Partially Observable Stochastic Contextual Bandits

by

Hongju Park

(Under the Direction of Mohamad Shirani Kazem Faradonbeh)

Abstract

Bandits constitute a classical framework for decision-making under uncertainty. Stochastic contextual bandits are a variant of bandits, which consist of multiple arms with each own stochastic context. In this setting, the goal is to learn the arms of highest reward subject to contextual information, while the unknown reward parameter of each arm needs to be learned. To maximize cumulative reward, an adaptive policy is required to manage the delicate trade-off between learning the best (i.e., exploration) and earning the most (i.e., exploitation). To study this problem, the existing literature mostly considers perfectly observed contexts. However, the setting of partial context observations remains unexplored to date, despite being theoretically more general and practically more versatile. Thus, we consider partial observations, which are noisy linear functions of the unobserved context vectors. Another important issue for contextual bandits is to find the optimal algorithm for the exploration-exploitation trade-off based on different reward structures. We suggest two different reward setups: shared across all arms and arm-specific. We study two different policies, which are Greedy and Thompson sampling algorithms, for these two different reward setups. This study shows that Greedy algorithm has the optimal rate performance in the shared parameter setup, while Thompson sampling successfully balances exploration and exploitation in the arm-specific reward parameter setup. Specifically, We establish the following primary results for these algorithms in two reward setups: (i) Greedy algorithm has a high-probability upper bound for regret in the shared parameter setup and (ii) Thompson sampling has poly-logarithmic high-probability upper bounds for regret in both parameter setups. Extensive numerical experiments with both real and synthetic data are presented as well, corroborating the efficacy of Thompson sampling and Greedy algorithm. To establish the results, we utilize martingale techniques and concentration inequalities for dependent data and also develop novel probabilistic bounds for time-varying suboptimality gaps, among others. These techniques pave the road towards studying other decision-making problems with contextual information.

Index words: [Contextual Bandits, Greedy Algorithm, Partial Observability, Reinforcement Learning, Thompson Sampling]

Partially Observable Stochastic Contextual Bandits

by

Hongju Park

B.S., Yonsei University, Republic of Korea, 2016

M.S., Yonsei University, Republic of Korea, 2018

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2024

Partially Observable Stochastic Contextual Bandits

by

Hongju Park

Major Professor:	Mohamad Kazem Shirani Faradonbeh
Committee:	Ray Bai
	Prashant Doshi
	Javad Mohammadpour Velni

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2024

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Mohamad Kazem Shirani Faradonbeh, for his invaluable guidance, unwavering support, and scholarly mentorship throughout the journey of completing this dissertation. Dr. Faradonbeh's expertise, insight, and dedication have been instrumental in shaping the direction of my research and fostering my academic growth. His passion for statistics, profound knowledge, and commitment to excellence have served as a constant source of inspiration and motivation.

I am indebted to Dr. Faradonbeh for his patience, encouragement, and constructive feedback at every stage of this dissertation. His advisorship extended beyond academic matters, encompassing professional development, research methodology, and career aspirations. I am truly grateful for the countless hours he devoted to reviewing drafts, discussing ideas, and providing invaluable insights that have significantly enriched the quality of this work.

I would also like to extend my appreciation to the members of my dissertation committee, Drs. Bai, Doshi, and Mohammadpour Velni, for their insightful feedback, valuable suggestions, and critical evaluation of my research. Furthermore, I would like to express my sincere gratitude to Dr. Liu, my graduate coordinator, for his invaluable assistance and support throughout the process of completing my dissertation and navigating the complexities of the graduation process.

My heartfelt thanks go to my friends, who have supported me throughout this journey with their encouragement, camaraderie, and intellectual exchange. Their camaraderie and encouragement have provided solace during challenging times and added joy to the process of scholarly inquiry.

Finally, I am deeply grateful to my family for their unwavering love, encouragement, and belief in my abilities. Their unconditional support, sacrifices, and understanding have been the cornerstone of my academic pursuits and personal growth.

In conclusion, I am profoundly grateful to Dr. Faradonbeh and all those who have contributed to this dissertation, directly or indirectly. Their support and encouragement have been instrumental in the realization of this academic milestone.

Publication List:

1. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh, "Thompson Sampling in Partially Observable Contextual Bandits." Under Review.

2. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "Sequentially Adaptive Experimentation for Learning Optimal Options Subject to Unobserved Contexts." *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*. 2023. (`link`)

3. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "Online Learning of Optimal Prescriptions under Bandit Feedback with Unknown Contexts." *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*. 2023. (`link`)

4. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "Balancing Exploration and Exploitation in Partially Observed Linear Contextual Bandits via Thompson Sampling." *ICML 2023 Workshop on New Frontiers in Learning, Control, and Dynamical Systems*. 2023. (`link`)

5. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "Worst-case Performance of Greedy Policies in Bandits with Imperfect Context Observations." *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022. (`link`)

6. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "A Regret Bound for Greedy Partially Observed Stochastic Contextual Bandits." *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*. 2022. (`link`)

7. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. "Efficient Algorithms for Learning to Control Bandits with Unobserved Contexts." *IFAC-PapersOnLine* 55.12 (2022): 383-388. (`link`)

8. **Hongju Park** and Mohamad Kazem Shirani Faradonbeh. “Analysis of Thompson Sampling for Partially Observable Contextual Multi-armed Bandits." *IEEE Control Systems Letters* 6 (2021): 2150-2155. (`link`)

9. **Hongju Park**, Taeyoung Park, and Yung-Seop Lee. “Partially Collapsed Gibbs Sampling for Latent Dirichlet Allocation." *Expert Systems with Applications* 131 (2019): 208-218. (`link`)

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

In the realm of personalized decision-making, contextual bandits stand out as powerful tools for sequential actions in dynamic environments. This model leverages contextual information to tailor decisions to individual users, striking a balance between exploration and exploitation to maximize long-term rewards. The range of applications for contextual bandits is extensive and encompasses various scenarios where time-varying and action-dependent information play a crucial role. These applications include personalized recommendation systems for news articles, interventions in healthcare settings, targeted advertising campaigns, and the optimization of clinical trials (Bouneffouf et al., 2012; Durand et al., 2018; Li et al., 2010; Nahum-Shani et al., 2018; Ren & Zhou, 2020; Tewari & Murphy, 2017; Varatharajah et al., 2018).

However, in many real-world applications such as robot control and image processing (Åström, 1965; Dougherty, 2020; Kaelbling et al., 1998; Kang et al., 2012; Lin et al., 2012; Nagrath, 2006), the full context may not be directly observable. leading to uncertainty and challenges in decision-making. Neglecting the imperfections in observations can compromise decisions, as seen in clinical scenarios where disregarding uncertainty in the medical profiles of septic patients can lead to adverse outcomes (Gottesman et al., 2019).

This dissertation investigates the application and enhancement of partially observable contextual bandit algorithms to address contextual uncertainty in personalized decision-making contexts.

The ubiquity of partially observable contexts arises in various domains, ranging from online advertising to healthcare interventions and recommender systems. In these settings, certain contextual features may be hidden or only partially observed, posing challenges for traditional contextual bandit algorithms that assume complete observability. Failure to account for partial observability can lead to suboptimal decisions and missed opportunities for learning from available data.

Decision-making policies for contextual bandits are extensively investigated in the literature, assuming that the context vectors are fully observed. One popular policy for contextual bandits is Upper-Confdent-Bounds (UCB) (Abbasi-Yadkori et al., 2011; Auer, 2002; Chu et al., 2011). The key idea behind UCB is to maintain a reward model, where the uncertainty about reward parameters is captured using confidence intervals. Thompson Sampling is another ubiquitous policy for contextual bandits (Agrawal & Goyal, 2013; Chapelle & Li, 2011; Modi & Tewari, 2020). Thompson sampling is a Bayesian approach that samples from a posterior distribution over the reward parameters, allowing for a more probabilistic exploration strategy. In recent findings, it has emerged that Greedy policies exhibit near optimality in specific contextual bandit scenarios such as the setup that a reward parameter is shared over all arms (Raghavan et al., 2023), as well as two-armed contextual bandits (Bastani et al., 2021).

The performance of policies are evaluated based on the measurement called regret, which is the expected decrease in cumulative reward of a policy as compared to the optimal policy. Upper bounds for regret of various algorithms for contextual bandits have been extensively studied in early literature about adversarial contexts, particularly focusing on high-probability problem-independent regret. First, square regret bounds of UCB-type algorithms have been established (Abbasi-Yadkori et al., 2011; Auer et al., 2002). Next, regret bounds that grow as the square root of time were established for adversarial contextual bandits (Abeille & Lazaric, 2017; Agrawal & Goyal, 2013; Russo & Van Roy, 2014).However, in the case of the stochastic context assumption, which represents a special case within adversarial contexts, the effectiveness of the aforementioned bandit policies remains uncertain. Our study addresses this uncertainty, leveraging the stochastic nature of contexts to derive tighter regret bounds.

1.2 Partially Observed Stochastic Contextual Bandits

In this dissertation, we study partially observable stochastic contextual bandits. The probabilistic structure of the problem under study, as time progresses, is as follows: At every time step, there are N available arms, each of which has an unobserved context denoted by $x_i(t)$ for arm i at time t. The context vectors are generated independently of the previous contexts and the other arms from a distribution. Moreover, the linear noisy transformed context of $x_i(t)$ is $y_i(t)$ generated based on

$$y_i(t) = Ax_i(t) + \xi_i(t), \tag{1.1}$$

where A is the sensing matrix capturing the relationship between $x_i(t)$ and $y_i(t)$ and $\xi_i(t)$ is an observation noise. The stochastic reward $r_i(t)$ of arm i is determined by the context vector and the unknown parameter as follows:

$$r_i(t) = f_r(x_i(t), i) + \varepsilon_i(t),$$

where f_r is a linear function of context $x_i(t)$ and $\varepsilon_i(t)$ is a reward noise. At each time t, the agent chooses an arm $a(t)$ given observations $\{y_i(t)\}_{i=1}^N$ and receives the reward $r_{a(t)}(t)$. Here, we consider two types of f_r: 1) $f_r(x_i(t), i) = x_i(t)^\top \mu_\star$, where the reward parameter $\mu_\star$ is *shared* over all arms; 2) $f_r(x_i(t), i) = x_i(t)^\top \mu_i$, where the parameter μ_i is the *arm-specific* reward parameter for the i-th arm.

It is worthwhile noting that contexts are not accessible for all policies including the optimal one. Accordingly, the optimal policy is the one maximizing the expected reward given the observation $y(t) =$

$\{y_i(t)\}_{i=1}^N$. The arm chosen by the optimal policy is the optimal arm, which is

$$a^\star(t) = \arg\max_i \mathbb{E}[r_i(t)|y(t)].$$

The primary objective of this dissertation is to evaluate the performance of policies that are designed to maximize the cumulative reward. To proceed, we consider a performance measure, regret, which is the expected decrease in cumulative reward caused by uncertainty as compared to the optimal policy. Regret of a policy is written as

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \mathbb{E}[r_{a^\star(t)}(t) - r_{a(t)}(t)|y(t)],$$

where $a(t)$ is the chosen arm at time t by the policy.

In this dissertation, we perform regret analysis under three different setups: 1) the sensing matrix A is known and invertible, in a shared parameter setup where contexts, reward noises, and observation noises have Gaussian distributions.; 2) Similar to the first setup, with the exception of a rectangular sensing matrix A; 3) the sensing matrix A is unknown and can be rectangular, in the arm-specific parameter setup where contexts, reward noises, and observation noises have sub-Gaussian distributions. Across these setups, assumptions become progressively more relaxed. In the first setup, we analyze the Thompson sampling algorithm, so-called the posterior sampling algorithm, which makes decisions as if samples from the (pseudo) posterior distribution are the truth. In the second setup, we consider the greedy algorithm taking action based on the current best estimates of reward for myopic reward without consideration of exploration. Lastly, in the third setup, we again employ the Thompson sampling algorithm, as the arm-specific parameter setup necessitates a delicate balance of exploration and exploitation.

The structure of this dissertation is outlined as follows: Chapter 2 presents an analysis indicating that the Thompson sampling algorithm exhibits a logarithmic upper bound for expected regret within the first setup. Following this, Chapter 3 illustrates how the Greedy algorithm possesses a high-probability upper bound for regret in the second setup. Additionally, Chapter 4 demonstrates that Thompson sampling yields a poly-logarithmic high-probability upper bound for regret in the third setup. Finally, Chapter 6 outlines avenues for future research.

Chapter 2

Analysis of Thompson Sampling for the Shared Parameter Setup

2.1 Introduction

Contextual Multi-Armed Bandits (CMAB) are canonical models in both theory and applications of Reinforcement Learning (RL). In this setting, there is a set of arms whose rewards depend on their multidimensional context vectors as well as the underlying parameter that reflects the weights of each context component. Thanks to their ability in modeling individual characteristics, CMAB models are widely used in different areas of automation and decision-making. For example, in personalized recommendation of news articles, CMAB models can raise the click rate by 12.5%, compared to context-free bandit algorithms (Li et al., 2010). In dynamic treatment of mice with skin tumors, adopting biological factors as contexts, leads to a 50% increase in life duration (Durand et al., 2018). CMAB can also provide a useful framework for sequential decision-making in precision health by incorporating contexts such as location, calendar busyness, and heart-rate (Tewari & Murphy, 2017).

The existing literature on bandit models for decision-making under uncertainty goes back at least to the seminal work of Lai and Robbins (Lai & Robbins, 1985) that introduces Upper Confidence Bound (UCB) algorithm. Broadly speaking, UCB prescribes acting based on optimism-based approximations of the unknown parameters, and is efficient in both discrete and continuous spaces (Abbasi-Yadkori et al., 2011; Faradonbeh et al., 2020c). Ensuing work establishes logarithmic regret bounds of UCB that hold uniformly over time (Auer et al., 2002). The sequence of papers focusing on CMAB models and theoretical performance guarantees of associated reinforcement learning policies continues by showing that the UCB algorithm appropriately addresses the exploitation-exploration trade-off (Auer, 2002), followed by a finer analysis that improves dependence on dimensions (Abbasi-Yadkori et al., 2011), and regret bounds for linear payoffs (Chu et al., 2011).

Another ubiquitous reinforcement learning policy that is usually faster than UCB, yet performs equally efficient, is Thompson Sampling (Chapelle & Li, 2011), (Russo & Van Roy, 2014). The main idea of Thompson Sampling is to select actions based on samples drawn from a posterior distribution over unknown parameters (Thompson, 1933). The posterior is updated by the observed rewards, and balances exploring for better options and more accurate learning, versus exploiting the available information to maximize earning. Theoretical analyses start with a regret bound for multi-armed bandits (Agrawal & Goyal, 2012), and continue to CMAB counterparts (Agrawal & Goyal, 2013). Moreover, Thompson Sampling has favorable performances in continuous spaces (Faradonbeh et al., 2020b) and large-scale problems (T. Hu et al., 2019). Other variants and more discussions can be found in a recent tutorial by Russo et al. (Russo et al., 2017).

Further adaptive policies for CMAB models include greedy-type algorithms that are efficient if the context distribution satisfies some diversity conditions (Raghavan et al., 2020), (Bastani et al., 2021). Moreover, the existing literature consists of studies on non-linear reward functions (of the contexts) under technical assumptions such as Lipschitz continuity. That includes, near-optimal regret bounds obtained by using partitioning techniques on the context and action space (Slivkins, 2011), and utilizing non-parametric regression techniques for unknown non-linear reward functions (Y. Hu et al., 2020). Finally, multi-agent settings and those with latent structure of users' reward functions are studied, as well as approaches aiming to provide personalized recommendations for new users (Hong et al., 2020; Maillard & Mannor, 2014; Zhou & Brunskill, 2016).

In many applications, context vectors are observed in a partial, transformed, or noisy manner. For example, it includes situations that inquiring the entire feature vector is too expensive, context variables correspond to physically distant stations, data is provided by a network of sensors, or privacy considerations restrict perfect context observations (Bensoussan, 2004). For restricted contexts, reinforcement learning algorithms together with combinatorial search algorithms demonstrate competitive empirical performance (Bouneffouf et al., 2017). In the presence of known side-information about unobserved parts of the contexts, ridge regression methods together with projections and UCB algorithms lead to improved efficiency (Tennenholtz et al., 2021). Another ubiquitous setting for studying control policies under partial observations is the state space model (Durbin & Koopman, 2012; Nagrath, 2006; Roesser, 1975). In this setting, unobserved states are estimated based on output observations using methods such as Kalman filter (Kalman, 1960; Stratonovich, 1959, 1960), and captures important applications such as robot navigation (Howard et al., 2008; Surmann et al., 2020).

When the number of control actions is finite, CMAB models are widely used for data-driven control. However, unlike the aforementioned frameworks with partial observations, proper designs and comprehensive analyses of decision-making algorithms in contextual bandits with imperfect observations are not currently available. Accordingly, we study (a slightly modified) Thompson Sampling reinforcement learning algorithm for CMAB models with partially observable contexts. Note that because contexts are the main factors in determining the optimal arm, additional learning procedures are needed to estimate unobserved contexts, and so modifications in the algorithm are inevitable.

Under minimal assumptions, we establish theoretical performance guarantees showing that the regret (i.e., the cumulative decrease in rewards due to uncertainty) scales as the logarithm of time, the logarithm of the number of arms, and the dimension. We present an effective method for estimating unobserved contexts based on transformed noisy outputs, and use them to form the posterior belief about the unknown parameter, which determines the optimal candidate arm at every time step. Furthermore, we specify the rates at which Thompson Sampling learns the unknown parameter. To obtain the results, certain technical tools from the theory of martingales are leveraged, and novel methods are developed for precisely specifying the behavior of the posterior distribution and its effect on the efficiency of the algorithm.

The remainder of this chapter is organized as follows. In Section 2.2, we formulate the problem and discuss preliminary results. In Section 2.3, we present the reinforcement learning algorithm that utilizes Thompson Sampling for partially observable CMAB models. Lastly, a theoretical analysis of the algorithm is provided in Section 2.4, followed by numerical illustrations in Section 2.5.

The following notation will be used throughout this chapter. For a matrix $A \in \mathbb{C}^{p \times q}$, $A^\top$ denotes its transpose, and the trace of A is denoted by $\mathbf{tr}(A)$. For a vector $v \in \mathbb{C}^d$, we use the Euclidean norm $||v|| = (\sum_{i=1}^{d} |v_i|^2)^{1/2}$, and for matrices, we use the operator norm; $||A|| = \sup_{||v||=1} ||Av||$. Further, $\overrightarrow{u} = u/||u||$ is the unit vector indicating the direction of u, and $C(A)$ denotes the column space of the matrix A. Finally, the sigma-field generated by random vectors $\{X_1, ..., X_n\}$ is denoted by $\sigma(X_1, \dots, X_n)$.

2.2 Problem Statement

We consider the following partially observed contextual multi-armed bandit (POCMAB) problem. Suppose that a slot machine with N arms is given, and each arm $i \in \{1, \cdots, N\}$ has the unobserved d-dimensional context $x_i(t)$, which is generated independently from $N(0_d, \Sigma_x)$, where Σ_x is the covariance matrix of $x_i(t)$. These contexts determine the rewards: At each time step $t = 1, 2, \dots$, the arm $a(t)$ is selected, which generates the reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_\star + \varepsilon_{a(t)}(t)$, where $x_{a(t)}(t)$ is the context of the selected arm, $\mu_\star$ is the unknown true parameter, and $\varepsilon_{a(t)}(t)$ is the reward observation noise with the distribution $N(0,\sigma^2)$. The observations at time t consist of the output vectors $\{y_i(t)\}_{1\le i\le N}$, generated according to $y_i(t) = Ax_i(t) + \xi_i(t)$, where $\xi_i(t)$ is the output observation noise that has the distribution $N(0_d, \Sigma_\xi)$ and Σ_ξ is the covariance matrix of $y_i(t)$ given $x_i(t)$. Further, the matrix $A \in \mathbb{R}^{d\times d}$ captures the relationship between the output and the context. For ease of presentation, we assume that A is a known non-singular square matrix.

The goal is to design a reinforcement learning policy to select an arm at every time step, such that the expected reward is maximized, based on the information available at the time. That is, at time t, the goal is to find the optimal arm $a^\star(t) = \mathrm{argmax}_{1\le i\le N}\,\mathbb{E}[r_i(t)|y_i(t)]$. The data available at time t, based on which we want to select $a^\star(t)$, consists of the outputs $\mathbf{y_t} = \{y_i(\tau)\}_{1\le i\le N,\,1\le\tau\le t}$, the rewards of the arms selected so far $\mathbf{r_{t-1}} = \{r_{a(\tau)}(\tau)\}_{1\le\tau\le t-1}$, and the previously selected arms $\mathbf{a_{t-1}} = \{a(\tau)\}_{1\le\tau\le t-1}$. Note that since the context vectors $x_i(t)$ are not observed, the optimal arm $a^\star(t)$ must be chosen according to a context estimate $\widehat{x}_i(t)$, based on the observations $\{y_i(t) : i = 1,\dots,N\}$. It is easy to see that it suffices to select

$$a^\star(t) = \underset{1\le i\le N}{\arg\max}\, \widehat{x}_i(t)^\top \mu_\star, \tag{2.1}$$

where $\widehat{x}_i(t)$ is the conditional expectation of $x_i(t)$ given $y_i(t)$ (the output observation of the ith arm at time t).

Due to uncertainty about the true parameter $\mu_\star$, a reinforcement learning algorithm incurs a performance degradation compared to the optimal policy that knows the true parameter $\mu_\star$, and selects the optimal arms $\{a^\star(t)\}_{t\ge1,}$ at every time step. Accordingly, the performance of reinforcement learning algorithms is commonly assessed by the cumulative decrease in rewards, which is called regret, and is defined as

$$\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_{a^\star(t)}(t) - r_{a(t)}(t)\right]. \tag{2.2}$$

Above $a(t)$ is the arm selected by the reinforcement learning policy under study. In the sequel, we present the Thompson Sampling algorithm for POCMAB models (Algorithm 1), and establish a regret bound for that based on d, N, T.

2.3 Reinforcement Learning Algorithm

Now, we explain a reinforcement learning algorithm that leverages Thompson Sampling to learn to maximize the reward in the POCMAB problem above, based on the output data at the time. At a high level, the main idea of the algorithm is that we maximize the expected value of the reward $r_i(t)$ given the output $y_i(t)$, because the contexts $\{x_i(t)\}_{1\leq i\leq N}$ are not observed. To do so, using conditional expectation with respect to the observations, the regret in (2.2) can be written as

$$\text{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[r_{a^\star(t)}(t) - r_{a(t)}(t) \middle| \{y_i(t)\}_{1\leq i\leq N}\right]\right]. \tag{2.3}$$

Note that depending on the problem understudy, technically different definitions of regret are considered in the literature (Bubeck & Cesa-Bianchi, 2012). The objective of the proposed reinforcement learning algorithm is to choose the arm $a(t)$ that minimizes the conditional expected reward gap given the observations $\{y_i(t)\}_{1\leq i\leq N,}$;

$$\mathbb{E}\left[\sum_{t=1}^{T} r_{a^\star(t)}(t) - r_{a(t)}(t) \middle| \{y_i(t)\}_{1\leq i\leq N}\right], \tag{2.4}$$

at each time t, and thereby aims to minimize the regret in (2.2).

Technically, to find $a(t)$ minimizing the conditional expected reward gap in (2.4), we use the conditional distribution of the reward $r_i(t)$ given $y_i(t)$, which is derived in Appendix. The conditional distribution of $r_i(t)$ given $y_i(t)$ is

$$N\left((Dy_i(t))^\top \mu_\star,\ \mu_\star^\top(A^\top\Sigma_\xi^{-1}A+\Sigma_x^{-1})^{-1}\mu_\star+\sigma^2\right), \tag{2.5}$$

where $D=(A^\top\Sigma_\xi^{-1}A+\Sigma_x^{-1})^{-1}A^\top\Sigma_\xi^{-1}$ is a matrix reflecting the average effect of $y_i(t)$ on $r_i(t)$. Next, let

$$\widehat{x}_i(t) \;=\; (A^\top\Sigma_\xi^{-1}A+\Sigma_x^{-1})^{-1}A^\top\Sigma_\xi^{-1}y_i(t)=Dy_i(t). \tag{2.6}$$

In fact, $\widehat{x}_i(t)$ is the conditional expectation $\mathbb{E}[x_i(t)|y_i(t)]$. Putting (2.5) and (2.6) together, the conditional expected reward gap in (2.4) can be written as

$$\begin{aligned}
\mathbb{E}\left[r_{a^\star(t)}(t)-r_{a(t)}(t)\middle|\,\{y_i(t)\}_{1\le i\le N}\right] &= \mathbb{E}\left[\mathbb{E}\left[r_{a^\star(t)}(t)-r_{a(t)}(t)|x_i(t)\right]\middle|\,\{y_i(t)\}_{1\le i\le N}\right]\\
&= \mathbb{E}\left[(x_{a^\star(t)}(t)-x_{a(t)}(t))^\top\mu_\star\middle|\,\{y_i(t)\}_{1\le i\le N}\right]\\
&= (\widehat{x}_{a^\star(t)}(t)-\widehat{x}_{a(t)}(t))^\top\mu_\star. \qquad (2.7)
\end{aligned}$$

Thus, a policy is designed to choose the arm maximizing $\widehat{x}_i(t)^\top\mu_\star$. To ensure that the algorithm performs enough exploration, we use the sample $\widetilde{\mu}(t)$ from the posterior distribution

$$N(\widehat{\mu}(t),B(t)^{-1}), \tag{2.8}$$

where the posterior mean $\widehat{\mu}(t)$ and the inverse of the covariance matrix $B(t)$ are as follows:

$$\begin{aligned}
B(t) &= \Sigma^{-1}+\sum_{\tau=1}^{t-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top, \qquad (2.9)\\
\widehat{\mu}(t) &= B(t)^{-1}\sum_{\tau=1}^{t-1}\widehat{x}_{a(\tau)}(\tau)r_{a(\tau)}(\tau). \qquad (2.10)
\end{aligned}$$

Based on the estimates of the contexts and the sample $\widetilde{\mu}(t)$, we select $a(t)$ such that

$$a(t) = \arg\max_{1 \leq i \leq N} \widehat{x}_i(t)^\top \widetilde{\mu}(t). \tag{2.11}$$

Then, we observe the reward $r_{a(t)}(t)$ of the arm $a(t)$, and update the posterior according to

$$B(t+1) = B(t) + \widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top, \tag{2.12}$$

$$\widehat{\mu}(t+1) = B(t+1)^{-1}(B(t)\widehat{\mu}(t) + \widehat{x}_{a(t)}(t) r_{a(t)}(t)). \tag{2.13}$$

The initial values are $\widehat{\mu}(1) = 0_d$ and $B(1) = \Sigma^{-1}$, where Σ is an arbitrary symmetric positive definite matrix.

Algorithm 1 : Thomson Sampling RL policy for POCMAB

```
Set $B(1) = \Sigma^{-1}$, $\widehat{\mu}(1) = \mathbf{0}_d$
for $t = 1, 2, \ldots,$ do
    for $i = 1, \ldots, N$ do
        Estimate context by $\widehat{x}_i(t)$ in (2.6)
    end for
    Sample $\widetilde{\mu}(t)$ from $N(\widehat{\mu}(t), B(t)^{-1})$
    Select arm $a(t) = \arg\max_{1 \leq i \leq N} \widehat{x}_i(t)^\top \widetilde{\mu}(t)$
    Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_\star + \epsilon_{r_{a(t)}}(t)$
    Update $B(t+1)$ and $\widehat{\mu}(t+1)$ by (2.12) and (2.13)
end for
```

The pseudo-code of Thompson sampling for POCMAB is provided in Algorithm 1. At every time and for each arm, Algorithm 1 calculates the context estimate $\widehat{x}_i(t)$ according to (2.6). Then, it chooses the arm $a(t)$ by (2.11), based on $\widetilde{\mu}(t)$ generated from the posterior in (2.8), and updates $\widehat{\mu}(t)$ and $B(t)$ according to (2.12) and (2.13). So, Algorithm 1 selects the arm maximizing $\widehat{x}_i(t)^\top \widetilde{\mu}(t)$ as a reliable estimate of the unknown expected reward at time t.

2.4 Analysis of Algorithm 1

In this section, we provide theoretical performance guarantees for the reinforcement learning policy in Algorithm 1, establishing that it efficiently learns optimal decisions from the data of partial observations. In the first result we show that Algorithm 1 learns the unknown parameter $\mu_\star$, fast and accurately. Then, in Theorem 2, we provide regret analysis, indicating that the regret of Algorithm 1 scales logarithmically with both the number of arms N, as well as the time of interaction with the environment T, and scales linearly with the dimension d.

The following result shows that $\widehat{\mu}(t)$ is a consistent estimator and its covariance matrix shrinks proportionally to the inverse of the time of interacting with the environment in Algorithm 1. Therefore, Theorem 1 provides sample efficiency for the Thompson Sampling reinforcement learning policy for POCMAB in Algorithm 1.

Theorem 1. *In Algorithm 1, let $\widehat{\mu}(t)$ be the parameter estimate at time t, defined by* (2.13). *Then, we have* $\lim_{t\to\infty} \widehat{\mu}(t) = \mu_\star$, *as well as* $\mathrm{Cov}\left(\widehat{\mu}(t)\right) = O(t^{-1})$.

Proof. First, for the prior $N(0_d, \Sigma)$ of $\mu_\star$, (2.9) and (2.10) imply that

$$\mathbb{E}\left[\widehat{\mu}(t)\right] = \mathbb{E}\left[B(t)^{-1} \sum_{\tau=1}^{t-1} \widehat{x}_{a(\tau)}(\tau) \widehat{x}_{a(\tau)}(\tau)^\top \mu_\star \right] = (I_d - \mathbb{E}[B(t)^{-1}]\Sigma^{-1})\mu_\star. \tag{2.14}$$

Further, let $\mathscr{F}_t = \sigma\left\{ \{y_i(\tau)\}_{1\le i\le N,\, 1\le\tau\le t}, \{a(\tau)\}_{1\le\tau\le t} \right\}$ be the sigma-field generated by the sequence of all observations and actions by time t. Given the sigma-field $\mathscr{F}_{t-1}$, we have

$$\begin{align}
\mathbb{E}\left[\widehat{\mu}(t) | \mathscr{F}_{t-1}\right] &= \mathbb{E}\left[B(t)^{-1} \sum_{\tau=1}^{t-1} \widehat{x}_{a(\tau)}(\tau) \widehat{x}_{a(\tau)}(\tau)^\top \mu_\star \middle| \mathscr{F}_{t-1} \right] = (I_d - B(t)^{-1}\Sigma^{-1})\mu_\star \tag{2.15} \\
\mathrm{Cov}\left(\widehat{\mu}(t) | \mathscr{F}_{t-1}\right) &= B(t)^{-1} \left(\sum_{\tau=1}^{t} \mathrm{Var}\left(r_{a(\tau)}(\tau) | \mathscr{F}_{t-1} \right) \widehat{x}_{a(\tau)}(\tau) \widehat{x}_{a(\tau)}(\tau)^\top \right) B(t)^{-1} \notag \\
&= B(t)^{-1}(B(t) - \Sigma^{-1}) B(t)^{-1} \sigma_{ry}^2, \tag{2.16}
\end{align}$$

where $\sigma_{ry}^2 = \mathrm{Var}(r_i(t)|y_i(t)) = \mu_\star^\top (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} \mu_\star + \sigma^2$ is derived in Appendix. Using (2.14), (2.15) and (2.16), we obtain

$$\begin{aligned}\mathrm{Cov}(\widehat{\mu}(t)) &= \mathrm{Cov}(\mathbb{E}[\widehat{\mu}(t)|\mathscr{F}_{t-1}]) + \mathbb{E}[\mathrm{Cov}(\widehat{\mu}(t)|\mathscr{F}_{t-1})] \\ &= \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}B(t)^{-1}\right] - \mathbb{E}\left[B(t)^{-1}\right]\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}\mathbb{E}\left[B(t)^{-1}\right] \\ &+ \mathbb{E}\left[B(t)^{-1}\right]\sigma_{ry}^2 - \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}B(t)^{-1}\right]\sigma_{ry}^2. \end{aligned} \tag{2.17}$$

Next, we show that $\lim\limits_{t\to\infty} t^{-1}B(t)$ is a positive definite matrix. It implies that $\mathrm{Cov}(\widehat{\mu}(t)) = O(t^{-1})$, since the other terms in (2.17) are $O(t^{-2})$, except $\mathbb{E}\left[B(t)^{-1}\right]\sigma_{ry}^2$. For this purpose, let $S = (D\Sigma_\xi D^\top)^{1/2}$, and define

$$\begin{aligned} X_t &= \sum_{\tau=1}^{t}\left(S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1} - \mathbb{E}[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}|\mathscr{F}_{\tau-1}]\right), \\ Y_t &= \sum_{\tau=1}^{t}\tau^{-1}(X_\tau - X_{\tau-1}). \end{aligned}$$

Then, X_t and Y_t are matrix valued martingales adapted to the filtration $\{\mathscr{F}_t\}_{t\geq 1}$. To see that, observe that the following two equivalences

$$\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{t-1}\right] = S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}, \tag{2.18}$$

$$\mathbb{E}\left[\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{\tau-1}\right]\middle|\mathscr{F}_{t-1}\right] = \mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{\tau-1}\right]. \tag{2.19}$$

lead to

$$\begin{aligned}
&\mathbb{E}\left[X_t|\mathscr{F}_{t-1}\right] \\
&=\sum_{\tau=1}^{t}\left(\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{t-1}\right]-\mathbb{E}\left[\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{\tau-1}\right]\middle|\mathscr{F}_{t-1}\right]\right) \\
&=\sum_{\tau=1}^{t}\left(\mathbb{E}[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}|\mathscr{F}_{t-1}]-\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{\tau-1}\right]\right)=X_{t-1},
\end{aligned} \tag{2.20}$$

for $\tau < t$ and

$$\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{t-1}\right]-\mathbb{E}\left[\mathbb{E}\left[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}\middle|\mathscr{F}_{\tau-1}\right]\middle|\mathscr{F}_{t-1}\right]=0_{d\times d},$$

for $\tau = t$. Further, since $\mathbb{E}[X_\tau|\mathscr{F}_{t-1}] = X_\tau$, for $\tau < t$, and we have $\mathbb{E}[X_t|\mathscr{F}_{t-1}] - \mathbb{E}[X_{t-1}|\mathscr{F}_{t-1}] = 0_{d\times d}$, it holds that

$$\mathbb{E}\left[Y_t|\mathscr{F}_{t-1}\right]=\sum_{\tau=1}^{t}\tau^{-1}\left(\mathbb{E}[X_\tau|\mathscr{F}_{t-1}]-\mathbb{E}[X_{\tau-1}|\mathscr{F}_{t-1}]\right)=\sum_{\tau=1}^{t-1}\tau^{-1}\left(X_\tau-X_{\tau-1}\right)=Y_{t-1}.$$

Now, define the martingale difference sequence $Z_t = X_t - X_{t-1}$, and let X_{tij} be the ijth entry of X_t, to get

$$\mathbb{E}\left[X_{tij}^2\right]=\mathbb{E}\left[\left(\sum_{\tau=1}^{t}Z_{\tau ij}\right)^2\right]=\sum_{\tau=1}^{t}\mathbb{E}\left[Z_{\tau ij}^2\right]+2\sum_{\tau_1<\tau_2}\mathbb{E}\left[Z_{\tau_1 ij}Z_{\tau_2 ij}\right]=\sum_{\tau=1}^{t}\mathbb{E}\left[Z_{\tau ij}^2\right],$$

using the fact that $\mathbb{E}\left[Z_{\tau_1 ij}Z_{\tau_2 ij}\right] = \mathbb{E}\left[Z_{\tau_1 ij}\mathbb{E}\left[Z_{\tau_2 ij}\middle|\mathscr{F}_{\tau_2-1}\right]\right] = 0$ for all $\tau_1 < \tau_2$.

