Online Learning of Optimal Prescriptions under Bandit Feedback with Unknown Contexts

Hongju Park
Department of Statistics
University of Georgia
Athens, GA 30602
hp97161@uga.edu

Mohamad Kazem Shirani Faradonbeh
Department of Mathematics
Southern Methodist University
Dallas, TX 75205
mohamadksf@mail.smu.edu

Abstract

Contextual bandits constitute a classical framework for decision-making under uncertainty. In this setting, the goal is to learn prescriptions of highest reward subject to the contextual information, while the unknown reward parameters of each prescription need to be learned by experimenting it. Accordingly, a fundamental problem is that of balancing exploration (i.e., prescribing different options to learn the parameters), versus exploitation (i.e., sticking with the best option to gain reward). To study this problem, the existing literature mostly considers perfectly observed contexts. However, the setting of partially observed contexts remains unexplored to date, despite being theoretically more general and practically more versatile. We study bandit policies for learning to select optimal prescriptions based on observations, which are noisy linear functions of the unobserved context vectors. Our theoretical analysis shows that the Thompson sampling policy successfully balances exploration and exploitation. Specifically, we establish (i) regret bounds that grow poly-logarithmically with time, (ii) square-root consistency of parameter estimation, and (iii) scaling with other quantities including dimensions and number of options. Extensive numerical experiments with both real and synthetic data are presented as well, corroborating the efficacy of Thompson sampling. To establish the results, we utilize 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 similar problems.

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 context(s) at each time. The data collected thus far is utilized, aiming to maximize cumulative rewards determined by both the context(s) and unknown parameters. So, any desirable policy needs to manage the delicate 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 strong performance 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 missing for imperfectly observed contexts, and it is adopted as the focus of this work.

Letting the time-varying components of the decision options (e.g., contexts) to be observed only partially, is known to be advantageous. More specifically, in various real-world problems including robot control and image processing [1–6], partial, transformed, or noisy signal-observation models

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

have been used traditionally to obtain better performance. On the other hand, overlooking imperfectness of observations can lead to compromised decisions. For example, if disregarding uncertainty in medical profiles of septic patients, clinical decisions end up with worse consequences [7]. Accordingly, partial observation models are studied in canonical settings such as linear systems [8], bandit monitoring [9–11], and Markov decision processes [12, 13]. The above have recently motivated some work on contextual bandit policies with partially observed contexts [14–16]. However, a reliable policy that can provably balance exploration and exploitation is not currently available, as will be elaborated shortly, after clarifying the technical setting and reviewing the literature.

The common bandit setting is the so-called linear one, where the expected reward of each arm is the inner product of context(s) and reward parameter(s). The latter can be either *arm-specific* [17, 18], or *shared* across all arms [19, 20]. We consider a framework for capturing both settings, with the focus being on the more general and challenging one of the former. Moreover, similar to the above references, we assume that there are finitely many arms to choose from. 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 [21, 20], as well as those with adversarial contexts [19, 22], together with non-linear or non-parametric reward functions [23–25]. Notably, all of these references assume fully observed contexts, in contradistinction to this work.

The discourse of efficient policies for contextual bandits has come a long way. Algorithms based on Optimism in the Face of Uncertainty (OFU) [26, 19, 21] held prominent positions and afterward was followed by Thompson sampling with its excelling empirical performance [27] and then supplemented with theoretical analysis [17, 28, 20]. More recently, Greedy policies have been shown to be nearly optimal under particular settings [29, 16], though, it is known that vanilla Greedy algorithms incur a linear regret under the arm-specific reward parameter setup [30]. That is caused, intuitively, by superior arms dominating some others, leaving them unexplored, and is also illustrated in our experiments at the end of this paper. Accordingly, the study of theoretical performance guarantees for Thompson sampling has gained much popularity and made significant progress in the recent literature. First, regret bounds growing as square-root of time were shown [17, 28, 20], succeeded by logarithmic regret in setting with a shared reward parameter under parameter sparsity [31] and partial observability of contexts [14, 15]. However, for the general case (that each arm has a possibly distinct reward parameter), the efficiency of bandit policies remains unanswered. Indeed, the analysis is more challenging in such settings as the policy needs to address the trade-off between exploration and exploitation, unlike the setting with a shared reward parameter. The problem constitutes the focus of this paper.

We analyze the Thompson sampling policy in contextual bandits with partially observable stochastic contexts. Our analysis indicates 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 the fourth power of the logarithm of time. For regret analysis in contextual bandits, it is crucial to examine the interdependencies of variables associated with regret. To address this issue, we delicately construct stochastic processes with self-normalized or martingale structures, and employ useful stochastic bounds for them.

