in Jun et al. (2019), and we will still present details here for completeness of our work. Furthermore, the proof of Theorem 5.4.4 in our work still relies on the same Lemma.

LEMMA D.6.0.1. (Wedin's $\sin\Theta$ Theorem) *Let the SVDs of matrices A and $\tilde{A}$ be defined as follows:*

$$\begin{pmatrix} U_1 & U_2 & U_3 \end{pmatrix}^\top A \begin{pmatrix} V_1 & V_2 \end{pmatrix} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} \tilde{U}_1 & \tilde{U}_2 & \tilde{U}_3 \end{pmatrix}^\top \tilde{A} \begin{pmatrix} \tilde{V}_1 & \tilde{V}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \\ 0 & 0 \end{pmatrix}.$$

Let $R = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma}_1$ and $S = A^\top\tilde{U}_1 - \tilde{V}_1\tilde{\Sigma}_1$, and define $U_{1\perp} = [U_2\ U_3]$ and $V_{1\perp} = [V_2\ V_3]$. Then suppose there is a number $q > 0$ such that

$$\min_{i,j} |\sigma_i(\tilde{\Sigma}_1) - \sigma_j(\Sigma_2)| \geq q, \quad \min_i \sigma_i(\tilde{\Sigma}_1) \geq q,$$

Then it holds that

$$\sqrt{\left\|U_{1\perp}^\top\tilde{U}_1\right\|_F^2 + \left\|V_{1\perp}^\top\tilde{V}_1\right\|_F^2} \leq \frac{\sqrt{\|R\|_F^2 + \|S\|_F^2}}{q}.$$

Based on Lemma D.6.0.1, we define $A = \widehat{\Theta}, U_1 = \widehat{U}, \Sigma_1 = \widehat{D}, V_1 = \widehat{V}, \tilde{A} = \Theta^*, \tilde{U}_1 = U, \tilde{\Sigma}_1 = D, \tilde{V}_1 = V, q = D_{rr}$. Therefore, according to Lemma D.6.0.1, we have that $R = (\widehat{\Theta} - \Theta^*)\widehat{V}$ and $S = -(\widehat{\Theta} - \Theta^*)^\top U$, and then it holds that

$$\sqrt{2\left\|\widehat{U}_\perp^\top U\right\|_{\mathrm{F}}\left\|\widehat{V}_\perp^\top V\right\|_{\mathrm{F}}} \leq \sqrt{\left\|\widehat{U}_\perp^\top U\right\|_{\mathrm{F}}^2 + \left\|\widehat{V}_\perp^\top V\right\|_{\mathrm{F}}^2} \leq \frac{\sqrt{\|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2}}{D_{rr}} \leq \frac{\sqrt{2}\cdot\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}}}{D_{rr}}.$$

And then by using the bound on $\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}}$ we can deduce that

$$\|\theta^*_{k+1:p}\|_2 = \left\|\widehat{U}_\perp^\top U D V^\top \widehat{V}_\perp\right\|_{\mathrm{F}} \leq \left\|\widehat{U}_\perp^\top U\right\|_{\mathrm{F}}\left\|\widehat{V}_\perp^\top V\right\|_{\mathrm{F}} \cdot \|D\|_{\mathrm{op}} \lesssim \frac{r\sigma^2 c^{\frac{2}{1+\delta}}}{c_l^2 D_{rr}^2}\left(\frac{d + \ln(1/\epsilon)}{|\mathcal{H}_2|}\right)^{\frac{2\delta}{1+\delta}}.$$

□

D.7. Proof of Theorem 5.4.3

We now prove Theorem 5.4.3 in this section. We first bring up the result shown in Eqn. (5.3) again: under Assumption 5.3.1, if we estimate Θ^* based on the exploration set $\mathcal{H}_2$ of size H, then our

estimator $\widehat{\Theta}$ satisfies the following property:

$$\|\theta^*_{k+1:p}\|_2 \lesssim \frac{rd^2c^{\frac{2}{1+\delta}}}{D_{rr}^2}\left(\frac{d+\ln(1/\epsilon)}{H}\right)^{\frac{2\delta}{1+\delta}},$$

under $\sigma^2 \asymp c_l \asymp 1/(d_1 d_2)$ with probability at least $1-\epsilon$. Our Algorithm 8 first randomly samples arms for the first T_1 rounds, and then for the rest of the time horizon it utilizes a doubling-trick-based idea. Based on line 3 of Algorithm 8, when we have that