Using the above, we show that Y_t is a square-integrable martingale. To that end, since $\{X_t - X_{t-1} : t \geq 1\}$ is a martingale difference sequence, we have

$$\mathbb{E}\left[Y_{tij}^2\right] = \sum_{\tau=1}^{t} \tau^{-2}\left(\mathbb{E}\left[X_{\tau ij}^2\right] - \mathbb{E}\left[X_{(\tau-1)ij}^2\right]\right) = \sum_{\tau=1}^{t} \tau^{-2}\mathbb{E}\left[Z_{\tau ij}^2\right],$$

where $X_0 = 0_{d\times d}$, and Y_{tij} is the ijth entry of Y_t. Since $\mathbb{E}\left[Z_{\tau ij}^2\right] \leq \mathbb{E}\left[||S^{-1}\widehat{x}_{a(t)}(t)||^4\right]$, for all τ, i, and j, the expectation $\mathbb{E}[Y_{tij}^2]$ is finite. So, by Martingale Convergence Theorem (Doob, 1953), the martingale Y_t converges almost surely to a limit Y, such that $\mathbb{E}[|Y|] < \infty$. It is straightforward to see that $t^{-1}X_t = Y_t - t^{-1}\sum_{\tau=1}^{t} Y_\tau$. Thus, since $\lim\limits_{t\to\infty} Y_t = Y$, the average of the sequence converges to the same limit as well; $\lim\limits_{t\to\infty} t^{-1}\sum_{\tau=1}^{t} Y_\tau = Y$. Thus, $t^{-1}X_t$ converges to $0_{d\times d}$. To show that $\lim_{t\to\infty} t^{-1}B(t)$ is a positive definite matrix, decompose X_t as follows:

$$X_t = S^{-1}(B(t) - \Sigma^{-1})S^{-1} - \sum_{\tau=1}^{t} \mathbb{E}[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}|\mathscr{F}_{\tau-1}].$$

Since $\lim_{t\to\infty} t^{-1}X_t = 0_{d\times d}$, we have

$$\lim_{t\to\infty} t^{-1}S^{-1}B(t)S^{-1} = \lim_{t\to\infty} t^{-1}\sum_{\tau=1}^{t} \mathbb{E}[S^{-1}\widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}(\tau)^\top S^{-1}|\mathscr{F}_{\tau-1}]. \tag{2.21}$$

To proceed, we express the following result about the matrix $M = \lim\limits_{t\to\infty}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\mathscr{F}_{t-1}]$, for which the proof is deferred to Appendix. Now, by (2.21), we have

$$\lim_{t\to\infty} t^{-1}S^{-1}B(t)S^{-1} = M,$$

which according to Lemma 1 is a positive definite matrix. Finally, the latter result, together with (2.17), implies that

$$\lim_{t\to\infty} t\mathrm{Cov}\left(\widehat{\mu}(t)\right) = \lim_{t\to\infty} t\mathbb{E}[B(t)^{-1}]\sigma_{ry}^2 = SM^{-1}S\sigma_{ry}^2,$$

which is the desired result. □

Lemma 1. *The matrix* $M = \lim_{t\to\infty} \mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\mathscr{F}_{t-1}]$ *is deterministic and positive definite.*

Theorem 1 establishes the square-root consistency of the parameter estimate $\widehat{\mu}(t)$, indicating the Algorithm 1 effectively learns the unknown true parameter $\mu_\star$. Here, the inverse of $\mathrm{Cov}(\widehat{\mu}(t))$ grows linearly with time t, only when the smallest eigenvalue of $A^\top A$ is non-zero. If A is singular, the maximum eigenvalue of $\mathrm{Cov}(\widehat{\mu}(t))$ does not decrease as t becomes larger. This also affects the consistency of learning the unknown parameter. A similar result holds for the samples $\widetilde{\mu}(t)$, as elaborated in the following corollary, for which the details are provided in Appendix.

Corollary 1. *For the samples* $\{\widetilde{\mu}(t)\}_{t\geq 1}$ *in Algorithm 1, we have*

$$\lim_{t\to\infty} \widetilde{\mu}(t) = \mu_\star, \qquad \mathrm{Cov}(\widetilde{\mu}(t)) = O(t^{-1}).$$

The following result provides a regret bound, and states that Algorithm 1 is able to efficiently learn optimal arms in POCMAB.

Theorem 2. *For the regret of Algorithm 1, we have*

$$\mathrm{Regret}(T) = O\left(d\sqrt{\log N}\log T\right).$$

Before proceeding towards the proof of Theorem 2, we discuss the intuition it provides. Since the regret at time t grows due to the difference between $\mu_\star$ and $\widetilde{\mu}(t)$, the growth rate of regret depends on the shrinkage rate of $||\mu_\star - \widetilde{\mu}(t)||^2$. According to Corollary 1, the shrinkage rate is $O(dt^{-1})$. Thus, aggregating the errors for the time period $1 \leq t \leq T$, the scaling with respect to T becomes logarithmic (see (2.31)), while the scaling with d is linear. On the other hand, the regret scales logarithmically *slow* with the number of arms N, because N has two opposite effects. On the one hand, since N is the total number of options, the probability of choosing a sub-optimal arms increases as N grows. On the other hand, the difference between the reward of the optimal arm and that of the chosen arm becomes smaller as N grows. The consequences of the two effects compensate each other, leading to the slow growth of the regret with respect to N. As mentioned, the suggested regret bound works for non-singular A. If A is singular and $\mu_\star \in C\left(A^\top\right)^\perp$, the regret grows linearly with T.

Proof. First, for the regret of Algorithm 1, it holds that $\text{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(\widehat{x}_{a^\star(t)} - \widehat{x}_{a(t)})^\top \mu_\star\right]$, according to (2.3) and (2.7). To proceed, we show that for an arbitrary $\mu_\star \in \mathbb{R}^d$, it holds that

$$\mathbb{E}\left[\underset{\widehat{x}_i(t), 1\leq i\leq N}{\operatorname{argmax}} \left\{\widehat{x}_i(t)^\top \mu_\star\right\}\right] = c_N \overrightarrow{S\mu_\star},$$

where the constant

$$c_N = \mathbb{E}\left[\max_{1\leq i\leq N}\{V_i : V_i \sim N(0,1)\}\right] \tag{2.22}$$

captures the magnitude, and the unit vector $\overrightarrow{S\mu_\star}$ indicates the direction of the expected value of the vector $\widehat{x}_i(t)$ that achieves the maximum value inside the expectation.

To show the above result, define

$$Z(\mu, N) = \underset{Z_i, 1\leq i\leq N}{\operatorname{argmax}} \left\{Z_i^\top \mu\right\}, \tag{2.23}$$

where Z_i are independent standard d-dimensional normally distributed random vectors. The vector Z_i can be decomposed as $Z_i = P_\mu Z_i + (I_d - P_\mu) Z_i$, where P_μ is the projection matrix onto

$C(\mu)$, which is the 1-dimensional subspace of the vectors inline with μ. Then, we have $Z(\mu, N) = \underset{Z_i(t), 1 \leq i \leq N}{\operatorname{argmax}} \left\{ (P_\mu Z_i(t))^\top \mu \right\}$, because $P_\mu \mu = \mu$. This implies that only the first term, $P_\mu Z_i$, affects the result of $\underset{Z_i, 1 \leq i \leq N}{\operatorname{argmax}} \left\{ Z_i^\top \mu \right\}$. This means that $Z(\mu, N)$ has the same distribution as $P_\mu Z(\mu, N) + (I_d - P_\mu) Z_i$, which means

$$Z(\mu, N) \overset{d}{=} P_\mu Z(\mu, N) + (I_d - P_\mu) Z_i, \tag{2.24}$$

where $\overset{d}{=}$ is used to denote equality of the probability distributions. Thus, since projection on a subspace is a linear operator, it interchanges with expectation, and so we have

$$\mathbb{E}[Z(\mu, N)] = \mathbb{E}\left[P_\mu Z(\mu, N) + (I_d - P_\mu) Z_i\right] = P_\mu \mathbb{E}[Z(\mu, N)] \in C(\mu). \tag{2.25}$$

Next, we claim that $\mathbb{E}[Z(\mu, N)] = c_N \overrightarrow{\mu}$, where c_N is defined in (2.22), for which it is known that (Cramér, 2016):

$$c_N = O\left(\sqrt{\log N}\right). \tag{2.26}$$

Because $Z_i^\top \overrightarrow{\mu}$ has the standard normal distribution $N(0,1)$, according to (2.22), we have $\mathbb{E}\left[\max_{1 \leq i \leq N} \left\{ Z_i^\top \overrightarrow{\mu} \right\}\right] = c_N$. Based on the definition in (2.23), it holds that $Z(\mu, N)^\top \overrightarrow{\mu} = \max_{1 \leq i \leq N} \left\{ Z_i^\top \overrightarrow{\mu} \right\}$. Moreover, because $\mathbb{E}[Z(\mu, N)] \in C(\mu)$ by (2.25), we have $c_N = \mathbb{E}[Z(\mu, N)]^\top \overrightarrow{\mu} = ||\mathbb{E}[Z(\mu, N)]|| \, ||\overrightarrow{\mu}|| = ||\mathbb{E}[Z(\mu, N)]||$. Putting the above together, we obtain

$$\mathbb{E}[Z(\mu, N)] = c_N \overrightarrow{\mu}. \tag{2.27}$$

Next, we apply the result in (2.27) to $\widehat{x}_{a^{\star}(t)}(t)$ and $\widehat{x}_{a(t)}(t)$. The definition of $Z(\mu, N)$ in (2.23) implies that $S^{-1}\widehat{x}_{a^{\star}(t)}(t)$ can be written as

$$\underset{S^{-1}\widehat{x}_i(t), 1\leq i\leq N}{\operatorname{argmax}} \left\{(S^{-1}\widehat{x}_i(t))^{\top} S\mu_{\star}\right\} = Z(S\mu_{\star}, N).$$

Similarly, it holds that

$$S^{-1}\widehat{x}_{a(t)}(t) = \underset{S^{-1}\widehat{x}_i, 1\leq i\leq N}{\operatorname{argmax}} \left\{(S^{-1}\widehat{x}_i(t))^{\top} S\widetilde{\mu}(t)\right\} = Z(S\widetilde{\mu}(t), N).$$

Using (2.27), we can find the expected values as follows:

$$\begin{aligned} \mathbb{E}[S^{-1}\widehat{x}_{a^{\star}(t)}] &= c_N \overrightarrow{S\mu_{\star}}, \\ \mathbb{E}[S^{-1}\widehat{x}_{a(t)}|\widetilde{\mu}(t)] &= c_N \overrightarrow{S\widetilde{\mu}(t)}. \end{aligned}$$

Using the above equations, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T}\left(S^{-1}\widehat{x}_{a^{\star}(t)}(t) - S^{-1}\widehat{x}_{a(t)}(t)\right)^{\top} S\mu_{\star}\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{T}\left(S^{-1}\widehat{x}_{a^{\star}(t)}(t) - S^{-1}\widehat{x}_{a(t)}(t)\right)^{\top} S\mu_{\star}\middle|\widetilde{\mu}(t)\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\left(c_N\overrightarrow{S\mu_{\star}} - c_N\overrightarrow{S\widetilde{\mu}(t)}\right)^{\top} S\mu_{\star}\right] \qquad (2.28) \end{aligned}$$

for the expected gap.

Now, let θ_t denote the angle between $S\mu_{\star}$ and $S\widetilde{\mu}(t)$, defined as

$$\theta_t = \cos^{-1}\frac{< S\mu_{\star}, S\widetilde{\mu}(t) >}{||S\mu_{\star}||\ ||S\widetilde{\mu}(t)||} \in [0, \pi]. \qquad (2.29)$$

Since the vectors $\overrightarrow{S\mu_\star}$ and $\overrightarrow{S\widetilde{\mu}(t)}$ are of the same length, the angle between $\overrightarrow{S\mu_\star} - \overrightarrow{S\widetilde{\mu}(t)}$ and $\overrightarrow{S\mu_\star}$ is $(\pi - \theta_t)/2$, which leads to $\left|\left|\overrightarrow{S\mu_\star} - \overrightarrow{S\widetilde{\mu}(t)}\right|\right| = 2\sin(\theta_t/2)$. Thus, we get

$$\begin{aligned}\left(\overrightarrow{S\mu_\star} - \overrightarrow{S\widetilde{\mu}(t)}\right)^\top S\mu_\star &= ||S\mu_\star|| \left|\left|\overrightarrow{S\mu_\star} - \overrightarrow{S\widetilde{\mu}(t)}\right|\right| \cos\left(\frac{\pi - \theta_t}{2}\right) \\ &= 2||S\mu_\star|| \sin\left(\frac{\theta_t}{2}\right) \cos\left(\frac{\pi - \theta_t}{2}\right) \\ &= 2||S\mu_\star|| \sin^2\left(\frac{\theta_t}{2}\right) = 2||S\mu_\star||(1 - \cos\theta_t).\end{aligned}$$

On the other hand, using (2.29), we obtain

$$1 - \cos\theta_t = \frac{||S\mu_\star - S\widetilde{\mu}(t)||^2 - (||S\mu_\star|| - ||S\widetilde{\mu}(t)||)^2}{2||S\mu_\star||\,||S\widetilde{\mu}(t)||} \leq \frac{||S\mu_\star - S\widetilde{\mu}(t)||^2}{2||S\mu_\star||\,||S\widetilde{\mu}(t)||}.$$

To proceed, define $\eta(t) = S\widetilde{\mu}(t) - S\mu_\star + S\mathbb{E}[B(t)^{-1}]\Sigma^{-1}\mu_\star$, and note that $\mathbb{E}[\eta(t)\eta(t)^T] = S\mathrm{Cov}(\widetilde{\mu}(t))S$. So, it holds that

$$\mathbb{E}[1 - \cos\theta_t] \leq \mathbb{E}\left[\frac{||\eta(t) - S\mathbb{E}[B(t)^{-1}]\Sigma^{-1}\mu_\star||^2}{2||S\mu_\star||\,||S\widetilde{\mu}(t)||}\right] \leq \mathbb{E}\left[\frac{||\eta(t)||^2 + ||S\mathbb{E}[B(t)^{-1}]\Sigma^{-1}\mu_\star||^2}{||S\mu_\star||\,||S\widetilde{\mu}(t)||}\right].$$

By Corollary 1, we have

$$\mathbb{E}[||\eta(t)||^2] = \mathbf{tr}\left(\mathbb{E}[\eta(t)\eta(t)^T]\right) = \mathbf{tr}(S\mathrm{Cov}(\widetilde{\mu}(t))S) = O(dt^{-1}). \tag{2.30}$$

Accordingly, we get

$$\mathbb{E}\left[\frac{||\eta(t)||^2 + ||S\mathbb{E}[B(t)^{-1}]\Sigma^{-1}\mu_\star||^2}{||S\mu_\star||\,||S\widetilde{\mu}(t)||}\right] = O(dt^{-1}),$$

because the expected value of the numerator is $O(t^{-1})$ by (2.30) and Theorem 1, while the denominator converges to $||S\mu_\star||^2$ as $t \to \infty$, by Corollary 1. Thus, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\frac{||\eta(t)||^2 + ||S\mathbb{E}[B(t)^{-1}]\Sigma^{-1}\mu_\star||^2}{||S\mu_\star||\,||S\widetilde{\mu}(t)||}\right] = O(d\log T). \tag{2.31}$$

Putting the latter result together with (2.26), it yields to the desired result, since c_N depends only on N, and $||S\mu_\star||$ is a constant:

$$\text{Regret}(T) = \sum_{t=1}^{T} c_N \mathbb{E}[2||S\mu_\star||(1-\cos\theta_t)] = O(d\sqrt{\log N}\log T).$$

□

2.5 Numerical Illustrations

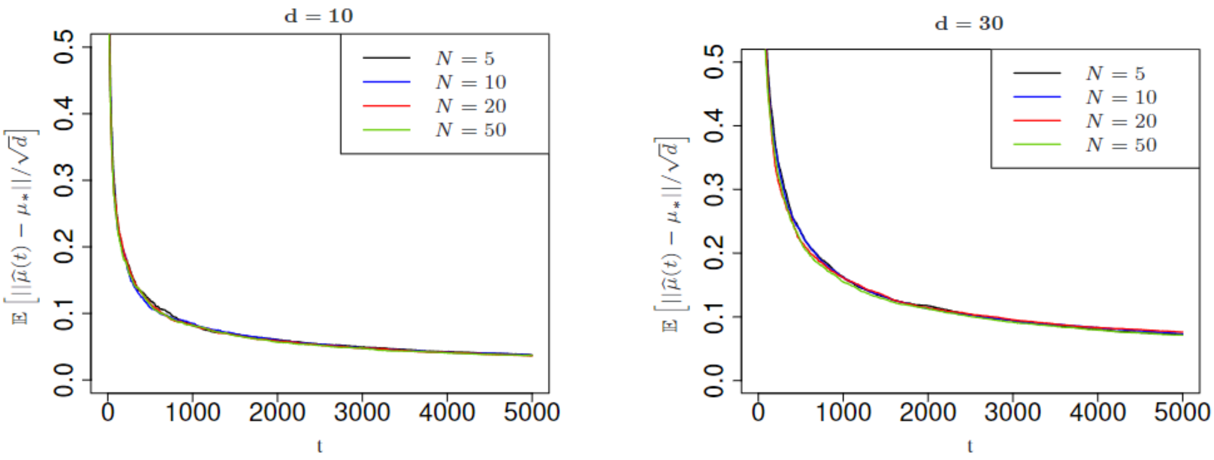

Figure 2.1: Plots of $\mathbb{E}\left[||\widehat{\mu}(t) - \mu_\star||/\sqrt{d}\right]$ over time for different number of arms $N = 5, 10, 20, 50$, and different dimensions of the contexts $d = 10, 30$.

We consider cases with different numbers of arms, $N = 5,\ 10,\ 20,\ 50$, and different dimensions of the contexts $d = 10,\ 30$, repeating 50 times for each case, for every time step. We report two quantities,

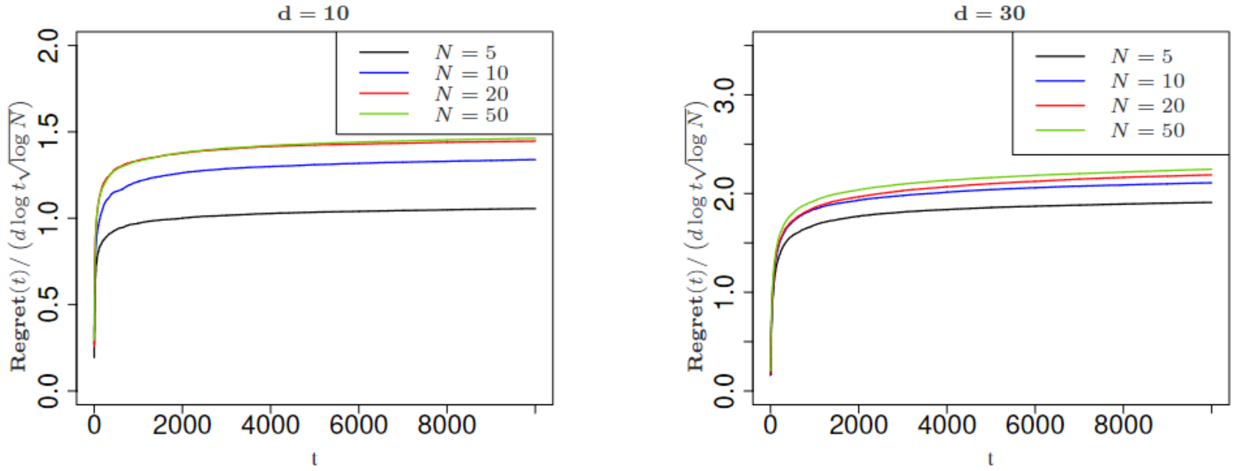

Figure 2.2: Plots of the regret normalized by $d \log t \sqrt{\log N}$, over time for different number of arms $N = 5, 10, 20, 50$, and the dimension of the context $d = 10, 30$.

$||\widehat{\mu}(t) - \mu_\star||$ and $\mathrm{Regret}(t)$, over time, and take averages of the quantities for 50 scenarios. The true parameter $\mu_\star$ as well as each row of A, are randomly generated. Further, we let $\Sigma_x = I_d$, $\Sigma_\xi = I_d$, and $\sigma^2 = 1$.

Figure 2.1 depicts the average norm of the normalized errors over time. We normalize the errors by $\sqrt{d}$, since $\mathrm{Cov}(\widetilde{\mu}(t)) = O(t^{-1})$, by Corollary 1, and so $\mathbf{tr}(\mathrm{Cov}(\widetilde{\mu}(t))) = O(dt^{-1})$. The curves in Figure 2.1 show that the errors decrease with the appropriate rates. Figure 2.2 illustrates the normalized regret over time. The regret is normalized by its bound $d \log t \sqrt{\log N}$ in Theorem 1. In Figure 2.2, the curves show that the normalized regret is constant over time, corroborating the regret bound in Theorem 2.

Chapter 3

Analysis of the Greedy Algorithm for the Shared Parameter Setup

3.1 Introduction

Contextual bandits are ubiquitous models sequential decision making in environments with finite action spaces. The range of applications is extensive and includes different problems that time-varying and action-dependent information are important, such as personalized recommendation of news articles, healthcare interventions, advertisements, and clinical trials (Bouneffouf et al., 2012; Durand et al., 2018; Li et al., 2010; Nahum-Shani et al., 2018; Ren & Zhou, 2020; Tewari & Murphy, 2017; Varatharajah et al., 2018).

In many applications, consequential variables for decision making are not perfectly observed. Technically, the context vectors are often observed in a partial, transformed, and/or noisy manner (Bensoussan, 2004; Bouneffouf et al., 2017; Tennenholtz et al., 2021). In general, sequential decision making algorithms under imperfect observations provide a richer class of models compared to those of perfect observations. Accordingly, they are commonly used in different problems, including space-state models for robot control and filtering (Kalman, 1960; Nagrath, 2006; Roesser, 1975; Stratonovich, 1960).

We study contextual bandits with imperfectly observed context vectors. The probabilistic structure of the problem under study as time proceeds, is as follows. At every time step, there are N available actions (also referred to as 'arms'), and the unobserved context of arm i at time t, denoted by $x_i(t) \in \mathbb{R}^{d_x}$, is generated according to a multivariate normal distribution $\mathcal{N}(0_{d_x}, \Sigma_x)$. Moreover, the corresponding observation (i.e., output) is $y_i(t) \in \mathbb{R}^{d_y}$, while the stochastic reward $r_i(t)$ of arm i is determined by the context and the unknown parameter $\mu_\star$. Formally, we have

$$y_i(t) = Ax_i(t) + \xi_i(t), \tag{3.1}$$

$$r_i(t) = x_i(t)^\top \mu_\star + \varepsilon_i(t), \tag{3.2}$$

where $\xi_i(t)$ and $\varepsilon_i(t)$ are the noises of observation and reward, which are identically distributed and independent following the distributions $\mathcal{N}(0_{d_y}, \Sigma_\xi)$ and $\mathcal{N}(0, \gamma_r^2)$, respectively. Further, the $d_y \times d_x$ sensing matrix A captures the relationship between $x_i(t)$ and the noiseless portion of $y_i(t)$. The above structure holds for all arm i and time t. From a dynamical system point of view, the setting can be understood as memoryless dynamical systems.

At each time, the goal is to learn to choose the optimal arm $a^\star(t)$ maximizing the reward, by utilizing the available information by time t. That is, the agent chooses an arm based on the data collected so far from the model in (3.1); $\{y_i(t)\}_{1\leq i\leq N}, \{a(\tau)\}_{1\leq\tau\leq t-1}, \{y_{a(\tau)}(t)\}_{1\leq\tau\leq t-1}, \{r_{a(\tau)}\}_{1\leq\tau\leq t-1}$. So, the resulting reward will be provided to the agent according to the equation in (3.2). Clearly, to choose high-reward arms, the agent needs accurate estimates of the unknown parameter $\mu_\star$, as well as those of the contexts $x_i(t)$, for $i = 1, \cdots, N$. However, because $x_i(t)$ is not observed, the estimation of $\mu_\star$ is available only through the output $y_i(t)$. Thereby, design of efficient reinforcement learning algorithms with guaranteed performance is challenging.

Bandits are thoroughly investigated in the literature, assuming that $\{x_i(t)\}_{1\leq i\leq N}$ are perfectly observed. Early papers focus on the method of Upper-Confident-Bounds (UCB) for addressing the exploitation-exploration trade-off (Abbasi-Yadkori et al., 2011; Abe & Long, 1999; Auer, 2002; Chu et al., 2011; Lai & Robbins, 1985). UCB-based methods take actions following optimistic estimations of the parameters, and are commonly studied in reinforcement learning (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2020c). Another popular and efficient family of policies use randomized exploration, usually in the Bayesian form of Thompson sampling (Agrawal & Goyal, 2013; Chapelle & Li, 2011; Faradonbeh et al., 2019, 2020a, 2020b; Modi & Tewari, 2020). For contextual bandits that contexts are generated under certain conditions, exploration-free policies with Greedy nature can expose efficient performance (Bastani et al., 2021).

Currently, theoretical results for bandits with imperfect context observations are incomplete. For contextual bandits with noisy observations with the same dimension as that of contexts, asymptotic analyses are available for a UCB-type algorithms (Yun et al., 2017), and Thompson sampling (Park & Faradonbeh, 2021), or in presence of additional information (Tennenholtz et al., 2021). Moreover, the relationship between the regret and gained information under the uncertainty of observations are analyzed (Lattimore, 2022; Lattimore & Gyorgy, 2021). However, analyses about contextual bandits with noisy transformed observations, whose dimension can be different from that of contexts, are scarce. Lastly, numerical analysis shows that Greedy algorithms outperform Thompson sampling under imperfect context observations in the suggested framework (Park & Faradonbeh, 2022a). Therefore, this work focuses on the non-asymptotic theoretical analysis of Greedy policies for imperfectly observed contextual bandits.

However, comprehensive analyses and non-asymptotic theoretical performance guarantees for general output observations are not currently available and are adopted as the focus of this work. We perform the *finite-time worst-case* analysis of Greedy reinforcement learning algorithms for imperfectly observed contextual bandits. We establish efficiency and provide high probability upper bounds for the regret that consists of poly-logarithmic factors of the time horizon and of the failure probability. Furthermore, the effects of other problem parameters such as the number of arms and the dimension are fully characterized. Illustrative numerical experiments showcasing the efficiency are also provided.

To study the performance of reinforcement learning policies, different technical difficulties arise in the high probability analyses. First, one needs to study the eigenvalues of the empirical covariance matrices, since the estimation accuracy depends on them. Furthermore, it is required to consider the number of times the algorithm selects sub-optimal arms. Note that both quantities are stochastic and so worst-case (i.e., high probability) results are needed for a statistically dependent sequence of random objects. To obtain the presented theoretical results, we employ advanced technical tools from martingale theory and random matrices. Indeed, by utilizing concentration inequalities for matrices with martingale difference structures, we carefully characterise the effects of order statistics and tail-properties of the estimation errors.

The remainder of this chapter is organized as follows. In Section 3.2, we formulate the problem and discuss the relevant preliminary materials. Next, a Greedy reinforcement learning algorithm for contextual bandits with imperfect context observations is presented in Section 3.3. Lastly, in Section 3.4, we provide theoretical performance guarantees for the proposed algorithm, followed by numerical experiments in Section 3.5.

We use $A^\top$ to refer to the transpose of the matrix $A \in \mathbb{C}^{p\times q}$. For a vector $v \in \mathbb{C}^d$, we denote the ℓ_2 norm by $\|v\| = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$. Additionally, $C(A)$ and $C(A)^\perp$ are employed to denote the column-space of the matrix A and its orthogonal subspace, respectively. Further, $P_{C(A)}$ is the projection operator onto $C(A)$. Moreover, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of the symmetric matrix A, respectively. Finally, $O(\cdot)$ denotes the order of magnitude, $\{X_i\}_{i\in E} = \{X_i : i \in E\}$ and $I(\cdot)$ is the indicator function.

3.2 Problem Formulation

First, we formally discuss the problem of contextual bandits with imperfect context observations. A bandit machine has N arms, each of which has its own unobserved context $x_i(t)$, for $i \in \{1, \cdots, N\}$. Equation (3.1) presents the observation model, where the observations $\{y_i(t)\}_{1\leq i\leq N}$ are linearly transformed functions of the contexts, perturbed by additive noise vectors $\{\xi_i(t)\}_{1\leq i\leq N}$. Equation (3.2) describes the process of reward generation for different arms, depicting that *if* the agent selects arm i, then the resulting reward is an *unknown* linear function of the unobserved context vector, subject to some additional randomness due to the reward noise $\varepsilon_i(t)$.

The agent aims to maximize the cumulative reward over time, by utilizing the sequence of observations. To gain the maximum possible reward, the agent needs to learn the relationship between the rewards $r_i(t)$ and the observations $y_i(t)$. For that purpose, we proceed by considering the conditional distribution of the reward $r_i(t)$ given the observation $y_i(t)$, i.e., $\mathbb{P}(r_i(t)|y_i(t))$, which is

$$\mathcal{N}(y_i(t)^\top D^\top \mu_\star, \gamma_{ry}^2), \tag{3.3}$$

where $D = (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_\xi^{-1}$ and $\gamma_{ry}^2 = \mu_\star^\top (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} \mu_\star + \gamma_r^2$.

Based on the conditional distribution in (3.3), in order to maximize the expected reward given the observation, we consider the conditional expectation of the reward given the observations, $y_i(t)^\top D^\top \mu_\star$. So, letting $\eta_\star = D^\top \mu_\star$ be the transformed parameter, we focus on the estimation of $\eta_\star$. The rationale is twofold; first, the conditional expected reward can be inferred with only knowing $\eta_\star$, regardless of the exact value of the true parameter $\mu_\star$. Second, $\mu_\star$ is not estimable when the rank of the sensing matrix A in the observation model is less than the dimension of $\mu_\star$. Indeed, estimability of $\mu_\star$ needs the restrictive assumptions of non-singular A and $d_y \geq d_x$.

The optimal policy that reinforcement learning policies need to compete against knows the true parameter $\mu_\star$. That is, to maximize the reward given the output observations, the optimal arm at time t, denoted by $a^\star(t)$, is

$$a^\star(t) = \underset{1\leq i\leq N}{\operatorname{argmax}}\, y_i(t)^\top \eta_\star. \tag{3.4}$$

Then, the performance degradation due to uncertainty about the environment that the parameter $\mu_\star$ represents, is the assessment criteria for reinforcement learning policies. So, we consider the following performance measure, which is commonly used in the literature, and is known as *regret* of the reinforcement learning policy that selects the sequence of actions $a(t), t = 1, 2, \cdots$:

$$\text{Regret}(T) = \sum_{t=1}^{T} \left(y_{a^\star(t)}(t) - y_{a(t)}(t) \right)^\top \eta_\star. \tag{3.5}$$

In other words, the regret at time T is the total difference in the obtained rewards, up to time T, where the difference at time t is between the optimal arms $a^\star(t)$ and the arm $a(t)$ chosen by the reinforcement learning policy based on the output observations by the time t. Note that this difference does not depends on the unknown contexts $\{x_i(t)\}_{1 \leq i \leq N}$. That is, the arm maximizing $x_i(t)^\top \mu_\star$ is not guaranteed to be $a^\star(t)$, since $y_i(t)^\top \eta_\star$ is a realized value of a random variable centered at $x_i(t)^\top \mu_\star$.

3.3 Reinforcement Learning Policy

In this section, we explain the details of the Greedy algorithm for contextual bandits with imperfect observations. Although inefficient in some reinforcement learning problems, Greedy algorithms are known to be efficient under certain conditions such as covariate diversity (Bastani et al., 2021). Intuitively, the latter condition expresses that the context vectors cover all directions in $\mathbb{R}^{d_x}$ with a non-trivial probability, so that additional exploration is not necessary.

As discussed in Section 3.2, it suffices for the policy to learn to maximize

$$\mathbb{E}[r_i(t) | y_i(t)] = y_i(t)^\top \eta_\star. \tag{3.6}$$

To estimate the quantity $y_i(t)^\top \eta_\star$, we use the least-squares estimate

$$\widehat{\eta}(t) = \underset{\eta}{\operatorname{argmax}} \sum_{\tau=1}^{t} (r_{a(\tau)}(\tau) - y_{a(\tau)}(\tau)^\top \eta)^2, \tag{3.7}$$

in lieu of the truth $\eta_\star$. So, the Greedy algorithm selects the arm $a(t)$ at time t, such that

$$a(t) = \underset{1 \le i \le N}{\operatorname{argmax}} \, y_i(t)^\top \widehat{\eta}(t). \tag{3.8}$$

The recursions to update the parameter estimate $\widehat{\eta}(t)$ and the empirical inverse covariance matrix $B(t)$ based on (3.7) are as follows:

$$B(t+1) = B(t) + y_{a(t)}(t) y_{a(t)}(t)^\top \tag{3.9}$$

$$\widehat{\eta}(t+1) = B(t+1)^{-1} \left(B(t)\widehat{\eta}(t) + y_{a(t)}(t) r_{a(t)}(t) \right), \tag{3.10}$$

where the initial values consist of $B(1) = \Sigma^{-1}$, for some arbitrary postitive definite matrix Σ, and $\widehat{\eta}(1) = \eta$ for an arbitrary vector η in $\mathbb{R}^{d_y}$. Algorithm 2 describes the pseudo-code for the Greedy algorithm.

Algorithm 2 : Greedy policy for contextual bandits with imperfect context observations

```
Set $B(1) = \Sigma^{-1}, \widehat{\eta}(1) = \eta$
for $t = 1, 2, \dots,$ do
    Select arm $a(t) = \arg\max_{1 \le i \le N} y_i(t)^\top \widehat{\eta}(t)$
    Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_\star + \varepsilon_{a(t)}(t)$
    Update $B(t+1)$ and $\widehat{\eta}(t+1)$ by (3.9) and (3.10)
end for
```

3.4 Theoretical Performance Guarantees

In this section, we present a theoretical result for Algorithm 2 presented in the previous section. The result provides a worst-case analysis and establishes a high probability upper-bound for the regret in (3.5).

Theorem 3. *Assume that Algorithm 2 is used in a contextual bandit with N arms and the output dimension d_y. Then, with probability at least $1-4\delta$, we have*

$$\text{Regret}(T) = \\ O\left(\frac{(\lambda_{a2}+\lambda_{y2})\gamma_{ry}}{\lambda_{a1}+\lambda_{y1}} N d_y^{3/2}\left(\log\frac{N d_y T}{\delta}\right)^{5/2}\log\frac{d_y T}{\delta}\right),$$

where $\lambda_{a1}=\lambda_{\min}(A\Sigma_x A^\top)$, $\lambda_{a2}=\lambda_{\max}(A\Sigma_x A^\top)$, $\lambda_{y1}=\lambda_{\min}(\Sigma_\xi)$, $\lambda_{y2}=\lambda_{\max}(\Sigma_\xi)$ and γ_{ry}^2 is the conditional variance in (3.3).

The regret bound above scales linearly with the number of arms N, with $d_y^{3/2}$ for the dimension of the observations d_y, and poly-logarithmically with time T. The dimension of unobserved context vectors does not affect the regret because the optimal policy in (3.4) does not have the exact values of the context vectors. So, similar to the reinforcement learning policy, the optimal policy needs to estimate the contexts as well, as $y_i(t)^\top\eta_\star$ in (3.4) is an estimate of $x_i(t)^\top\mu_\star$ for the optimal policy to find the optimal arm.

The rationale of the linear growth of the regret with N is that a policy is more likely to choose one of sub-optimal arms, when more sub-optimal arms exist, incurring more regret. In addition, the quadratic term of d_y and the maximum eigenvalue λ_{a2} are generated by the use of truncation for the ℓ_2 norm of vector $(\|y_i(t)\|_2^2 = O(\lambda_{a2} d_y \upsilon_T(\delta)^2))$ as well as the matrix Azuma's inequality. Further, the poly-logarithmic terms of T, N, d_y and δ, $(\log(Nd_yT/\delta))^{5/2}\log(d_yT/\delta)$, are originated in the truncation event and the Azuma's inequality. Lastly, the minimum eigenvalue λ_{a1} and the conditional reward variance γ_{ry}^2 are associated with the variance of the estimator $\widehat{\eta}(t)$, whose larger value causes a greater regret.

Proof. We use the following intermediate results, whose proofs are delegated to Appendices. For simplicity, let $\widehat{\eta}(1)$ be a random variable with $\mathbb{E}[\widehat{\eta}(1)] = \eta_\star$ and $\text{Cov}(\widehat{\eta}(1)) = \Sigma^{-1}\gamma_{ry}^2$ so that $\mathbb{E}[\widehat{\eta}(t)] = \eta_\star$ and $\text{Cov}(\widehat{\eta}(t)|B(t)) = B(t)^{-1}\gamma_{ry}^2$ for all t. First, for $T>0$ and $0<\delta<0.25$, we define

$$W_T = \left\{\max_{\{1\le\tau\le t \; and \; 1\le i\le N\}} ||S_y^{-1/2} y_i(\tau)||_\infty \le \upsilon_T(\delta)\right\}, \tag{3.11}$$

where $\upsilon_T(\delta) = (2\log(Nd_yT/\delta))^{1/2}$.

Lemma 2. *For the event W_T defined in* (3.11)*, we have $\mathbb{P}(W_T) \geq 1 - \delta$.*

Lemma 2 guarantees that all the observation up to time T are generated in the truncation event W_T with the probability at least $1 - \delta$.

Lemma 3. *Let $\sigma\{X_1, \ldots, X_n\}$ be the sigma-field generated by random vectors $X_1, \ldots, X_n$. For the observation of chosen arm $y_{a(t)}(t)$ at time t, the estimator $\widehat{\eta}(t)$ defined in* (3.10)*, and the filtration $\{\mathscr{F}_t\}_{1\leq t\leq T}$ defined according to*

$$\mathscr{F}_t = \sigma\{\{a(\tau)\}_{1\leq\tau\leq t}, \{y_i(\tau)\}_{1\leq\tau\leq t, 1\leq i\leq N}, \{r_{a(\tau)}(\tau)\}_{1\leq\tau\leq t}\},$$

we have

$$\mathbb{E}[V_t|\mathscr{F}_{t-1}] = P_{C(S_y^{1/2}\widehat{\eta}(t))}(k_N - 1) + I_{d_y},$$

where $V_t = S_y^{-1/2} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-1/2}$ and $k_N = \mathbb{E}\left[\left(\max_{1\leq i\leq N}\{Z_i\}\right)^2\right]$ for N independent Z_i with the standard normal distribution and $S_y = \mathrm{Cov}(y_i(t))$. That is, k_N is the expected maximum of N independent standard normal random variables.

Lemma 3 sets the stage for analysis of the (unnormalized) empirical inverse covariance $B(t)$ in (3.9)

Lemma 4. *(Matrix Azuma Inequality Tropp, 2012) Consider the sequence $\{M_k\}_{1\leq k\leq K}$ of symmetric $d \times d$ random matrices adapted to some filtration $\{\mathscr{G}_k\}_{1\leq k\leq K}$, such that $\mathbb{E}[M_k|\mathscr{G}_{k-1}] = 0$. Assume that there is a deterministic sequence of symmetric matrices $\{A_k\}_{1\leq k\leq K}$ that satisfy $M_k^2 \preceq A_k^2$, almost surely. Let $\sigma^2 = \|\sum_{1\leq k\leq K} A_k^2\|$. Then, for all $\varepsilon \geq 0$, it holds that*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{K} M_k\right) \geq \varepsilon\right) \leq d \cdot e^{-\varepsilon^2/8\sigma^2}.$$

Lemma 5 provides a high probability lower bound for the minimum eigenvalue of $B(t)$. Then, Lemma 6 bounds the estimation error.