The organization is outlined below. In Section 2, we formulate the problem and discuss preliminaries. Next, Thompson sampling policy for partially observable contextual bandits is presented in Section 3. We provide its theoretical performance guarantees in Section 4, followed by real-data experiments in Section 5. The paper is wrapped by final remarks and future directions.

2 Problem Formulation

In this section, we express the technicalities of the partially observable linear contextual bandit problem. The decision-maker tries to maximize their cumulative reward by selecting from N arms, the reward of arm $i \in \{1, \cdots, N\}$ being

$$r_i(t) = x(t)^\top J_i \mu_\star + \varepsilon_i(t), \tag{1}$$

where $x(t)$ is the **unobserved** d_x dimensional stochastic context generated independently at time t with $\mathbb{E}\left[x(t)\right] = \mathbf{0}_{d_x}$ and $\mathrm{Cov}(x(t)) = \Sigma_x$, $\mu_\star$ is the reward parameter in $\mathbb{R}^{d_\mu}$, J_i is the $(d_x \times d_\mu)$ dimensional weight matrix of the ith arm, and $\varepsilon_i(t)$ is the reward sub-Gaussian noise satisfying there

exist a fixed positive constant R_1 such that

$$\mathbb{E}\left[e^{\lambda\varepsilon_i(t)}\right] \leq \exp\left(\frac{\lambda^2 R_1^2}{2}\right),\ \forall\,\lambda > 0. \tag{2}$$

The decision-making policy observes $y(t)$; a transformed noisy function of the context

$$y(t) = Ax(t) + \xi(t), \tag{3}$$

where A is a $d_y \times d_x$ sensing matrix, and $\xi(t)$ is the sensing (or measurement) noise, its covariance matrix being denoted by Σ_y. We suppose that each element of $\xi(t)$ is sub-Gaussian as well, and (without loss of generality) satisfies an inequality just similar to (2). At each time t, the decision-maker chooses an arm, denoted by $a(t)$, given the history of actions $\{a(\tau)\}_{1\leq\tau\leq t-1}$, rewards $\{r_{a(\tau)}(\tau)\}_{1\leq\tau\leq t-1}$, and past observations $\{y(\tau)\}_{1\leq\tau\leq t-1}$, as well as the current one $y(t)$. Once choosing the arm $a(t)$, the decision-maker gets a reward $r_{a(t)}(t)$ according to (1). Note that rewards of other arms are *not* realized.

Technically, the weight matrices are employed to represent different types of association between the parameter $\mu_\star$ and context $x(t)$ for the reward of each arm. For the sake of simplicity in our presentation and model design, we use the notation $\mu_\star$ to represent the entire original parameter related to all arms, whereas the parameter associated with arm i is denoted as $J_i\mu_\star$.

Now, we look into the optimal arm identification. Based on (1), the optimal arm i maximizes $x(t)^\top J_i\mu_\star$. To find an approximate value of $x(t)^\top J_i\mu_\star$, we utilize the information through $y(t)$. First, we estimate $x(t)^\top J_i\mu_\star$ with known $\mu_\star$ from the perspective of the optimal policy. We consider an arbitrary linear combination of $x(t)$, denoted by $x(t)^\top\mu$, for a known vector $\mu \in \mathbb{R}^{d_x}$ and a linear prediction $b^\top y(t)$ of it. The linear prediction $y(t)^\top b$ should be chosen so as to minimize the variance of prediction error, $\mathrm{Var}(x(t)^\top\mu - y(t)^\top b)$, subject to the condition that the predictor is unbiased $\mathbb{E}\left[x(t)^\top\mu - y(t)^\top b\right] = 0$ based on (3). This linear prediction is called the Best Linear Unbiased Prediction (BLUP) [32, 33], satisfying $b = D^\top\mu$, which is invariant of the value of $x(t)$, where $D = (A^\top\Sigma_y^{-1}A + \Sigma_x^{-1})^{-1}A^\top\Sigma_y^{-1}$. Accordingly, by plugging $J_i\mu_\star$ into μ, we get the BLUP of $x(t)^\top J_i\mu_\star$, written as $y(t)^\top D^\top J_i\mu_\star$. Similarly, $Dy(t)$ is the BLUP of $x(t)$, denoted by

$$\widehat{x}(t) := Dy(t). \tag{4}$$

Next, we examine the estimation of $x(t)^\top J_i\mu_\star$ from the perspective of a decision-maker, who does not know the true value of $\mu_\star$. From (1), we get