$$\left[\frac{d^{2+4\delta}r^{1+\delta}}{D_{rr}^{2+2\delta}}2^{i(1+\delta)}\right]^{\frac{1}{1+3\delta}} \geq 2^i \implies i \leq \left\lfloor \log_2\left(\frac{d^{\frac{1+2\delta}{\delta}}r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}}\right)\right\rfloor := L,$$

then in the first L batches, we will run out of time to do random exploration. Since we have that

$$\frac{2d^{\frac{1+2\delta}{\delta}}r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}} \geq \sum_{j=1}^{L}2^j = 2^{L+1}-2 \geq \frac{d^{\frac{1+2\delta}{\delta}}r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}}-2,$$

we know before the batch $L+1$, we already repeat random sampling for T_{init} rounds, with

$$T_1 + \frac{d^{\frac{1+2\delta}{\delta}}r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}} - 2 \leq T_{\text{init}} \leq T_1 + \frac{2d^{\frac{1+2\delta}{\delta}}r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}}.$$

For the sake of simplicity in our proof, we assume that our algorithm terminates exactly at the end of some batch, i.e. the M-th batch. And otherwise, our proof will be the same by using the index of the last batch. In other words, it holds that

$$\sum_{i=L+1}^{M}2^i + T_{\text{init}} = T \iff 2^{M+1} = T + 2^{L+1} - T_{\text{init}}.$$

Therefore, if we set ϵ as $\epsilon/2^{i+1}$ in both β_t of Algorithm 9 and λ, τ in the matrix estimation for the i-th batch, then based on Theorem 5.4.2, with probability at least $1-\epsilon$ it holds that

$$\begin{aligned}
R(T) = \widetilde{O}\Bigg(T_{\text{init}} + \sum_{i=L+1}^{M}\Bigg[&C\left(2^{\frac{1+\delta}{1+3\delta}}\right)^i + \sqrt{d^3r}\left(2^{\frac{1}{1+\delta}}\right)^i + \sqrt{dr2^i} \\
&+2^i\cdot\frac{d^{\frac{2+4\delta}{1+\delta}}r}{D_{rr}^2}\cdot\left(\frac{1}{T_{\text{init}}+\sum_{j=L+1}^{i}C\left(2^{\frac{1+\delta}{1+3\delta}}\right)^j}\right)^{\frac{2\delta}{1+\delta}}\Bigg]\Bigg) \\
= \widetilde{O}\left(A_1 + \sum_{i=L+1}^{M}(A_{i,2}+A_{i,3}+A_{i,4}+A_{i,5})\right),
\end{aligned}$$

with $C = \left(\frac{d^{2+4\delta} r^{1+\delta}}{D_{rr}^{2+2\delta}}\right)^{\frac{1}{1+3\delta}}$. For A_1, it naturally holds that $A_1 \lesssim T_{\text{init}}$. For $A_{i,2}$, we have that

$$\sum_{i=L+1}^{M} A_{i,2} \lesssim C \cdot \frac{1}{2^{\frac{1+\delta}{1+3\delta}} - 1} \cdot T^{\frac{1+\delta}{1+3\delta}}.$$

For $A_{i,3}$, we have that

$$\sum_{i=L+1}^{M} A_{i,3} \lesssim \sqrt{d^3 r} \frac{1}{2^{\frac{1}{1+\delta}} - 1} \cdot (T - T_{\text{init}})^{\frac{1}{1+\delta}} \lesssim \sqrt{d^3 r} \cdot T^{\frac{1}{1+\delta}}.$$

For $A_{i,3}$, it holds that

$$\sum_{i=L+1}^{M} A_{i,4} \lesssim \sqrt{dr}\sqrt{2^i} \lesssim \sqrt{dr} \cdot \frac{1}{\sqrt{2} - 1} \cdot (T - T_{\text{init}})^{\frac{1}{2}} \lesssim \sqrt{drT}.$$