Lemma 5. *For $B(t)$ in* (3.9) *and $t \leq T$, on the event W_T defined in* (3.11)*, by Lemma 3 and 3, with the probability at least $1-\delta$, we have*

$$\lambda_{\min}(B(t)) \geq \lambda_{s1}(t-1)\left(1-\sqrt{\frac{32\upsilon_T(\delta)^4}{t-1}\log\frac{d_yT}{\delta}}\right).$$

Lemma 6. *In Algorithm 2, let $\widehat{\eta}(t)$ be the parameter estimate, as defined in* (3.10)*. Then, for t $\leq T$, on the event W_T defined* (3.11)*, we have*

$$\mathbb{P}\left(\|\widehat{\eta}(t)-\eta_\star\| > \varepsilon | B(t)\right) \leq 2e^{-\frac{\varepsilon^2}{2d_y\lambda_{\max}(B(t)^{-1})\gamma_{ry}^2}}.$$

Next, Lemma 7 gives an upper bound for the probability that Algorithm 2 does not choose the optimal arm at time t. Finally, Lemma 8 studies the weighted sum of indicator functions $I(a^\star(t) \neq a(t))$ that count the effective number of times that the algorithm chooses sub-optimal arms.

Lemma 7. *Given $B(t)$, an upper bound of probability of choosing a sub-optimal arm is bounded as follows:*

$$\mathbb{P}(a^\star(t) \neq a(t)|B(t)) \leq \frac{2N\lambda_{s2}^{1/2}d_y\upsilon_T(\delta)\gamma_{ry}}{\sqrt{\eta_\star^T S_y \eta_\star}}\lambda_t^{1/2},$$

where $\lambda_t = \lambda_{\max}(B(t)^{-1})$.

Lemma 8. *For $I(a^\star(t) \neq a(t))$, on the event W_T, with the probability at least $1-\delta$, we have*

$$\sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} I(a^\star(t) \neq a(t)) \leq \sqrt{32 \log T \log(T\delta^{-1})} + \sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} \mathbb{P}(a^*(\tau) \neq a(\tau) | B(t)),$$

where $t^* = 128 \upsilon_T(\delta)^4 \log \frac{d_y T}{\delta} + 1$.

Note that $\text{Regret}(T)$ is the sum of the conditional expected reward difference $\left(y_{a^\star(t)}(t) - y_{a(t)}(t)\right)^\top \eta_\star$ for $1 \leq t \leq T$. The difference $\left(y_{a^\star(t)}(t) - y_{a(t)}(t)\right)^\top \eta_\star$ at time t is greater than 0, only when $a^\star(t) \neq a(t)$. Thus, the regret can be rewritten as $\text{Regret}(T) = \sum_{t=1}^T \left(y_{a^\star(t)}(t) - y_{a(t)}(t)\right)^\top \eta_\star I(a^\star(t) \neq a(t))$. To find an upper bound of the regret, we find high probability upper bounds for $\left(y_{a^\star(t)}(t) - y_{a(t)}(t)\right)^\top \eta_\star$ and $I(a^\star(t) \neq a(t))$, respectively. For both upper bounds, the inverse of the (unnormalized) empirical covariance matrix $B(t)$ in (3.9) matters in that the matrix determines the size of estimation error $\|\widehat{\eta}(t) - \eta_\star\|$.

By, Lemma 5, we have

$$\lambda_{\min}(B(t)) \geq \lambda_{s1}(t-1)\left(1 - \sqrt{\frac{32\upsilon_T(\delta)^4}{t-1} \log \frac{d_y T}{\delta}}\right), \tag{3.12}$$

for all $1 \leq t \leq T$ with the probability at least $1 - 2\delta$. This implies that $B(t)$ grows linearly with the horizon almost surely. Next, we investigate the estimation error $\|\eta_\star - \widehat{\eta}(t)\|$ based on the above result of the minimum eigenvalue of $B(t)$. Using $\|y_i(t)\|_\infty \leq \lambda_{s2}^{1/2} \upsilon_T(\delta)$ on the event W_T, we have

$$(y_{a^\star(t)}(t) - y_{a(t)}(t))^\top \eta_\star \leq \lambda_{s2}^{1/2} \upsilon_T(\delta) \|\widehat{\eta}(t) - \eta_\star\|, \tag{3.13}$$

where $\lambda_{s2} = \lambda_{\max}(S_y)$. So, we write the regret in the following form:

$$\text{Regret}(T) \leq \sum_{t=1}^T \lambda_{s2}^{1/2} \upsilon_T(\delta) \|\widehat{\eta}(t) - \eta_\star\| I(a^\star(t) \neq a(t)). \tag{3.14}$$

Here, we denote $\lambda_t = \lambda_{\max}(B(t)^{-1}) = (\lambda_{\min}(B(t)))^{-1}$. By (3.12), we can find $t^* = 128 v_T(\delta)^4 \log \frac{d_y T}{\delta} + 1$, such that

$$\lambda_t \leq \frac{2}{\lambda_{s1}(t-1)}, \tag{3.15}$$

with the probability at least $1-\delta$, for all $t^* < t \leq T$. By Lemma 6 and (3.15), for all $t^* < t \leq T$, with the probability at least $1 - 3\delta$, we have

$$\lambda_{s2}^{1/2} v_T(\delta) \|\widehat{\eta}(t) - \eta_\star\| \leq a_1 (t-1)^{-1/2}, \tag{3.16}$$

where $a_1 = 4(\lambda_{s2}/\lambda_{s1})^{1/2} v_T(\delta) \sqrt{2 d_y \log(2T\delta^{-1})}$. Thus, with $(y_{a^\star(t)} - y_{a(t)})^\top \eta_\star \leq 2\lambda_{s2}^{1/2} v_T(\delta) \|\eta_\star\|$ for $t < t^*$, the regret can be represented

$$\mathrm{Regret}(T) \leq \sum_{t<t^*} 2\lambda_{s2}^{1/2} v_T(\delta)\|\eta_\star\| + \sum_{t^* \leq t \leq T} a_1 (t-1)^{-1/2} I(a^\star(t) \neq a(t)), \tag{3.17}$$

with the probability at least $1 - 3\delta$. Now, we consider the probability to choose the optimal arm at time t. By Lemma 7, we have

$$\sum_{t^* \leq t \leq T} \frac{\mathbb{P}(a^\star(t) \neq a(t) | B(t))}{\sqrt{t-1}} \leq \frac{2^{3/2} N \lambda_{s2}^{1/2} d_y v_T(\delta) \gamma_{ry}}{\|\eta_\star\| \lambda_{s1}^{1/2}} \log T. \tag{3.18}$$

Now, we construct an upper bound about the indicator function $I(a^*(t) \neq a(t))$ in (3.14), by Lemma 8.

$$\sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} I(a^\star(t) \neq a(t)) \leq \sqrt{32 \log T \log(T\delta^{-1})} + \sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} \mathbb{P}(a^*(\tau) \neq a(\tau) | B(\tau)), \tag{3.19}$$

with the probability at least $1-\delta$. Therefore, by (3.18) and (3.19), with the probability at least $1-4\delta$, the following inequalities hold for the regret of the algorithm, which yield to the desired result:

$$\begin{aligned}\text{Regret}(T) &= \sum_{t=1}^{T}(y_{a^\star(t)}(t) - y_{a(t)}(t))^\top \eta_\star I(a^\star(t) \neq a(t)) \\ &\leq 2\lambda_{s2}^{1/2} v_T(\delta)\|\eta_\star\| t^* + \sum_{t^* \leq t \leq T} a_1 \frac{1}{\sqrt{t-1}} I(a^\star(t) \neq a(t)) \\ &= O\left(\frac{\lambda_{s2}}{\lambda_{s1}} \gamma_{ry} N d_y^{2/3} \left(\log \frac{N d_y T}{\delta}\right)^{5/2} \log \frac{d_y T}{\delta}\right). \end{aligned} \tag{3.20}$$

Finally, using $S_y = A\Sigma_x A^\top + \Sigma_\xi$, $\lambda_{s2}/\lambda_{s1} = O((\lambda_{a2} + \lambda_{y2})/(\lambda_{a1} + \lambda_{y1}))$, with the probability at least $1-4\delta$, we have

$$\text{Regret}(T) = O\left(\frac{(\lambda_{a2} + \lambda_{y2})\gamma_{ry}}{\lambda_{a1} + \lambda_{y1}} N d_y^{3/2} \left(\log \frac{N d_y T}{\delta}\right)^{5/2} \log \frac{d_y T}{\delta}\right). \tag{3.21}$$

This bound is relatively looser in terms of $N,\ d_y$ and tighter in terms of T as compared to the bound $O(polylog(N)\sqrt{d_x T})$ for fully observable contexts Agrawal and Goyal, 2013; Chu et al., 2011. But, this looser bound in terms of $N,\ d_y$ is created to improve the regret bound in terms of T. □

3.5 Numerical Illustrations

In this section, we perform numerical analyses for the theoretical result in the previous section. We simulate cases for $N = 10, 20, 50$ and different dimensions of the observations $d_y = 5, 20, 50$ with a fixed context dimension $d_x = 20$. Each case is repeated 100 times and the average and worst quantities of 100 scenarios are reported.

For Figure 3.1, the left plot depicts the average (solid) and worst-case (dashed) regret among all scenarios, normalized by $\log t$. The number of arms N varies as shown in the graph, while the dimension is fixed to $d_y = 10$. Next, the right one illustrates that the normalized regrets increase over time for different d_y at the fixed number of arms $N = 5$. For both plots, the worst-case regret curves are well above the average ones, but the slopes of curves for both cases become flat as time goes on, implying that the worst-case regret grows logarithmically in terms of t as well. Figure 3.2 presents the average and worst-case regret (non-normalized) at time $T = 2000$ for different $N = 10, 20, 50$ and $d_y = 5, 20, 50$. The plot shows that the regret at $T = 2000$ increase as N and d_y become larger. In addition, it shows that the dimension of observations d_y has a greater effect on the regret than that of the number of arms N.

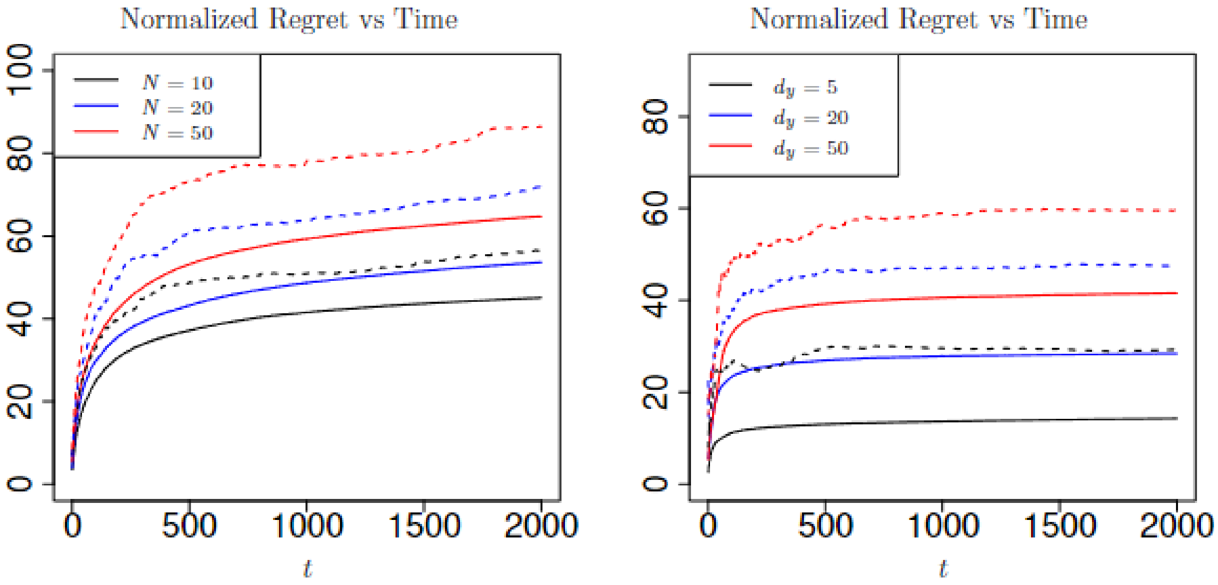

Figure 3.1: Plots of $\text{Regret}(t)/\log t$ over time for the different number of arms $N = 10, 20, 100$ and $d_y = 5, 20, 50$. The solid and dashed lines represent average and worst regret curves, respectively.

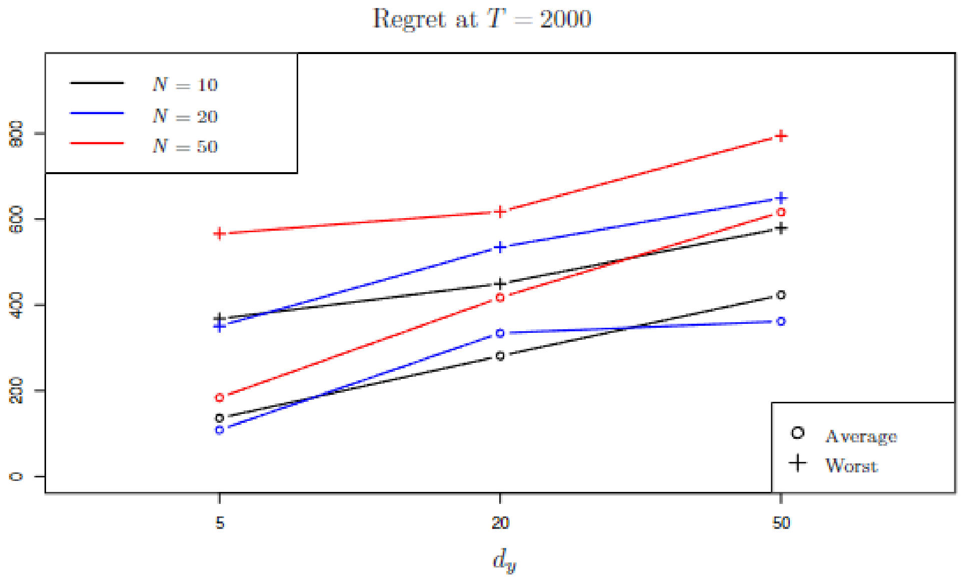

Figure 3.2: Plot of average and worst-case $\text{Regret}(T)$ at $T = 2000$ for different number of arms $N = 10, 20, 50$ and dimension of observations $d_y = 5, 20, 50$.

Chapter 4

Analysis of Thompson Sampling for the Arm-specific Parameter Setup

4.1 Introduction

Contextual bandits have emerged in the recent literature as widely-used decision-making models involving time-varying information. In this setup, a policy takes action after (perfectly or partially) observing the contexts at each time. The data collected by the time is utilized, aiming to maximize cumulative rewards determined by both the contexts and unknown parameters. So, any desirable policy needs to manage the trade-off between learning the best (i.e., exploration) and earning the most (i.e., exploitation). For this purpose, Thompson sampling stands-out among various competitive algorithms, thanks to its performance guarantees as well as computationally favorable implementations. Its main idea is to explore based on samples from a data-driven posterior belief about the unknown parameters. However, comprehensive studies are currently unavailable for imperfectly observed contexts, and this is adopted as the focus of this work.

Letting the time-varying components of the decision options (e.g., contexts) be observed partially only, is known to be advantageous in various real-world problems (Åström, 1965; Dougherty, 2020; Kaelbling et al., 1998; Kang et al., 2012; Lin et al., 2012; Nagrath, 2006). On the other hand, overlooking imperfections in observations can lead to compromised decisions, for example in clinical treatment of septic patients (Gottesman et al., 2019). The study of partial observation models includes linear systems (Kargin et al., 2023), Markov decision processes (Bensoussan, 2004; Krishnamurthy & Wahlberg, 2009), and partial monitoring (Kirschner et al., 2020; Lattimore, 2022; Tsuchiya et al., 2023). Note that in the latter setting, partiality pertains to the bandit feedback of the rewards, whereas in this work partiality relates to the context observations. Further, partial observability has recently motivated some work on contextual bandits that do *not* involve the exploration-exploitation dilemma (Kim et al., 2023; Park & Faradonbeh, 2021, 2022a, 2022b, 2022c, 2024, n.d.-a, n.d.-b, n.d.-c).

The common bandit setting is the so-called linear one, where the *expected* reward of each arm is the inner product of (adversarial or stochastic) context(s) and reward parameter(s). The latter in stochastic contextual bandits can be either *arm-specific* (Bastani & Bayati, 2020; Goldenshluger & Zeevi, 2013), or *shared* across all arms (Chakraborty et al., 2023; Dani et al., 2008). We analyze both settings, with the focus being on the more general and challenging one of the former. For the sake of completeness, the authors also refer to a (non-exhaustive) variety of extant approaches in the realm of contextual bandits. That includes (possibly infinite but bounded) action sets in a Euclidean space (Abbasi-Yadkori et al., 2011; Abeille & Lazaric, 2017), as well as those with adversarial contexts (Agarwal et al., 2014; Dani et al., 2008), together with non-linear or non-parametric reward functions (Dumitrascu et al., 2018; Guan & Jiang, 2018; Wanigasekara & Yu, 2019). Notably, all of these references assume fully observed contexts, in contradistinction to this work.

Efficient policies for contextual bandits are diverse, including the popular family of policies based on Optimism in the Face of Uncertainty (OFU) that have theoretical guarantees for balancing exploration and exploitation (Abbasi-Yadkori et al., 2011; Auer, 2002; Dani et al., 2008). Besides, Thompson sampling is recognized as a pioneer, first via excelling empirical performance (Chapelle & Li, 2011), and then supplemented with theoretical analyses (Abeille & Lazaric, 2017; Agrawal & Goyal, 2013; Russo & Van Roy, 2014). More recently, it is shown that Greedy policies can be nearly optimal in contextual bandits with one or two reward parameter(s) (Bastani et al., 2021; Park & Faradonbeh, 2022c; Raghavan et al., 2023). In contrast, for contextual bandits with multiple arm-specific reward parameters, it is known that vanilla Greedy is non-optimal (Bastani et al., 2021). That is caused by some arms dominating the rest, leaving them unexplored, and is also illustrated in our numerical experiments.

The study of theoretical performance guarantees for Thompson sampling made significant progress in the recent literature with an emphasis on instance-independent regret analysis. First, minimax regret bounds growing as square-root of time were shown for adversarial contextual bandits (Abeille & Lazaric, 2017; Agrawal & Goyal, 2013; Russo & Van Roy, 2014) and for settings with an Euclidean action set (Hamidi & Bayati, 2020). Logarithmic regret bounds for stochastic contextual bandits with a *shared reward parameter* are established as well (Chakraborty et al., 2023). However, for the arm-specific reward parameters, efficient bandit policies remain unavailable. In this case, the analysis is significantly more challenging since the policy needs to delicately address the trade-off between exploration and exploitation, (unlike the setting with a shared reward parameter).

Instance-dependent regret analysis for classical (non-contextual) bandits has been extensively studied (Garivier et al., 2016, 2019; Kaufmann et al., 2016; Lattimore, 2018). In contrast, in contextual bandits it remains largely unexplored, albeit with a few positive results under specific conditions, improving from square-root to logarithmic regret bounds (Abbasi-Yadkori et al., 2011; Dani et al., 2008). However, these studies have two limitations: the assumption that a suboptimality gap (i.e., expected reward difference between the best and second best arms) is greater than a positive constant over time and instance-dependent algorithms. To address these challenges, we propose a novel approach that utilizes a probabilistic suboptimality gap, which is a suboptimality gap attained with a positive probability, for analyzing instance-independent Thompson sampling, working without knowledge of the reward parameter, the covariance matrices, and the observation structure.

We analyze the Thompson sampling policy in partially observable contextual bandits focusing on the high-probability frequentist regret. Our analysis demonstrates that the error in estimating the reward parameters decays with square-root of time, and the worst-case regret grows at most as fast as a polylogarithmic function of time. Further, the effect of the ambient dimension d is of the order of $\sqrt{d}$ on the estimation, while it exacerbates the regret bound as d^4. Lastly, scalings of the above two quantities with the number of arms N are $\mathcal{O}(\sqrt{N})$ and $\mathcal{O}(N)$, respectively.

For regret analysis in partially observed contextual bandits, it is crucial to examine the effects of the partiality of information, the suboptimality gaps, and the probabilities of pulling such suboptimal arms. The existing technical approaches often fail to provide useful results mainly due to the inter-dependencies of the numbers of arm pulls, which is referred to as the *sample size* of an arm in this chapter. This challenge is addressed in this work by developing novel technical tools, as follows. First, we take into account an instance-dependent probabilistic suboptimality gap. Next, incorporating this with the aspect of partial observability, we analyze the suggested Thompson sampling with an appropriate exploration mechanism ensuring that the probability of pulling suboptimal arms decays as $\mathcal{O}\left(t^{-1/2}\right)$ as time proceeds. Given the interdependence of sample sizes across arms, we delicately construct some martingale sequences and employ useful stochastic bounds for them, in order to derive our regret bound.

The organization is as follows. In Section 4.2, we formulate the problem and discuss preliminaries. Then, the Thompson sampling policy for partially observable contextual bandits is presented in Section 4.3. We provide its theoretical performance guarantees in Section 4.4, followed by real-data experiments in Section 4.5. Further technical discussions and intermediate lemmas are delegated to appendices.

For an integer i, $[i]$ represents $\{1, 2, \ldots, i\}$. $M^\top$ is the transpose of the matrix $M \in \mathbb{C}^{p \times q}$, and $C(M)$ denotes the column space of M. In addition, I_d represents the identity matrix with dimension d, where d is an integer. For a vector $v \in \mathbb{C}^d$, we let the ℓ_2 norm denote $\|v\| = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$ and the weighted ℓ_2 norm with a positive definite matrix A by $\|v\|_A = \sqrt{v^\top A v}$. Finally, $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues.

4.2 Problem Formulation

In this section, we state the partially observable contextual bandit problem and provide its formulation. A policy aims to maximize cumulative reward by selecting one from N arms, the reward of arm $i \in [N]$ at time t being

$$r_i(t) = x_i(t)^\top \mu_i + \varepsilon_i(t). \tag{4.1}$$

Here, $x_i(t)$ is the *unobserved* d_x-dimensional stochastic context of arm i, independently generated over time and across arms, from a distribution with $\mathbb{E}\left[x_i(t)\right] = \mathbf{0}_{d_x}$ and unknown covariance $\mathrm{Cov}(x_i(t)) = \Sigma_x$. Further, $\mu_i \in \mathbb{R}^{d_x}$ is the *unknown arm-specific* reward parameter of the i-th arm, and $\varepsilon_i(t)$ is the noise in the realization of the reward value. We assume that each coordinate of a context and the reward noise, both have sub-Gaussian tails. That is, there exists a fixed constant $R_1 > 0$, that for all real λ, we have

$$\mathbb{E}\left[e^{\lambda \varepsilon_i(t)}\right] \leq \exp\left(\frac{\lambda^2 R_1^2}{2}\right). \tag{4.2}$$

The policy observes the following transformed noisy version $y_i(t)$ of the context $x_i(t)$:

$$y_i(t) = A x_i(t) + \xi_i(t), \tag{4.3}$$

where A is the unknown $d_y \times d_x$ sensing matrix, and $\xi_i(t)$ is the sensing (or measurement) noise, its *unknown* covariance matrix being denoted by Σ_ξ. We assume that each element of $\xi_i(t)$ is sub-Gaussian as well. At each time t, the policy chooses an arm, denoted by $a(t)$, given the history of actions $\{a(\tau)\}_{\tau \in [t-1]}$, rewards $\{r_{a(\tau)}(\tau)\}_{\tau \in [t-1]}$, past observations $\{y_i(\tau)\}_{\tau \in [t-1], i \in [N]}$, and the current ones $\{y_i(t)\}_{i \in [N]}$. Once choosing the arm $a(t)$, we obtain a reward $r_{a(t)}(t)$ according to (4.1), whereas rewards of other arms are *not* realized.

Note that (fully observable) contextual bandits consider a class of policies $\{\pi : \mathcal{X} \to [N]\}$ for the space of contexts $\mathcal{X}$. However, in our case when the context vectors $\{x_i(t)\}_{i \in [N]}$ are unknown, we take into account a class of policies $\{\pi : \mathcal{Y} \to [N]\}$ for the space of observations $\mathcal{Y}$. Policies of this class account for $y_i(t)$ to infer $x_i(t)^\top \mu_i$ for decision.

Remark 1 (Best Linear Unbiased Prediction (BLUP) (Harville, 1976; Robinson, 1991))**.** *Let $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ be stochastic observation and unobserved context related as in* (4.3)*. Then, for a linear function $x^\top \mu$ of x, its BLUP is $y^\top b$, where b is chosen to minimize the prediction error* $\mathrm{Var}(x^\top \mu - y^\top b)$ *subject to unbiasedness* $\mathbb{E}[x^\top \mu - y^\top b] = 0$*. This linear prediction is invariant to x, and is given by $b = D^\top \mu$ for $D = (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_\xi^{-1}$. Accordingly, $y_i(t)^\top D^\top \mu_i$ is the BLUP of $x_i(t)^\top \mu_i$.*

So, BLUP is the best reward estimator even for a policy that has access to $A, \Sigma_\xi, \Sigma_x, \{\mu_i\}_{i \in [N]}$. This reflects the optimal policy to compete against, as will be discussed in detail shortly in Remark 2.

For ease of presentation, we set

$$\eta_i = D^\top \mu_i, \tag{4.4}$$

for $i \in [N]$, which is a *transformed parameter* corresponding to the model parameters one can (at best) hope to learn by using the partial observations of the contexts. The transformed parameter η_i represents the projected information of the original reward parameter μ_i that we can learn through the lens of $y_i(t)$. Thus, the optimal policy that fully knows D and μ_i for $i \in [N]$ and uses this knowledge to select the arm of highest expected reward, is

$$a^\star(t) = \text{argmax}_{i\in[N]}\ y_i(t)^\top \eta_i,$$

where $a^\star(t)$ is referred to as an optimal arm at time t.

Similarly to other problems in sequential decision-making, regret is the performance measure, which is the loss of cumulative reward compared to the optimal policy. So, at time T, the regret of the policy that pulls $a(t) \in [N]$ at round t is

$$\text{Regret}(T) = \sum_{t=1}^{T} \left(y_{a^\star(t)}(t)^\top \eta_{a^\star(t)} - y_{a(t)}(t)^\top \eta_{a(t)} \right).$$

Remark 2. *As above, we aim to compete against an optimal policy $a^\star(t)$ that knows A, Σ_ξ, Σ_x, as well as $\{\mu_i\}_{i\in[N]}$, which* all are unknown *to the bandit policy.*

However, exact knowledge of contexts changes the setting essentially and nullifies the problem, because the regret with respect to such policy, *cannot* grow sublinearly with time. This relies on the fact that even with fully known reward parameters, bandit policies might select sub-optimal arms due to their uncertainty about the contexts. So, since contexts vary with time, such suboptimal pulls persist as we proceed, causing a linear regret (with positive probability). Moreover, the above-mentioned optimal policy aligns with the existing literature of partially observed contextual bandits (Jose & Moothedath, 2024; Kim et al., 2023).

Now, we describe the assumptions for the upcoming theoretical analyses. Note that the bandit algorithm in Section 4.3 does not need knowledge of the quantities introduced below. First, we define exhaustive and exclusive events in the observation space that correspond to each arm being optimal.

Definition 1 (Optimality Regions). *Concatenate the observations in $y(t) = \left(y_1(t)^\top, \dots, y_N(t)^\top\right)^\top$ and let $A_i^\star \subset \mathbb{R}^{Nd_y}$ be the region in the space of $y(t)$ that makes arm i optimal. That is, as long as $y(t) \in A_i^\star$,* *it holds that $a^\star(t) = i$. Further, denote the optimality probability of arm i by*

$$p_i = \mathbb{P}(y(t) \in A_i^\star) = \mathbb{P}(a^\star(t) = i).$$

The assumption below states a margin condition and properly modifies a similar assumption in (Bastani et al., 2021) to the setting of partially observable contextual bandits.

Assumption 1 (Margin Condition). *Consider the normalized observation vectors $\dot{y}_i(t) = y_i(t)/\|y(t)\|$ for $i \in [N]$ and the transformed parameters $\{\eta_i\}_{i\in[N]}$ in (4.4). There is $C > 0$ such that for all $u > 0$ and all $i \in [N]$ of positive optimality probability p_i in Definition 1, it holds that*

$$\max_{j\in[N],j\neq i} \mathbb{P}\left(0 < \dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j \leq u \middle| y(t) \in A_i^\star\right) \leq Cu.$$

This expression bounds the conditional probability of $\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j$, which is the suboptimality gap of the j-th arm if arm i is optimal. The assumption states that optimal arms are highly likely to be distinguishable. More precisely, it expresses that the likelihood of suboptimality gaps being smaller than u is proportional to u. The above inequality holds, for example, if the sensing noise or the context vectors have bounded probability density functions all over their Euclidean spaces (Faradonbeh et al., 2018; Wong et al., 2020). By Assumption 1, for all $i,\ j \in [N]$, there exists a subset $A_i^\kappa \subseteq A_i^\star$ for $\kappa = 2/C > 0$ such that

$$\mathbb{P}(y(t) \in A_i^\kappa) > \frac{p_i}{2} \quad \text{and} \quad \mathbb{P}(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j > \kappa | y(t) \in A_i^\kappa) = 1. \tag{4.5}$$

Following the previous paragraph, (4.5) implies that κ is the minimum value that the suboptimality gap can have given the event $y(t) \in A_i^\kappa$, which happens with probability 1/2 given $y(t) \in A_i^\star$. Here, κ is an instance-dependent constant, which is referred to as a *probabilistic suboptimality gap* in this chapter, as it is probabilistically satisfied. Moreover, the role of κ in the analysis of Algorithm 3 for partially observable contextual bandits of this work, is intrinsically similar to the role of the well-known *gap* in multi-armed bandits[1] (Lattimore & Szepesvári, 2020). However, this differs from the instance-dependent gap discussed in the literature on contextual bandits (Abbasi-Yadkori et al., 2011; Dani et al., 2008), where it serves as a positive hard threshold, above which the suboptimality gap consistently remains over time. Note that our policy in the next section, does *not* need any information about κ.

The assumption above is typical in the bandit literature. On the other hand, the next assumption is adopted for simplifying expressions in the probabilistic analysis of how the reward values are affected by the information lost in the sensing process (i.e., the imperfectness of context observations).

Assumption 2 (Sub-Gaussianity). *For the context vectors $x_i(t)$ and the corresponding observations $y_i(t)$ in* (4.3)*, there is a constant $R_2 > 0$ such that for all $\mu_i \in \mathbb{R}^{d_x}$, $\eta_i \in \mathbb{R}^{d_y}$, and $\lambda \in \mathbb{R}$, it holds that*

$$\max_{i\in[N]} \ \mathbb{E}\left[\exp\left(\lambda\left(x_i(t)^\top \mu_i - y_i(t)^\top \eta_i\right)\right) \middle| y_i(t)\right] \leq \exp\left(\frac{\lambda^2 R_2^2}{2}\right).$$

The above expression can be interpreted as conditional sub-Gaussianity of the error $x_i(t)^\top \mu_i - y_i(t)^\top \eta_i$, given the observation $y_i(t)$. Therefore, its role is similar to the degree of tail heaviness in finite sample analysis of estimation accuracy (Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013). This assumption holds for a general class of stochastic measurement errors and contexts, including Gaussian and bounded random vectors, and rules out heavy-tailed distributions and those that the covariance of $x_i(t)\big|y_i(t)$ grows with $\|y_i(t)\|$.

Remark 3. *Equivalents of the presented algorithms and results for perfectly observed contextual bandits can be obtained by simply letting $A = I_{d_x}$ and $\Sigma_\xi \to \mathbf{0}_{d_y \times d_y}$.*

[1] Technically, it is the difference between the expected reward of the best arm, and that of the second best arm.

4.3 Thompson Sampling with Partial Contextual Observations

In this section, we outline a version of the well-known Thompson sampling algorithm that can be implemented using *only* the observation vectors. Originially Thompson sampling is built on posterior distributions resulting from prior and reward distributions. However, it is recognized for its robust performance in absence of exact probability distributions and performs effectively if there exist mismatches between actual distributions and hypothetical ones.

Now, we introduce Thompson sampling for partially observable contextual bandits. It is noteworthy that this algorithm does *not* require any information about the parameters such as A, D, Σ_x, and Σ_ξ. In Section 4.2, we show that the best conditional prediction of the reward of the i-th arm given $y_i(t)$, is $y_i(t)^\top \eta_i$, regardless of all the probability distributions. Similarly, Thompson sampling learns to choose an optimal arm *as if* the rewards are generated from a Gaussian distribution with the variance v^2. The policy fixes the posterior dispersion parameter v^2 subject to $v^2 \geq R^2$, where

$$R^2 = R_1^2 + R_2^2, \tag{4.6}$$

and R_1 and R_2 are introduced in (4.2) and Assumption 2, respectively. Intuitively, the above constraint is to guarantee sufficient exploration.

The policy starts with the initial values $\widehat{\eta}_i(1) = \mathbf{0}_{d_y}$ and $B_i(1) = I_{d_y}$, which represent the mean and (unscaled) inverse covariance matrix of a hypothetical prior distribution of η_i for arm $i \in [N]$, respectively. Then, we sample from the following posterior belief about η_i:

$$\widetilde{\eta}_i(t) \sim \mathcal{N}(\widehat{\eta}_i(t), v^2 B_i(t)^{-1}), \;\; i = 1, 2 \ldots, N. \tag{4.7}$$

Algorithm 3 : Thompson sampling for contextual bandits with partial context observations

```
Set $B_i(1) = I_{d_y}$, $\widehat{\eta}_i(1) = \mathbf{0}_{d_y}$ for $i = 1, 2, \ldots, N$
for $t = 1, 2, \ldots,$ do
    for $i = 1, 2, \ldots, N$ do
        Sample $\widetilde{\eta}_i(t)$ from $\mathcal{N}(\widehat{\eta}_i(t), v^2 B_i(t)^{-1})$
    end for
    Select arm $a(t) = \text{argmax}_{i \in [N]} y_i(t)^\top \widetilde{\eta}_i(t)$
    Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_{a(t)} + \varepsilon_{a(t)}(t)$
    Update $B_i(t+1)$ and $\widehat{\eta}_i(t+1)$ by (4.8) and (4.9) for $i = 1, 2, \ldots, N$
end for
```

Subsequently, the policy pulls the arm $a(t)$ as if the samples above are the true values of $\{\eta_i\}_{i \in [N]}$;

$$a(t) = \underset{1 \leq i \leq N}{\text{argmax}}\, y_i(t)^\top \widetilde{\eta}_i(t).$$

Once the algorithm gains the reward of the chosen arm $a(t)$, it updates the posterior parameters:

$$B_i(t+1) = B_i(t) + y_i(t) y_i(t)^\top \mathbb{I}(a(t) = i), \tag{4.8}$$

$$\widehat{\eta}_i(t+1) = B_i(t+1)^{-1} \left(B_i(t) \widehat{\eta}_i(t) + y_i(t) r_{a(t)}(t) \mathbb{I}(a(t) = i) \right). \tag{4.9}$$

Note that based on the bandit feedback, the quantities are updated only for the chosen arm $a(t)$, and those of the other arms remain unaltered. The pseudo-code is provided in Algorithm 3.

4.4 Theoretical Performance Analyses

In this section, we establish the theoretical results of Algorithm 3 for partially observable contextual bandits with arm-specific parameters. The following results provide estimation error bounds of the estimators defined in (4.9) and a high-probability regret bound for Algorithm 3. It is worth noting that the accuracy of parameter estimation and regret growth are closely related because higher estimation accuracy leads to lower regret. Thus, we build the estimation accuracy first and then construct a regret bound based on it. The first theorem presents the estimation error bound, which scales with the rate of the inverse of the square root of t.

To proceed, we present auxiliary lemmas that serve as building blocks for the main results, Theorem 4 and 5. To begin, Lemma 9 establishes a truncation bound for the following steps of proofs. Next, Lemma 10, supported by Lemma 11, guarantees the sub-Gaussian tail property for the reward prediction error given an observation. Additionally, Lemma 12 demonstrates that the minimum eigenvalue of $B_i(t)$ grows linearly with the sample size $n_i(t)$ with a high probability. Furthermore, Lemma 13 is Azuma's inequality. Lastly, Lemma 14 and 15 provide upper bounds for the estimation error and sample bias, respectively.

First, we define the parameter space, where the norm of an element is bound. This bounded parameter space is commonly adapted in the antecedent literature (Bastani et al., 2021; Dani et al., 2008; Goldenshluger & Zeevi, 2013; Kargin et al., 2023).

Definition 2 (Parameter Bounds). *For the transformed reward parameters $\{\eta_i\}_{i\in[N]}$, there exists a positive constant c_η such that $\|\eta_i\| \leq c_\eta$, for all $i = 1, \ldots, N$.*

Note that according to the transformed parameter space above, a similar bound also holds for the parameters $\{\mu_i\}_{i\in[N]}$ so that their norms are bounded by a positive constant c_μ. This expresses that the unknown reward parameters live in an unknown bounded region. Intuitively, this enables us to control the effect of parameter sizes on regret growth.