$$r_i(t) = y(t)^\top D^\top J_i\mu_\star + \zeta_i(t), \tag{5}$$

where $\zeta_i(t) = (x(t)^\top J_i\mu_\star - y(t)^\top D^\top J_i\mu_\star) + \varepsilon_i(t)$ is a noise centered at 0. $\zeta_i(t)$ is independent of others because of the independence of the prediction error, $x(t)^\top J_i\mu_\star - y(t)^\top D^\top J_i\mu_\star$. Here, $J_i\mu_\star$ is not estimable based on the equation (5), since the space spanned by $\{Dy(\tau)\}_{\tau=1:a(\tau)=i}^{t}$ does not generally include $J_i\mu_\star$, if $d_y < d_x$. Thus, instead of $J_i\mu_\star$, we estimate $D^\top J_i\mu_\star$ defined as the transformed parameter of the arm i, denoted by

$$\eta_i := D^\top J_i\mu_\star. \tag{6}$$

Thus, using (5) and (6), we get

$$r_i(t) = y(t)^\top\eta_i + \zeta_i(t). \tag{7}$$

Despite the inestimability of $J_i\mu_\star$, η_i is always guaranteed to be estimable because $\{y(\tau)\}_{\tau=1:a(\tau)=i}^{t}$ span $\mathbb{R}^{d_y}$, thanks to the full rank $\mathrm{Var}(y(t))$. Given that even the optimal policy cannot make a better unbiased prediction of $x(t)^\top J_i\mu_\star$ than the BLUP $y(t)^\top\eta_i$ by taking advantage of any other information, the optimal arm at time t is given as

$$a^\star(t) = \underset{1\leq i\leq N}{\mathrm{argmax}}\ y(t)^\top\eta_i. \tag{8}$$

Regret is a performance measure, quantifying the cumulative reward decrease by the actions of a decision-maker as compared to the actions taken by the optimal policy. In accordance with the optimal arm in (8), regret is expressed as

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} y(t)^\top\left(\eta_{a^\star(t)} - \eta_{a(t)}\right), \tag{9}$$

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

Now, we describe contextual bandits with arm-specific parameters in terms of weight matrices $\{J_i\}_{i=1}^N$. For the presentation of arm-specific parameters and contexts, we use the notations $\mu_\star = [\mu_{\star 1}^\top, \mu_{\star 2}^\top, \ldots, \mu_{\star N}^\top]^\top$ for arm-specific parameters, and $x(t) = [x_1(t)^\top, x_2(t)^\top, \ldots, x_N(t)^\top]^\top$ for arm-specific contexts. Herein, 'parameters' indicate transformed parameters $\{\eta_i\}_{i=1}^N$ defined in (6), while 'original parameter' represents $\mu_\star$. Reward functions can have two different types of contexts: shared and arm-specific type.

Shared Context: First, we illustrate the canonical model with arm-specific parameters and a shared context. For this framework, the weight matrices have the following form

$$J_i = \left[\mathbf{0}_{d_x \times d_x} \ \cdots \ \mathbf{0}_{d_x \times d_x} \ \underbrace{I_{d_x}}_{i\text{th}} \ \mathbf{0}_{d_x \times d_x} \ \cdots \ \mathbf{0}_{d_x \times d_x} \right], \ i \in [N]. \tag{10}$$

J_i selects $\mu_{\star i}$ from $\mu_\star = \left[\mu_{\star 1}^\top, \mu_{\star 2}^\top, \ldots, \mu_{\star N}^\top\right]^\top$ by a linear transformation, which means $J_i \mu_\star = \mu_{\star i}$. Accordingly, J_i satisfies the following equations: $x(t)^\top J_i \mu_\star = x(t)^\top \mu_{\star i}$.

Arm-specific Contexts: Second, the other canonical model with arm-specific parameters consists of arm-specific contexts. For this model, J_i is a block diagonal matrix such that

$$J_i = \mathrm{diag}\left(\mathbf{0}_{d_{\mu_0} \times d_{\mu_0}}, \ \cdots, \mathbf{0}_{d_{\mu_0} \times d_{\mu_0}}, \ \underbrace{I_{d_{\mu_0}}}_{i\text{th}}, \ \mathbf{0}_{d_{\mu_0} \times d_{\mu_0}}, \ \cdots \ \mathbf{0}_{d_{\mu_0} \times d_{\mu_0}} \right), \ i \in [N], \tag{11}$$

where $d_{\mu_0} = d_\mu / N$ is the dimension of an arm-specific parameter, $\mu_{\star i}$. Here, $J_i \mu_\star$ are 0 except for the ith d_{μ_0} elements, which is $\mu_{\star i}$. Correspondingly, J_i satisfies $x(t)^\top J_i \mu_\star = x_i(t)^\top \mu_{\star i}$.