And finally for $A_{i,5}$ we can show that

$$\begin{aligned}
\sum_{i=L+1}^{M} A_{i,5} &= \sum_{i=L+1}^{M} 2^i \cdot \frac{d^{\frac{2+4\delta}{1+\delta}} r}{D_{rr}^2} \cdot \left(\frac{1}{T_{\text{init}} + \sum_{j=L+1}^{i} C \left(2^{\frac{1+\delta}{1+3\delta}}\right)^j} \right)^{\frac{2\delta}{1+\delta}} \\
&\lesssim \sum_{i=L+1}^{M} 2 \cdot C \cdot \left(\frac{\left(2^{\frac{1+\delta}{2\delta}}\right)^i}{\frac{T_1 - 2}{C} + \frac{d^{\frac{1+2\delta}{\delta}} r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}} C} + \sum_{j=L+1}^{i} \left(2^{\frac{1+\delta}{1+3\delta}}\right)^j} \right)^{\frac{2\delta}{1+\delta}} \\
&\lesssim 2 \cdot C \cdot \sum_{L+1}^{M} \left[\frac{\left(2^{\frac{1+\delta}{1+3\delta}} - 1\right)^{\frac{2\delta}{1+\delta}}}{2^{\frac{(1+\delta)(2\delta)}{(1+3\delta)(1+\delta)}}} \cdot 2^{\left(\frac{1+\delta}{1+3\delta}\right) i} \right] \lesssim C \cdot T^{\frac{1+\delta}{1+3\delta}},
\end{aligned}$$

given that

$$T_1 \geq 2 - \frac{d^{\frac{1+2\delta}{\delta}} r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}} + \left(\frac{d^{\frac{1+2\delta}{\delta}} r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}} \right)^{\frac{1+\delta}{1+3\delta}} \cdot \frac{1}{2^{\frac{1+\delta}{1+3\delta}} - 1} \cdot C \geq 2 + \left(\frac{2}{\sqrt{2} - 1} \right) \cdot \frac{d^{\frac{1+2\delta}{\delta}} r^{\frac{1+\delta}{2\delta}}}{D_{rr}^{\frac{1+\delta}{\delta}}}.$$

Therefore, with the above condition on T_1 satisfied, the following result holds with probability at least $1 - \epsilon$

$$R(T) = \widetilde{O}\left(\frac{d^{\frac{2+4\delta}{1+3\delta}} r^{\frac{1+\delta}{1+3\delta}}}{D_{rr}^{\frac{2+2\delta}{1+3\delta}}} \cdot T^{\frac{1+\delta}{1+3\delta}} + d^{\frac{3}{2}} r^{\frac{1}{2}} T^{\frac{1}{1+\delta}} \right).$$

□

D.8. Proof of Theorem 5.4.4

The proof of Theorem 5.4.4 is adapted from that of Theorem 5.4.3 presented in the above Appendix D.7. According to Li (1998), it holds that

$$|\sigma_i(\widehat{\Theta}) - \sigma_i(\Theta^*)| \leq \left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}}, \quad \forall i \in [d].$$

Denote H as the size of the exploration buffer set $\mathcal{H}_2$ at the end of the exploration phase for the $i-$th batch, then according to Theorem 5.4.1 we know that

$$\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}} \leq C_1 \frac{\sigma\sqrt{r}}{c_l} \left(\frac{d + \ln\left(2^{i+1}/\epsilon\right)}{H}\right)^{\frac{\delta}{1+\delta}} \cdot c^{\frac{1}{1+\delta}} := E, \quad C_1 > 0, \tag{D.9}$$

with probability at least $1 - \epsilon/2^{i+1}$. We define the useful rank $\hat{r}$ as:

$$\hat{r} = \min\left\{i \in [d+1] : \hat{D}_{ii} \leq C_1 \frac{\sigma\sqrt{i}}{c_l} \left(\frac{d + \ln\left(2^{i+1}/\epsilon\right)}{H}\right)^{\frac{\delta}{1+\delta}} \cdot c^{\frac{1}{1+\delta}} := R(i)\right\} - 1 \wedge 1,$$

We will first show that $\hat{D}_{(r+1)(r+1)} \leq R(r+1)$ and hence $\hat{r} \leq r$ holds if we have Eqn. (D.9). This is because that $\hat{D}_{(r+1)(r+1)} \leq E = R(r) < R(r+1)$. Furthermore, we will illustrate that all the subspaces we remove based on our estimated $\hat{r}$ are sufficiently minimal. Specifically, we know that

$$D_{(\hat{r}+1)(\hat{r}+1)} \leq \widehat{D}_{(\hat{r}+1)(\hat{r}+1)} + |\widehat{D}_{(\hat{r}+1)(\hat{r}+1)} - D_{(\hat{r}+1)(\hat{r}+1)}| \leq R(\hat{r}+1) + E \leq 2R(r+1).$$