Since each element of a context and observation noise is sub-Gaussian and the sum of two sub-Gaussian random variables is sub-Gaussian as well, based on (4.3), a positive number c_y exists such that

$$\mathbb{E}\left[e^{\lambda y_{ij}(t)}\right] \leq \exp\left(\frac{\lambda^2 c_y^2}{2}\right), \tag{4.10}$$

for all real $\lambda > 0$, where $y_{ij}(t)$ is the $j-$th element of $y_i(t)$. Next, we find a high-probability upper bound for the norm of observations for the following steps. To find the high-probability bound for $\|y_i(t)\|$ for a confidence level $\delta > 0$, we define W_T such that

$$W_T = \left\{ \max_{\{i\in[N],\tau\in[T]\}} ||y_i(\tau)||_\infty \leq v_T(\delta) \right\}, \tag{4.11}$$

where $v_T(\delta) = c_y\sqrt{2\log(2TNd_y/\delta)}$. In the next lemma, we show that the event W_T happens with probability at least $1-\delta$.

Lemma 9. *For the event W_T defined in* (4.11)*, we have $\mathbb{P}(W_T) \geq 1-\delta$.*

Proof. By (4.10) and the properties of sub-Gaussian random variables,

$$\mathbb{P}\left(|y_{ij}(t)| \geq \varepsilon\right) \leq 2 \cdot e^{-\frac{\varepsilon^2}{2c_y^2}},$$

is satisfied for given $i, j \in [N]$. Accordingly, we have

$$\mathbb{P}\left(\|y_i(t)\|_\infty \geq \varepsilon\right) \leq 2d_y \cdot e^{-\frac{\varepsilon^2}{2c_y^2}}.$$

By taking the union of the events over time and arms, we get

$$\mathbb{P}\left(\max_{i\in[N],\tau\in[T]} \|y_i(t)\|_\infty \geq \varepsilon \right) \leq 2TNd_y \cdot e^{-\frac{\varepsilon^2}{2c_y^2}}$$

By plugging $c_y(2\log(2TNd_y/\delta))^{1/2}$ in ε, we have

$$\begin{aligned} &\mathbb{P}\left(\max_{i\in[N],\tau\in[T]} \|y_i(t)\|_\infty \geq c_y(2\log(2TNd_y/\delta))^{1/2} \right) \\ \leq\ & 2TNd_y \cdot \exp\left(-\frac{2c_y^2\log(2TNd_y/\delta)}{2c_y^2} \right) = \delta. \end{aligned}$$

Thus,

$$\mathbb{P}(W_T) \geq 1 - \mathbb{P}\left(\max_{i\in[N],\tau\in[T]} \|y_i(t)\| \geq v_T(\delta)\right) \geq 1 - \delta.$$

□

By Lemma 9, we have a positive constant L such that

$$\|y_i(t)\| \leq \sqrt{d_y} v_T(\delta) := L = \mathcal{O}\left(\sqrt{d_y \log(TNd_y/\delta)}\right), \tag{4.12}$$

for all $1 \leq i \leq N$ and $1 \leq t \leq T$ with probability at least $1-\delta$.

The next lemma presents that reward prediction errors given observations have the sub-Gaussian property when observations and rewards have sub-Gaussian distributions, and thereby, a confidence ellipsoid is constructed for the estimator in (4.9). This result is built on Theorem 1 in the work of Abbasi-Yadkori et al., 2011 with proper modifications.

Lemma 10. *Let $w_t = r_{a(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}$ and $\mathscr{F}_{t-1} = \sigma\{\{y(\tau)\}_{\tau=1}^{t}, \{a(\tau)\}_{\tau=1}^{t}, \{r_{a(\tau)}(\tau)\}_{\tau=1}^{t-1}\}$. Then, w_t is $\mathscr{F}_{t-1}$-measurable and conditionally R-sub-Gaussian for some $R > 0$ such that*

$$\mathbb{E}[e^{\nu w_t} | \mathscr{F}_{t-1}] \leq \exp\left(\frac{\nu^2 R^2}{2}\right).$$

In addition, for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\|\widehat{\eta}_i(t) - \eta_i\|_{B_i(t)} = \left\| \sum_{\tau=1}^{t-1} y_i(\tau) w_\tau \mathbb{I}(a(\tau) = i) \right\|_{B_i(t)^{-1}} \leq R\sqrt{d_y \log\left(\frac{1 + L^2 n_i(t)}{\delta}\right)} + c_\eta.$$

This lemma provides the sub-Gaussianity for the reward prediction error w_t given $y_i(t)$, and shows a self-normalized bound for a vector-valued martingale $\sum_{\tau=1}^{t-1} y_i(\tau) w_\tau \mathbb{I}(a(\tau) = i)$. The reward estimation error w_t can be decomposed into two parts. The one is the reward error $\varepsilon_i(t)$ given (4.1) due to the randomness of rewards. This error is created even if the context $x_i(t)$ is known. The other is the reward mean prediction error $x_i(t)^\top \mu_i - y_i(t)\eta_i$ caused by unknown contexts. The first step of the proof for this lemma is to show the sub-Gaussian property of w_t based on the decomposition. Next, using the sub-Gaussian property of reward prediction errors, we construct a confidence ellipsoid for the transformed reward estimator in (4.9) with some martingale techniques.

Proof. To show the sub-Gaussianity of w_t given the observation $y(t)$, we use the following decomposition of $r_i(t) - y_i(t)^\top D^\top \mu_i$:

$$r_i(t) - y_i(t)^\top D^\top \mu_i = (r_i(t) - x_i(t)^\top \mu_i) + (x_i(t)^\top \mu_i - y_i(t)^\top D^\top \mu_i). \tag{4.13}$$

The first and second terms on the RHS are R_1 and R_2-sub-Gaussian by (4.2) and Assumption 2, respectively. Because the two terms are independent of each other, we have

$$\begin{aligned} \mathbb{E}[e^{\nu(r_i(t)-y_i(t)^\top D^\top \mu_i)}|y(t)] &= \mathbb{E}[e^{\nu\varepsilon_i(t)}]\mathbb{E}[e^{\nu(x_i(t)^\top \mu_i - y_i(t)^\top D^\top \mu_i)}|y(t)] \\ &\leq \exp\left(-\frac{\nu^2 R_1^2}{2}\right)\exp\left(-\frac{\nu^2 R_2^2}{2}\right). \end{aligned}$$

Thus, we have

$$\mathbb{E}[e^{\nu(r_i(t)-y_i(t)^\top D^\top \mu_i)}|y(t)] \leq \exp\left(-\frac{\nu^2 R^2}{2}\right), \tag{4.14}$$

where $R^2 = R_1^2 + R_2^2$. Now, we construct a confidence ellipsoid of the transformed reward parameter based on the sub-Gaussian property of the reward prediction error.

Lemma 11. *Let*

$$D^{\eta}_{it} = \exp\left(\frac{(r_{a(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)})y_{a(t)}(t)^\top \eta_{a(t)}}{R} - \frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right)^{\mathbb{I}(a(t)=i)},$$

$M^{\eta}_{it} = \prod_{\tau=1}^{t} D^{\eta}_{i\tau}$ *and* $t^\star$ *be a stopping time. Then,* $\mathbb{E}[M^{\eta}_{it^\star}] \leq 1$.

Proof. First, we take the expected value of D^{η}_{it} conditioned on $\mathscr{F}_{t-1}$ and arrange it as follows:

$$\begin{aligned}
&\mathbb{E}[D^{\eta}_{it}|\mathscr{F}_{t-1}] \\
&= \mathbb{E}\left[\exp\left(\frac{(r_{a(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)})y_{a(t)}(t)^\top \eta_{a(t)}}{R} - \frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right)^{\mathbb{I}(a(t)=i)} \middle| y(t), a(t)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\zeta_{a(t)}(t) y_{a(t)}(t)^\top \eta_{a(t)}}{R}\right)^{\mathbb{I}(a(t)=i)} \middle| y(t), a(t)\right] \exp\left(-\frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right)^{\mathbb{I}(a(t)=i)}.
\end{aligned}$$

Then, by (4.14), we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\zeta_{a(t)}(t) y_{a(t)}(t)^\top \eta_{a(t)}}{R}\right)^{\mathbb{I}(a(t)=i)} \middle| y(t), a(t)\right] \exp\left(-\frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right)^{\mathbb{I}(a(t)=i)} \\
\leq \; &\left(\exp\left(\frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right) \exp\left(-\frac{1}{2}(y_{a(t)}(t)^\top \eta_{a(t)})^2\right)\right)^{\mathbb{I}(a(t)=i)} = 1.
\end{aligned}$$

Thus, we have

$$\mathbb{E}[M^{\eta}_{it}|\mathscr{F}_{t-1}] = \mathbb{E}[M^{\eta}_{i1} D^{\eta}_{i2} \cdots D^{\eta_i}_{i(t-1)} D^{\eta}_{it}|\mathscr{F}_{t-1}] = D^{\eta}_{1} \cdots D^{\eta}_{i(t-1)} \mathbb{E}[D^{\eta}_{it}|\mathscr{F}_{t-1}] \leq M^{\eta}_{i(t-1)},$$

showing that $\{M^{\eta}_{i\tau}\}_{\tau=1}^{\infty}$ is a supermartingale and accordingly

$$\mathbb{E}[M^{\eta}_{it}] = \mathbb{E}[\mathbb{E}[M^{\eta}_{it}|\mathscr{F}_{t-1}]] \leq \mathbb{E}[M^{\eta}_{i(t-1)}] \leq \cdots \leq \mathbb{E}[\mathbb{E}[D^{\eta}_{i1}|\mathscr{F}_{1}]] \leq 1.$$

Next, we examine the quantity $M_{it^\star}^\eta$. Since M_{it}^η is a nonnegative supermartingale, by Doob's martingale convergence theorems (Doob, 1953), M_{it}^η converges to a random variable, which is denoted by M_i^η. Let $Q_{it}^\eta = M_{i\min(t,t^\star)}^\eta$ be the stopping time version of $\{M_{it}^\eta\}_t$. Then, by Fatou's Lemma (Rudin et al., 1976), we have

$$\mathbb{E}[M_{it^\star}^\eta] = \mathbb{E}[\text{liminf}_{t\to\infty} Q_{it}^\eta] \leq \text{liminf}_{t\to\infty}\mathbb{E}[Q_{it}^\eta] \leq 1.$$

□

Now, we continue the proof of Lemma 10. Let ϕ_{η_i} be the probability density function of multivariate Gaussian distribution of η_i with the mean $\mathbf{0}_{d_y}$ and the covariance matrix $v^2 I_{d_y}$. By Lemma 9 in the work of Abbasi-Yadkori et al., 2011, we have

$$\mathbb{P}_{\phi_{\eta_i}}\left(\|S_{it^\star}\|^2_{B_i(t^\star)^{-1}} > 2R^2 \log\left(\frac{\det(B_i(t^\star))^{1/2}}{\delta}\right)\right) \leq \delta, \tag{4.15}$$

where $\mathbb{P}_{\phi_{\eta_i}}$ denotes the probability measure associated with ϕ_{η_i} representing the distribution of η_i and $S_{it} = \sum_{\tau=1}^{t-1} y_{a(\tau)}(\tau) w_\tau \mathbb{I}(a(\tau) = i)$. Lemma 11 and (4.15) are sufficient conditions for the use of Theorem 1 in the work of Abbasi-Yadkori et al., 2011, thus we get

$$\mathbb{P}_{\phi_{\eta_i}}\left(\exists t^\star < \infty \; s.t. \; \|S_{it^\star}\|^2_{B_i(t^\star)^{-1}} > 2R^2 \log\left(\frac{\det(B_i(t^\star))^{1/2}}{\delta}\right)\right) \leq \delta. \tag{4.16}$$

By Lemma 10 in the work of Abbasi-Yadkori et al., 2011, we have

$$\det(B_i(t)) \leq (1 + n_i(t)L^2/d_y)^{d_y},$$

and subsequently, we have

$$2\log\left(\frac{\det(B_i(t))^{1/2}}{\delta}\right) \leq d_y \log\left(\frac{1+L^2 n_i(t)}{\delta}\right).$$

Thus, with probability at least $1-\delta$, we get

$$\|S_{it}\|^2_{B_i(t)^{-1}} < R\sqrt{d_y \log\left(\frac{1+L^2 n_i(t)}{\delta}\right)} + c_\eta,$$

for all $t > 0$. Because S_{it} can be written as

$$S_{it} = \sum_{\tau=1:a(\tau)=i}^{t-1} y_{a(\tau)}(\tau)(r_{a(\tau)}(\tau) - y_{a(\tau)}(\tau)^\top \eta_{a(\tau)})\mathbb{I}(a(\tau)=i) = B_i(t)(\widehat{\eta}_i(t) - \eta_i),$$

we have

$$\|\widehat{\eta}_i(t) - \eta_i\|_{B_i(t)} = \|S_{it}\|_{B_i(t)^{-1}}.$$

Therefore, with probability of at least $1-\delta$, for all $t > 0$, we have

$$\|\widehat{\eta}_i(t) - \eta_i\|_{B_i(t)} \leq R\sqrt{d_y \log\left(\frac{1+L^2 n_i(t)}{\delta}\right)} + c_\eta,$$

which is a similar result to Theorem 2 in the work of Abbasi-Yadkori et al., 2011. □

Lemma 10, together with Lemma 12 and 14, provides theoretical foundations for the square-root estimation accuracy, which will be showcased in Theorem 4. The next lemma guarantees the linear growth of eigenvalues of covariance matrices $\{B_i(t)\}_{i\in[N]}$ defined in (4.8) with respect to the number of samples of each arm.

Lemma 12. *Let $n_i(t)$ be the count of i-th arm chosen up to the time t. For $B_i(t)$ in* (4.8)*, with probability at least $1-\delta$, if $\nu_{(1)} \leq n_i(t) \leq T$, we have*

$$\lambda_{\max}\left(B_i(t)^{-1}\right) \leq \frac{2}{\lambda_m} n_i(t)^{-1},$$

where $\nu_{(1)} = 8L^4 \log(TN/\delta)/\lambda_m^2$.

Proof. We investigate the (unscaled) inverse covariance matrix $B_i(t)$, whose eigenvalues are closely related to estimation accuracy. It is worth noting that the matrix $B_i(t)$ is the sum of mutually dependent rank 1 matrices. Due to the dependence of the matrices, classical techniques for independent random variables cannot be applied to them. To address this issue, we construct a martingale sequence and use the next lemma (Azuma's inequality), which provides a high-probability bound for a sum of martingale sequences.

Lemma 13. *(Azuma's Inequality) Consider the sequence $\{X_t\}_{1\leq t\leq T}$ random variables adapted to some filtration $\{\mathscr{G}_t\}_{1\leq t\leq T}$, such that $\mathbb{E}[X_t|\mathscr{G}_{t-1}] = 0$. Assume that there is a deterministic sequence $\{c_t\}_{1\leq t\leq T}$ that satisfies $X_t^2 \leq c_t^2$, almost surely. Let $\sigma^2 = \sum_{t=1}^{T} c_t^2$. Then, for all $\varepsilon \geq 0$, it holds that*

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t \geq \varepsilon\right) \leq e^{-\varepsilon^2/2\sigma^2}.$$

The proof of Lemma 13 is provided in the work of Azuma, 1967. We use the above lemma and construct a martingale via its difference sequence. Then, we establish a lower bound for the smallest eigenvalue of $B_i(t)$, which we show is crucial in the analysis of the worst-case estimation error. Let the sigma-field generated by the contexts and chosen arms up to time t be

$$\mathscr{G}_{t-1} = \sigma\{\{x_i(\tau)\}_{\tau\in[t],i\in[N]}, \{a(\tau)\}_{\tau\in[t]}\}.$$

Consider $V_t^i = y_{a(t)}(t)y_{a(t)}(t)^\top \mathbb{I}(a(t)=i)$ in order to study the behavior of $B_i(t)$. Since

$$\begin{aligned}\mathrm{Var}(y_i(t)|\mathscr{G}_{t-1}) &= \mathbb{E}[y_i(t)y_i(t)^\top|\mathscr{G}_{t-1}] - \mathbb{E}[y_i(t)|\mathscr{G}_{t-1}]\mathbb{E}[y_i(t)|\mathscr{G}_{t-1}]^\top \\ &= \mathbb{E}[V_t|\mathscr{G}_{t-1}] - Ax_i(t)x_i(t)^\top A^\top,\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}[V_t^i|\mathscr{G}_{t-1}] &= \left(\mathrm{Var}(y_i(t)|\mathscr{G}_{t-1}) + Ax_i(t)x_i(t)^\top A^\top\right)\mathbb{I}(a(t)=i) \\ &\succeq \Sigma_\xi \mathbb{I}(a(t)=i) \succeq \lambda_m I_{d_y}\mathbb{I}(a(t)=i), \qquad (4.17)\end{aligned}$$

where $M_1 \succeq M_2$ for square matrices M_1 and M_1 represents that $M_1 - M_2$ is a semi-positive definite matrix and $\lambda_m = \lambda_{\min}(\Sigma_\xi)$, i.e., for all $t > 0$ and $\|z\| = 1$, it holds that

$$z^\top \left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau^i|\mathscr{G}_{\tau-1}]\right) z \geq \lambda_m n_i(t). \qquad (4.18)$$

Now, we focus on a high-probability lower bound for the smallest eigenvalue of $B_i(t)$. To proceed, define the martingale difference X_t^i and martingale Y_t^i such that

$$\begin{aligned}X_t^i &= V_t^i - \mathbb{E}[V_t|\mathscr{G}_{t-1}], \qquad (4.19) \\ Y_t^i &= \sum_{\tau=1}^{t} \left(V_\tau^i - \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}]\right). \qquad (4.20)\end{aligned}$$

Then, $X_t^i = Y_t^i - Y_{t-1}^i$ and $\mathbb{E}\left[X_t^i|\mathscr{G}_{t-1}\right] = 0$. Thus, $z^\top X_t^i z$ is a martingale difference sequence. Here, we are interested in the minimum eigenvalue of $\sum_{\tau=1}^{t-1} V_\tau^i$. Because $(z^\top X_t^i z)^2 \leq \|y_i(t)\|^4 \mathbb{I}(a(t) = i) \leq L^4 \mathbb{I}(a(t) = i)$ and thereby $\sum_{\tau=1}^{t-1}\left(z^\top X_\tau^i z\right)^2 \leq n_i(t)L^4$, using Lemma 13, we get the following inequality

$$\mathbb{P}\left(z^\top\left(\sum_{\tau=1}^{t-1} X_\tau^i\right) z \leq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2n_i(t)L^4}\right),$$

for $\varepsilon \leq 0$. By plugging $n_i(t)\varepsilon$ into ε above, we have

$$\mathbb{P}\left(z^\top\left(\sum_{\tau=1}^{t-1} X_\tau^i\right) z \leq n_i(t)\varepsilon\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2L^4}\right) \tag{4.21}$$

for $\varepsilon \leq 0$. Because

$$z^\top\left(\sum_{\tau=1}^{t-1}\left(V_\tau^i - \mathbb{E}[V_\tau^i|\mathscr{G}_{\tau-1}]\right)\right) z \leq z^\top\left(\sum_{\tau=1}^{t-1}\left(V_\tau^i - \lambda_m I_{d_y}\mathbb{I}(a(\tau)=i)\right)\right) z$$

based on (4.17), we have the following inequality

$$\begin{aligned} &\mathbb{P}\left(z^\top\left(\sum_{\tau=1}^{t-1}\left(V_\tau^i - \mathbb{E}[V_\tau^i|\mathscr{G}_{\tau-1}]\right)\right) z \leq n_i(t)\varepsilon\right) \\ \geq\ &\mathbb{P}\left(z^\top\left(\sum_{\tau=1}^{t-1}\left(V_\tau^i - \lambda_m I_{d_y}\mathbb{I}(a(\tau)=i)\right)\right) z \leq n_i(t)\varepsilon\right). \end{aligned} \tag{4.22}$$

Putting (4.21) and (4.22) together, we obtain

$$\mathbb{P}\left(z^\top\left(\sum_{\tau=1}^{t-1} V_\tau^i\right) z \leq n_i(t)(\lambda_m+\varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2L^4}\right), \tag{4.23}$$

where $-\lambda_m \leq \varepsilon \leq 0$ is arbitrary. Because $z^\top B_i(t) z \geq z^\top\left(\sum_{\tau=1}^{t-1} V_\tau^i\right) z$ based on $B_i(t) = I_{d_y} + \sum_{\tau=1}^{t-1} V_\tau^i$, we have

$$\mathbb{P}\left(z^\top B_i(t) z \leq n_i(t)(\lambda_m+\varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2L^4}\right), \tag{4.24}$$

for $-\lambda_m \le \varepsilon \le 0$. By putting $\exp\left(-n_i(t)\varepsilon^2/(2L^4)\right) = \delta/(TN)$, (4.24) can be written as

$$z^\top B_i(t) z \ge n_i(t)\left(\lambda_m - \sqrt{\frac{2L^4}{n_i(t)}\log\frac{TN}{\delta}}\right), \tag{4.25}$$

for any $z \in \mathbb{R}^{d_y}$ such that $\|z\| = 1$ and all $1 \le t \le T$ with probability at least $1-\delta$. That is, we have

$$n_i(t)\left(\lambda_m - \sqrt{\frac{2L^4}{n_i(t)}\log\frac{TN}{\delta}}\right) \le \lambda_{\min}(B_i(t)),$$

because the inequality (4.25) is achieved for any $z \in \mathbb{R}^{d_y}$. If $n_i(t) \ge \nu_{(1)} := 8L^4\log(TN/\delta)/\lambda_m^2 = \mathcal{O}(L^4\log(TN/\delta))$, we have

$$\lambda_{\max}\left(B_i(t)^{-1}\right) \le \frac{2}{\lambda_m} n_i(t)^{-1}.$$

□

In the lemma above, the minimum sampling size $\nu_{(1)}$ is required to guarantee the linear growth of the eigenvalues of $B_i(t)$ based on (4.25). The next lemma shows that the estimate in (4.9) has the square-root estimation accuracy regarding $n_i(t)$.

Lemma 14. *Let $\widehat{\eta}_i(t)$ be the estimate in* (4.9)*. Then, if $\nu_{(1)} < n_i(t) \le T$, with probability at least $1-\delta$, for all $i \in [N]$, we have*

$$\|\widehat{\eta}_i(t) - \eta_i\| \le \sqrt{\frac{2}{\lambda_m}}\left(R\sqrt{d_y\log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) n_i(t)^{-1/2}.$$

Proof. First, it is given that

$$\|\widehat{\eta}_i(t) - \eta_i\|_{B_i(t)} = \|B_i(t)^{1/2}(\widehat{\eta}_i(t) - \eta_i)\| \leq R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta$$

by Lemma 10. Then, because $\sqrt{\lambda_m n_i(t)/2} \leq \lambda_{\min}(B_i(t)^{1/2})$ for $n_i(t) \geq \nu_{(1)}$ by Lemma 12, we have

$$\sqrt{\frac{\lambda_m n_i(t)}{2}}\|\widehat{\eta}_i(t) - \eta_i\| \leq \lambda_{\min}(B_i(t)^{1/2})\|\widehat{\eta}_i(t) - \eta_i\| \leq \|B_i(t)^{1/2}(\widehat{\eta}_i(t) - \eta_i)\|.$$

Therefore, putting the two inequalities above together, we have

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{2}{\lambda_m}}\left(R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) n_i(t)^{-1/2},$$

if $n_i(t) \geq \nu_{(1)}$. □

The next lemma provides an upper bound for the norm of sample bias, $\widetilde{\eta}_i(t) - \eta_i$, which is represented as the sum of the degree of exploration $\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)$ and estimation error $\widehat{\eta}_i(t) - \eta_i$. This lemma is used to find the bound for the contribution of sample bias to the regret growth. This lemma is built on the linear growth of eigenvalues of $B_i(t)$ along with the confidence ellipsoid of the estimates, $\{\widehat{\eta}_i(t)\}_{i\in[N]}$, in Lemma 10.

Lemma 15. *Consider $\widetilde{\eta}_i(t)$, a sample of the i-th arm in* (4.7). *Then, if $\nu_{(1)} < n_i(t) \leq T$, with probability at least $1-\delta$, for all $i \in [N]$, we have*

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{2}{\lambda_m}}\left(v\sqrt{2d_y \log\frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) n_i(t)^{-1/2}.$$

Proof. First, we consider the distribution of $\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)$. Note that we sample $\widetilde{\eta}_i(t)$ from $\mathcal{N}\left(\widehat{\eta}_i(t), v^2 B_i(t)^{-1}\right)$. Using $\mathbb{P}\left(\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| > \epsilon | B_i(t)\right) \leq \mathbb{P}\left(\sqrt{d_y} Z > \epsilon | B_i(t)\right)$ for $Z|B_i(t) \sim$

$\mathcal{N}\left(0, v^2 \lambda_{\max}(B_i(t)^{-1})\right)$, we have

$$\mathbb{P}\left(\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| > \epsilon | B_i(t)\right) < 2 \cdot \exp\left(-\frac{\epsilon^2}{2 d_y v^2 \lambda_{\max}(B_i(t)^{-1})}\right).$$

By putting $2 \cdot \exp\left(-\epsilon^2/(2v^2\lambda_{\max}(B_i(t)^{-1}))\right) = \delta/(TN)$, we have

$$\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| < v\sqrt{2 d_y \lambda_{\max}(B_i(t)^{-1}) \log \frac{2TN}{\delta}}.$$

If $n_i(t) > \nu_{(1)}$, by Lemma 12, we have $\lambda_{\max}(B_i(t)^{-1}) \leq \sqrt{2/(\lambda_m n_i(t))}$ and subsequently

$$\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| < v\sqrt{\frac{2}{\lambda_m}}\sqrt{2 d_y \log \frac{2TN}{\delta}} n_i(t)^{-1/2}.$$

Therefore, by putting the above inequality together with Lemma 14, for $\nu_{(1)} < n_i(t) \leq T$, we have

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{2}{\lambda_m}}\left(v\sqrt{2 d_y \log \frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1 + TL^2}{\delta}\right)} + c_\eta\right) n_i(t)^{-1/2}.$$

□

Now, we are ready to prove the following theorem that guarantees a square-root estimation accuracy of transformed parameters.

Theorem 4 (Estimation Accuracy). *For all arms $i \in [N]$ such that $p_i > 0$, let η_i and $\widehat{\eta}_i(t)$ be the true parameter in* (4.4) *and its estimate in* (4.9)*, respectively. Then, with probability at least $1 - \delta$, Algorithm 3 guarantees*

$$\|\widehat{\eta}_i(t) - \eta_i\|^2 = \mathcal{O}\left(\frac{R^2 d_y}{t\, p_i} \ \log \frac{TN d_y}{\delta}\right),$$

if $\tau_i^{(1)} < t \leq T$ *and* $v \geq R$*, where* R *is defined in* (4.6) *and* $\tau_i^{(1)} = \mathcal{O}(v^2 p_i^{-2} N d_y^{3.5} \kappa^{-5} \log^5(TNd_y/\delta))$ *is the minimum time the algorithm is run.*

The theorem above indicates that the estimation error bound decreases in t. This is not straightforward since the estimation accuracy of the parameter for the i-th arm generally increases with the sample size of the i-th arm, $n_i(t)$, instead of the overall horizon t. Further, R^2 reflects the total variance including reward and observation errors exacerbating the estimation accuracy. Lastly, the term p_i^{-1} implies that an arm with a larger p_i is more likely to be chosen, resulting in more pulls and a better estimation accuracy.

Proof sketch. To prove the theorem, we first find the high-probability upper bound for the observation vectors and employ martingale concentration inequalities. Next, leveraging sub-Gaussianity, we construct confidence ellipsoids for the transformed parameters via self-normalized martingales. Then, based on the martingale property of the eigenvalues of $B_i(t)$, we show that the minimum eigenvalue of it grows linearly with the *random* number of pulls of i-th arm, $n_i(t)$, which itself is proven to have a linear growth with time. In the proof process for the linear growth of $n_i(t)$, the minimum time $\tau_i^{(1)}$ is required to remove sub-linear terms of t. Putting together the above steps, we obtain the square-root consistency for the estimates of the transformed parameters. □

Before starting the proof, remind the constants described in the statement in Theorem 4. L is the bound for the ℓ_2-norm of observations. p_i is the probability of optimality of the i-th arm, as defined in Definition 1. κ is the minimum value of suboptimality gap with a positive probability (0.5) defined in (4.5).

Proof. First, we show that the number of selections of each arm scales linearly with a high probability. We utilize the inequality below to find a high-probability upper bound for $n_i(t)$.

$$n_i(t) \geq \sum_{\tau=1}^{t} \mathbb{I}(a(\tau) = i, A_{i\tau}^{\kappa}).$$

We construct a martingale sequence $\mathbb{I}(a(t) = i, A^{\kappa}_{it}) - \mathbb{P}(a(\tau) = i, A^{\kappa}_{it}|G^{\star}_{t-1})$ with respect to a filtration $\{G^{\star}_t\}_{t=1}^{\infty}$, where $G^{\star}_t = \sigma\{\{a(\tau)\}_{\tau=1}^{t}\}$ and $A^{\kappa}_{it} = \{y(t) \in A^{\kappa}_i\}$. By Azuma's inequality, we have

$$\sum_{\tau=1}^{t} \mathbb{I}(a(t) = i, A^{\kappa}_{it}) \geq -\sqrt{2t \log \delta^{-1}} + \sum_{\tau=1}^{t} \mathbb{P}(a(\tau) = i|G^{\star}_{\tau-1}, A^{\kappa}_{i\tau})\mathbb{P}(A^{\kappa}_{i\tau}). \tag{4.26}$$

Since $\mathbb{P}(a(t) = i|G^{\star}_{t-1}, A^{\kappa}_{it})$ can be written as $\mathbb{P}(a(t) = i|G^{\star}_{t-1}, A^{\kappa}_{it}) = 1 - \sum_{j \neq i} \mathbb{P}(a(t) = j|G^{\star}_{t-1}, A^{\kappa}_{it})$, we focus on an upper bound for $\sum_{\tau=1}^{t} \sum_{j \neq i} \mathbb{P}(a(\tau) = j|G^{\star}_{\tau-1}, A^{\kappa}_{i\tau})$. To proceed, we rewrite the probability $\mathbb{P}(a(t) = j|G^{\star}_{t-1}, A^{\kappa}_{it})$ as follows:

$$\begin{aligned}
& \mathbb{P}(a(t) = j|G^{\star}_{t-1}, A^{\kappa}_{it}) \\
= \ & \mathbb{P}(a(t) = j, E^1_{jt}, E^2_{jt}|G^{\star}_{t-1}, A^{\kappa}_{it}) + \mathbb{P}(a(t) = j, (E^1_{jt})^c, E^2_{jt}|G^{\star}_{t-1}, A^{\kappa}_{it}) \\
+ \ & \mathbb{P}(a(t) = j, (E^2_{jt})^c|G^{\star}_{t-1}, A^{\kappa}_{it}),
\end{aligned} \tag{4.27}$$

where $E^1_{jt} = \{y_j(t)^\top \widetilde{\eta}_j(t) < y_j(t)^\top \eta_j + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j)\}$ and $E^2_{jt} = \{y_j(t)^\top \widehat{\eta}_j(t) \leq y_j(t)^\top \eta_j + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j)\}$. Based on the decomposition above, we will show the upper bound for $\sum_{\tau=1}^{t} \mathbb{P}(a(\tau) = j|A^{\kappa}_{i\tau}, F^{\star}_{\tau-1})$ by establishing upper bounds of the above three terms in Lemmas 19, 20, and 21. Moving forward, we will find an upper bound for the first term in (4.27).

Lemma 16. *For all* $1 \leq t \leq T$ *and instantiations of* $F^{\star}_{t-1} = \sigma\{\{y(\tau)\}_{\tau=1}^{t}, \{a(\tau)\}_{\tau=1}^{t-1}, \{r_{a(\tau)}(\tau)\}_{\tau=1}^{t-1}\}$, *we have*

$$\mathbb{P}(a(t) = j, E^1_{jt}, E^2_{jt}|A^{\kappa}_{it}, F^{\star}_{t-1}) \leq \frac{1 - p_{ijt}}{p_{ijt}} \mathbb{P}(a(t) = i, E^1_{jt}, E^2_{jt}|A^{\kappa}_{it}, F^{\star}_{t-1}),$$

where $p_{ijt} = \mathbb{P}(y_i(t)^\top \widetilde{\eta}_i(t) > 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i)|A^{\kappa}_{it}, F^{\star}_{t-1})$.

Proof. We consider upper and lower bounds of the probabilities $\mathbb{P}\left(a(t) = j|A^{\kappa}_{it}, E^1_{jt}, F^{\star}_{t-1}\right)$ and $\mathbb{P}\left(a(t) = i|A^{\kappa}_{it}, E^1_{jt}, F^{\star}_{t-1}\right)$, respectively. First, we aim to find an upper bound for

$\mathbb{P}\left(a(t) = j | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right)$. Given E_{jt}^{1}, if arm j is selected, $y_k(t)^\top \widetilde{\eta}_k(t) \leq 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i)$ for all k including j. Using this fact, we get

$$\mathbb{P}\left(a(t) = j | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right) \leq \mathbb{P}\left(y_k(t)^\top \widetilde{\eta}_k(t) \leq 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i), \forall k | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right).$$

Since the sample of each arm is generated independently given $F_{t-1}^{\star}$, the term on the RHS above can be written as

$$\begin{aligned} &\mathbb{P}\left(y_k(t)^\top \widetilde{\eta}_k(t) \leq 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i), \forall k \neq i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right) \\ =\ &(1 - p_{ijt}) \cdot \mathbb{P}\left(y_k(t)^\top \widetilde{\eta}_k(t) \leq 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i), \forall k \neq i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right). \end{aligned} \tag{4.28}$$

Similarly, we have an upper bound for $\mathbb{P}\left(a(t) = i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right)$ as follows.

$$\begin{aligned} &\mathbb{P}\left(a(t) = i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right) \\ \geq &\mathbb{P}\left(y_i(t)^\top \widetilde{\eta}_i(t) > 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i) \geq y_k(t)^\top \widetilde{\eta}_k(t), \forall k \neq i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right) \\ = &p_{ijt} \cdot \mathbb{P}\left(y_j(t)^\top \widetilde{\eta}_k(t) \leq 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i), \forall k \neq i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right). \end{aligned} \tag{4.29}$$

Putting the two inequalities (4.28) and (4.29) together, we have

$$\mathbb{P}\left(a(t) = j | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right) \leq \frac{1 - p_{ijt}}{p_{ijt}} \mathbb{P}\left(a(t) = i | A_{it}^{\kappa}, E_{jt}^{1}, F_{t-1}^{\star}\right).$$

Since whether E_{jt}^{2} is true is determined by $F_{t-1}^{\star}$, we get

$$\mathbb{P}(a(t) = j, E_{jt}^{1}, E_{jt}^{2} | A_{it}^{\kappa}, F_{t-1}^{\star}) \leq \frac{1 - p_{ijt}}{p_{ijt}} \mathbb{P}(a(t) = i, E_{jt}^{1}, E_{jt}^{2} | A_{it}^{\kappa}, F_{t-1}^{\star}).$$

□

By Lemma 16, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(a(t)=j, E_{jt}^1, E_{jt}^2 | G_{t-1}^\star, A_{it}^\kappa) \\
=\ & \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(a(t)=j, E_{jt}^1, E_{jt}^2 | A_{it}^\kappa, F_{t-1}^\star) | G_{t-1}^\star, A_{it}^\kappa] \\
\leq\ & \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \mathbb{P}(a(t)=i, E_{jt}^1, E_{jt}^2 | A_{it}^\kappa, F_{t-1}^\star) \middle| G_{t-1}^\star, A_{it}^\kappa \right].
\end{aligned}$$

By simple calculation, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \mathbb{P}(a(t)=i, E_{jt}^1, E_{jt}^2 | A_{it}^\kappa, F_{t-1}^\star) \middle| G_{t-1}^\star, A_{it}^\kappa \right] \\
=\ & \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \mathbb{I}(a(t)=i, E_{jt}^1, E_{jt}^2) \middle| A_{it}^\kappa, F_{t-1}^\star \right] \middle| G_{t-1}^\star, A_{it}^\kappa \right] \\
=\ & \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \mathbb{I}(a(t)=i, E_{jt}^1, E_{jt}^2) \middle| G_{t-1}^\star, A_{it}^\kappa \right] \\
\leq\ & \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \middle| G_{t-1}^\star, A_{it}^\kappa \right].
\end{aligned}$$

Thus, we have

$$\sum_{t=1}^{T} \mathbb{P}(a(t)=j, E_{jt}^1, E_{jt}^2 | G_{t-1}^\star, A_{it}^\kappa) \leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}} \middle| G_{t-1}^\star, A_{it}^\kappa \right]. \tag{4.30}$$

The next lemma provides a lower and upper bound for probabilities about normal distribution, which will be used to find a lower bound for p_{ijt}.

Lemma 17. *For a Gaussian distributed random variable Z with mean m and variance σ^2, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi}z} e^{-z^2/2} \leq \mathbb{P}(|Z-m| > z\sigma) \leq \frac{1}{\sqrt{\pi}z} e^{-z^2/2}.$$

The lemma above is Lemma 5 in the work of Agrawal and Goyal, 2013, which can be derived from Formula 7.1.13 in the work of Abramowitz and Stegun, 1964. The next lemma suggests an upper bound for the term on RHS in (4.30).