3 Thompson Sampling Policy

In this section, we outline the Thompson sampling algorithm for partially observable contextual bandits. Thompson sampling takes action as if samples generated from a posterior distribution given the data thus far are the true values. In order to calculate a (hypothetical) posterior distribution, a decision-maker assumes that the reward of the ith arm at time t is generated as follows: $r_i(t) = y(t)^\top D^\top J_i \mu_\star + \psi_i(t)$, where $\psi_i(t)$ has the normal distribution with the mean 0 and variance $v^2 = R_1^2$ for R_1 defined in (2). In the beginning, the decision-maker starts with the initial value $\widehat{\mu}(1) = \mathbf{0}_{d_\mu}$ and $B(1) = I_{d_\mu}$, which are the mean and (unscaled) covariance matrix of a prior distribution of $\mu_\star$, respectively. The posterior distribution of $\mu_\star$ at time t is given as $\mathcal{N}(\widehat{\mu}(t), v^2 B(t)^{-1})$. By taking advantage of the posterior distribution of $\mu_\star$ and (6), we can derive the posterior distribution of η_i, which is $\mathcal{N}(\widehat{\eta}_i(t), v^2 B_i^+(t))$, where

$$\widehat{\eta}_i(t) = D^\top J_i \widehat{\mu}(t), \tag{12}$$

$$B_i^+(t) = D^\top J_i B(t)^{-1} J_i^\top D. \tag{13}$$

Then, we sample from the following posterior distribution of the transformed parameters η_i:

$$\widetilde{\eta}_i(t) \sim \mathcal{N}(\widehat{\eta}_i(t), v^2 B_i^+(t)), \tag{14}$$

for $i = 1, 2 \ldots, N$. Accordingly, the decision-maker pulls the arm $a(t)$ such that $a(t) = \underset{1 \le i \le N}{\mathrm{argmax}} \ y(t)^\top \widetilde{\eta}_i(t)$. Then, once the decision-maker gains the reward of the chosen arm $a(t)$, it can update $\widehat{\mu}(t)$ and $B(t)$ based on the recursions below:

$$B(t+1) = B(t) + J_{a(t)}^\top D y(t) y(t)^\top D^\top J_{a(t)}, \tag{15}$$

$$\widehat{\mu}(t+1) = B(t+1)^{-1} \left(B(t) \widehat{\mu}(t) + J_{a(t)}^\top D y(t) r_{a(t)}(t) \right). \tag{16}$$

Algorithm 1 : Thompson sampling algorithm for contextual bandits with imperfect context observations

Set $B(1) = I_{d_\mu}$, $\widehat{\mu}(1) = \mathbf{0}_{d_\mu}$
for $t = 1, 2, \ldots,$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Set $\widehat{\eta}_i(t) = D^\top J_i \widehat{\mu}(t)$ and $B_i(t) = D^\top J_i B(t) J_i^\top D$
    Sample $\widetilde{\eta}_i(t)$ from $\mathcal{N}(\widehat{\eta}_i(t), v^2 B_i^+(t))$
  end for
  Select arm $a(t) = \text{argmax}_{i \in [N]} y(t)^\top \widetilde{\eta}_i(t)$
  Gain reward $r_{a(t)}(t) = x(t)^\top J_{a(t)} \mu_\star + \varepsilon_{a(t)}(t)$
  Update $B(t+1)$ and $\widehat{\mu}(t+1)$ by (15) and (16)
end for

4 Theoretical Performance Analyses

In this section, we establish the theoretical results of Algorithm 1 for partially observable contextual bandits with arm-specific parameters. For the following results, without loss of generality, we set $\|\mu_\star\| \leq h$ and $c_\mu = 1$. To proceed, we define optimal probabilities.

Definition 1 (Optimal Probability)**.** *Let $A_i^\star \in \mathbb{R}^{d_y}$ be the region in the space of $y(t)$ that makes arm i optimal: $a^\star(t) = i$. Then, denote the optimality probability arm i by $p_i = \mathbb{P}(y(t) \in A_i^\star) = \mathbb{P}(a^\star(t) = i)$.*

Based on the above definition and Assumption 2 in Appendix B, we define a set $A_i^\star$ and $\kappa > 0$ such that, for $A_i^\star$, there exist a subset $A_i \subseteq A_i^\star$ and $\kappa > 0$ such that

$$\mathbb{P}(y(t) \in A_i) > \frac{1}{2}\mathbb{P}(y(t) \in A_i^\star) \qquad \text{and} \qquad \mathbb{P}(\dot{y}(t)^\top (\eta_i - \eta_j) > \kappa | y(t) \in A_i) = 1, \tag{17}$$

where κ is referred to a suboptimality gap with a positive probability, which is dependent on problems.

The following results provide estimation error bounds of the estimators defined in (12) and a high probability regret upper bound for Algorithm 1. 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 accuracy of estimation 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.