To abuse the notation, we rewrite the SVD of $\widehat{\Theta}$ and Θ^* as

$$\widehat{\Theta} = \begin{pmatrix}\widehat{U} & \widehat{U}_r & \widehat{U}_\perp\end{pmatrix} \cdot \begin{pmatrix}\widehat{D}_{\hat{r}} & 0 & 0 \\ 0 & \widehat{D}_{r-\hat{r}} & 0 \\ 0 & 0 & \widehat{D}_0\end{pmatrix} \cdot \begin{pmatrix}\widehat{V}^\top \\ \widehat{V}_r^\top \\ \widehat{V}_\perp^\top\end{pmatrix}$$

$$\Theta^* = \begin{pmatrix}\tilde{U} & \tilde{U}_r & \tilde{U}_\perp\end{pmatrix} \cdot \begin{pmatrix}\tilde{D}_{\hat{r}} & 0 & 0 \\ 0 & \tilde{D}_{r-\hat{r}} & 0 \\ 0 & 0 & 0\end{pmatrix} \cdot \begin{pmatrix}\tilde{V}^\top \\ \tilde{V}_r^\top \\ \tilde{V}_\perp^\top\end{pmatrix}.$$

And by making sure that H is sufficiently large such that $R(r+1) \leq D_{rr}/2$, we have that

$$\min|\sigma_i(D_{\hat{r}}) - \sigma_j(D_{r-\hat{r}})| \geq \frac{D_{rr}}{2}, \quad \min\sigma_i(D_{\hat{r}}) \geq D_{rr}.$$

In Lemma D.6.0.1, with $A = \widehat{\Theta}, U_1 = \widehat{U}, U_{1\perp} = [\widehat{U}_r, \widehat{U}_\perp], \Sigma_1 = \widehat{D}, V_1 = \widehat{V}, V_{1\perp} = [\widehat{V}_r, \widehat{V}_\perp], \tilde{A} = \Theta^*, \tilde{U}_1 = \tilde{U}, \tilde{\Sigma}_1 = D, \tilde{V}_1 = \tilde{V}, q = D_{rr}/2$, we can show that

$$\left\|\widehat{U}_{1\perp}^\top \tilde{U}\right\|_{\mathrm{F}} \left\|\widehat{V}_{1\perp}^\top \tilde{V}\right\|_{\mathrm{F}} \leq \frac{4\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}}^2}{D_{rr}^2}.$$

After we do the same transformation in Algorithm 9, we know the effective dimension (denoted by $\hat{k}$) satisfies that $\hat{k} = d_1 d_2 - (d_1 - \hat{r})(d_2 - \hat{r}) \leq d_1 d_2 - (d_1 - r)(d_2 - r) = k$. And it holds that

$$\begin{aligned}
\|\theta^*_{\hat{k}+1:p}\|_2 &= \left\| U_{1\perp}^\top \begin{pmatrix} \tilde{U} & \tilde{U}_r \end{pmatrix} \cdot \begin{pmatrix} D_{\hat{r}} & 0 \\ 0 & D_{r-\hat{r}} \end{pmatrix} \cdot \begin{pmatrix} \tilde{V}^\top \\ \tilde{V}_r^\top \end{pmatrix} V_{1\perp} \right\|_{\mathrm{F}} \\
&= \left\| U_{1\perp}^\top \tilde{U} D_{\hat{r}} \tilde{V}^\top V_{1\perp} + U_{1\perp}^\top \tilde{U}_r D_{r-\hat{r}} \tilde{V}_r^\top V_{1\perp} \right\|_{\mathrm{F}} \\
&\leq \left\| U_{1\perp}^\top \tilde{U} \right\|_{\mathrm{F}} \left\| \tilde{V}^\top V_{1\perp} \right\|_{\mathrm{F}} \cdot \|D_{\hat{r}}\|_{\mathrm{op}} + \left\| U_{1\perp}^\top \tilde{U}_r \right\|_{\mathrm{F}} \left\| \tilde{V}_r^\top V_{1\perp} \right\|_{\mathrm{F}} \cdot \|D_{r-\hat{r}}\|_{\mathrm{op}} \\
&\leq \|\Theta^*\|_{\mathrm{op}} \cdot \frac{4\left\|\widehat{\Theta} - \Theta^*\right\|_{\mathrm{F}}^2}{D_{rr}^2} + \sqrt{r - \hat{r}}^2 \cdot 2R(r+1) \\
&\widetilde{O}\left(\frac{rd^2}{D_{rr}^2} \left(\frac{d}{H}\right)^{\frac{2\delta}{1+\delta}} + r^{\frac{3}{2}} d \left(\frac{d}{H}\right)^{\frac{\delta}{1+\delta}} \right) \asymp \widetilde{O}\left(r^{\frac{3}{2}} d \left(\frac{d}{H}\right)^{\frac{\delta}{1+\delta}} \right).
\end{aligned}$$