Lemma 18. *For p_{ijt} defined in Lemma 16 and A_{it}^{κ} in* (4.5)*, we have*

$$\mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}}\middle| G_{t-1}^{\star}, A_{it}^{\kappa}\right] \leq \frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_{\eta}\sqrt{1+L^2 n_i(t)}\right)+63.$$

Proof. First, we rewrite p_{ijt} as follows:

$$p_{ijt} = \mathbb{P}(y_i(t)^{\top}(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > y_i(t)^{\top}(\eta_i-\widehat{\eta}_i(t))+0.5(y_j(t)^{\top}\eta_j - y_i(t)^{\top}\eta_i)|A_{it}^{\kappa}, F_{t-1}^{\star}).$$

Let $\theta_{it} = y_i(t)^{\top}(\eta_i-\widehat{\eta}_i(t))$. Note that $y_i(t)^{\top}(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) \sim \mathcal{N}(0, v^2 y_i(t)^{\top}B_i(t)^{-1}y_i(t))$ given $y_i(t)$ and $B_i(t)$. Let $\Phi_{it}(\cdot)$ be the CDF of the normal distribution with mean 0 and variance $v^2 y_i(t)^{\top}B_i(t)^{-1}y_i(t)$. For ease of presentation, let $\kappa_{ijt} = -0.5(y_j(t)^{\top}\eta_j - y_i(t)^{\top}\eta_i)$ and $\sigma_{it}^2 = y_i(t)^{\top}B_i(t)^{-1}y_i(t)$. Then, we write $\mathbb{P}(y_i(t)^{\top}(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > \theta_{it}-\kappa_{ijt}|A_{it}^{\kappa}, F_{t-1}^{\star}) = \Phi_{it}(-\theta+\kappa_{ijt})$. Since $\Phi_{it}(-\theta+\kappa_{ijt}) \geq (v\sigma_{it}/2\sqrt{\pi}(\theta-\kappa_{ijt}))\exp\left(-(\theta-\kappa_{ijt})^2/(2v^2\sigma_{it}^2)\right)$ for $\theta > \kappa_{ijt}+v\sigma_{it}$ by Lemma 17, we have

$$\begin{aligned} p_{ijt}^{-1} &= \Phi_{it}(-\theta+\kappa_{ijt})^{-1} \\ \leq\ & \mathbb{I}(\kappa_{ijt}+v\sigma_{it}<\theta)\frac{2\sqrt{\pi}(\theta-\kappa_{ijt})}{v\sigma_{it}}\exp\left(\frac{(\theta-\kappa_{ijt})^2}{2v^2\sigma_{it}^2}\right)+\mathbb{I}(\kappa_{ijt}+v\sigma_{it}\geq\theta)\Phi_{it}(v\sigma_{it})^{-1}. \end{aligned}$$

Thus, we get

$$\mathbb{E}[p_{ijt}^{-1}|A_{it}^{\kappa}, F_{t-1}^{\star}] \leq \int_{\kappa_{ijt}+v\sigma_{it}}^{\infty} \frac{2\sqrt{\pi}}{v\sigma_{it}} \left(1 + \frac{(\theta - \kappa_{ijt})^2}{v^2\sigma_{it}^2}\right) \exp\left(\frac{(\theta - \kappa_{ijt})^2}{2v^2\sigma_{it}^2}\right) S_{it}(\theta) d\theta + \Phi(-1)^{-1}, \tag{4.31}$$

where $\Phi(\cdot)$ is the CDF of standard normal distribution and S_{it} is the survival function of θ_{it}. Note that

$$y_i(t)^\top(\widehat{\eta}_i(t) - \eta_i) = y_i(t)^\top B_i(t)^{-1} \sum_{\tau=1:a(\tau)=i}^{t} y_a(\tau)(\tau)(r_a(\tau)(t) - y_{a(\tau)}(\tau)^\top \eta_{a(\tau)})$$

and

$$\begin{aligned} & \mathrm{Var}(y_i(t)^\top(\widehat{\eta}_i(t) - \eta_i)|\{y(\tau)\}_{1\leq\tau\leq t}, \{a(\tau)\}_{1\leq\tau\leq t-1}) \\ = \;\; & y_i(t)^\top B_i(t)^{-1} y_i(t) \mathrm{Var}(r_{a(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}|\{y(\tau)\}_{1\leq\tau\leq t}, \{a(\tau)\}_{1\leq\tau\leq t-1}) \\ \leq \;\; & y_i(t)^\top B_i(t)^{-1} y_i(t) R^2. \end{aligned}$$

Since $r_a(\tau)(t) - y_{a(\tau)}(\tau)^\top \eta_{a(\tau)}$ is R-sub-Gaussian by Lemma 10 and $v^2 \geq R^2$, we have

$$\begin{aligned} S_{\theta_{it}}(\theta) &= \mathbb{P}(y_i(t)^\top(\widehat{\eta}_i(t) - \eta_i) > \theta|\{y(\tau)\}_{1\leq\tau\leq t}, \{a(\tau)\}_{1\leq\tau\leq t-1}) \\ &\leq \exp\left(-\frac{\theta^2}{2y_i(t)^\top B_i(t)^{-1} y_i(t) v^2}\right), \end{aligned} \tag{4.32}$$

for $\theta > 0$. Then, we have

$$\begin{aligned}
&\int_{\kappa_{ijt}+v\sigma_{it}}^{\infty} \frac{2\sqrt{\pi}}{v\sigma_{it}} \left(1+\frac{(\theta-\kappa_{ijt})^2}{v^2\sigma_{it}^2}\right) \exp\left(\frac{(\theta-\kappa_{ijt})^2}{2v^2\sigma_{it}^2}\right) S_{it}(\theta) d\theta \\
\leq\ & \int_{\kappa_{ijt}+v\sigma_{it}}^{\infty} \frac{2\sqrt{\pi}}{v\sigma_{it}} \left(1+\frac{(\theta-\kappa_{ijt})^2}{v^2\sigma_{it}^2}\right) \exp\left(\frac{(\theta-\kappa_{ijt})^2}{2v^2\sigma_{it}^2}\right) \exp\left(-\frac{\theta^2}{2v^2\sigma_{it}^2}\right) d\theta \\
=\ & \int_{\kappa_{ijt}+v\sigma_{it}}^{\infty} \frac{2\sqrt{\pi}}{v\sigma_{it}} \left(1+\frac{(\theta-\kappa_{ijt})^2}{v^2\sigma_{it}^2}\right) \exp\left(\frac{-2\kappa_{ijt}\theta+\kappa_{ijt}^2}{2v^2\sigma_{it}^2}\right) d\theta
\end{aligned}$$

Using the first and second moments of the shifted exponential distribution with the scale parameter $v^2\sigma_{it}^2/\kappa_{ijt}$ and location parameter $\kappa_{ijt}/2$, we have

$$\begin{aligned}
&\frac{2\sqrt{\pi}}{v\sigma_{it}} \int_{\kappa_{ijt}+v\sigma_{it}}^{\infty} \left(1+\frac{(\theta-\kappa_{ijt})^2}{v^2\sigma_{it}^2}\right) \exp\left(\frac{-2\kappa_{ijt}(\theta-(\kappa_{ijt}/2))}{2v^2\sigma_{it}^2}\right) d\theta \\
\leq\ & \frac{2\sqrt{\pi}v\sigma_{it}}{\kappa_{ijt}} \int_{\kappa_{ijt}/2}^{\infty} \left(\frac{(\theta-\kappa_{ijt}/2)^2-\kappa_{ijt}(\theta-\kappa_{ijt}/2)+\kappa_{ijt}^2/4+v^2\sigma_{it}^2}{v^2\sigma_{it}^2}\right) \times \\
& \qquad \frac{1}{v^2\sigma_{it}^2/\kappa_{ijt}} \exp\left(\frac{-(\theta-(\kappa_{ijt}/2))}{v^2\sigma_{it}^2/\kappa_{ijt}}\right) d\theta \\
=\ & \frac{2\sqrt{\pi}}{\kappa_{ijt}v\sigma_{it}} (2(v^2\sigma_{it}^2/\kappa_{ijt})^2-\kappa_{ijt}(v^2\sigma_{it}^2/\kappa_{ijt})+\kappa_{ijt}^2/4+v^2\sigma_{it}^2). \qquad (4.33)
\end{aligned}$$

Note that $\sigma_{it}^2 = y_i(t)^\top B_i(t)^{-1} y_i(t)$ and $\kappa_{ijt} = 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j)$ given A_{it}^κ. The term $\sigma_{it}^3/\kappa_{ijt}^3$ can be bounded as follows:

$$\frac{\sigma_{it}^3}{\kappa_{ijt}^3} = \frac{(y_i(t)^\top B_i(t)^{-1} y_i(t))^{3/2}/\|y(t)\|^3}{(0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j))^3/\|y(t)\|^3} \leq \frac{8}{\kappa^3}, \qquad (4.34)$$

because $\lambda_{\max}(B_i(t)^{-1}) \leq 1$ and $(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j)/\|y(t)\| \geq \kappa$ given A_{it}^κ by Assumption 1. In addition, because $1/(1+L^2 n_i(t)) \leq \lambda_{\min}(B_i(t)^{-1})$ for all t and $\|\eta_i\| \leq c_\eta$ for all i by Definition 2, we have

$$\frac{\kappa_{ijt}}{\sigma_{it}} = \frac{(0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j))/\|y(t)\|}{(y_i(t)^\top B_i(t)^{-1} y_i(t))^{0.5}/\|y(t)\|} \leq 2c_\eta \sqrt{1 + L^2 n_i(t)}. \tag{4.35}$$

Then, putting (4.33) together with (4.34) and (4.35), we have

$$\begin{aligned} & \frac{2\sqrt{\pi}}{\kappa_{ijt} v \sigma_{it}} (2(v^2\sigma_{it}^2/\kappa_{ijt})^2 - \kappa_{ijt}(v^2\sigma_{it}^2/\kappa_{ijt}) + \kappa_{ijt}^2/4 + v^2\sigma_{it}^2) \\ = \quad & \frac{2\sqrt{\pi}}{v}\left(2v^4 \frac{\sigma_{it}^3}{\kappa_{ijt}^3} + \frac{\kappa_{ijt}}{4\sigma_{it}}\right) \leq \frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3} + \frac{1}{2} c_\eta \sqrt{1 + L^2 n_i(t)}\right). \end{aligned}$$

By (4.31) and the intermediate result above, we get

$$\mathbb{E}\left[\frac{1 - p_{ijt}}{p_{ijt}} \middle| G_{t-1}^\star, A_{it}^\kappa\right] \leq \frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3} + \frac{1}{2} c_\eta \sqrt{1 + L^2 n_i(t)}\right) + 63, \tag{4.36}$$

because $\Phi(-1) > 1/8$. □

Lemma 19. *For the events E_{jt}^1 and E_{jt}^2 defined in* (4.27) *and A_{it}^κ in* (4.5)*, we have*

$$\begin{aligned} & \sum_{t=1}^T \mathbb{P}(a(t) = j, E_{jt}^1, E_{jt}^2 | G_{t-1}^\star, A_{it}^\kappa) \\ & \qquad \leq \nu_{(2)} \left((2\sqrt{\pi})/v \left(16v^4/\kappa^3 + 0.5c_\eta \sqrt{1 + L^2 \nu_{(2)}}\right) + 63\right) + 4, \end{aligned}$$

where $\nu_{(2)} = \max(\nu_{(1)}, 64v^2/(\kappa^2\lambda_m)\log T)$.

Proof. We rewrite p_{ijt} to decompose the components of it into independent terms as follows:

$$\begin{aligned} p_{ijt} &= \mathbb{P}(y_i(t)^\top \widetilde{\eta}_i(t) > 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i) | A_{it}^\kappa, F_{t-1}^\star) \\ &= \mathbb{P}(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -y_i(t)^\top (\widehat{\eta}_i(t) - \eta_i) - 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | A_{it}^\kappa, F_{t-1}^\star). \end{aligned}$$

Let $A_{it}^{\eta} = \{|y_i(t)^\top(\widehat{\eta}_i(t) - \eta_i)| < (1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)\}$, which represents an event where the estimator of a transformed parameter is close to the parameter. Then, we have

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -y_i(t)^\top(\widehat{\eta}_i(t) - \eta_i) - \frac{1}{2}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)\middle| A_{it}^{\kappa}, F_{t-1}^{\star}, A_{it}^{\eta}\right) \times \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbb{P}(A_{it}^{\eta}|A_{it}^{\kappa}, F_{t-1}^{\star}) \\
&\geq \mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -(1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)|A_{it}^{\eta}, A_{it}^{\kappa}, F_{t-1}^{\star})\mathbb{P}(A_{it}^{\eta}|A_{it}^{\kappa}, F_{t-1}^{\star}) \\
&\geq \left(1 - \exp\left(-\frac{((1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j))^2}{2v^2\sigma_{it}^2}\right)\right)\mathbb{P}(A_{it}^{\eta}|A_{it}^{\kappa}, F_{t-1}^{\star}),
\end{aligned}$$

using $y_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) \sim \mathcal{N}(0, v^2\sigma_{it}^2)$. By Lemma 12, if $n_i(t) > \nu_{(1)}$, we have

$$\sigma_{it}^2/\|y(t)\|^2 = y_i(t)^\top B_i(t)^{-1}y_i(t)/\|y(t)\|^2 \leq \frac{2}{\lambda_m}n_i(t)^{-1} \tag{4.37}$$

In addition, if $n_i(t) > \nu_{(2)} := \max(\nu_{(1)}, 64v^2/(\kappa^2\lambda_m)\log T) = \mathcal{O}\left(\kappa^{-2}L^4\log(TN/\delta)\right)$, we have

$$\frac{((1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)/\|y(t)\|)^2}{2v^2\sigma_{it}^2/\|y(t)\|^2} \geq \frac{\lambda_m n_i(t)\kappa^2}{64v^2} \geq \log T, \tag{4.38}$$

and thereby we have $\exp\left(-((1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j))^2/(2v^2\sigma_{it}^2)\right) \leq T^{-1}$. Accordingly, if $n_i(t) > \nu_{(2)}$, we get

$$\left(1 - \exp\left(-\frac{((1/4)(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j))^2}{2v^2\sigma_{it}^2}\right)\right)\mathbb{P}(A_{it}^{\eta}) \geq \left(1 - \frac{1}{T}\right)\mathbb{P}(A_{it}^{\eta}|A_{it}^{\kappa}, F_{t-1}^{\star}).$$

Thus, we get

$$\mathbb{E}\left[\frac{1}{p_{ijt}}\middle| G^{\star}_{t-1}, A^{\kappa}_{it}\right] - 1 \leq \frac{1}{\left(1-\frac{1}{T}\right)\mathbb{P}(A^{\eta}_{it}|A^{\kappa}_{it}, F^{\star}_{t-1})} - 1,$$

for $n_i(t) > \nu_{(2)}$. Note that

$$\begin{aligned}\mathbb{P}(A^{\eta}_{it}|A^{\kappa}_{it}, F^{\star}_{t-1}) &= \mathbb{P}(|y_i(t)^{\top}(\widehat{\eta}_i(t) - \eta_i)| < (1/4)(y_i(t)^{\top}\eta_i - y_j(t)^{\top}\eta_j)|F^{\star}_{t-1}) \\ &\leq 1 - \exp\left(-\frac{((1/4)\kappa)^2}{2v^2\sigma^2_{it}}\right).\end{aligned}$$

Since $\mathbb{P}(A^{\eta}_{it}|A^{\kappa}_{it}, F^{\star}_{t-1}) > 1 - T^{-1}$ for $n_i(t) > \nu_{(2)}$ by (4.38), we have

$$\mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}}\middle| G^{\star}_{t-1}, A^{\kappa}_{it}\right] \leq \frac{1}{\left(1-\frac{1}{T}\right)^2} - 1 \leq \frac{4}{T}. \tag{4.39}$$

Thus, putting (4.36) and (4.39) together, we have

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{P}(a(t) = j, E^1_{jt}, E^2_{jt}|G^{\star}_{t-1}, A^{\kappa}_{it}) \\ &\leq \sum_{t:n_i(t)\leq\nu_{(2)}}\mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}}\mathbb{I}(a(t)=i)\middle| A^{\kappa}_{it}, F^{\star}_{t-1}\right] + \sum_{t:n_i(t)>\nu_{(2)}}\mathbb{E}\left[\frac{1-p_{ijt}}{p_{ijt}}\mathbb{I}(a(t)=i)\middle| A^{\kappa}_{it}, F^{\star}_{t-1}\right] \\ &\leq \nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3} + \frac{1}{2}c_{\eta}\sqrt{1+L^2\nu_{(2)}}\right) + 63\right) + 4.\end{aligned}$$

□

We showed an upper bound for the first term in (4.27). Now, we aim to establish an upper bound for the second term in (4.27).

Lemma 20. *For the events E_{jt}^1 and E_{jt}^2 defined in* (4.27) *and A_{it}^{κ} in* (4.5)*, for $t \in [T]$ with probability at least $1-\delta$, we have*

$$\sum_{\tau=1}^{t} \mathbb{P}\left(a(\tau)=j, (E_{j\tau}^1)^c, E_{j\tau}^2 \middle| G_{t-1}^{\star}, A_{it}^{\kappa}\right) \le \nu_{(2)} + 2.$$

Proof. To start, we decompose the summation of $\mathbb{P}(a(\tau)=j, (E_{j\tau}^1)^c, E_{j\tau}^2 | G_{\tau-1}^{\star}, A_{i\tau}^{\kappa})$ into two based on the sample size $n_j(t)$ as follows:

$$\begin{aligned}
&\sum_{\tau=1}^{t} \mathbb{P}(a(\tau)=j, (E_{j\tau}^1)^c, E_{j\tau}^2 | G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}) \\
&= \sum_{\tau=1}^{t} \mathbb{E}[\mathbb{I}(a(\tau)=j, n_j(\tau) < \nu_{(2)}, (E_{j\tau}^1)^c, E_{j\tau}^2) \\
&\qquad\qquad + \mathbb{I}(a(\tau)=j, n_j(\tau) \ge \nu_{(2)}, (E_{j\tau}^1)^c, E_{j\tau}^2) | G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}] \\
&\le \nu_{(2)} + \sum_{\tau: n_j(\tau) \ge \lceil \nu_{(2)} \rceil}^{t} \mathbb{E}[\mathbb{E}[\mathbb{I}(a(\tau)=j, n_j(\tau) \ge \nu_{(2)}, (E_{j\tau}^1)^c, E_{j\tau}^2) | A_{i\tau}^{\kappa}, F_{\tau-1}^{\star}] | G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}]. \qquad (4.40)
\end{aligned}$$

Now, we investigate the case with $n_j(t) \ge \nu_{(2)}$. We consider $\mathbb{P}\left((E_{jt}^1)^c \middle| n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}, F_{t-1}^{\star}\right)$ to find an upper bound for the second term in (4.40). To do so, we rewrite

$$\begin{aligned}
& \mathbb{P}\left((E_{jt}^1)^c \middle| n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}, F_{t-1}^{\star}\right) \\
= \; & \mathbb{P}\left((E_{jt}^1)^c, (E_{jt}^2)^c \middle| n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}, F_{t-1}^{\star}\right) + \mathbb{P}\left((E_{jt}^1)^c, E_{jt}^2 \middle| n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}, F_{t-1}^{\star}\right) \\
\le \; & \mathbb{P}\left((E_{jt}^2)^c \middle| F_{t-1}^{\star}, n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}\right) + \mathbb{P}\left((E_{jt}^1)^c, E_{jt}^2 \middle| F_{t-1}^{\star}, n_j(t) \ge \nu_{(2)}, A_{it}^{\kappa}\right).
\end{aligned}$$

By (4.32), (4.37), and (4.38), if $n_j(t) \ge \nu_{(2)}$, we have

$$\begin{aligned}
& \mathbb{P}\left((E_{jt}^2)^c, n_j(t) \geq \nu_{(2)} \middle| F_{t-1}^\star, A_{it}^\kappa \right) \\
\leq \ & \mathbb{P}(y_j(t)^\top \widehat{\eta}_j(t) > y_j(t)^\top \eta_j + (1/4)(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa) \\
\leq \ & \exp\left(-\frac{n_j(t)\lambda_m \kappa^2}{64 v^2} \right) \leq \frac{1}{T}.
\end{aligned} \tag{4.41}$$

Similarly, we have

$$\begin{aligned}
& \mathbb{P}((E_{jt}^1)^c, E_{jt}^2 | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa) \\
= \ & \mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) > 0.5(y_j(t)^\top \eta_j + y_i(t)^\top \eta_i), E_{jt}^2 | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa) \\
\leq \ & \mathbb{P}(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > (1/4)(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa) \\
\leq \ & \exp\left(-\frac{n_j(t)\lambda_m \kappa^2}{64 v^2} \right) \leq \frac{1}{T}.
\end{aligned} \tag{4.42}$$

Putting (4.41) and (4.42) together, we have

$$\begin{aligned}
& \mathbb{P}\left((E_{jt}^1)^c \middle| F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa \right) \\
\leq \ & \mathbb{P}\left((E_{jt}^2)^c | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa \right) + \mathbb{P}\left((E_{jt}^1)^c, E_{jt}^2 | F_{t-1}^\star, n_j(t) \geq \nu_{(2)}, A_{it}^\kappa \right) \\
\leq \ & \frac{2}{T}.
\end{aligned} \tag{4.43}$$

Since the part of summation for $n_i(t) < \nu_{(2)}$ is bounded by $\nu_{(2)}$, it suffices to show a bound for the other part. Based on the fact that whether $E_{j\tau}^2$ is true determined by $F_{t-1}^\star$, we have

$$\mathbb{E}[\mathbb{I}(a(\tau) = j, n_j(\tau) \geq \nu_{(2)}, (E_{j\tau}^1)^c, E_{j\tau}^2) | A_{i\tau}^\kappa, F_{\tau-1}^\star] = \mathbb{I}(E_{j\tau}^2)\mathbb{P}(a(\tau) = j, (E_{j\tau}^1)^c | A_{i\tau}^\kappa, F_{\tau-1}^\star).$$

Using the equation above, we have

$$\begin{aligned}
&\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}[\mathbb{E}[\mathbb{I}(a(\tau)=j, n_j(\tau)\geq\nu_{(2)}, (E_{j\tau}^1)^c, E_{j\tau}^2)|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star}]|G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}] \\
= &\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}[\mathbb{I}(E_{j\tau}^2)\mathbb{P}(a(\tau)=j, (E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star})|G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}].
\end{aligned}$$

Because $\mathbb{I}(E_{j\tau}^2) = \mathbb{I}(E_{j\tau}^2, n_j(\tau) \geq \nu_{(2)})$ given $n_j(\tau) \geq \lceil\nu_{(2)}\rceil$ and $\mathbb{P}(a(\tau) = j, (E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star}) \leq \mathbb{P}((E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star})$, we have

$$\begin{aligned}
&\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}[\mathbb{I}(E_{j\tau}^2)\mathbb{P}(a(\tau)=j, (E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star})|G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}] \\
\leq &\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}[\mathbb{I}(E_{j\tau}^2, n_j(\tau)\geq\nu_{(2)})\mathbb{P}((E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star})|G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}].
\end{aligned}$$

If $\tau \geq \nu_{(2)}$, by (4.43), we have $\mathbb{P}((E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star}) \leq 2/T$. Accordingly, we get

$$\begin{aligned}
&\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}[\mathbb{I}(E_{j\tau}^2, n_j(\tau)\geq\nu_{(2)})\mathbb{P}((E_{j\tau}^1)^c|A_{i\tau}^{\kappa}, F_{\tau-1}^{\star})|G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}] \\
\leq &\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}\left[\mathbb{I}(E_{j\tau}^2, n_j(\tau)\geq\nu_{(2)})\left(\frac{2}{T}\right)\middle| G_{\tau-1}^{\star}, A_{i\tau}^{\kappa}\right].
\end{aligned}$$

Because $\mathbb{I}(E_{j\tau}^2, n_j(\tau) \geq \nu_{(2)}) \leq 1$, we have

$$\begin{aligned}
& \sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} \mathbb{E}\left[\mathbb{I}(E^2_{j\tau}, n_j(\tau) \geq \nu_{(2)})\left(\frac{2}{T}\right)\middle| G^\star_{\tau-1}, A^\kappa_{i\tau}\right] \\
\leq \; & \frac{2}{T}\sum_{\tau:n_j(\tau)\geq\lceil\nu_{(2)}\rceil}^{t} 1 \\
\leq \; & 2.
\end{aligned}$$

Therefore, we have

$$\sum_{\tau=1}^{t} \mathbb{P}(a(\tau) = j, (E^1_{j\tau})^c, E^2_{j\tau} | G^\star_{\tau-1}, A^\kappa_{i\tau}) \leq \nu_{(2)} + 2.$$

□

Now, we show an upper bound for the sum of third term in (4.27).

Lemma 21. *For $t \in [T]$, with probability at least $1-\delta$, we have*

$$\sum_{\tau=1}^{t} \mathbb{P}(a(\tau) = j, (E^2_{j\tau})^c | G^\star_{\tau-1}, A^\kappa_{i\tau}) \leq \nu_{(2)} + 1.$$

Proof. We consider $\mathbb{E}[\mathbb{I}(a(t) = j, (E^2_{j\tau})^c)|G^\star_{t-1}, A^\kappa_{it}]$.

$$\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(a(t) = j, (E^2_{j\tau})^c)|G^\star_{t-1}, A^\kappa_{it}] \\
= \; & \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(a(t) = j, (E^2_{j\tau})^c, n_i(t) < \nu_{(2)}) + \mathbb{I}(a(t) = j, (E^2_{j\tau})^c, n_i(t) \geq \nu_{(2)})|G^\star_{t-1}, A^\kappa_{it}] \\
\leq \; & \nu_{(2)} + \sum_{t=\lceil\nu_{(2)}\rceil}^{T} \mathbb{P}(a(\tau) = j, (E^2_{j\tau})^c, n_i(t) \geq \nu_{(2)})|G^\star_{\tau-1}, A^\kappa_{i\tau})
\end{aligned}$$

By (4.41), we have

$$\begin{aligned}\mathbb{P}(a(\tau)=j,(E_{j\tau}^{2})^{c},n_{i}(t)\geq\nu_{(2)})|G_{\tau-1}^{\star},A_{i\tau}^{\kappa}) &\leq \mathbb{P}((E_{j\tau}^{2})^{c},n_{i}(t)\geq\nu_{(2)})|G_{\tau-1}^{\star},A_{i\tau}^{\kappa}) \\ &\leq \frac{1}{T}.\end{aligned}$$

Therefore, we have

$$\sum_{\tau=1}^{t}\mathbb{P}(a(\tau)=j,(E_{j\tau}^{2})^{c}|G_{\tau-1}^{\star},A_{i\tau}^{\kappa})\leq\nu_{(2)}+1.$$

□

Now, we are ready to show an upper bound for (4.27) with Lemma 19, 20, and 21.

Lemma 22. *For* $t\in[T]$*, with probability at least* $1-\delta$*, we have*

$$\begin{aligned}&\sum_{\tau=1}^{t}\mathbb{P}(a(\tau)=i|G_{t-1}^{\star},A_{it}^{\kappa})\mathbb{P}(A_{it}^{\kappa}) \\ &\qquad \geq \frac{p_i}{2}\left(t-N\left(\nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_\eta\sqrt{1+L^2\nu_{(2)}}\right)+65\right)+7\right)\right),\end{aligned}$$

where $\nu_{(2)}$ *is the sample size defined in Lemma 19.*

Proof. Note that $\mathbb{P}(a(t)=i|G_{t-1}^{\star},A_{it}^{\kappa})=1-\sum_{j\neq i}\mathbb{P}(a(t)=j|G_{t-1}^{\star},A_{it}^{\kappa})$. To find an upper bound for (4.27), we showed upper bounds of the three terms in (4.27) in Lemma 19, 20, and 21. Putting them together, we have

$$\sum_{t=1}^{t}\mathbb{P}(a(\tau)=j|G_{t-1}^{\star},A_{it}^{\kappa})\leq\nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_\eta\sqrt{1+L^2\nu_{(2)}}\right)+63\right)+4+2\nu_{(2)}+3.$$

By summing the probabilities above over all arms except for i, we have

$$\sum_{j\neq i}\mathbb{P}(a(\tau)=j|G^{\star}_{t-1},A^{\kappa}_{it})\leq N\left(\nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_\eta\sqrt{1+L^2\nu_{(2)}}\right)+65\right)+7\right).$$

Therefore, we get

$$\begin{aligned}\sum_{\tau=1}^{t}\mathbb{P}(a(\tau)=i|G^{\star}_{t-1},A^{\kappa}_{it})\mathbb{P}(A^{\kappa}_{it})&=\sum_{\tau=1}^{t}\left(1-\sum_{j\neq i}\mathbb{P}(a(\tau)=i|G^{\star}_{t-1},A^{\kappa}_{it})\right)\mathbb{P}(A^{\kappa}_{it})\\ \geq\ \frac{p_i}{2}\left(t-N\left(\nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_\eta\sqrt{1+L^2\nu_{(2)}}\right)+65\right)+7\right)\right).\end{aligned}$$

□

Lemma 23. *If $t>\tau_i^{(1)}$ for given i,*

$$n_i(t)\geq\frac{p_i}{4}t.$$

where the minimum time $\tau_i^{(1)}$ is

$$\tau_i^{(1)}=\max(4\ell(\delta,T,N,\kappa),(32/p_i^2)\log\delta^{-1})$$

and

$$\ell(\delta,T,N,\kappa)=N\left(\nu_{(2)}\left(\frac{2\sqrt{\pi}}{v}\left(\frac{16v^4}{\kappa^3}+\frac{1}{2}c_\eta\sqrt{1+L^2\nu_{(2)}}\right)+65\right)+7\right).$$

Proof. Consider a martingale sequence $\mathbb{I}(a(t)=i,A^{\kappa}_{it})-\mathbb{P}(a(t)=i,A^{\kappa}_{it}|G^{\star}_{t-1})$ with respect to the filtration $\{G^{\star}_{t-1}\}_{t=1}^{\infty}$ defined in . By Azuma's inequality, with probability at least $1-\delta$

$$\sum_{\tau=1}^{t}\mathbb{I}(a(t)=i,A^{\kappa}_{it})\geq-\sqrt{2t\log\delta^{-1}}+\sum_{\tau=1}^{t}\mathbb{P}(a(t)=i|G^{\star}_{t-1},A^{\kappa}_{it})\mathbb{P}(A^{\kappa}_{it}).$$

By Lemma 22, we have

$$\sum_{\tau=1}^{t} \mathbb{I}(a(t) = i, A_{it}^{\kappa})$$
$$\geq -\sqrt{2t \log \delta^{-1}} + \frac{p_i}{2} \left(t - N \left(\nu_{(2)} \left(\frac{2\sqrt{\pi}}{v} \left(\frac{16v^4}{\kappa^3} + \frac{1}{2} c_\eta \sqrt{1 + L^2 \nu_{(2)}} \right) + 65 \right) + 7 \right) \right).$$

For ease of presentation, let

$$\ell(\delta, T, N, \kappa) = N \left(\nu_{(2)} \left(\frac{2\sqrt{\pi}}{v} \left(\frac{16v^4}{\kappa^3} + \frac{1}{2} c_\eta \sqrt{1 + L^2 \nu_{(2)}} \right) + 65 \right) + 7 \right).$$

Because $\sqrt{2t \log \delta^{-1}} \leq (p_i t)/8$ for $t \geq (128/p_i^2) \log \delta^{-1}$ and $(p_i/2)\,(t - \ell(\delta, T, N, \kappa)) \geq 3p_i t/8$ for $t \geq 4\ell(\delta, T, N, \kappa)$, we have

$$-\sqrt{2t \log \delta^{-1}} + \frac{p_i}{2} (t - \ell(\delta, T, N, \kappa)) \geq \frac{p_i t}{4},$$

if $t \geq \tau_i^{(1)} := \max(4\ell(\delta, T, N, \kappa), (128/p_i^2) \log \delta^{-1}) = \mathcal{O}(p_i^{-2} N \kappa^{-5} L^7 \log^{1.5}(TNd_y/\delta))$. Therefore, if $t \geq \tau_i^{(1)}$, we have

$$n_i(t) \geq \frac{p_i t}{4}, \tag{4.44}$$

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

Now, we are ready to prove Theorem 4. From Lemma 10, we have

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq R \sqrt{\frac{2}{\lambda_m}} \left(\sqrt{d_y \log \left(\frac{1 + TL^2}{\delta} \right)} + c_\eta \right) n_i(t)^{-1/2}, \tag{4.45}$$

if $n_i(t) > \nu_{(1)}$. Since we have $n_i(t) \geq (p_i t)/4$ by (4.44) for $t > \tau_i^{(1)}$, we get

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq R\sqrt{\frac{8}{\lambda_m p_i}}\left(\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) t^{-1/2}.$$

if $n_i(t) > \nu_{(1)}$ and $t > \tau_i^{(1)}$. Thus, putting the two sample conditions together, if $t > \tau_i := \max(4p_i^{-1}\nu_{(1)}, \tau_i^{(1)})$, we have

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq R\sqrt{\frac{8}{\lambda_m p_i}}\left(\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) t^{-1/2}.$$

Thus, if $t > \tau_M := \max_{i\in[N]} \tau_i = \mathcal{O}(v^2 {p_{\min}^{+}}^{-2} N\kappa^{-5} L^7 \log^{1.5}(TNd_y/\delta))$, with probability at least $1-\delta$, we have the following estimation accuracy

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq R\sqrt{\frac{8}{\lambda_m p_i}}\left(\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) t^{-1/2}.$$

□

Corollary 2. *For all arms $i \in [N]$ such that $p_i > 0$, let $\widetilde{\eta}_i(t)$ be a sample generated by Algorithm 3. Then, with probability at least $1-\delta$, if $t > \tau_i$, Algorithm 3 satisfies*

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{8}{p_i\lambda_m}}\left(v\sqrt{2d_y \log\frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) t^{-1/2}.$$

Proof. By Lemma 15, with probability $1-\delta$, if $n_i(t) \geq \nu_{(1)}$, we have

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{2}{\lambda_m}}\left(v\sqrt{2d_y \log\frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) n_i(t)^{-1/2} \qquad (4.46)$$

As $n_i(t) > p_i t/4$ if $t \geq \tau_i$ with probability $1-\delta$ by (4.44), we have

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{8}{p_i \lambda_m}} \left(v\sqrt{2d_y \log \frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta \right) t^{-1/2}.$$

□

Built on Theorem 4, the next theorem expresses that the regret scales poly-logarithmically with time.

Theorem 5 (Regret Bound)**.** *Suppose that $v \geq R$ and let $p_{\min}^+ = \min_{i\in[N]:p_i>0} p_i$. The regret of Algorithm 3 satisfies the following with probability at least $1-\delta$:*

$$\mathrm{Regret}(T) = \mathcal{O}\left(\frac{v^2 N d_y^4}{{p_{\min}^+}^2 \kappa^5} \log^{5.5}\left(\frac{TNd_y}{\delta}\right)\right).$$

This theorem demonstrates that the regret scales at most $\log^{5.5} T$ with time and proportionally to at most inverse square with $p_{\min}^+$. In addition, the term N is caused by the derivation of the union bound of the events that suboptimal arms are chosen over time. Next, scaling d_y^4 is imposed by the high-probability magnitude $\mathcal{O}(\sqrt{d_y \log TNd_y/\delta})$ of sub-Gaussian observation vectors, which excludes cases, where regret increases due to scaling of the observation magnitude. Furthermore, the regret increases at rate κ^{-5} as the probabilistic suboptimality gap κ shrinks, because the minimum time τ_M is required for the application of Theorem 4. Finally, an excessive exploration expectedly exacerbates the regret, which corresponds to the posterior dispersion v^2 being overly large.