Theorem 1. *Let η_i and $\widehat{\eta}_i(t)$ be the transformed true parameter in* (6) *and its estimate in* (12)*, respectively. Then, with probability at least $1 - \delta$, Algorithm 1 guarantees*

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

for all times t in the range $\tau_i < t \leq T$, where $\tau_i = \mathcal{O}(p_i^{-1} \kappa^{-2} N d_\mu d_y^2 \log^3(T N d_y / \delta))$ is the minimum sample size.

The order of the minimum sample size is primarily due to the gap analysis involving κ and the truncation of observations, which is necessary for the application of Azuma's inequality. The estimation accuracy above is established based on the result that Thompson sampling guarantees linear growth of the number of selections of each arm over time with a high probability. Moving forward, the following theorem demonstrates that the regret upper bound scales at most $\log^4 T$ with respect to the time thanks to the linear growths of square-root estimation accuracy.

Theorem 2. *The regret of Algorithm 1 satisfies the following with probability at least $1 - \delta$:*

$$\text{Regret}(T) = \mathcal{O}\left(\frac{N d_\mu d_y^3}{(p_{\min}^+)^{3/2} \kappa^2} \log^4\left(\frac{T N d_y}{\delta}\right)\right),$$

where $p_{\min}^+ = \min_{i \in [N]: p_i > 0} p_i$.

This theorem demonstrates that the regret scales at most $\log^4 T$ with time, and with the rate $(p^+_{\min})^{-3/2}$ with respect to the optimal probabilities defined in Definition 1. In addition, the term N is caused by the use of the inclusion-exclusion formula to find the bound of the probability that the optimal arm is not chosen. Next, the regret bound increases quadratically as the suboptimality gap κ decreases. In addition, the truncation of observations incurs $\sqrt{d_y \log(TNd_y/\delta)}$, which subsequently leads to additional $d_y^2 \log^2(TNd_y/\delta)$ by increasing the minimum sample size. Lastly, the estimation error contributes to the regret growth with $\sqrt{d_\mu \log(TNd_y/\delta)}$.

The above results are unprecedented to the best of our knowledge. Especially, a high probability poly-logarithmic regret bound of Thompson sampling with respect to the time horizon has not been shown for stochastic contextual bandits with arm-specific parameters, even though the previously available regret bounds are shown for Thompson sampling for adversarial contextual bandits [17] and the greedy first algorithm for the stochastic contextual bandits [30].

5 Numerical Experiments

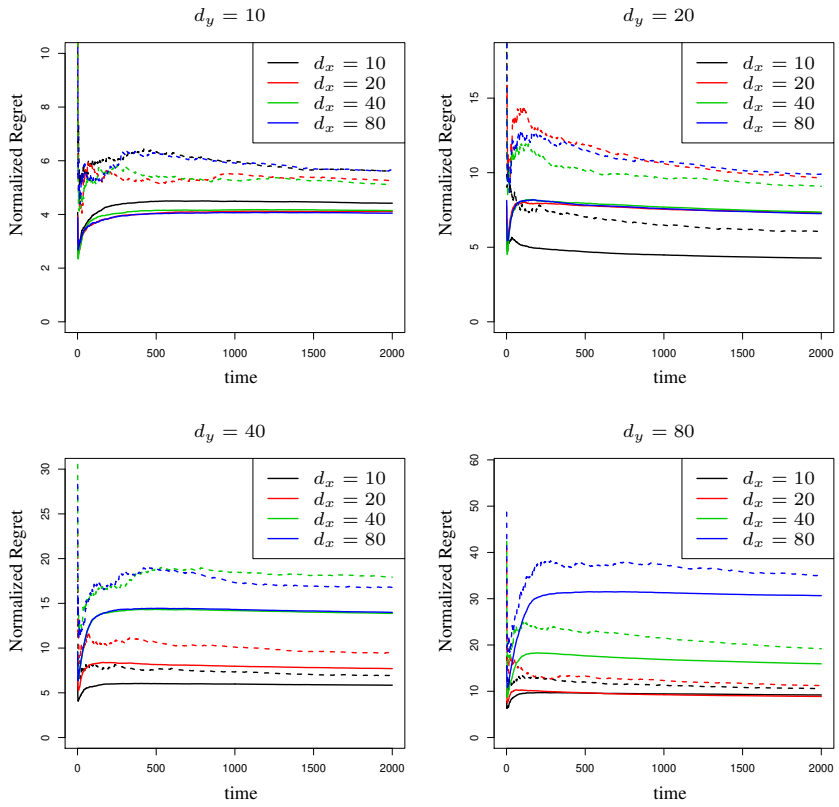

Figure 1: Plots of $\mathrm{Regret}(t)/(\log t)^2$ over time for the different dimensions of context at $N = 5$ and $d_y = 10, 20, 40, 80$. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