Note the second term will be dominant for large H, s.t. $H \geq \frac{d^{\frac{1+2\delta}{\delta}}}{r^{\frac{1+\delta}{2\delta}} D_{rr}^{\frac{2+2\delta}{\delta}}}$.

By using $T_1 = \min\left\{ d \cdot 2^{\frac{i(1+\delta)}{1+2\delta}}, 2^i \right\}$ at each batch in line 3 of Algorithm 8, we can identically prove Theorem 5.4.4 with the same procedure as the proof of Theorem 5.4.3. And the only slight difference lies in the control of the term $A_{i,5}$. Therefore, we will omit the redundant details here.

□

D.9. Proof of Theorem 5.5.1

In this section, we will present a regret lower bound for the LowHTR. Our proof relies on the following Lemma for the MAB with heavy-tailed rewards:

Lemma D.9.0.1. (Xue et al., 2020) *For any multi-armed bandit algorithm $\mathcal{B}$ with $T \geq K \geq 4$ where K is the number of arms, an arm $a^* \in \{1, \ldots, K\}$ is chosen uniformly at random, this arm pays $1/\gamma$ with probability $p(a^*) = 2\gamma^{1+\delta}$ and the rest pays $1/\gamma$ with probability $\gamma^{1+\delta}$ ($2\gamma^{1+\delta} < 1$). If we set $\gamma = (K/(T+2K))^{\frac{1}{1+\delta}}$, and denote $r_{t,a}$ as the observed reward of arm a at round t under*

algorithm $\mathcal{B}$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} r_{t,a^*} - \sum_{t=1}^{T} r_{t,a_t}\right] \geq \frac{1}{8} T^{\frac{1}{1+\delta}} K^{\frac{\delta}{1+\delta}}.$$

Therefore, we can naturally consider the LowHTR problem with a finite and fixed arm set of size K. For simplicity, we set $d_1 = d_2 = d$ and set $K = (d-1)r \geq 4$. To adapt the results from Lemma D.9.0.1, we make the reward function of an arm $X_{t,a} \in \mathbb{R}^{d^2}$ as

$$r_{t,a} = \begin{cases} \frac{1}{\gamma}, & \text{with probability } \gamma \cdot \langle X_{t,a}, \Theta^* \rangle \\ 0, & \text{with probability } 1 - \gamma \cdot \langle X_{t,a}, \Theta^* \rangle \end{cases},$$

and then we only need to make $\langle X_{t,a^*}, \Theta^* \rangle = 2\gamma^\delta$ and $\langle X_{t,a}, \Theta^* \rangle = \gamma^\delta$ for any other arm a where a^* is uniformly chosen from $[K]$.

The contextual matrices are designed in the following way. For the first column, the first r entries are set to be $\left[\sqrt{\frac{1}{r(r+1)}}, \sqrt{\frac{2}{r(r+1)}}, \ldots, \sqrt{\frac{r}{r(r+1)}}\right]$. And for the rest $(d-1)r$ entries in the first r rows, we flatten them and set the i-th entry as $\frac{1}{\sqrt{2}}$ for the i-th arm matrix. All the other elements in the last $(d-k)$ rows are set to null for all arm matrices. We can easily check that the Frobenious norm of all arm matrices are bounded by 1.