Proof sketch. For this proof outline, we focus on the effects of T, N, L, and τ_M. First, we show that the regret grows with $\log^2 T$ over time when T is greater than the minimum sample size $\tau_M = \mathcal{O}\left(L^7 N \log^{1.5}(TNd_y/\delta)\right)$. Note that regret is the sum of reward gaps, which is incurred when a suboptimal arm is chosen. For $t > \tau_M$, the reward gap is $\mathcal{O}\left(LNd_y t^{-1/2}\log(TNd_y/\delta)\right)$ and the probability of choosing a suboptimal arm decreases at rate $t^{-1/2}$, resulting in their product decreasing at rate t^{-1} over time. Accordingly, the sum of product terms diminishing as t^{-1} is $\mathcal{O}\left(LNd_y \log^2(TNd_y/\delta)\right)$. Then, we split the regret as:

$$\text{Regret}(T) = \sum_{t=1}^{\lfloor \tau_M \rfloor} \text{gap}(t)\mathbb{I}(a^\star(t) \neq a(t)) + \sum_{t=\lceil \tau_M \rceil}^{T} \text{gap}(t)\mathbb{I}(a^\star(t) \neq a(t)), \tag{4.47}$$

where $\text{gap}(t) = y_{a^\star(t)}(t)^\top \eta_{a^\star(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}(t)$. By the intermediate result above, we have a bound for the second term, which is $\mathcal{O}(LNd_y \log^2(TNd_y/\delta))$. As the first term is bounded by $\tau_M L \max_{i\in[N]} \|\eta_i\|$, which is $\mathcal{O}\left(L^8 N \log^{1.5}(TNd_y/\delta)\right)$, by applying $L = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$, we get the order of the first term $\mathcal{O}(Nd_y^4 \log^{5.5}(TNd_y/\delta))$, which dominates the second one. Consequently, we get the result of Theorem 5. □

Proof. Note that the regret can be written as

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{T} (y_{a^\star(t)}(t)^\top \eta_{a^\star(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}(t)) \\
&\leq 2c_\eta L \tau_M + \sum_{t=\lceil \tau_M \rceil}^{T} (y_{a^\star(t)}(t)^\top \eta_{a^\star(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}(t))\mathbb{I}(a^\star(t) \neq a(t)),
\end{aligned}$$

because $\|y_i(t)\| \leq L$ for all $i \in [N]$ and $t \in [T]$ and $\|\eta_i\| \leq c_\eta$ for all $i \in [N]$. The order of first term is $\mathcal{O}({p_{\min}^+}^{-2} NL^8 \kappa^{-5} \log^{1.5}(TNd_y/\delta))$. Now, we aim to show an upper bound for the second term. The second term can be written as

$$\begin{aligned}
&\sum_{t=\lceil \tau_M \rceil}^{T} (y_{a^\star(t)}(t)^\top \eta_{a^\star(t)}(t) - y_{a(t)}(t)^\top \eta_{a(t)}(t))\mathbb{I}(a^\star(t) \neq a(t)) \\
\leq &\sum_{t=\lceil \tau_M \rceil}^{T} (y_{a^\star(t)}(t)^\top (\eta_{a^\star(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)) - y_{a(t)}(t)^\top (\eta_{a(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)))\mathbb{I}(a^\star(t) \neq a(t)),
\end{aligned}$$

because $y_{a(t)}(t)^\top \widetilde{\eta}_{a(t)} - y_{a^\star(t)}(t)^\top \widetilde{\eta}_{a^\star(t)}(t) \geq 0$. Since $\|y_i(t)\| \leq L$ for all $t \in [T]$, we have

$$\begin{aligned}
&\sum_{t=\lceil \tau_M \rceil}^{T} (y_{a^\star(t)}(t)^\top(\eta_{a^\star(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)) - y_{a(t)}(t)^\top(\eta_{a(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)))\mathbb{I}(a^\star(t) \neq a(t)) \\
\leq \;& L \sum_{t=\lceil \tau_M \rceil}^{T} (\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\|)\mathbb{I}(a^\star(t) \neq a(t)).
\end{aligned}$$

By Corollary 2, if $t > \tau_M$, we have

$$\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\| \leq g^{(1)}(\delta) t^{-1/2},$$

where

$$\begin{aligned}
g^{(1)}(\delta) &= 2\sqrt{\frac{8}{p^+_{\min}\lambda_m}}\left(v\sqrt{2d_y \log\frac{2TN}{\delta}} + R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta\right) \\
&= \mathcal{O}\left(v{p^+_{\min}}^{-1/2}\sqrt{d_y \log(TNd_y/\delta)}\right).
\end{aligned}$$

Accordingly, the regret can be written as

$$\text{Regret}(T) \leq 2c_\eta L\tau_M + Lg^{(1)}(\delta)\sum_{t=\lceil \tau_M \rceil}^{T} t^{-1/2}\mathbb{I}(a^\star(t) \neq a(t)).$$

Thus, it suffices to show an upper bound for $\sum_{t=\lceil \tau_M \rceil}^{T} t^{-1/2}\mathbb{I}(a^\star(t) \neq a(t))$. To proceed, we apply Lemma 13 to find a high-probability upper-bound for the summation of the martingale difference sequence $\{t^{-0.5}\mathbb{I}(a^\star(t) \neq a(t)) - t^{-0.5}\mathbb{P}(a^\star(t) \neq a(t)|G^\star_{t-1})\}_{t=\lceil \tau_M \rceil}^{T}$ with respect to the filtration $\{G^\star_{t-1}\}_{t=1}^{\infty}$, and get the following inequality with probability at least $1-\delta$:

$$\sum_{t=\lceil \tau_M \rceil}^{T} \frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t) \neq a(t)) \leq \sqrt{4\log T \log \delta^{-1}} + \sum_{t=\lceil \tau_M \rceil}^{T} \frac{1}{\sqrt{t}}\mathbb{P}(a^\star(t) \neq a(t)). \tag{4.48}$$

To find a bound for $\mathbb{P}(a^\star(t) \neq a(t))$, we consider $\mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t)) > 0 | A^\star_{it})$. With

$$\{y_i(t)^\top \widetilde{\eta}_i(t) < y_j(t)^\top \widetilde{\eta}_j(t)\} \subset$$
$$\left\{ y_j(t)^\top (\widetilde{\eta}_j(t) - \eta_j) > \frac{1}{2}(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \right\} \cup \left\{ y_i(t)^\top (\widetilde{\eta}_i(t) - \eta_i) < -\frac{1}{2}(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \right\},$$

we decompose the following probability as follows:

$$\begin{aligned}
&\mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t)) > 0 | G^\star_{t-1}, A^\star_{it}) \\
&\leq \ \mathbb{P}(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -y_i(t)^\top (\widehat{\eta}_i(t) - \eta_i) + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | G^\star_{t-1}, A^\star_{it}) \\
&+ \ \mathbb{P}(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y_j(t)^\top (\widehat{\eta}_j(t) - \eta_j) + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | G^\star_{t-1}, A^\star_{it}).
\end{aligned} \tag{4.49}$$

By Theorem 4, with probability of at least $1 - \delta$, we have

$$y_i(t)^\top (\widehat{\eta}_i(t) - \eta_i) \leq \frac{h(\delta, T)\|y(t)\|}{t^{1/2}},$$

for all $i \in [N]$, if $\tau_M < t \leq T$ and $i \in [N]$, where

$$h(\delta, T) = R\sqrt{\frac{8}{p^+_{\min}\lambda_m}} \left(\sqrt{d_y \log\left(\frac{1 + TL^2}{\delta}\right)} + c_\eta \right) = \mathcal{O}\left(R\sqrt{d_y \log(TNd_y/\delta)} \right).$$

Accordingly, we have

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top \widetilde{\eta}_j(t) - y_j(t)^\top \widetilde{\eta}_i(t) > 0 \middle| G^\star_{t-1}, A^\star_{it}\right) \\
&\leq \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -h(\delta, T)\|y(t)\|t^{-1/2} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| G^\star_{t-1}, A^\star_{it}\right) \\
&+ \mathbb{P}\left(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -h(\delta, T)\|y(t)\|t^{-1/2} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| G^\star_{t-1}, A^\star_{it}\right).
\end{aligned} \tag{4.50}$$

Now, let $E_{ijt} = \{h(\delta, T)t^{-1/2} < 0.25(\dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j)\} \cap A^\star_{it}$, where $\dot{y}_i(t) = y_i(t)/\|y(t)\|$. Then, we can decompose the first term on the RHS in (47) as follows:

$$\begin{aligned}
&\mathbb{P}\left(\dot{y}_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T)}{t^{1/2}} + (\dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j) \middle| G^\star_{t-1}, A^\star_{it}\right) \\
&= \mathbb{P}\left(\dot{y}_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T)}{t^{1/2}} + 0.5(\dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j) \middle| G^\star_{t-1}, E_{ijt}, A^\star_{it}\right) \\
&\qquad\qquad\qquad\qquad \times \mathbb{P}(E_{ijt} | G^\star_{t-1}, A^\star_{it}) \\
&+ \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T)}{t^{1/2}} + 0.5(\dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j) \middle| G^\star_{t-1}, E^c_{ijt}, A^\star_{it}\right) \\
&\qquad\qquad\qquad\qquad \times \mathbb{P}(E^c_{ijt} | G^\star_{t-1}, A^\star_{it}).
\end{aligned} \tag{4.51}$$

We aim to show that the above probability is $\mathcal{O}(t^{-0.5})$ by showing each term in the RHS of (48) is $\mathcal{O}(t^{-0.5})$. By Assumption 1, if $t > \tau_M$, we have

$$\mathbb{P}(E^c_{ijt} | G^\star_{t-1}, A^\star_{it}) = \mathbb{P}\left(4h(\delta, T)t^{-1/2} > (\dot{y}_i(t)^\top \eta_i - \dot{y}_j(t)^\top \eta_j) \middle| G^\star_{t-1}, A^\star_{it}\right) \leq \frac{4h(\delta, T)C}{t^{1/2}}. \tag{4.52}$$

Thus, the second term in (48) is $\mathcal{O}(t^{-0.5})$. Now, we aim to show that the first term in (48) is $\mathcal{O}(t^{-0.5})$. Note that

$$\begin{aligned}
& \mathbb{P}\left(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T)}{t^{1/2}} + 0.5(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j) \middle| G^\star_{t-1}, E_{ijt}, A^\star_{it}\right) \\
\leq \; & \mathbb{P}\left(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j) \middle| G^\star_{t-1}, A^\star_{it}\right).
\end{aligned}$$

Now it suffices to show that the first term in the RHS of the inequality above is $\mathcal{O}(t^{-1/2})$. Using $\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) \sim \mathcal{N}(0, v^2\dot{y}_i(t)^\top B_i(t)^{-1}\dot{y}_i(t))$ given $y_i(t)$ and $\lambda_{\min}(B_i(t)^{-1}) \leq 8/(\lambda_m p_i t)$ by Lemma 12 and (4.44), we have

$$\begin{aligned}
& \mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)|y(t), G^\star_{t-1}, A^\star_{it}) \\
\leq \; & \exp\left(-\frac{tp_i\lambda_m(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)^2}{256v^2}\right).
\end{aligned}$$

Thus, the first term on the RHS of the above inequality can be written as

$$\begin{aligned}
& \mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)|G^\star_{t-1}, A^\star_{it}) \\
= \; & \mathbb{E}[\mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}.(t)) > 0.25(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)|y(t), G^\star_{t-1}, A^\star_{it})|G^\star_{t-1}, A^\star_{it}] \\
\leq \; & \mathbb{E}\left[\exp\left(-\frac{tp_i\lambda_m(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)^2}{256v^2}\right)\middle| G^\star_{t-1}, A^\star_{it}\right].
\end{aligned}$$

By integration by part, we have

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(-\frac{tp_i\lambda_m(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j)^2}{256v^2}\right)\middle| G^\star_{t-1}, A^\star_{it}\right] \\
= \; & \int_0^\infty \frac{2tp_i\lambda_m u}{256v^2}\exp\left(-\frac{tp_i\lambda_m u^2}{256v^2}\right)\mathbb{P}(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j < u|G^\star_{t-1}, A^\star_{it})du.
\end{aligned}$$

Since $\mathbb{P}(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j < u|G^\star_{t-1}, A^\star_{it}) = \mathbb{P}(\dot{y}_i(t)^\top\eta_i - \dot{y}_j(t)^\top\eta_j < u|A^\star_{it}) \leq Cu$ for $C > 0$ based on Assumption 1, the term above can be written as

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\frac{tp_i\lambda_m((\dot{y}_i(t)^\top\eta_i-\dot{y}_j(t)^\top\eta_j)^2}{256v^2}\right)\middle| G^\star_{t-1},A^\star_{it}\right]\\ \leq\ &\int_0^\infty\frac{2u}{256v^2/(tp_i\lambda_m)}\exp\left(-\frac{u^2}{256(v^2/tp_i\lambda_m)}\right)\mathbb{P}(\dot{y}_i(t)^\top\eta_i-\dot{y}_j(t)^\top\eta_j<u|A^\star_{it})du\\ \leq\ &\frac{\sqrt{\pi}}{\sqrt{256v^2/(tp_i\lambda_m)}}\int_0^\infty\frac{2u}{\sqrt{256\pi v^2/(tp_i\lambda_m)}}\exp\left(-\frac{u^2}{256v^2/(tp_i\lambda_m)}\right)Cudu=8vC\sqrt{\frac{\pi}{\lambda_m p_i t}},\end{aligned}$$

where we used the following result about one-sided Gaussian integrals

$$\int_0^\infty x^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}dx=\sigma^2/2.$$

Accordingly, we have

$$\begin{aligned}&\mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t))>-y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i)+0.5(y_i(t)^\top\eta_i-y_j(t)^\top\eta_j)|G^\star_{t-1},A^\star_{it})\\ \leq\ &C\sqrt{\frac{4}{p_i t}}\left(8v\sqrt{\frac{\pi}{\lambda_m}}+4h(\delta,T)\right),\end{aligned}\tag{4.53}$$

where

$$h(\delta,T)=\sqrt{\frac{2}{\lambda_m}}\left(R\sqrt{d_y\log\left(\frac{1+TL^2}{\delta}\right)}+c_\eta\right)=\mathcal{O}\left(R\sqrt{d_y\log(TNd_y/\delta)}\right).$$

Similarly, we get

$$\begin{aligned}&\mathbb{P}(y_j(t)^\top(\widetilde{\eta}_j(t)-\widehat{\eta}_j(t))>-y_j(t)^\top(\widehat{\eta}_j(t)-\eta_j)+0.5(y_i(t)^\top\eta_i-y_j(t)^\top\eta_j)|G^\star_{t-1},A^\star_{it})\\ \leq\ &C\sqrt{\frac{4}{p_j t}}\left(8v\sqrt{\frac{\pi}{\lambda_m}}+4h(\delta,T)\right),\end{aligned}\tag{4.54}$$

if $t > \tau_M$. Accordingly, based on (4.49), (4.53), and (4.54), we obtain the following bounds for the probabilities

$$\begin{aligned}
&\mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t)) > 0|G^\star_{t-1}, A^\star_{it}) \\
\leq\ & \frac{2C}{\sqrt{p^+_{\min}}}\left(v\left(8\sqrt{\frac{\pi}{\lambda_m}} + 8\sqrt{\frac{\pi}{\lambda_m}}\right) + 4h(\delta, T) + 4h(\delta, T)\right) t^{-1/2}.
\end{aligned}$$

By simple calculations, we get

$$\mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t) > 0|G^\star_{t-1}, A^\star_{it}) \leq \frac{16C}{\sqrt{p^+_{\min}}}\left(2v\sqrt{\frac{\pi}{\lambda_m}} + h(\delta, T)\right) t^{-1/2}.$$

By summing the above probability up over $i, j \in [N]$, if $t > \tau_M$, we get an upper bound for the probability of choosing a sub-optimal arm at time t

$$\begin{aligned}
\mathbb{P}(a^\star(t) \neq a(t)|G^\star_{t-1}) &= \sum_{i=1}^{N}\sum_{j\neq i}\mathbb{P}(a(t) = j|G^\star_{t-1}, A^\star_{it})\mathbb{P}(A^\star_{it}) \\
&\leq \sum_{i=1}^{N}\sum_{j\neq i}\mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t) > 0|G^\star_{t-1}, A^\star_{it})\mathbb{P}(A^\star_{it}) \\
&\leq N\left(\frac{16C}{\sqrt{p^+_{\min}}}\left(2v\sqrt{\frac{\pi}{\lambda_m}} + h(\delta, T)\right) t^{-1/2}\right).
\end{aligned}$$

By plugging the inequality above to (4.48), with probability at least $1 - \delta$, we have

$$\begin{aligned}
&\sum_{t=\lceil\tau_M\rceil}^{T}\frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t) \neq a(t)) \\
\leq\ & \sqrt{4\log T \log\delta^{-1}} + \sum_{t=\lceil\tau_M\rceil}^{T} N\left(\frac{16C}{\sqrt{p^+_{\min}}}\left(2v\sqrt{\frac{\pi}{\lambda_m}} + h(\delta, T)\right) t^{-1}\right) \\
\leq\ & \sqrt{4\log T \log\delta^{-1}} + \frac{16CN}{\sqrt{p^+_{\min}}}c_M(\delta, T)\log T,
\end{aligned}$$

where

$$c_M(\delta, T) = 8v\sqrt{\pi/\lambda_m} + 4h(\delta, T) = \mathcal{O}\left(v\sqrt{d_y \log(Td_y/\delta)}\right).$$

Therefore, putting together $L = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$, $g^{(1)}(\delta) = \mathcal{O}\left(v\sqrt{{p_{\min}^{+}}^{-1} d_y \log(TNd_y/\delta)}\right)$, $c_M(\delta, T) = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$, and $\tau_M = \mathcal{O}(v^2 {p_{\min}^{+}}^{-2} N L^7 \kappa^{-5} \log^{1.5}(TNd_y/\delta))$,

$$\begin{aligned}
\mathrm{Regret}(T) &\leq 2c_\eta L \tau_M + L g^{(1)}(\delta)\left(\sqrt{4\log T \log \delta^{-1}} + \frac{16 c_M(\delta, T) C N}{\sqrt{p_{\min}^{+}}}\log T\right) \\
&= \mathcal{O}\left((p_{\min}^{+})^{-2} L^8 \log^{1.5}\left(\frac{TNd_y}{\delta}\right) + vLNd_y {p_{\min}^{+}}^{-0.5} \log^2(TNd_y/\delta)\right) \\
&= \mathcal{O}\left(\frac{v^2 N d_y^4}{(p_{\min}^{+})^2 \kappa^5} \log^{5.5}\left(\frac{TNd_y}{\delta}\right)\right).
\end{aligned}$$

This regret bound is inflated by $d_y^3 \log^{3.5}(TNd_y/\delta)$ due to the order of maximum magnitude of observation norm L. If the support of observations is bounded by a positive constant so that L is a positive constant unrelated to other factors (N, d_y, T, and δ), the upper bound can be reduced to $\mathcal{O}\left(v^2 N d_y (p_{\min}^{+})^{-2} \kappa^{-5} \log^2(TNd_y/\delta)\right)$.

□

As discussed in the proof sketch of Theorem 5, the high-probability bound L for an observation norm significantly affects the regret bound. As such, in the next corollary, we suggest a tighter regret bound for observations with bounded support.

Corollary 3. *If the observations are assumed to be generated from a distribution with bounded support, the regret of Algorithm 3 is*

$$\mathrm{Regret}(T) = \mathcal{O}\left(v^2 N d_y {p_{\min}^{+}}^{-2} \kappa^{-5} \log^2(TNd_y/\delta)\right).$$

If observations are generated from bounded support, the orders of first and second terms in the above proof sketch in (4.47) are reduced to $\mathcal{O}\left(N\log^{1.5}(TNd_y/\delta)\right)$ and $\mathcal{O}\left(Nd_y\log^2(TNd_y/\delta)\right)$, respectively. Subsequently, we get the result of the corollary above from Theorem 5.

The above results are unprecedented even for fully observable contextual bandits as well as partially observed ones with the arm-specific parameter setup to the best of our knowledge. Especially, a high-probability poly-logarithmic regret bound for Thompson sampling with respect to the time horizon has not been shown for contextual bandits, even though the previously available minimax regret bound for Thompson sampling has a square-root order with respect to time for the adversarially chosen contexts (Agrawal & Goyal, 2013). Meanwhile, the suggested regret bound is a special case of the minimax regret bound, where a positive probabilistic suboptimality gap is not guaranteed due to the existence of an adversary. A logarithmic regret bound for contextual bandits is shown to be achieved by the greedy-first algorithm that takes a greedy action if a criterion is met and explores otherwise (Bastani et al., 2021). However, the above regret bound is valuable in that the greedy first algorithm needs another standard algorithm such as Thompson sampling and OFU-type algorithms for exploration.

4.5 Numerical Experiments

Simulation Experiments: In this sub-section, we numerically show the results in Section 4.4 with synthetic data. First, to explore the relationships between the regret and dimension of observations and contexts, we simulate various scenarios for the model with arm-specific parameters with $N = 5$ arms and different dimensions of the observations $d_y = 10,\ 20,\ 40,\ 80$ and context dimension $d_x = 10,\ 20,\ 40,\ 80$. Each case is repeated 50 times and the average and worst quantities amongst all 50 scenarios are reported. Figure 4.1 illustrates regret normalized by $(\log t)^2$, which is the regret growth that the minimum time effect is removed. Second, Figure 4.2 showcases the average estimation errors of the estimates in (4.9) for five different arm-specific parameters defined in (4.4), changing dimensions of observations and contexts.

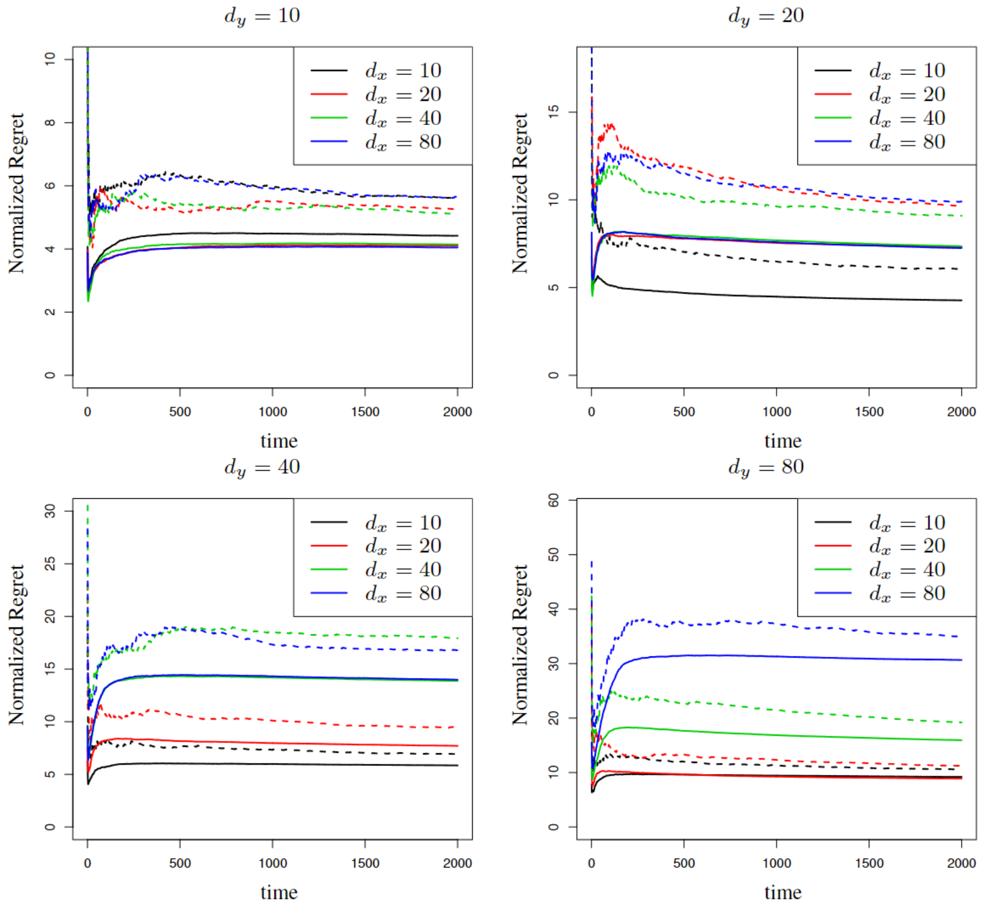

Figure 4.1: Plots of $\text{Regret}(t)/(\log t)^2$ over time for the different dimensions of context at $N = 5$ and $d_y = 10, 20, 40, 80$. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

These errors are normalized by $t^{-1/2}$ based on Theorem 4. Since the error decreases with a rate $t^{-1/2}$, the normalized errors for all the arms are flattened over time. This demonstrates that the square-root accuracy estimations of $\{\eta_i\}_{i=1}^N$ are available regardless of whether the dimension of observations is greater or less than that of contexts.

Moving on, Figure 4.3 provides insights into the average and worst-case regrets of Thompson sampling compared to the Greedy algorithm, with variations in the number of arms ($N = 10, 20, 30$). It is worth noting that the Greedy algorithm is considered optimal for the model with a shared parameter, but the

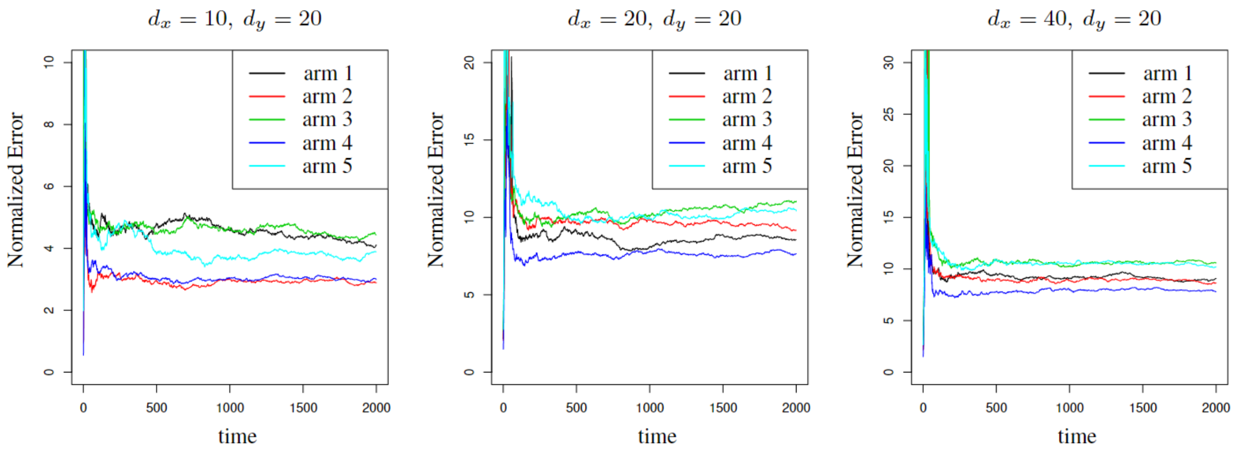

Figure 4.2: Plots of normalized estimation errors $\sqrt{t}\|\widehat{\eta}_i(t) - \eta_i\|$ of Algorithm 3 over time for partially observable stochastic contextual bandits with five arm-specific parameters and dimensions of observations and contexts $d_y = 20$, $d_x = 10,\ 20,\ 40$.

worst-case regret of it exhibits linear growth in the model with arm-specific parameters. The worst-case linear regret growth of the greedy algorithm can occur when some arms, which are totally dominated by other arms, are missing in potential action because of no explicit exploration scheme. In Figure 4.3, the plots represent the average and worst-case regrets of the models with arm-specific parameters, showing that the greedy algorithm has greater worst-case regret for the model with arm-specific parameters, especially for the case with a large number of arms.

Real Data Experiments: In this sub-section, we assess the performance of the proposed algorithm using two healthcare datasets: Eye Movement and EEG[2]. These two datasets are presented in previous studies by (Bastani & Bayati, 2020; Bietti et al., 2021) using contextual bandits with arm-specific parameters and shared context. These datasets involve classification tasks based on patient information. The Eye Movement and EEG data sets are comprised of 26 and 14-dimensional (shared) contexts, respectively, with the corresponding patient class. Also, the number of class for Eye Movement and EEG datasets are 3 and 2, respectively, and each category of patient class is considered an arm in the perspective of the

[2] The datasets are publicly available at: https://www.openml.org/

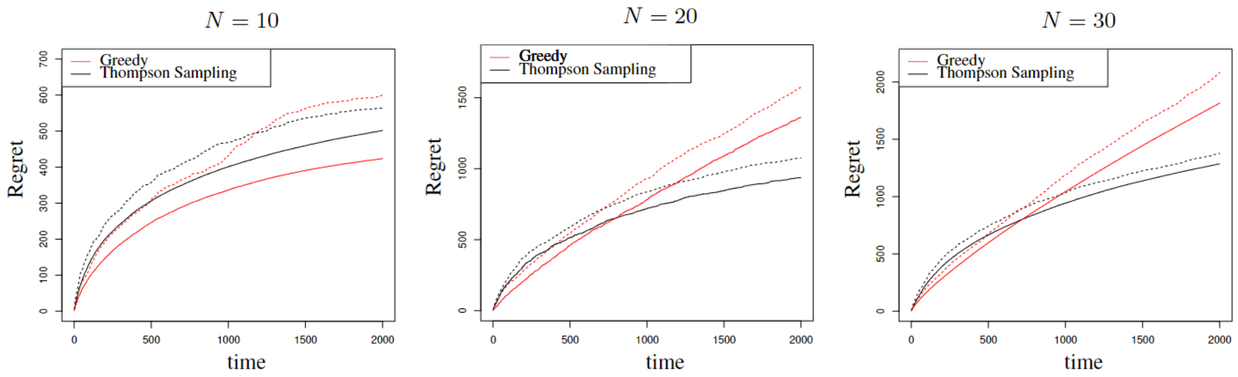

Figure 4.3: Plots of regrets over time with the different number of arms $N = 10,\ 20,\ 30$ for Thomson sampling versus the Greedy algorithm. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

bandit problem. We analyze these datasets under the logistic linear regression assumption, where reward is assumed to be generated based on (4.1) and

$$\log \frac{\mathbb{P}(l(t) = i)}{1 - \mathbb{P}(l(t) = i)} = x(t)^\top \mu_i = \mathbb{E}[r_i(t)], \tag{4.55}$$

where $l(t)$ is the true label of the patient at time t. Because the datasets do not have rewards based on this setup, we generated rewards based on (4.1) and (4.55) with artificial noises.

For evaluation, we generate 100 scenarios for each dataset. We calculate the average correct decision rate defined as $t^{-1} \sum_{\tau=1}^{t} \mathbb{I}(a(\tau) = l(\tau))$. We compare the suggested algorithm against the regression oracle with the estimates trained on the entire data in hindsight, which are not updated over time. We artificially create observations of the patients' contexts based on the structure given in (4.3) with a sensing matrix A consisting of 0 and 1 only. We reduce the dimension of the patient contexts from 26 to 13 for the Eye movement dataset and from 14 to 10 for the EEG dataset. Figure 4.4 displays the average correct decision rates of the regression oracle and Thompson sampling for the two real datasets. We evaluate the mean correct decision rates over every 100 patients and then average them across 100 scenarios. Accordingly, each

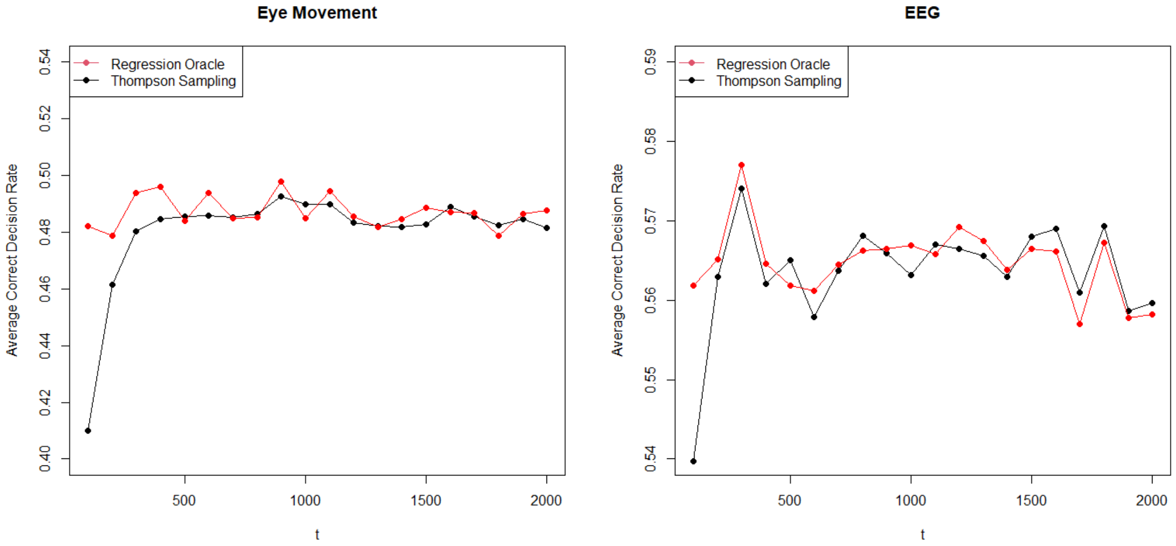

Figure 4.4: Plots of average correct decision rates of the regression oracle and Thompson sampling for Eye movement (left) and EEG dataset (right).

dot represents the sample mean of 10,000 results. For both data sets, the correct decision rate of Thompson sampling converges to that of the regression oracle over time. More results of real data experiments are provided in Appendix .3.2.

Chapter 5

Concluding Remarks

In this dissertation, we introduce partially observable stochastic contextual bandits with two different reward parameter setups and investigate the performance of Thompson sampling and Greedy algorithm for them.

First, in the shared parameter setup, we show that Thompson sampling learns the unknown true parameter accurately. Further, we establish theoretical performance guarantees showing that the regret of the proposed algorithm scales linearly with dimension, and logarithmically with time and the number of arms. Next, we construct a high-probability regret bound for Greedy algorithms, which grows poly-logarithmically with the horizon T.

Subsequently, in the arm-specific parameter setup, we focused exclusively on Thompson sampling, as the numerical experiments presented in Chapter 4 illustrated that Greedy algorithm can incur a worst-case linear regret. We show that Thompson sampling with an appropriate exploration scheme guarantees the square-root consistency for partial estimation of reward parameters. Further, we prove regret bounds that grow poly-logarithmically with time, linearly with N, and d_y^4 with the dimension of observations. Our analysis techniques can be applied to analogous reinforcement learning problems involving partial observations, thanks to generality of technical assumptions and tight quantification of the exploration Thompson sampling performs by leveraging the partial observations.

Chapter 6

Future Works

In partially observable stochastic contextual bandits, the uncertainty of contexts is modeled as stochasticity in the contexts associated with different actions. On the other hand, in partially observable adversarial contextual bandits, where transformed noisy contexts are observed, the uncertainty of contexts is modeled as an adversary that actively tries to counteract the learning algorithm by selecting the most unfavorable contexts. In adversarial contexts, we do not have any probabilistic assumptions for context distributions as the adversary has control over context generation. Thus, we need different estimation methods for this adversarial setup, whereas an agent takes advantage of the distribution of contexts to estimate the mean reward of each arm. We used BLUP to estimate mean rewards in Chapter 4, but we need to consider a different estimation method for the adversarial setup.

We consider the same reward and observation model for arm i as that of our previous study as follows:

$$\begin{aligned} r_i(t) &= x_i(t)^\top \mu_\star + \varepsilon_i(t), && (6.1) \\ y_i(t) &= A x_i(t) + \xi_i(t), && (6.2) \end{aligned}$$

where $\mu_\star$ is the reward parameter, $x_i(t)$ is an adversarial unobserved context of arm i, $y_i(t)$ is the noisy-transformed observation of $x_i(t)$ of arm i, and $\xi_i(t)$ is the noise of observation of arm i. Here, we do not consider arm-specific parameters $\{\mu_i\}_{i=1}^N$, because the adversary can create a setup identical to the arm-specific parameter setup described in the previous section based on the given setup above by artificially generating contexts in a particular way. Since the context $x_i(t)$ is not observed, we aim to find the optimal policy such that

$$a^\star(t) = \underset{i\in[N]}{\operatorname{argmax}}\, \mathbb{E}[x_i(t)^\top \mu_\star | y(t)], \tag{6.3}$$

where $y(t)$ is a concatenation of observations such that $y(t) = (y_1(t), y_2(t), \dots, y_N(t))$. Thus, we focus on the estimation of $\mathbb{E}[x_i(t)^\top \mu_\star | y(t)]$. To do so, by multiplying $\Sigma_\xi^{-0.5}$ on both side of (6.2), we get

$$\Sigma_\xi^{-0.5} y_i(t) = \Sigma_\xi^{-0.5} A x_i(t) + \Sigma_\xi^{-0.5} \xi_i(t), \tag{6.4}$$

where each element of noise $\xi_i(t)$ has the same finite variance. Considering $x_i(t)$ as an unknown constant, we can find the least square estimator of $x_i(t)^\top \mu_\star$, the close form of which is is as follows:

$$\widehat{x_i(t)^\top \mu_\star} := ((A^\top \Sigma_\xi^{-1} A)^- A^\top \Sigma_\xi^{-1} y_i(t))^\top \mu_\star = y_i(t)^\top D^\top \mu_\star,$$

where $D = (A^\top \Sigma_\xi^{-1} A)^- A^\top \Sigma_\xi^{-1}$ and M^- represents a pseudo-inverse matrix of a square matrix M. Then, by the Gauss-Markov theorem, $\widehat{x_i(t)^\top \mu_\star}$ is the best linear unbiased estimate (BLUE) of $x_i(t)^\top \mu_\star$ if $\mu_\star$ is in $C(A^\top \Sigma_\xi^{-1} A)$, which is the column space of $A^\top \Sigma_\xi^{-1} A$.

Here, $x_i(t)^\top \mu_\star$ is not estimable if $\mu_\star$ is not in $C(A^\top \Sigma_\xi^{-1} A)$. In other words, the value of this estimator is not invariant with the choice of the pseudo-inverse matrix $(A^\top \Sigma_\xi^{-1} A)^-$. In this case, the adversary can manipulate contexts without providing the agent with any reward information, rendering it an unfair game for the agent and resulting in any policies unable to attain sub-linear regret. Consequently, we focus only on cases where sub-linear regret may be attainable for potentially effective policies.

Similarly to the previous discussion, we consider the transformed parameter $\eta_\star$ such that

$$\eta_\star = D^\top \mu_\star \tag{6.5}$$

Because we consider $\mu_\star \in C(A^\top \Sigma_\xi^{-1} A)$, we have

$$A^\top \eta_\star = A^\top D^\top \mu_\star = \mu_\star. \tag{6.6}$$

Thus, using (6.2) and (6.6), we get

$$r_i(t) = y_i(t)^\top \eta_\star + \zeta_i(t)$$

where $\zeta_i(t) = (Ax_i(t) - y_i(t))^\top \eta_\star + \varepsilon_i(t) = \xi_i(t)^\top \eta_\star + \varepsilon_i(t)$ is a noise independent from the others. In fact, given the assumption that the observation $y_i(t)$ is of a positive definite covariance matrix Σ_ξ, the estimation of $\eta_\star$ is guaranteed to be available. Accordingly, the optimal arm in (6.3) can be written as

$$a^\star(t) = \operatorname*{argmax}_{i \in [N]} y_i(t)^\top \eta_\star.$$

In future work, we plan to develop the Thompson sampling algorithm, designed to choose the action $a^\star(t)$. In addition, we analyze the algorithm with a focus on regret minimization.