5.1 Simulation Experitments

In this sub-section, we numerically show the results in Section 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 1 illustrates regret normalized by $(\log t)^2$, which is the actual regret growth because $(\log t)^2$

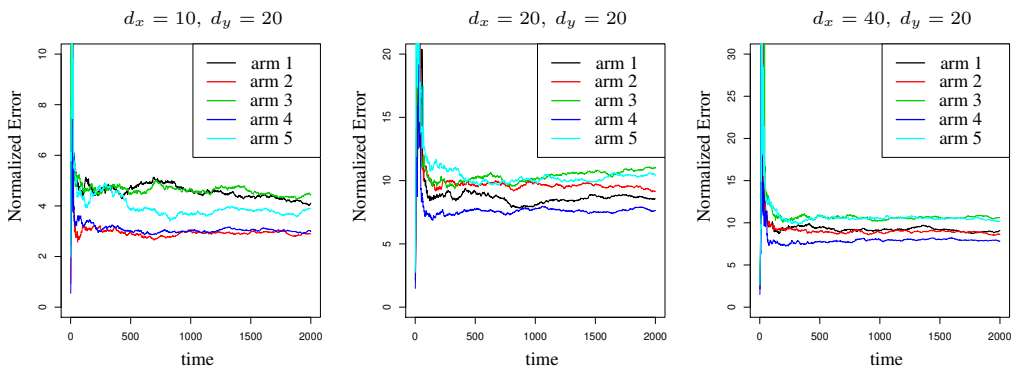

Figure 2: Plots of normalized estimation errors $\sqrt{t}\|\widehat{\eta}_i(t) - \eta_i\|$ of Algorithm 1 over time for partially observable stochastic contextual bandits with five arm-specific parameters and dimensions of observations and contexts $d_y = 20$, $d_x = 10,\ 20,\ 40$.

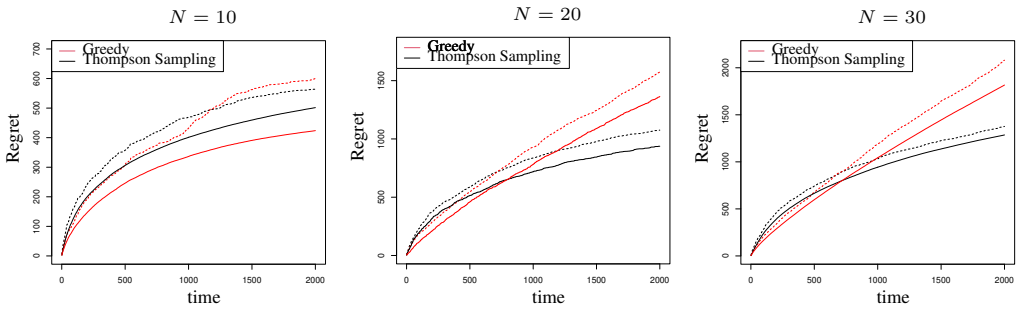

Figure 3: Plots of regrets over time with the different number of arms $N = 10,\ 20,\ 30$ for Thomson sampling versus the Greedy algorithm. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

of $(\log t)^4$ in the regret bound in Theorem 2 is caused by the minimum sample size. Second, Figure 2 showcases the average estimation errors of the estimates in (12) for five different arm-specific parameters defined in (6), changing dimensions of observations and contexts. These errors are normalized by $t^{-0.5}$ based on Theorem 1. Since the error decreases with a rate $t^{-0.5}$, 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 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 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 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.

5.2 Real Data Experiments

In this sub-section, we assess the performance of the proposed algorithm using two healthcare datasets: Eye movement and EGG[1]. These two datasets are presented in previous studies by [18, 34] using contextual bandits with arm-specific parameters and shared context. These datasets involve