Next, we consider the parameter matrix Θ^* of rank r. For the first column, the first r entries are set to be $\left[\sqrt{\frac{4}{r(r+1)}}\gamma^\delta, \sqrt{\frac{8}{r(r+1)}}\gamma^\delta, \ldots, \sqrt{\frac{4r}{r(r+1)}}\gamma^\delta\right]$. And similarly for the rest $(d-1)r$ entries in the first r rows, we flatten them and uniformly choose an index from $[(d-1)r]$, then the corresponding entry is $\sqrt{2}\gamma^\delta$ and all the rest elements in Θ^* are 0. The norm of Θ^* can also be bounded with large T. By using the feature matrices and the parameter matrix described above, we can recover the scenario in Lemma D.9.0.1, and thus we have that

$$\mathbb{E}R(T) \geq \frac{1}{8} T^{\frac{1}{1+\delta}} (d-1)^{\frac{\delta}{1+\delta}} r^{\frac{\delta}{1+\delta}} \asymp T^{\frac{1}{1+\delta}} d^{\frac{\delta}{1+\delta}} r^{\frac{\delta}{1+\delta}} \gtrsim T^{\frac{1}{1+\delta}}.$$

□

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9. PMLR, 2012.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Kaito Ariu, Kenshi Abe, and Alexandre Proutière. Thresholded lasso bandit. In *International Conference on Machine Learning*, pp. 878–928. PMLR, 2022.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158. PMLR, 2019.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Djallel Bouneffouf and Emmanuelle Claeys. Hyper-parameter tuning for the contextual bandit. *arXiv preprint arXiv:2005.02209*, 2020.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier, 2003.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 418–427. PMLR, 2019.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Kani Chen, Inchi Hu, Zhiliang Ying, et al. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Annals of Statistics*, 27(4): 1155–1163, 1999.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

Pin-Yu Chen and Cho-Jui Hsieh. *Adversarial robustness for machine learning*. Academic Press, San Diego, CA, August 2022.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Wei Chu, Seung-Taek Park, Todd Beaupre, Nitin Motgi, Amit Phadke, Seinjuti Chakraborty, and Joe Zachariah. A case study of behavior-driven conjoint analysis on yahoo! front page today module. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1104, 2009.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Rama Cont and Jean-Philipe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2):170–196, 2000.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. pp. 355–366, 2008.

Persi Diaconis, Charles Stein, Susan Holmes, and Gesine Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pp. 1–25. Institute of Mathematical Statistics, 2004.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 1585–1593. PMLR, 2021.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123. PMLR, 2022a.

Qin Ding, Yue Kang, Yi-Wei Liu, Thomas Chun Man Lee, Cho-Jui Hsieh, and James Sharpnack. Syndicated bandits: A framework for auto tuning hyper-parameters in contextual bandit algorithms. *Advances in Neural Information Processing Systems*, 35:1170–1181, 2022b.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46(2):814, 2018.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics*, 212(1):177–202, 2019.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239, 2021.

Yasong Feng, Tianyu Wang, et al. Lipschitz bandits with batched feedback. *Advances in Neural Information Processing Systems*, 35:19836–19848, 2022.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

Aditya Gopalan, Odalric-Ambrym Maillard, and Mohammadi Zaki. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Ole-Christoffer Granmo. Solving two-armed bernoulli bandit problems using a bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2):207–234, 2010.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

Yi Han and Thomas CM Lee. Uncertainty quantification for sparse estimation of spectral lines. *IEEE Transactions on Signal Processing*, 70:6243–6256, 2022.

Botao Hao, Jie Zhou, Zheng Wen, and Will Wei Sun. Low-rank tensor bandits. *arXiv preprint arXiv:2007.15788*, 2020.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

David C Hoaglin, Frederick Mosteller, and John W Tukey. *Understanding robust and exploratory data analysis*, volume 76. John Wiley & Sons, 2000.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

Kyoungseok Jang, Kwang-Sung Jun, Se-Young Yun, and Wanmo Kang. Improved regret bounds of bilinear bandits using action space analysis. In *International Conference on Machine Learning*, pp. 4744–4754. PMLR, 2021.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Nicholas Johnson, Vidyashankar Sivakumar, and Arindam Banerjee. Structured stochastic linear bandits. *arXiv preprint arXiv:1606.05693*, 2016.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

Minhyun Kang and Gi-Soo Kim. Heavy-tailed linear bandit with huber regression. In *Uncertainty in Artificial Intelligence*, pp. 1027–1036. PMLR, 2023.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35:19971–19983, 2022.

Yue Kang, Cho-Jui Hsieh, and Thomas Lee. Online continuous hyperparameter optimization for generalized linear contextual bandits. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=lQE2AcbYge.