.1 Appendices for Chapter 2

.1.1 Derivation of the conditional distribution $\mathbb{P}(x_i(t)|y_i(t))$

Note that $y_i(t) = Ax_i(t) + \xi_i(t)$, where the distributions of $\xi_i(t)$ and $x_i(t)$ are $N(0_d, \Sigma_\xi)$ and $N(0_d, \Sigma_x)$, respectively. The conditional distribution of $x_i(t)$ given $y_i(t)$ can be calculated as follows.

$$\begin{aligned}
\mathbb{P}(x_i(t)|y_i(t)) &\propto \mathbb{P}(y_i(t)|x_i(t))\mathbb{P}(x_i(t)) \\
&\propto \exp\left((y_i(t) - Ax_i(t))^\top \Sigma_\xi^{-1}(y_i(t) - Ax_i(t))\right) \exp\left(x_i(t)^\top \Sigma_x^{-1} x_i(t)\right) \\
&\propto N((A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_\xi^{-1} y_i(t), (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1}) \qquad (7)
\end{aligned}$$

.1.2 Derivation of the conditional distribution $\mathbb{P}(r_i(t)|y_i(t))$

Let $\Sigma_{xy} = (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1}$ and recall $\widehat{x}_i(t) = (A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_\xi^{-1} y_i(t) = Dy_i(t)$.

$$\begin{aligned}
&\mathbb{P}(r_i(t)|\mu, y_i(t)) \\
&= \int_{\mathbb{R}^d} \mathbb{P}(r_i(t)|\mu, x_i(t))\mathbb{P}(x_i(t)|y_i(t)) dx_i(t) \\
&\propto \int_{\mathbb{R}^d} \exp\left(-\frac{(r_i(t) - x_i(t)^\top \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}(x_i(t) - \widehat{x}_i(t))^\top \Sigma_{xy}^{-1}(x_i(t) - \widehat{x}_i(t))\right) dx_i(t) \\
&\propto \exp\left(-\frac{\left(r_i(t) - ((A^\top \Sigma_\xi^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_\xi^{-1} y_i(t))^\top \mu\right)^2}{2(\mu^\top \Sigma_{xy} \mu + \sigma^2)}\right) \\
&\propto N\left(\widehat{x}_i(t)^\top \mu, \sigma_{ry}^2\right). \qquad (8)
\end{aligned}$$

.1.3 Derivation of the posterior $\mathbb{P}(\mu|\mathbf{r}_{t-1}, \mathbf{y}_{t-1})$

Let $\mathbb{P}(\mu)$, the pdf of $N(0, \sigma_{ry}^2\Sigma)$, be the prior of $\mu_\star$. We can decompose the posterior as follows.

$$\begin{aligned}\mathbb{P}(\mu|\mathbf{r}_{t-1}, \mathbf{y}_{t-1}) &\propto \mathbb{P}(\mathbf{r}_{t-1}, \mathbf{y}_{t-1}|\mu)\mathbb{P}(\mu)\\ &\propto \mathbb{P}(\mathbf{r}_{t-1}|\mathbf{y}_{t-1}, \mu)\mathbb{P}(\mu).\end{aligned}$$

Using the prior and the conditional distribution in (8), we have

$$\begin{aligned}\mathbb{P}(\mu|\mathbf{r}_{t-1}, \mathbf{y}_{t-1}) &\propto \prod_{\tau=1}^{t-1} \exp\left(-\frac{(r_{a(\tau)}(\tau) - \widehat{x}_{a(\tau)}(\tau)^\top \mu)^2}{2\sigma_{ry}^2}\right) \exp\left(-\frac{1}{2\sigma_{ry}^2}\mu^\top \Sigma^{-1}\mu\right)\\ &\propto \exp\left(-\frac{1}{2\sigma_{ry}^2}\left(\mu - \widehat{\mu}(t)\right)^\top B(t)\left(\mu - \widehat{\mu}(t)\right)\right), \qquad (9)\end{aligned}$$

which is the kernel of the pdf of $N(\widehat{\mu}(t), \sigma_{ry}^2 B(t)^{-1})$, where $\widehat{\mu}(t) = B(t)^{-1}\sum_{\tau=1}^{t-1}\widehat{x}_{a(t)}(t) r_{a(t)}(t)$ and

$$B(t) = \sum_{\tau=1}^{t-1} \widehat{x}_{a(\tau)}(\tau)\widehat{x}_{a(\tau)}^\top(\tau) + \Sigma^{-1}.$$

Thus, the posterior distribution is $N(\widehat{\mu}(t), \sigma_{ry}^2 B(t)^{-1})$. But, to allow for the possibility that σ_{ry}^2 is unknown, we use a re-scaled posterior distribution, $N(\widehat{\mu}(t), B(t)^{-1})$, which does not depend on σ_{ry}^2.

.1.4 Derivation of the recursion formula to update the parameter.

Note that we can decompose the posterior as follows.

$$\begin{aligned}\mathbb{P}(\mu|\mathbf{r}_t, \mathbf{y}_t) &\propto \mathbb{P}(\mathbf{r}_t, \mathbf{y}_t, \mu)\\ &\propto \mathbb{P}(r_{a(t)}(t)|y_{a(t)}(t), \mu)\mathbb{P}(\mu|\mathbf{r}_{t-1}, \mathbf{y}_{t-1}).\end{aligned}$$

Using the conditional distribution (8) and the posterior in (9), we get

$$\begin{aligned}\mathbb{P}(\mu|\mathbf{r}_t, \mathbf{y}_t) &\propto \mathbb{P}(r_{a(t)}(t)|y_{a(t)}(t), \mu)\mathbb{P}(\mu|\mathbf{r}_{t-1}, \mathbf{y}_{t-1}) \\ &\propto \exp\left(-\frac{(r_{a(t)}(t) - \widehat{x}_{a(t)}(t)^\top \mu)^2}{2\sigma_{ry}^2}\right) \exp\left(-\frac{1}{2\sigma_{ry}^2}\left(\mu - \widehat{\mu}(t)\right)^\top B(t)^{-1}\left(\mu - \widehat{\mu}(t)\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_{ry}^2}\left(\mu - \widehat{\mu}(t+1)\right)^\top B(t+1)^{-1}\left(\mu - \widehat{\mu}(t+1)\right)\right),\end{aligned}$$

where $\widehat{\mu}(t+1) = B(t+1)^{-1}\left(B(t)\widehat{\mu}(t) + \widehat{x}_{a(t)}(t)r_{a(t)}(t)\right)$ and $B(t+1) = B(t) + \widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top$.

.1.5 Proof of Lemma 1

Proof. Recall that we used the notation $S = \mathrm{Var}(\widehat{x}_i(t))^{0.5} = (D\Sigma_\xi D^\top)^{0.5}$ and $Z(\mu, N) = \mathrm{argmax}_{Z_i, 1\le i\le N}\{Z_i^\top \mu\}$. Note that $S^{-1}\widehat{x}_i(t)$ has the distribution $N(0_d, I_d)$ and $S^{-1}\widehat{x}_{a(t)}(t) = Z(S\widetilde{\mu}(t), N)$. $S^{-1}\widehat{x}_i(t)$ can be decomposed as

$$S^{-1}\widehat{x}_i(t) = P_{S\widetilde{\mu}(t)}S^{-1}\widehat{x}_i(t) + P_{S\widetilde{\mu}(t)^\perp}S^{-1}\widehat{x}_i(t),$$

where $P_{S\widetilde{\mu}(t)^\perp}$ denotes the projection matrix onto a subspace orthogonal to the column-space $C(S\widetilde{\mu}(t))$, which we denote $C(S\widetilde{\mu}(t))^\perp$. As shown in (2.24), we have

$$S^{-1}\widehat{x}_{a(t)}(t) \overset{d}{=} P_{S\widetilde{\mu}(t)}S^{-1}\widehat{x}_{a(t)}(t) + P_{S\widetilde{\mu}(t)^\perp}S^{-1}\widehat{x}_i(t),$$

where $\overset{d}{=}$ expresses that the two quantities have an identical distribution. Further, based on the fact that the function $Z(\mu, N)$ defined in (2.23) is affected only by $\{P_\mu Z_i\}_{1\le i\le N}$, but not by $\{(I_d - P_\mu)Z_i\}_{1\le i\le N}$, we established that $P_{S\widetilde{\mu}(t)}S^{-1}\widehat{x}_{a(t)}(t)$ and $P_{S\widetilde{\mu}(t)^\perp}S^{-1}\widehat{x}_i(t)$ are statistically independent. Now, consider the following decomposition.

$$
\begin{aligned}
&\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]\\
&= P_{S\widetilde{\mu}(t)}E[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}]P_{S\widetilde{\mu}(t)} + P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp}\\
&+ P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp} + P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)}.
\end{aligned}
$$

By replacing $P_{S\widetilde{\mu}(t)}S^{-1}\widehat{x}_{a(t)}(t)$ with $P_{S\widetilde{\mu}(t)^\perp}S^{-1}\widehat{x}_i(t)$ based on the independence and the equivalence of the distribution, we get

$$
\begin{aligned}
P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp} &= P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_i(t)\widehat{x}_i(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp}\\
&= P_{S\widetilde{\mu}(t)^\perp}, \qquad (10)
\end{aligned}
$$

and

$$
\begin{aligned}
&P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_i(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp} + P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_i(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)}\\
&= P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)|\widetilde{\mu}(t)]\mathbb{E}[\widehat{x}_i(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)^\perp}\\
&\qquad\qquad + P_{S\widetilde{\mu}(t)^\perp}\mathbb{E}[S^{-1}\widehat{x}_i(t)|\widetilde{\mu}(t)]\mathbb{E}[\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)}\\
&= 0, \qquad (11)
\end{aligned}
$$

because $\mathbb{E}[\widehat{x}_i(t)|\widetilde{\mu}(t)] = 0$. Thus, by putting (10) and (11) together, we have

$$
\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)] = P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)} + P_{S\widetilde{\mu}(t)^\perp}.
$$

On the other hand, $P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)}$ can be written as

$$\begin{aligned}
& P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)} \\
= \ & \frac{S\widetilde{\mu}(t)\widetilde{\mu}(t)^\top S}{\widetilde{\mu}(t)^\top S^2\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]\frac{S\widetilde{\mu}(t)\widetilde{\mu}(t)^\top S}{\widetilde{\mu}(t)^\top S^2\widetilde{\mu}(t)} \\
= \ & \frac{S\widetilde{\mu}(t)}{\widetilde{\mu}(t)^\top S^2\widetilde{\mu}(t)}\mathbb{E}[(\widetilde{\mu}(t)^\top SS^{-1}\widehat{x}_{a(t)}(t))^2|\widetilde{\mu}(t)]\frac{\widetilde{\mu}(t)^\top S}{\widetilde{\mu}(t)^\top S^2\widetilde{\mu}(t)} \\
= \ & P_{S\widetilde{\mu}(t)}\mathbb{E}\left[\left(\left(S^{-1}\widehat{x}_{a(t)}(t)\right)^\top \overrightarrow{S\widetilde{\mu}(t)}\right)^2\middle|\widetilde{\mu}(t)\right]. \qquad (12)
\end{aligned}$$

Since $\widehat{x}_i(t)^\top S^{-1}\overrightarrow{S\widetilde{\mu}(t)}$ has a standard normal distribution, we have

$$\mathbb{E}\left[\left(\widehat{x}_{a(t)}^\top(t)S^{-1}\overrightarrow{S\widetilde{\mu}(t)}\right)^2\middle|\widetilde{\mu}(t)\right] = \mathbb{E}\left[\left(\underset{1\le i\le N}{max}(\{V_i : V_i \sim N(0,1)\}\right)^2\right]. \qquad (13)$$

We define the quantity in (13) as k_N,

$$k_N = \mathbb{E}\left[\left(\underset{1\le i\le N}{max}(\{V_i : V_i \sim N(0,1)\}\right)^2\right], \qquad (14)$$

which is greater than 1 and grows as N gets larger, because $\mathbb{E}[V_i^2] = 1 < \mathbb{E}\left[\left(\underset{1\le i\le N}{max}(\{V_i : V_i \sim N(0,1)\}\right)^2\right]$. Thus, $\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]$ can be written as

$$\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)] = P_{S\widetilde{\mu}(t)}k_N + P_{S\widetilde{\mu}(t)^\perp} = P_{S\widetilde{\mu}(t)}(k_N - 1) + I_d. \qquad (15)$$

Because the column-spaces of the matrices $P_{S\widetilde{\mu}(t)}$ and $P_{S\widetilde{\mu}(t)^\perp}$ are orthogonal, the non-zero eigenvalues of $P_{S\widetilde{\mu}(t)}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]P_{S\widetilde{\mu}(t)}$ and $P_{S\widetilde{\mu}(t)^\perp}$ are the eigenvalues of $\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]$. That is, $(d-1)$ eigenvalues of $\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]$ are 1, and the other eigenvalue is k_N. This means that $\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]$ is positive definite, since $k_N > 1$.

Next, for the true parameter $\mu_\star$, we claim $\lim_{t\to\infty}\widetilde{\mu}(t) = \mu_\star$. With (2.14), (2.17), and the fact that $\widetilde{\mu}(t)$ is generated from the posterior $N(\widehat{\mu}(t), B(t)^{-1})$, we have

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\mu}(t)\right] &= \mathbb{E}\left[\mathbb{E}[\widetilde{\mu}(t)|\mathscr{F}_{t-1}]\right] = \mathbb{E}\left[\widehat{\mu}(t)\right] = (I_d - \mathbb{E}[B(t)^{-1}]\Sigma^{-1})\mu_\star, && (16)\\
\mathrm{Cov}(\widetilde{\mu}(t)) &= \mathrm{Cov}(\mathbb{E}[\widetilde{\mu}(t)|\mathscr{F}_{t-1}]) + \mathbb{E}[\mathrm{Cov}(\widetilde{\mu}(t)|\mathscr{F}_{t-1})]\\
&= \mathrm{Cov}(\widehat{\mu}(t)) + \mathbb{E}[B(t)^{-1}]\\
&= \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}B(t)^{-1}\right] - \mathbb{E}\left[B(t)^{-1}\right]\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}\mathbb{E}\left[B(t)^{-1}\right]\\
&+ \mathbb{E}\left[B(t)^{-1}\right]\sigma_{ry}^2 - \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}B(t)^{-1}\right]\sigma_{ry}^2 + \mathbb{E}\left[B(t)^{-1}\right]\\
&= \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}B(t)^{-1}\right] - \mathbb{E}\left[B(t)^{-1}\right]\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}\mathbb{E}\left[B(t)^{-1}\right]\\
&+ \mathbb{E}\left[B(t)^{-1}\right](\sigma_{ry}^2 + 1) - \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}B(t)^{-1}\right]\sigma_{ry}^2. && (17)
\end{aligned}$$

Since $\lim_{t\to\infty} B(t)^{-1} = 0_{d\times d}$ and thereby $\lim_{t\to\infty}\mathrm{Cov}(\widetilde{\mu}(t)) = 0_{d\times d}$, $\widetilde{\mu}(t)$ is a consistent estimator of $\mu_\star$. That is,

$$\lim_{t\to\infty}\widetilde{\mu}(t) = \mu_\star. \qquad (18)$$

Thus, $\lim_{t\to\infty} P_{S\widetilde{\mu}(t)} = P_{S\mu_\star}$. Using

$$\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\widetilde{\mu}(t)]|\mathscr{F}_{t-1}]$$

and (15), we get

$$\lim_{t\to\infty}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\mathscr{F}_{t-1}] = \lim_{t\to\infty}\mathbb{E}[P_{S\widetilde{\mu}(t)}(k_N - 1) + I_d|\mathscr{F}_{t-1}] = P_{S\mu_\star}(k_N - 1) + I_d.$$

Because the eigenvalues $P_{S\mu_\star}(k_N - 1) + I_d$ are $(d-1)$ 1s and k_N, which is greater than 1, $P_{S\mu_\star}(k_N - 1) + I_d$ is positive definite. Therefore, $\lim_{t\to\infty}\mathbb{E}[S^{-1}\widehat{x}_{a(t)}(t)\widehat{x}_{a(t)}(t)^\top S^{-1}|\mathscr{F}_{t-1}]$ is positive definite.

□

.1.6 Proof of Corollary 1

Proof. Recall $\mathrm{Cov}(\widetilde{\mu}(t))$ in (17)

$$\begin{aligned}\mathrm{Cov}(\widetilde{\mu}(t)) &= \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}B(t)^{-1}\right] - \mathbb{E}\left[B(t)^{-1}\right]\Sigma^{-1}\mu_\star\mu_\star^\top\Sigma^{-1}\mathbb{E}\left[B(t)^{-1}\right] \\ &+ \mathbb{E}\left[B(t)^{-1}\right](\sigma_{ry}^2+1) - \mathbb{E}\left[B(t)^{-1}\Sigma^{-1}B(t)^{-1}\right]\sigma_{ry}^2.\end{aligned}$$

Since $B(t)^{-1} = O(t^{-1})$ by Lemma 1 and the other terms are negligible except $\mathbb{E}\left[B(t)^{-1}\right](\sigma_{ry}^2+1)$ in above terms, we have $\mathrm{Cov}(\widetilde{\mu}(t)) = O(t^{-1})$. In addition, we already showed that $\lim_{t\to\infty}\widetilde{\mu}(t) = \mu_\star$ in (18).Therefore,

$$\lim_{t\to\infty}\widetilde{\mu}(t) = \mu_\star, \quad \mathrm{Cov}(\widetilde{\mu}(t)) = O(t^{-1}).$$

□

.2 Appendices for Chapter 3

.2.1 Proof of Lemma 2

Note that $S_y^{-0.5}y_i(t)$ has the normal distribution $N(0, I_{d_y})$. Then, we have

$$\mathbb{P}\left(|y_{ij}(t)| \geq \varepsilon\right) \leq 2\cdot e^{-\frac{\varepsilon^2}{2}} \tag{19}$$

where $y_{ij}(t)$ is the jth component of $y_i(t)$. By plugging $v_T(\delta)$ to ε, we have

$$\mathbb{P}\left(|y_{ij}(t)| \geq v_T(\delta)\right) \leq 2\cdot e^{-\frac{v_T(\delta)^2}{2}} = 2\cdot e^{-\log\frac{2Nd_yT}{\delta}} = \frac{\delta}{Nd_yT}. \tag{20}$$

Thus,

$$\mathbb{P}(W_T) \geq 1 - \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d_y} \mathbb{P}\left(|y_{ij}(t)| \geq v_T(\delta)\right) \geq 1 - \delta. \tag{21}$$

.2.2 Proof of Lemma 3

We use the following decomposition

$$S_y^{-0.5} y_{a(t)}(t) = P_{C(S_y^{0.5}\widehat{\eta}(t))} S_y^{-0.5} y_{a(t)}(t) + P_{C(S_y^{0.5}\widehat{\eta}(t))^{\perp}} S_y^{-0.5} y_{a(t)}(t). \tag{22}$$

We claim that $P_{C(S_y^{0.5}\widehat{\eta}(t))} S_y^{-0.5} y_{a(t)}$ and $P_{C(S_y^{0.5}\widehat{\eta}(t))^{\perp}} S_y^{-0.5} y_i(t)$ are statistically independent. To show it, define

$$Z(\nu, N) = \underset{Z_i, 1 \leq i \leq N}{\operatorname{argmax}} \left\{ Z_i^{\top} \nu \right\}, \tag{23}$$

where Z_i has the distribution $N(\mathbf{0}_{d_y}, I_{d_y})$ and ν is an arbitrary vector in $\mathbb{R}^{d_y}$. The vector Z_i can be decomposed as $Z_i = P_{C(\nu)} Z_i + (I_d - P_{C(\nu)}) Z_i$.Then, we have $Z(\nu, N) = \underset{Z_i, 1 \leq i \leq N}{\operatorname{argmax}} \left\{ (P_{C(\nu)} Z_i)^{\top} \nu \right\}$, because $P_{C(\nu)} \nu = \nu$. This implies that only the first term of the decomposed terms, $P_{C(\nu)} Z_i$, affects the result of $\underset{Z_i, 1 \leq i \leq N}{\operatorname{argmax}} \left\{ Z_i^{\top} \nu \right\}$. This means that $Z(\nu, N)$ has the same distribution as $P_{C(\nu)} Z(\nu, N) + (I_d - P_{C(\nu)}) Z_i$, which means

$$Z(\nu, N) \overset{d}{=} P_{C(\nu)} Z(\nu, N) + (I_d - P_{C(\nu)}) Z_i, \tag{24}$$

where $\overset{d}{=}$ is used to denote the equality of the probability distributions. Note that

$$S_y^{-0.5} y_{a(t)} = \underset{S_y^{-0.5} y_i, 1 \leq i \leq N}{\operatorname{argmax}} (S_y^{-0.5} y_i(t))^\top S_y^{0.5} \widehat{\eta}(t).$$

Thus, $S_y^{-0.5} y_{a(t)}$ has the same distribution as $P_{C(S_y^{0.5}\widehat{\eta}(t))} S_y^{-0.5} y_{a(t)} + P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp} S_y^{-0.5} y_i(t)$, where $P_{C(S_y^{0.5}\widehat{\eta}(t))} S_y^{-0.5} y_{a(t)}$ and $P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp} S_y^{-0.5} y_i(t)$ are statistically independent. By the decomposition (22) and the independence,
$\mathbb{E}\left[S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} | \mathscr{F}_{t-1}\right]$ can be written as

$$\begin{aligned}
&\mathbb{E}\left[S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} | \mathscr{F}_{t-1}\right] \\
&= \mathbb{E}\left[(P_{C(S_y^{0.5}\widehat{\eta}(t))} + P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp}) S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} (P_{S_y^{0.5}\widehat{\eta}(t)} + P_{S_y^{0.5}\widehat{\eta}(t)^\perp}) | \mathscr{F}_{t-1}\right] \\
&= \mathbb{E}\left[P_{C(S_y^{0.5}\widehat{\eta}(t))} S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} P_{C(S_y^{0.5}\widehat{\eta}(t))} | \mathscr{F}_{t-1}\right] + P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp}. \qquad (25)
\end{aligned}$$

To proceed, we show that the first term above, $\mathbb{E}[P_{S_y^{0.5}\widehat{\eta}(t)} S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} P_{S_y^{0.5}\widehat{\eta}(t)} | \widehat{\eta}(t)] = aP_{S_y^{0.5}\widehat{\eta}(t)}$ for some constant $a > 1$. Using $P_{C(\nu)} = \nu\nu^\top / \nu^\top \nu$ for an arbitrary vector $\nu \in \mathbb{R}^{d_y}$, we have

$$\begin{aligned}
& P_{S_y^{0.5}\widehat{\eta}(t)} \mathbb{E}[S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} | \widehat{\eta}(t)] P_{S_y^{0.5}\widehat{\eta}(t)} \\
= \; & \frac{S_y^{0.5}\widehat{\eta}(t)\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y \widehat{\eta}(t)} \mathbb{E}[S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} | \widehat{\eta}(t)] \frac{S_y^{0.5}\widehat{\eta}(t)\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y \widehat{\eta}(t)} \\
= \; & \frac{S_y^{0.5}\widehat{\eta}(t)}{\widehat{\eta}(t)^\top S_y \widehat{\eta}(t)} \mathbb{E}[(\widehat{\eta}(t)^\top S_y^{0.5} S_y^{-0.5} y_{a(t)}(t))^2 | \widehat{\eta}(t)] \frac{\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y \widehat{\eta}(t)} \\
= \; & P_{S_y^{0.5}\widehat{\eta}(t)} \mathbb{E}\left[\left.\left(\left(\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^\top S_y^{-0.5} y_{a(t)}(t)\right)^2 \right| \widehat{\eta}(t)\right], \qquad (26)
\end{aligned}$$

where $\overrightarrow{S_y^{0.5}\widehat{\eta}(t)} = S_y^{0.5}\widehat{\eta}(t)/\|S_y^{0.5}\widehat{\eta}(t))\|$ is the unit vector aligned linearly with $S_y^{0.5}\widehat{\eta}(t)$. Now, it suffices to prove that

$$\mathbb{E}\left[\left.\left(\left(\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^\top\left(S_y^{-0.5}y_{a(t)}(t)\right)\right)^2\right|\widehat{\eta}(t)\right] > 1.$$

Note that $\left(\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^\top S_y^{-0.5}y_i(t)$ has the standard normal distribution, since $S_y^{-0.5}y_i(t)$ has the distribution $N(0, I_{d_y})$. Thus, $\left(\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^\top S_y^{-0.5}y_{a(t)}(t)$ is the maximum variable of N variables with the standard normal density. Thus, using

$$a(t) = \underset{1\leq i\leq N}{\operatorname{argmax}}\{y_i(t)^\top\widehat{\eta}(t)\} = \underset{1\leq i\leq N}{\operatorname{argmax}}\left\{y_i(t)^\top S_y^{-0.5}\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right\},$$

we have

$$y_{a(t)}(t)^\top S_y^{-0.5}\overrightarrow{S_y^{0.5}\widehat{\eta}(t)} \overset{d}{=} \max_{1\leq i\leq N}\{V_i : V_i \sim N(0,1)\}. \tag{27}$$

where $\overset{d}{=}$ denotes the equality in terms of distribution. As such, we have

$$\mathbb{E}\left[\left.\left(y_{a(t)}(t)^\top S_y^{-0.5}\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^2\right|\widehat{\eta}(t)\right] = \mathbb{E}\left[\left(\max_{1\leq i\leq N}(\{V_i : V_i \sim N(0,1)\}\right)^2\right]. \tag{28}$$

We define the quantity in (28) as k_N,

$$k_N = \mathbb{E}\left[\left(\max_{1\leq i\leq N}(\{V_i : V_i \sim N(0,1)\}\right)^2\right], \tag{29}$$

which is greater than 1 for $N \geq 2$ and grows as N gets larger, because $\mathbb{E}[V_i^2] = 1 < \mathbb{E}\left[\left(\max_{1\leq i\leq N}(\{V_i : V_i \sim N(0,1)\}\right)^2\right]$. Therefore,

$$\begin{aligned}\mathbb{E}[S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}|\widehat{\eta}(t)] &= P_{C(S_y^{0.5}\widehat{\eta}(t))}k_N + P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp} \\ &= P_{C(S_y^{0.5}\widehat{\eta}(t))}(k_N - 1) + I_{d_y}. \end{aligned} \tag{30}$$

.2.3 Proof of Lemma 5

Consider $V_t = S_y^{-1/2}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-1/2}$ defined in Lemma 3 to identify the behavior of $B(t)$. By Lemma 3, the minimum eigenvalue of $\mathbb{E}[V_t|\mathscr{F}_{t-1}]$ is greater than 1 for all t. Thus, for all $t > 0$, it holds that

$$\lambda_{\min}\left(\sum_{\tau=1}^{t-1}\mathbb{E}[V_\tau|\widehat{\eta}(\tau)]\right) \geq t-1. \tag{31}$$

Now, we focus on a high probability lower-bound for the smallest eigenvalue of $B(t)$. On the event W_T, the matrix $v_T^2(\delta)I - V_t$ is positive semidefinite for all i and t. Let

$$\begin{aligned} X_\tau &= V_\tau - \mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}], \\ Y_\tau &= \sum_{j=1}^{\tau}\left(V_j - \mathbb{E}[V_j|\mathscr{F}_{j-1}]\right). \end{aligned} \tag{32}$$

Then, $X_\tau = Y_\tau - Y_{\tau-1}$ and $\mathbb{E}\left[X_\tau|\mathscr{F}_{\tau-1}\right] = 0$. Thus, X_τ is a martingale difference sequence. Because $v_T^2(\delta)I - V_t \succeq 0$ for all $t \leq T$, $4v_T^4(\delta)I - X_\tau^2 \succeq 0$, for all $\tau \leq T$, on the event W_T. By Lemma 4, we get

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t-1}X_\tau\right) \leq (t-1)\varepsilon\right) \leq d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32v_T^4(\delta)}\right), \tag{33}$$

for $\varepsilon \leq 0$. Now, using $\sum_{\tau=1}^{t-1} X_\tau = \sum_{\tau=1}^{t-1} V_\tau - \sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau | \mathscr{F}_{\tau-1}]$, together with

$$\begin{aligned} &\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau - \sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}]\right) \\ &\leq \lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) - \lambda_{\min}\left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}]\right) \end{aligned} \tag{34}$$

and (31), we obtain

$$\begin{aligned} &P\left(\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) \leq (t-1)(1+\varepsilon)\right) \\ &\leq d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32 v_T^4(\delta)}\right), \end{aligned} \tag{35}$$

where $-1 \leq \varepsilon \leq 0$ is arbitrary, and we used the fact that $\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) \geq 0$. Indeed, using $\sum_{\tau=1}^{t-1} V_\tau = S_y^{-0.5} B(t) S_y^{-0.5}$, on the event W_T defined in (3.11), for $-1 \leq \varepsilon \leq 0$ we have

$$\begin{aligned} &\mathbb{P}\left(\lambda_{\min}(B(t)) \leq \lambda_{s1}(t-1)(1+\varepsilon)\right) \\ &\leq\ d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32 v_T^4(\delta)}\right), \end{aligned} \tag{36}$$

where $\lambda_{s1} = \lambda_{\min}(S_y)$. In other words, by equating $d_y \cdot \exp\left(-(t-1)\varepsilon^2/(32 v_T^4(\delta)\right)$ to δ/T, (36) can be written as

$$\lambda_{\min}(B(t)) \geq \lambda_{s1}(t-1)\left(1 - \sqrt{\frac{32 v_T(\delta)^4}{t-1} \log\frac{d_y T}{\delta}}\right),$$

for all $1 \leq t \leq T$ with the probability at least $1 - 2\delta$.

.2.4 Proof of Lemma 6

Note that $\widehat{\eta}(t)$ has the distribution $N\left(\mathbb{E}[\widehat{\eta}(t)|\mathscr{F}_{t-1}], \mathrm{Cov}(\widehat{\eta}(t)|\mathscr{F}_{t-1})\right)$ given the observations up to time t, where

$$\begin{aligned} \mathbb{E}[\widehat{\eta}(t)|\mathscr{F}_{t-1}] &= B(t)^{-1}\left(\Sigma^{-1}+\sum_{\tau=1}^{t-1} y_{a(\tau)}(\tau)y_{a(\tau)}(\tau)^\top\right)\eta_\star = \eta_\star \\ \mathrm{Cov}(\widehat{\eta}(t)|\mathscr{F}_{t-1}) &= B(t)^{-1}\gamma_{ry}^2. \end{aligned} \tag{37}$$

For $Z \sim N(0, \lambda_{\max}(B(t)^{-1})\gamma_{ry}^2)$, using the Chernoff bound, we get

$$\begin{aligned} \mathbb{P}\left(\|\widehat{\eta}(t)-\eta_\star\| > \varepsilon|B(t)\right) &\leq \mathbb{P}\left(d_y Z^2 > \varepsilon^2\right) \\ &\leq 2\cdot\exp\left(-\frac{\varepsilon^2}{2d_y\lambda_{\max}(B(t)^{-1})\gamma_{ry}^2}\right), \end{aligned} \tag{38}$$

where $\varepsilon \geq 0$.

.2.5 Proof of Lemma 7

Let $a^{\cdot}(t)$ be the arm with the second largest expected reward at time t and $\eta_{\cdot}$ be a vector such that $y_{a^\star(t)}(t)^\top\eta_{\cdot} = y_{a^{\cdot}(t)}(t)^\top\eta_{\cdot}$ and $\theta(y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t), \eta_\star - \eta_{\cdot}) = 0$, where $\theta(x,y)$ is the angle between two vectors x and y. Then,

$$\begin{aligned} (y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t))^\top\eta_\star &= (y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t))^\top\eta_{\cdot} + (y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t))^\top(\eta_\star - \eta_{\cdot}) \\ &= \|y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t)\|\,\|\eta_\star - \eta_{\cdot}\|\cos\theta(y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t), \eta_\star - \eta_{\cdot}) \\ &= \|y_{a^\star(t)}(t) - y_{a^{\cdot}(t)}(t)\|\,\|\eta_\star - \eta_{\cdot}\|. \end{aligned} \tag{39}$$

If $\|y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t)\| \, \|\eta_\star - \widehat{\eta}(t)\| \le (y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star$, we can guarantee $a^\star(t) = a(t)$. Thus, the probability not to choose the optimal arm at time t given the observations and $B(t)$ is

$$\begin{aligned}
&\mathbb{P}(a^\star(t) \neq a(t) | \{y_i(t)\}_{1 \le i \le N}, B(t)) \\
&= \mathbb{P}\left(\|\widehat{\eta}(t) - \eta_\star\| > \frac{(y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star}{\|y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t)\|} \middle| \{y_i(t)\}_{1 \le i \le N}, B(t) \right) \\
&\le 2 \cdot \exp\left(- \frac{\left(\frac{(y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star}{\|y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t)\|} \right)^2}{2 d_y \lambda_{\max}(B(t)^{-1}) \gamma_{ry}^2} \right).
\end{aligned} \tag{40}$$

Using $\|y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t)\|^2 \le \lambda_{a2} d_y v_T(\delta)^2$ on the event W_T, we have

$$2 \cdot \exp\left(- \frac{\left(\frac{(y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star}{\|y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t)\|} \right)^2}{2 d_y \lambda_t \sigma_{ry}^2} \right) \le 2 \cdot \exp\left(- \frac{((y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star)^2}{2 d_y^2 v_T(\delta)^2 \lambda_{a2} \lambda_t \sigma_{ry}^2} \right). \tag{41}$$

Let $X_1 \ldots, X_N$ be the order statistics of variables with the standard normal density. The joint distribution of the maximum, X_N, and the second maximum variable, X_{N-1}, of N independent ones with the standard normal density is

$$f_{X_{(N-1)}, X_{(N)}}(x_{N-1}, x_N) = N(N-1)\phi(x_N)\phi(x_{N-1})\Phi(x_{N-1})^{N-2}, \tag{42}$$

where ϕ and Φ are the pdf and cdf of the standard normal distribution, respectively. The density of $D = X_N - X_{N-1}$, which is the difference between the maximum and second largest variable, can be bounded by $N\phi(0)$ as follows:

$$\begin{aligned} f_D(d) &= \int f_{D,X_{N-1}}(d, x_{N-1}) dx_{N-1} \\ &= \int N(N-1)\phi(x_{N-1}+d)\phi(x_{N-1})\Phi(x_{N-1})^{N-2} dx_{N-1} \\ &\leq N\phi(0). \end{aligned} \tag{43}$$

Thus, the density γD is bounded by $N\phi(0)/\gamma = N/\sqrt{2\pi\gamma^2}$.

We denote $\Delta_t = (y_{a^\star(t)}(t) - y_{a^\cdot(t)}(t))^\top \eta_\star$. The term on the right hand side is the upper bound $\mathbb{P}(a^\star(t) \neq a(t)|B(t), \Delta_t)$. Thus, by marginalizing Δ_t from it, we have

$$\begin{aligned} \mathbb{P}(a^\star(t) \neq a(t)|B(t)) &= \int_{-\infty}^{\infty} \mathbb{P}(a^\star(t) \neq a(t)|B(t), \Delta_t) f_{\Delta_t}(\Delta_t) d\Delta_t \\ &\leq 2\int_{-\infty}^{\infty} \exp\left(-\frac{\Delta_t^2}{2d_y^2\lambda_{a2}\upsilon_T(\delta)^2\lambda_t\sigma_{ry}^2}\right) f_{\Delta_t}(\Delta_t) d\Delta_t \\ &\leq 2Nd_y\lambda_{a2}^{1/2}\upsilon_T(\delta)\lambda_t^{1/2}\gamma_{ry}/\sqrt{\eta_\star^T S_y \eta_\star}, \end{aligned}$$

where the density of Δ_t, $f_{\Delta_t}(\Delta_t)$, is bounded by $N/\sqrt{2\pi\eta_\star^\top S_y \eta_\star}$ by (43).

.2.6 Proof of Lemma 8

We construct a martingale difference sequence that satisfies the conditions in Lemma 4. To that end, let $G_1 = H_1 = 0$,

$$G_\tau = (t-1)^{-1/2} I(a^\star(t) \neq a(t)) - (t-1)^{-1/2}\mathbb{P}(a^\star(t) \neq a(t)|\mathscr{F}_{t-1}^*),$$

and $H_t = \sum_{\tau=1}^t G_\tau$, where

$$\mathscr{F}_{t-1}^* = \sigma\{\{B(\tau)\}_{1 \leq \tau \leq t-1}\}.$$

Since $\mathbb{E}[G_\tau|\mathscr{F}^*_{\tau-1}] = 0$, the above sequences $\{G_\tau\}_{\tau\geq 0}$ and $\{H_\tau\}_{\tau\geq 0}$ are a martingale difference sequence and a martingale with respect to the filtration $\{\mathscr{F}^*_\tau\}_{1\leq\tau\leq T}$, respectively. Let $c_\tau = 2(\tau-1)^{-1/2}$. Since $\sum_{\tau=1}^T |G_\tau| \leq \sum_{\tau=2}^T c_\tau^2 \leq 4\log T$, by Lemma 4, we have

$$\mathbb{P}(H_T - H_1 > \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{8\sum_{t=1}^T c_t^2}\right) \leq \exp\left(-\frac{\varepsilon^2}{32\log T}\right).$$

Thus, with the probability at least $1-\delta$, it holds that

$$\sum_{t_T^*\leq t\leq T} \frac{1}{\sqrt{t-1}} I(a^\star(t) \neq a(t)) \leq \sqrt{32\log T \log\delta^{-1}} + \sum_{t_T^*\leq t\leq T} \frac{1}{\sqrt{t-1}} \mathbb{P}(a^*(\tau) \neq a(\tau)|\mathscr{F}^*_{\tau-1}).$$

.3 Appendices for Chapter 4

.3.1 Organization of Appendices

The appendices are organized as follows. First, we provide additional experiments with real datasets, which are not shown in Section 4.4. Second, Appendix .3.3 describes the shared parameter setup and Thompson sampling algorithm for it. Lastly, Appendix .3.4 provides the worst-case regret upper bounds for the model with a shared parameter, accompanied by its detailed proof.