[1]The datasets can be found at: `https://www.openml.org/`

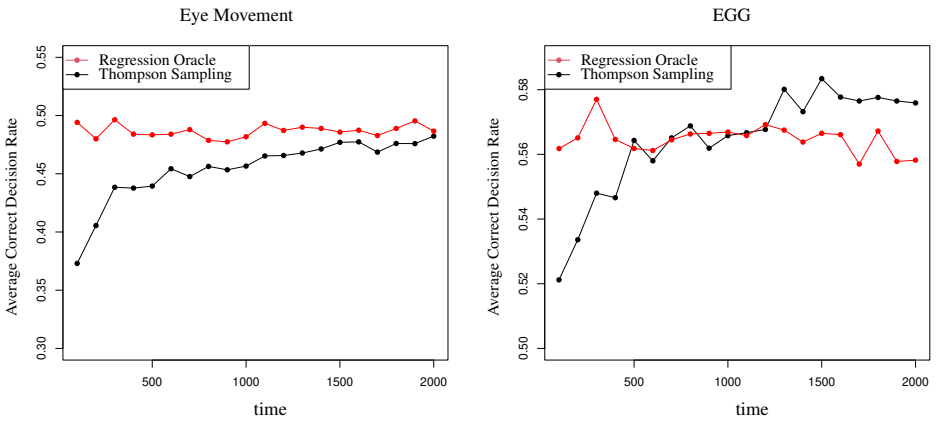

Figure 4: Plots of average correction decision rates of the regression oracle and Thompson sampling for Eye movement (left) and EGG dataset (right).

classification tasks based on patient information. The Eye movement and EGG data sets are comprised of 26 and 14-dimensional contexts with the corresponding patient class categories of 3 and 2, respectively. Each category of patient class is considered an arm in the perspective of the bandit problem, where a decision-maker gets a reward of 1 for successful classification and 0 otherwise. We calculate the average correct decision rate of 100 scenarios defined as $t^{-1}\sum_{\tau=1}^{t}\mathbb{I}(a(\tau)=l(\tau))$, where $l(t)$ is the true label of the patient randomly chosen at time t. We compare the suggested algorithm against the regression oracle with the estimates trained on the entire data in hindsight. We artificially create observations of the patients' contexts based on the structure given in (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 EGG dataset.

Figure 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 dot represents a sample mean of 10,000 results. For the Eye movement data set, the correct decision rate of Thompson sampling converges to that of the regression oracle over time. In addition, for the EGG dataset, Thompson sampling outperforms the regression oracle over time. To the best of our knowledge, this can be caused by complex reasons with non-linearity in the data and potential arm selection bias incurred by actions with higher optimal probabilities.

6 Concluding Remarks and Future Work

We studied Thompson sampling for partially observable stochastic contextual bandits under relaxed assumptions with a particular focus on the arm-specific parameter setup. Indeed, the suggested model is versatile, encompassing a wide range of possible observation structures and offering estimation methods suitable for stochastic contexts. Further, we showed that Thompson sampling guarantees the square-root consistency of parameter estimation for reward parameters. Finally, we proved regret bounds for Thompson sampling with a poly-logarithmic rate for the most common two cases of parameter setups. Our techniques for the analysis hold for other analogous reinforcement learning problems such as a Markov Decision Process thanks to the inclusive assumptions and comprehensive approaches.

A topic of prospective research involves proposing and examining algorithms designed for partially observable contextual bandits, where both the sensing matrix and observation covariance matrix are unknown. Additionally, there is an opportunity to explore the introduction of non-linear structures into both the observation and reward models. Lastly, investigating this framework in the presence of an adversary presents a fascinating challenge for future investigations.

Contents

Appendices

The appendices are organized as follows. First, Appendix A and B explain the notations and the necessary assumptions for the theoretical analysis, respectively. Second, in Appendix C, we discuss

fully observed contextual bandits with a shared parameter in terms of weight matrices. Then, in Appendix D, we present the theoretical results for the general model, with the comprehensive proofs found in Appendix F, G, and H. Following this, Appendix E provides insights into the estimation accuracy and worst-case regret upper bounds for the model with a shared parameter, accompanied by their proofs detailed in Appendix I and J. Lastly, the complete proofs for Theorem 1 and 2 can be found in Appendix K and L, respectively.

A Notations

The following notation will be used. We use $M^\top$ to refer to the transpose of the matrix $M \in \mathbb{C}^{p\times q}$, and $C(M)$ is employed to denote the column space of M. For a vector $v \in \mathbb{C}^d$, we denote the ℓ_2 norm by $\|v\| = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$, and its unit vector by $\dot{v} = v/\|v\|$. Finally, $P_{C(M)}$ is projection on $C(M)$, and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues.

B Technical Assumptions

We describe two assumptions for the theoretical analyses in Section 4. These assumptions, which are commonly adopted in regret analyses, are presented in the antecedent literature [19, 35, 30, 8]. The first assumption is about the boundedness of the parameter space.

Assumption 1 (Parameter Set)**.** *For a parameter and weight matrix J_i, there exists a positive constant c_μ such that $\|J_i\mu_\star\| \le c_\mu$, for all $i = 1, \dots, N$.*

The next assumption is the margin condition of normalized observations, which is slightly modified based on Definition 2 and Assumption 2 in the work of [30].