Yue Kang, Cho-Jui Hsieh, and Thomas Lee. Low-rank matrix bandits with heavy-tailed rewards. *arXiv preprint arXiv:2404.17709*, 2024b.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. *Advances in Neural Information Processing Systems*, 36, 2024c.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Bernoulli rank-1 bandits for click feedback. *arXiv preprint arXiv:1703.06513*, 2017a.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, pp. 392–401. PMLR, 2017b.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 2004.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

Branislav Kveton, Csaba Szepesvári, Anup Rao, Zheng Wen, Yasin Abbasi-Yadkori, and S Muthukrishnan. Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*, 2017.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Ren-Cang Li. Relative perturbation theory: I. eigenvalue and singular value variations. *SIAM Journal on Matrix Analysis and Applications*, 19(4):956–982, 1998.

Wenjie Li, Adarsh Barik, and Jean Honorio. A simple unified framework for high dimensional bandit problems. In *International Conference on Machine Learning*, pp. 12619–12655. PMLR, 2022.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 1306–1314. PMLR, 2021.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pp. 4154–4163. PMLR, 2019.

Xiuyuan Lu, Zheng Wen, and Branislav Kveton. Efficient online recommendation via low-rank ensemble sampling. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 460–464, 2018.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 460–468. PMLR, 2021.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Haixu Ma, Yufeng Liu, and Guorong Wu. Elucidating multi-stage progression of neuro-degeneration process in alzheimer's disease. *Alzheimer's & Dementia*, 18:e068774, 2022a.

Haixu Ma, Donglin Zeng, and Yufeng Liu. Learning individualized treatment rules with many treatments: A supervised clustering approach using adaptive fusion. *Advances in Neural Information Processing Systems*, 35:15956–15969, 2022b.

Haixu Ma, Donglin Zeng, and Yufeng Liu. Learning optimal group-structured individualized treatment rules with many treatments. *Journal of Machine Learning Research*, 24(102):1–48, 2023.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pp. 975–999. PMLR, 2014.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of statistics*, 39(2):1069–1097, 2011.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pp. 8271–8280. PMLR, 2021.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural*

Information Processing Systems, 33:10328–10337, 2020.

Chara Podimata and Alex Slivkins. Adaptive discretization for adversarial lipschitz bandits. In *Conference on Learning Theory*, pp. 3788–3805. PMLR, 2021.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Jaxk Reeves, Jien Chen, Xiaolan L Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology*, 46(6):900–915, 2007.

Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

Yevgeny Seldin, François Laviolette, Nicolo Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093, 2012.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

Amr Sharaf and Hal Daumé III. Meta-learning for contextual bandit exploration. *arXiv preprint arXiv:1901.08159*, 2019.

Yinan Shen, Jingyang Li, Jian-Feng Cai, and Dong Xia. Computationally efficient and statistically optimal robust low-rank matrix estimation. *arXiv preprint arXiv:2203.00953*, 2022.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 679–702. JMLR Workshop and Conference Proceedings, 2011.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Yitzchak Solomon, Alexander Wagner, and Paul Bendich. A fast and robust method for global topological functional optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 109–117. PMLR, 2021.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Gilbert W Stewart. Matrix perturbation theory. 1990.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, pp. 1–11, 2022.

Cindy Trinh, Emilie Kaufmann, Claire Vernade, and Richard Combes. Solving bernoulli rank-one bandits with unimodal thompson sampling. In *Algorithmic Learning Theory*, pp. 862–889. PMLR, 2020.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pp. 19–27. PMLR, 2013.

Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

Sofía S Villar, Jack Bowden, and James 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.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Yizhou Wang, Yue Kang, Can Qin, Huan Wang, Yi Xu, Yulun Zhang, and Yun Fu. Momentum is all you need for data-driven adaptive optimization. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1385–1390. IEEE, 2023.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. *Advances in neural information processing systems*, 29, 2016.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. *arXiv preprint arXiv:2004.13465*, 2020.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pp. 3851–3860. PMLR, 2017.

Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, and Inderjit S Dhillon. Parallel matrix factorization for recommender systems. *Knowledge and Information Systems*, 41(3):793–819, 2014.

Myeonghun Yu, Qiang Sun, and Wenxin Zhou. Low-rank matrix recovery under heavy-tailed errors. *Bernoulli*, 2023.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pp. 392–401. PMLR, 2016.

Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*, 2021.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 746–755. PMLR, 2020.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International conference on algorithmic learning theory*, pp. 3–18. Springer, 2015.

Xuehu Zhu, Yue Kang, and Junmin Liu. Estimation of the number of endmembers via thresholding ridge ratio criterion. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):637–649, 2019.

Zheqing Zhu and Benjamin Van Roy. Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3636–3646, 2023.