.3.2 Real Data Experiments

In this section, we analyze two realdata sets used in Section 4.4 under two different assumptions. The first assumption is a simple linear regression, where a decision-maker gets a reward of 1 for successful classification and 0 otherwise. The reward is assumed to be generated as follows:

$$r_i(t) = x(t)^\top \mu_i + \varepsilon_i(t) = \mathbb{I}(l(t) = i),$$

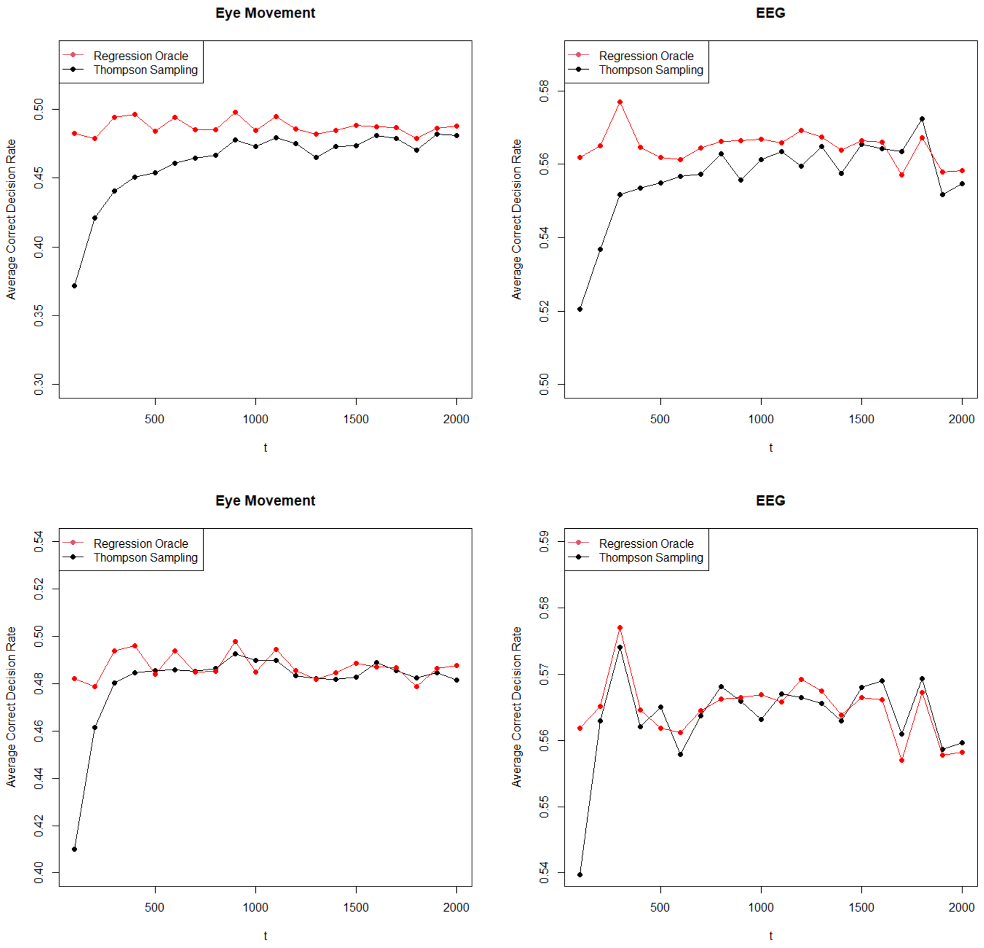

Figure 1: Plots of average correct decision rates of the regression oracle and Thompson sampling for Eye Movement (top left) and EEG dataset (top right) under the simple linear regression setup and Eye Movement (bottom left) and EEG dataset (bottom right) under the logistic linear regression setup.

where $l(t)$ is the true label of the patient randomly chosen at time t and $x(t)^\top \mu_i$ represents $\mathbb{P}(l(t) = i|x(t))$. This assumption is prone to a reward model misspecification since the expected value $\mathbb{P}(l(t) = i|x(t))$ is constrained to be between 0 and 1. Next, the second assumption is a logistic linear regression, introduced in Section 4.4.

For evaluation, we generate 100 scenarios for each dataset. We calculate the (average and worst case) regret as well as the average correct decision rate introduced in Section 4.4 for Thompson sampling versus

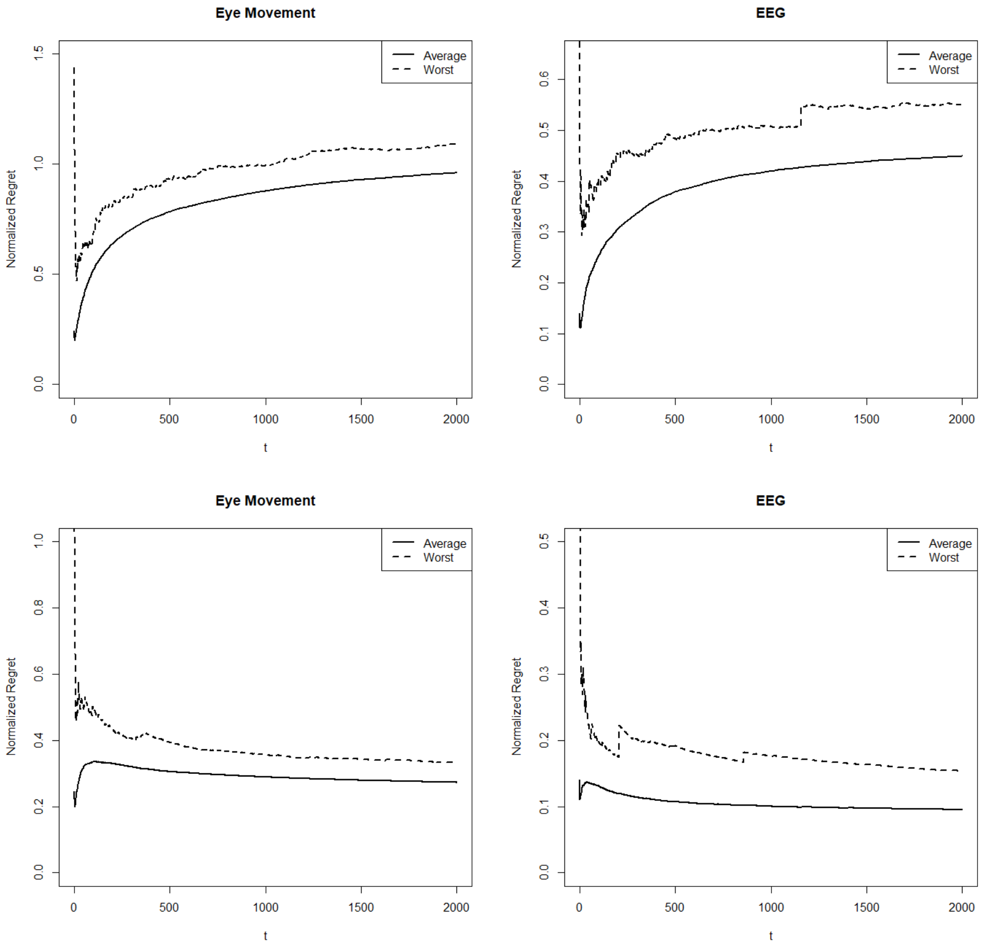

Figure 2: Plots of normalized regret of the regression oracle and Thompson sampling for Eye Movement (top left) and EEG dataset (top right) under the simple linear regression setup and Eye Movement (bottom left) and EEG dataset (bottom right) under the logistic linear regression setup.

regression oracle. We consider the estimates of regression oracle as the truth for regret evaluation for Thompson sampling. The observations are generated in the same manner introduced in Section 4.4.

Figure 1 displays the average correct decision rates of the regression oracle and Thompson sampling for the two real datasets under the two assumptions. We evaluate the mean correct decision rates over every 100 patients and then average them across 100 scenarios. Accordingly, each dot represents a sample mean of 10,000 results. For both data sets, the correct decision rate of Thompson sampling converges to that of the regression oracle over time. In addition, Figure 2 illustrates the average and worst cases of normalized regret of Thompson sampling against the regression oracle. Regret grows slightly faster than $\log^2 t$ for the simple linear model, but seems to scale with at most $\log^2 t$ in the logistic regression model. The over growth of regret in the first model can be caused by a potential model misspecification (Foster et al., 2020).

.3.3 Shared Parameter Setup

In Section 4.2, we describe the arm-specific parameter setup. In this section, we introduce the shared parameter setup, where the reward is generated

$$r_i(t) = x_i(t)^\top \mu_\cdot + \varepsilon_i(t).$$

In this setup, the parameter $\mu_\cdot$ is shared across all arms. Accordingly, the parameter can be learned regardless of the actions taken. Thus, the transformed parameter can be written as

$$\eta_\cdot = D^\top \mu_\cdot.$$

The optimal arm is

$$a^\star(t) = \text{argmax}_{i \in [N]} \, y_i(t)^\top \eta_\cdot$$

and subsequently regret is

$$\text{Regret}(T) = \sum_{t=1}^{T} (y_{a^\star(t)}(t) - y_{a(t)}(t))^\top \eta_\cdot.$$

The Thompson sampling algorithm for the shared parameter setup is basically the same as Algorithm 3, but simpler than it in that the estimate of transformed parameters and (unscaled) inverse covariance matrices are the same for all arms. Thus, we use the notation $\widehat{\eta}_\cdot(t)$ and $B_\cdot(t)$ for the estimate and (unscaled) inverse covariance, respectively. The update procedure for estimators is

$$\begin{aligned}
B_\cdot(t+1) &= B_\cdot(t) + y_{a(t)}(t) y_{a(t)}(t)^\top, && (44)\\
\widehat{\eta}_\cdot(t+1) &= B_\cdot(t+1)^{-1} \left(B_\cdot(t) \widehat{\eta}_\cdot(t) + y_{a(t)}(t) r_{a(t)}(t) \right), && (45)
\end{aligned}$$

Algorithm 4 : Thompson sampling algorithm for partially observable contextual bandits with a shared parameter

```
Set $B_{\cdot}(1) = I_{d_y}, \widehat{\eta}_{\cdot}(1) = \mathbf{0}_{d_y}$ for $i = 1, 2, \ldots, N$
for $t = 1, 2, \ldots, T$ do
    for $i = 1, 2, \ldots, N$ do
        Sample $\widetilde{\eta}_i(t)$ from $\mathcal{N}(\widehat{\eta}_{\cdot}(t), v^2 B_{\cdot}(t)^{-1})$
    end for
    Select arm $a(t) = \text{argmax}_{i \in [N]} y_i(t)^\top \widetilde{\eta}_i(t)$
    Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_{\cdot} + \varepsilon_{a(t)}(t)$
    Update $B_{\cdot}(t+1)$ and $\widehat{\eta}_{\cdot}(t+1)$ by (44) and (45)
end for
```

which is similar to (4.8) and (4.9), but there is not the indicator function $\mathbb{I}(a(t) = i)$. This implies that a decision-maker can learn the entire reward parameter regardless of the chosen arm in the shared parameter setup, while it learns an arm-specific parameter only given the arm chosen in the arm-specific parameter setup. The pseudo-code for the Thompson Sampling algorithm is described in Algorithm 4.

.3.4 Results for the shared parameter setup

In this section, we present the theoretical result of the model with a shared parameter described in Appendix .3.3. For this setup, we have $n_i(t) = t$ for all $i \in [N]$, which means that a decision-maker can learn the shared parameter regardless of the chosen arm. The next theorem provides a high probability regret upper bound for Thompson sampling for partially observable contextual bandits with a shared parameter.

Theorem 6. *Assume that Algorithm 4 is used in partially observable contextual bandits with a shared parameter. Then, the following regret bound holds with probability at least* $1 - \delta$*:*

$$\mathrm{Regret}(T) = \mathcal{O}\left(vNd_y^{2.5}\log^{3.5}\left(\frac{TNd_y}{\delta}\right)\right).$$

The regret bound scales at most $\log^{3.5} T$ with respect to the time horizon and linearly with N. $\sqrt{d_y \log(TNd_y/\delta)}$ and $d_y^2 \log^3(TNd_y/\delta)$ are incurred by the estimation errors and the minimum time, respectively. Lastly, N is resulted by the use of the inclusion-exclusion formula to find the bound for the sum of probabilities that the optimal arms are not chosen over time.

Note that a high probability logarithmic (with respect to time) upper bound for regret for the greedy algorithm under the normality assumption has been found for the model with a shared parameter by Park and Faradonbeh, 2022c. As compared to the setting in Park and Faradonbeh, 2022c, the result above is constructed based on less strict assumptions, in which contexts, observation noise, and reward noise have sub-Gaussian distributions for observation noise, contexts, and reward noise.

Proof. To begin with, as mentioned in Appendix .3.3, we have the following equalities in the shared parameter setting:

$$n_i(t) = t, \eta_i = \eta_\cdot, \widehat{\eta}_i(t) = \widehat{\eta}_\cdot(t), \text{ and } B_i(t) = B_\cdot(t),$$

for all $i \in [N]$. So, we decompose the regret as follows:

$$\begin{aligned}
&\mathrm{Regret}(T) = \sum_{t=1}^{T} (y_{a^\star(t)}(t) - y_{a(t)}(t))^\top \eta_{\blacksquare} \\
&\leq \sum_{t=1}^{\lfloor \nu_{(1)} \rfloor} 2c_\eta L + \sum_{t=\lceil \nu_{(1)} \rceil}^{T} \left((y_{a^\star(t)}(t) - y_{a(t)}(t))^\top \eta_{\blacksquare} + (y_{a(t)}(t) - y_{a^\star(t)}(t))^\top \widetilde{\eta}_{a(t)}(t) \right) \mathbb{I}(a^\star(t) \neq a(t)).
\end{aligned}$$

Now, because $\|y_i(t)\| \leq L$ for all $i \in [N]$ and $t \in [T]$, the above regret bound leads to

$$\mathrm{Regret}(T) \leq 2L \left(c_\eta \nu_{(1)} + \sum_{t=\lceil \nu_{(1)} \rceil}^{T} \|\widetilde{\eta}_\star(t) - \eta_\star\| \mathbb{I}(a^\star(t) \neq a(t)) \right).$$

By Lemma 15, if $t > \nu_{(1)}$, with probability at least $1 - \delta$, we have

$$\|\widetilde{\eta}_\star(t) - \eta_{\blacksquare}\| \leq g(\delta) t^{-1/2},$$

where

$$\begin{aligned}
g(\delta) &= 2 \left(v \sqrt{2d_y \log \frac{2TN}{\delta}} + R \sqrt{d_y \log \left(\frac{1 + TL^2/\lambda}{\delta} \right)} + c_\eta \right) \\
&= \mathcal{O} \left(\sqrt{d_y \log(TNd_y/\delta)} \right).
\end{aligned}$$

Now, we use Azuma's inequality to find a high probability upper bound for $\sum_{t=\lceil \nu_{(1)} \rceil}^{T} t^{-1/2} \mathbb{I}(a^\star(t) \neq a(t))$. For that purpose, consider the martingale sequence $\sum_{\tau=1}^{t} (\tau^{-1/2} \mathbb{I}(a^\star(\tau) \neq a(\tau)) - \tau^{-1/2} \mathbb{P}(a^\star(\tau) \neq a(\tau)))$ with respect to a filtration $\{\sigma\{\varnothing\}\}_{\tau=1}^{t-1}$, where $\varnothing$ is the empty set. Since $t^{-1/2} \mathbb{I}(a^\star(t) \neq a(t)) \leq t^{-1/2}$ and $\sum_{t=\lceil \nu_{(1)} \rceil}^{T} 2t^{-1} \leq 4 \log T$ (assuming $\lceil \nu_{(1)} \rceil \geq 2$), we have

$$\mathbb{P} \left(\sum_{t=\lceil \nu_{(1)} \rceil}^{T} \frac{1}{\sqrt{t}} \mathbb{I}(a^\star(t) \neq a(t)) - \sum_{t=\lceil \nu_{(1)} \rceil}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(t) \neq a(t)) > \varepsilon \right) \leq \exp \left(-\frac{\varepsilon^2}{4 \log T} \right).$$

By putting $\delta = \exp\left(-\varepsilon^2/(4\log T)\right)$, with probability at least $1-\delta$, we have

$$\sum_{t=\lceil \nu_{(1)}\rceil}^{T} \frac{1}{\sqrt{t}} \mathbb{I}(a^\star(t) \neq a(t)) \leq \sqrt{4\log T \log \delta^{-1}} + \sum_{t=\lceil \nu_{(1)}\rceil}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(t) \neq a(t)). \tag{46}$$

Now, we proceed towards establishing an upper bound for the second term on the right side in (46). Denote $A_{it}^\star = \{y(t) \in A_i^\star\}$, where $A_i^\star$ is defined in Definition 1. By using the fact that

$$\begin{aligned}
&\{y_i(t)^\top \widetilde{\eta}_i(t) < y_j(t)^\top \widetilde{\eta}_j(t)\} \\
&\subset \left\{ y_j(t)^\top (\widetilde{\eta}_j(t) - \eta_\cdot) > \frac{1}{2}(y_i(t) - y_j(t))^\top \eta_\cdot \right\} \bigcup \left\{ y_i(t)^\top (\widetilde{\eta}_i(t) - \eta_\cdot) < -\frac{1}{2}((y_i(t) - y_j(t))^\top \eta_\cdot) \right\},
\end{aligned}$$

we get

$$\begin{aligned}
& \mathbb{P}(y_j(t)^\top \widetilde{\eta}_j(t) - y_i(t)^\top \widetilde{\eta}_i(t) > 0 | G_{t-1}^\star, A_{it}^\star) \\
\leq\ & \mathbb{P}(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_\cdot(t)) > -y_j(t)^\top (\widehat{\eta}_\cdot(t) - \eta_\cdot) + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot | G_{t-1}^\star, A_{it}^\star) \\
+\ & \mathbb{P}(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -y_i(t)^\top (\widehat{\eta}_\cdot(t) - \eta_\cdot) + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot | G_{t-1}^\star, A_{it}^\star).
\end{aligned}$$

By Lemma 14, with probability of at least $1-\delta$, we have

$$y_i(t)^\top (\widehat{\eta}_\cdot(t) - \eta_\cdot) \leq \frac{h(\delta, T)\|y(t)\|}{t^{1/2}},$$

for all $\nu_{(1)} < t \leq T$ and $i \in [N]$, where

$$h(\delta, T) = \sqrt{\frac{2}{\lambda_m}} \left(R\sqrt{d_y \log\left(\frac{1+TL^2}{\delta}\right)} + c_\eta \right) = \mathcal{O}\left(R\sqrt{d_y \log(TNd_y/\delta)} \right).$$

Accordingly, we have

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top \widetilde{\eta}_j(t) - y_j(t)^\top \widetilde{\eta}_i(t) > 0 \middle| A^\star_{it}\right) \\
\leq\ & \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -h(\delta, T)\|y(t)\| t^{-1/2} + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot \middle| A^\star_{it}\right) \\
+\ & \mathbb{P}\left(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_\cdot(t)) > -h(\delta, T)\|y(t)\| t^{-1/2} + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot \middle| A^\star_{it}\right). \qquad (47)
\end{aligned}$$

Now, let $E_{ijt} = \{h(\delta, T)\|y(t)\|t^{-1/2} < 0.25(y_i(t) - y_j(t))^\top \eta_\cdot\}$. Then, we can decompose the first term on the RHS in (47) as follows:

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -\frac{h(\delta, T)\|y(t)\|}{t^{1/2}} + (y_i(t) - y_j(t))^\top \eta_\cdot \middle| A^\star_{it}\right) \\
&= \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -\frac{h(\delta, T)\|y(t)\|}{t^{1/2}} + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot \middle| E_{ijt}, A^\star_{it}\right) \mathbb{P}(E_{ijt}|A^\star_{it}) \\
&+ \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -\frac{h(\delta, T)\|y(t)\|}{t^{1/2}} + 0.5(y_i(t) - y_j(t))^\top \eta_\cdot \middle| E^c_{ijt}, A^\star_{it}\right) \mathbb{P}(E^c_{ijt}|A^\star_{it}).
\end{aligned}$$
$$(48)$$

We aim to show that the above probability is $\mathcal{O}(t^{-0.5})$ by showing each term in the RHS of (48) is $\mathcal{O}(t^{-0.5})$. By Assumption 1, if $t > \nu_{(1)}$, we have

$$\mathbb{P}(E^c_{ijt}|A^\star_{it}) = \mathbb{P}\left(4h(\delta, T)t^{-1/2} > (y_i(t) - y_j(t))^\top \eta_\cdot / \|y(t)\| \middle| A^\star_{it}\right) \leq \frac{4h(\delta, T)C}{t^{1/2}}. \qquad (49)$$

Thus, we showed that the second term in (48) is $\mathcal{O}(t^{-0.5})$. Now, we aim to show that the first term in (48) is $\mathcal{O}(t^{-0.5})$. Note that

$$\begin{aligned}
&\mathbb{P}\left(\dot{y}_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -\frac{h(\delta, T)}{t^{1/2}} + 0.5(\dot{y}_i(t) - \dot{y}_j(t))^\top \eta_\cdot \middle| E_{ijt}, A^\star_{it}\right) \\
\leq\ & \mathbb{P}\left(\dot{y}_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > 0.25(\dot{y}_i(t) - \dot{y}_j(t))^\top \eta_\cdot \middle| A^\star_{it}\right), \qquad (50)
\end{aligned}$$

where $\dot{y}_i(t) = y_i(t)/\|y(t)\|$. Now it suffices to show that the first term in the RHS of the inequality above is $\mathcal{O}(t^{-1/2})$. Using $\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) \sim \mathcal{N}(0, v^2\dot{y}_i(t)^\top B_\cdot(t)^{-1}\dot{y}_i(t))$ given $y(t)$ and $\mathscr{G}^\star_{t-1}$ and $\lambda_{\min}(B_\cdot(t)^{-1}) \leq 2/(\lambda_m t)$ by Lemma 12, we have

$$\mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > 0.25(\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot|y(t), A^\star_{it}) \leq \exp\left(-\frac{t\lambda_m((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot)^2}{64v^2}\right).$$

Thus, the first term on the RHS of the above inequality can be written as

$$\begin{aligned} & \mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > 0.25(\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot|A^\star_{it}) \\ = \; & \mathbb{E}[\mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > 0.25(\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot|y(t), A^\star_{it})|A^\star_{it}] \\ \leq \; & \mathbb{E}\left[\exp\left(-\frac{t\lambda_m((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot)^2}{64v^2}\right)\middle| A^\star_{it}\right]. \end{aligned}$$

By integration by part, we have

$$\begin{aligned} & \mathbb{E}\left[\exp\left(-\frac{t\lambda_m((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot)^2}{64v^2}\right)\middle| A^\star_{it}\right] \\ = \; & \int_0^\infty \frac{2t\lambda_m u}{64v^2}\exp\left(-\frac{t\lambda_m u^2}{64v^2}\right)\mathbb{P}((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot < u|A^\star_{it})du. \end{aligned}$$

Since $\mathbb{P}((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot < u|A^\star_{it}) \leq Cu$ for $C > 0$ based on Assumption 1, the term above can be written as

$$\begin{aligned} & \mathbb{E}\left[\exp\left(-\frac{t\lambda_m((\dot{y}_i(t) - \dot{y}_j(t))^\top\eta_\cdot)^2}{64v^2}\right)\middle| A^\star_{it}\right] \\ \leq \; & \frac{2\sqrt{\pi}}{\sqrt{64v^2/(\lambda_m t)}}\int_0^\infty \frac{2t\lambda_m u}{\sqrt{64\pi v^2/(\lambda_m t)}}\exp\left(-\frac{u^2}{64v^2/(t\lambda_m)}\right)Cudu = 8vC\sqrt{\frac{\pi}{\lambda_m t}}, \end{aligned}$$

where we used the following result about one-sided Gaussian integrals

$$\int_0^\infty x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \sigma^2/2.$$

Thus, combining (49) and (50), we get

$$\begin{aligned} & \mathbb{P}(\dot{y}_i(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_\cdot(t)) > -\dot{y}_i(t)^\top(\widehat{\eta}_\cdot(t) - \eta_\cdot) + 0.5(\dot{y}_i(t) - \dot{y}_j(t))^\top \eta_\cdot | A_{it}^\star) \\ \leq \; & Ct^{-1/2}\left(8v\sqrt{\frac{\pi}{\lambda_m}} + 4h(\delta, T)\right). \end{aligned}$$

Similarly, we have

$$\begin{aligned} & \mathbb{P}(\dot{y}_j(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_\cdot(t)) > -\dot{y}_j(t)^\top(\widehat{\eta}_\cdot(t) - \eta_\cdot) + 0.5(\dot{y}_i(t) - \dot{y}_j(t))^\top \eta_\cdot | A_{it}^\star) \\ \leq \; & Ct^{-1/2}\left(8v\sqrt{\frac{\pi}{\lambda_m}} + 4h(\delta, T)\right). \end{aligned}$$

Using (47), we have

$$\begin{aligned} & \mathbb{P}(\dot{y}_j(t)^\top \widetilde{\eta}_j(t) - \dot{y}_i(t)^\top \widetilde{\eta}_i(t)) > 0 | A_{it}^\star) \\ \leq \; & LCt^{-1/2}\left(v\left(8\sqrt{\frac{\pi}{\lambda_m}} + 8\sqrt{\frac{\pi}{\lambda_m}}\right) + 4h(\delta, T) + 4h(\delta, T)\right). \end{aligned} \tag{51}$$

By summing the probabilities in (51) over $i, j \in [N]$, we have

$$\begin{aligned} & \sum_{i=1}^N \sum_{j=1}^N \mathbb{P}(\dot{y}_j(t)^\top \widetilde{\eta}_j(t) - \dot{y}_i(t)^\top \widetilde{\eta}_i(t) > 0 | A_{it}^\star) \mathbb{P}(A_{it}^\star) \\ \leq \; & \frac{2C}{\sqrt{t}} \sum_{i=1}^N \sum_{j=1}^N \mathbb{P}(A_{it}^\star)\left(8v\sqrt{\frac{\pi}{\lambda_m}} + 4h(\delta, T)\right) \\ \leq \; & \frac{2c_M(\delta, T)CN}{\sqrt{t}}, \end{aligned} \tag{52}$$

where

$$c_M(\delta, T) = 8v\sqrt{\pi/\lambda_m} + 4h(\delta, T) = \mathcal{O}\left(v\sqrt{d_y \log(Td_y/\delta)}\right). \tag{53}$$

Note that by using the inclusion-exclusion formula, we can bound the probability of pulling a sub-optimal arm as follows

$$\mathbb{P}(a^\star(t) \neq a(t)) \leq \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0 | A_{it}^\star)\mathbb{P}(A_{it}^\star). \tag{54}$$

Putting (52), (54), and the minimal sample size $\nu_{(1)}$ together, we obtain the following inequality

$$\sum_{t=\lceil \nu_{(1)} \rceil}^{T} \frac{1}{\sqrt{t}}\mathbb{P}(a^\star(t) \neq a(t)) \quad \leq \quad 2c_M(\delta, T)CN \log T.$$

Then, according to (46), with probability at least $1 - \delta$, it holds that

$$\sum_{t=\lceil \nu_{(1)} \rceil}^{T} \frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t) \neq a(t)) \leq \sqrt{4 \log T \log \delta^{-1}} + 2c_M(\delta, T)CN \log T.$$

Therefore, using $\nu_{(1)} = 8L^4 \log(T/\delta)/\lambda_m^2$ and $L = c_y\sqrt{2d_y \log(TNd_y/\delta)}$, with probability at least $1 - \delta$, the regret bound below holds true

$$\begin{aligned} \text{Regret}(T) \quad &\leq \quad 2c_\eta L\nu_{(1)} + Lg(\delta)\left(\sqrt{4 \log T \log \delta^{-1}} + 2c_M(\delta, T)CN \log T\right) \\ &= \quad \mathcal{O}\left(vNd_y^{2.5} \log^{3.5}\left(\frac{TNd_y}{\delta}\right)\right). \end{aligned}$$

□

Bibliography

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, *24*, 2312–2320.

Abbasi-Yadkori, Y., & Szepesvári, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. *Proceedings of the 24th Annual Conference on Learning Theory*, 1–26.

Abe, N., & Long, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. *ICML*, 3–11.

Abeille, M., & Lazaric, A. (2017). Linear thompson sampling revisited. *Artificial Intelligence and Statistics*, 176–184.

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (Vol. 55). US Government printing office.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., & Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. *International Conference on Machine Learning*, 1638–1646.

Agrawal, S., & Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. *Conference on learning theory*, 39–1.

Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. *International Conference on Machine Learning*, 127–135.

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, *10*(1), 174–205.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*(Nov), 397–422.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*(2), 235–256.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, *19*(3), 357–367.

Bastani, H., & Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, *68*(1), 276–294.

Bastani, H., Bayati, M., & Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science*, *67*(3), 1329–1349.

Bensoussan, A. (2004). *Stochastic control of partially observable systems*. Cambridge University Press.

Bietti, A., Agarwal, A., & Langford, J. (2021). A contextual bandit bake-off. *The Journal of Machine Learning Research*, *22*(1), 5928–5976.

Bouneffouf, D., Bouzeghoub, A., & Gançarski, A. L. (2012). A contextual-bandit algorithm for mobile context-aware recommender system. *International conference on neural information processing*, 324–331.

Bouneffouf, D., Rish, I., Cecchi, G. A., & Féraud, R. (2017). Context attentive bandits: Contextual bandit with restricted context. *arXiv preprint arXiv:1705.03821*.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Chakraborty, S., Roy, S., & Tewari, A. (2023). Thompson sampling for high-dimensional sparse linear contextual bandits. *International Conference on Machine Learning*, 3979–4008.

Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, *24*, 2249–2257.

Chu, W., Li, L., Reyzin, L., & Schapire, R. (2011). Contextual bandits with linear payoff functions. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

Cramér, H. (2016). *Mathematical methods of statistics (pms-9), volume 9*. Princeton university press.

Dani, V., Hayes, T. P., & Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Doob, J. L. (1953). *Stochastic processes* (Vol. 10). New York Wiley.

Dougherty, E. R. (2020). *Digital image processing methods*. CRC Press.

Dumitrascu, B., Feng, K., & Engelhardt, B. (2018). Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, *31*.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., & Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. *Machine learning for healthcare conference*, 67–82.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica*, *96*, 342–353.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2019). On applications of bootstrap in continuous space reinforcement learning. *2019 IEEE 58th Conference on Decision and Control (CDC)*, 1977–1984.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2020a). Input perturbations for adaptive control and learning. *Automatica*, *117*, 108950.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2020b). On adaptive linear–quadratic regulators. *Automatica*, *117*, 108982.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2020c). Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, *66*(4), 1802–1808.

Foster, D. J., Gentile, C., Mohri, M., & Zimmert, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, *33*, 11478–11489.

Garivier, A., Lattimore, T., & Kaufmann, E. (2016). On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, *29*.

Garivier, A., Ménard, P., & Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, *44*(2), 377–399.

Goldenshluger, A., & Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems*, *3*(1), 230–261.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, *25*(1), 16–18.

Guan, M., & Jiang, H. (2018). Nonparametric stochastic contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Hamidi, N., & Bayati, M. (2020). On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*.

Harville, D. (1976). Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, *4*(2), 384–395.

Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., & Boutilier, C. (2020). Latent bandits revisited. *arXiv preprint arXiv:2006.08714*.

Howard, T. M., Green, C. J., Kelly, A., & Ferguson, D. (2008). State space sampling of feasible motions for high-performance mobile robot navigation in complex environments. *Journal of Field Robotics*, *25*(6-7), 325–345.

Hu, T., Laber, E. B., Li, Z., Meyer, N. J., & Pacifici, K. (2019). Note on thompson sampling for large decision problems. *arXiv preprint arXiv:1905.04735*.

Hu, Y., Kallus, N., & Mao, X. (2020). Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *Conference on Learning Theory*, 2007–2010.

Jose, S. T., & Moothedath, S. (2024). Thompson sampling for stochastic bandits with noisy contexts: An information-theoretic regret analysis. *arXiv preprint arXiv:2401.11565*.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99–134.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kang, Y., Roh, C., Suh, S.-B., & Song, B. (2012). A lidar-based decision-making method for road boundary detection using multiple kalman filters. *IEEE Transactions on Industrial Electronics*, *59*(11), 4360–4368.

Kargin, T., Lale, S., Azizzadenesheli, K., Anandkumar, A., & Hassibi, B. (2023). Thompson sampling for partially observable linear-quadratic control. *2023 American Control Conference (ACC)*, 4561–4568.

Kaufmann, E., Cappé, O., & Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, *17*(1), 1–42.

Kim, J.-h., Yun, S.-Y., Jeong, M., Nam, J., Shin, J., & Combes, R. (2023). Contextual linear bandits under noisy features: Towards bayesian oracles. *International Conference on Artificial Intelligence and Statistics*, 1624–1645.

Kirschner, J., Lattimore, T., & Krause, A. (2020). Information directed sampling for linear partial monitoring. *Conference on Learning Theory*, 2328–2369.

Krishnamurthy, V., & Wahlberg, B. (2009). Partially observed markov decision process multiarmed bandits—structural results. *Mathematics of Operations Research*, *34*(2), 287–302.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22.

Lattimore, T. (2018). Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, *19*(20), 1–32.

Lattimore, T. (2022). Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *Conference on Learning Theory*, 1547–1575.

Lattimore, T., & Gyorgy, A. (2021). Mirror descent and the information ratio. *Conference on Learning Theory*, 2965–2992.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*, 661–670.

Lin, J.-W., Chen, C.-W., & Peng, C.-Y. (2012). Kalman filter decision systems for debris flow hazard assessment. *Natural hazards*, *60*(3), 1255–1266.

Maillard, O.-A., & Mannor, S. (2014). Latent bandits. *International Conference on Machine Learning*, 136–144.

Modi, A., & Tewari, A. (2020). No-regret exploration in contextual reinforcement learning. *Conference on Uncertainty in Artificial Intelligence*, 829–838.

Nagrath, I. (2006). *Control systems engineering*. New Age International.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462.

Park, H., & Faradonbeh, M. K. S. (2021). Analysis of thompson sampling for partially observable contextual multi-armed bandits. *IEEE Control Systems Letters*, *6*, 2150–2155.

Park, H., & Faradonbeh, M. K. S. (2022a). Efficient algorithms for learning to control bandits with unobserved contexts. *IFAC-PapersOnLine*, *55*(12), 383–388.

Park, H., & Faradonbeh, M. K. S. (2022b). A regret bound for greedy partially observed stochastic contextual bandits. *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*.

Park, H., & Faradonbeh, M. K. S. (2022c). Worst-case performance of greedy policies in bandits with imperfect context observations. *arXiv preprint arXiv:2204.04773*.

Park, H., & Faradonbeh, M. K. S. (2024). Thompson sampling in partially observable contextual bandits. *arXiv preprint arXiv:2402.10289*.

Park, H., & Faradonbeh, M. K. S. (n.d.-a). Balancing exploration and exploitation in partially observed linear contextual bandits via thompson sampling. *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*.

Park, H., & Faradonbeh, M. K. S. (n.d.-b). Online learning of optimal prescriptions under bandit feedback with unknown contexts. *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*.

Park, H., & Faradonbeh, M. K. S. (n.d.-c). Sequentially adaptive experimentation for learning optimal options subject to unobserved contexts. *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*.

Raghavan, M., Slivkins, A., Vaughan, J. W., & Wu, Z. S. (2020). Greedy algorithm almost dominates in smoothed contextual bandits. *arXiv preprint arXiv:2005.10624*.

Raghavan, M., Slivkins, A., Vaughan, J. W., & Wu, Z. S. (2023). Greedy algorithm almost dominates in smoothed contextual bandits. *SIAM Journal on Computing*, *52*(2), 487–524.

Ren, Z., & Zhou, Z. (2020). Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical science*, 15–32.

Roesser, R. (1975). A discrete state-space model for linear image processing. *IEEE Transactions on Automatic Control*, *20*(1), 1–10.

Rudin, W., et al. (1976). *Principles of mathematical analysis* (Vol. 3). McGraw-hill New York.

Russo, D., & Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, *39*(4), 1221–1243.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2017). A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*.

Slivkins, A. (2011). Contextual bandits with similarity information. *Proceedings of the 24th annual Conference On Learning Theory*, 679–702.

Stratonovich, R. L. (1959). Optimum nonlinear systems which bring about a separation of a signal with constant parameters from noise. *Radiofizika*, *2*(6), 892–901.

Stratonovich, R. L. (1960). Application of the markov processes theory to optimal filtering. *Radio Engineering and Electronic Physics*, *5*, 1–19.

Surmann, H., Jestel, C., Marchel, R., Musberg, F., Elhadj, H., & Ardani, M. (2020). Deep reinforcement learning for real autonomous mobile robot navigation in indoor environments. *arXiv preprint arXiv:2005.13857*.

Tennenholtz, G., Shalit, U., Mannor, S., & Efroni, Y. (2021). Bandits with partially observable confounded data. *Conference on Uncertainty in Artificial Intelligence. PMLR*.

Tewari, A., & Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile health* (pp. 495–517). Springer.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, *12*(4), 389–434.

Tsuchiya, T., Ito, S., & Honda, J. (2023). Best-of-both-worlds algorithms for partial monitoring. *International Conference on Algorithmic Learning Theory*, 1484–1515.

Varatharajah, Y., Berry, B., Koyejo, S., & Iyer, R. (2018). A contextual-bandit-based approach for informed decision-making in clinical trials. *arXiv preprint arXiv:1809.00258*.

Wanigasekara, N., & Yu, C. (2019). Nonparametric contextual bandits in metric spaces with unknown metric. *Advances in Neural Information Processing Systems*, *32*.

Wong, K. C., Li, Z., & Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *Annals of statistics*, *48*(2).

Yun, S.-Y., Nam, J. H., Mo, S., & Shin, J. (2017). Contextual multi-armed bandits under feature uncertainty. *arXiv preprint arXiv:1703.01347*.

Zhou, L., & Brunskill, E. (2016). Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*.