Assumption 2 (Margin Condition)**.** *Consider the normalized observation $\dot{y}(t) = y(t)/\|y(t)\|$, and the transformed parameters $\{\eta_i\}_{i\in[N]}$ as defined in* (6)*. Then, given the event $\{y(t) \in A_i^\star\}$, we assume that there is $C' > 0$, such that for all $u > 0$;*

$$\forall i \neq j,\ \ \mathbb{P}\left(0 < \dot{y}(t)^\top(\eta_i - \eta_j) \le u \middle| y(t) \in A_i^\star\right) \le C'u.$$

C Weight matrices for the Shared Parameter

We consider the canonical model with a shared parameter and N arm-specific contexts. In this case, contexts must be arm-specific ones because all arms are indistinct if there is a shared context. To this end, J_i has the form as follows:

$$J_i = \left[\mathbf{0}_{d_\mu\times d_\mu}\ \cdots\ \mathbf{0}_{d_\mu\times d_\mu}\ \underbrace{I_{d_\mu}}_{i\text{th}}\ \mathbf{0}_{d_\mu\times d_\mu}\ \cdots\ \mathbf{0}_{d_\mu\times d_\mu}\right]^\top, \tag{18}$$

which makes the context $x(t)$ vanish except for the ith d_μ elements by multiplication. That is, J_i satisfies the following equations

$$f(x(t), i) = x(t)^\top J_i\mu_\star = x_i(t)^\top\mu_\star.$$

In the model with a shared parameter described above, a decision maker can learn the parameter, regardless of a chosen arm. Thus, explicit exploration schemes are not needed for this framework. For this reason, analyses of this model are easier than those of arm-specific parameters.

D Results for the general model

We show the results for the general model with any cases of weight matrices. Lemma 1 presents that reward 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 (16). This result came from Theorem 1 of [21] with some modifications.

Lemma 1. *Let $w_t = r_{a(t)}(t) - \widehat{x}(t)^\top J_{a(t)}\mu$ and $\mathscr{F}_{t-1} = \sigma\{\{y(\tau)\}_{\tau=1}^t, \{a(\tau)\}_{\tau=1}^t\}$. 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 $\delta > 0$, assuming that $\|\mu_\star\| \leq h$ and $B(1) = \lambda I$, $\lambda > 0$, with probability at least $1-\delta$, we have

$$\|\widehat{\mu}(t) - \mu_\star\|_{B(t)} = \left\|\sum_{\tau=1}^{t-1} J_{a(\tau)}^\top D y(\tau) w_\tau\right\|_{B(t)} \leq R\sqrt{d_\mu \log\left(\frac{1+L^2 t/\lambda}{\delta}\right)} + v^{-1}h,$$

where $L = \sqrt{d_y} v_T(\delta)$, $v_T(\delta) = (2\lambda_M \log(2d_y T/\delta))^{1/2}$ and $\lambda_M = \lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$.

The next lemma guarantees the linear growth of eigenvalues of covariance matrices $\{B_i^+(t)\}_{i\in[N]}$ defined in (13). This is a cornerstone for the results presented in the remaining part of this section.

Lemma 2. *Let $n_i(t)$ be the count of ith arm chosen up to the time t. For $B_i^+(t)$ in* (13)*, on the event W_T defined in* (21)*, with probability at least $1-\delta$, if $N_i^{(1)}(\delta, T) \leq n_i(t) \leq T$ for given $T > 0$, we have*

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

where $N_i^{(1)}(\delta, T) = 8d_y \nu_{iM}^2 v_T(\delta)^4 \log(T/\delta)/(\lambda_m^2 \nu_{im+}^2)$; ν_{im+} and ν_{iM} be the non-zero minimum and maximum eigenvalues of $J_i^\top D D^\top J_i$, respectively.

The next lemma provides a piece of theoretical evidence that the frequency of the i arm of being chosen scales linearly with the time horizon when the arm has a positive probability of being the optimal arm. As a consequence, the estimation errors of arm-specific transformed parameters decrease with the rate $t^{-0.5}$ for all arms with non-zero $\mathbb{P}(a^\star(t) = i)$.

Lemma 3. *Let the minimum sample size be*

$$N_i^{(2)}(\delta, T, \kappa) = \max\left(N_i^{(1)}(\delta, T),\ 16\lambda_m^{-1}\nu_{iM}\nu_{im+}^{-1}\left(R\sqrt{d_\mu \log\left(1 + L^2 T/\delta\right)} + vh\right)^2 \kappa^{-2}\right).$$

If $n_i(t) > N_i^{(2)}(\delta, T, \kappa)$ and $n_j(t) > N_j^{(2)}(\delta, T, \kappa)$ for $j \neq i$,

$$\begin{aligned} &\mathbb{P}(a(t) = i|\mathscr{F}_{t-1}) \\ \geq\ & \frac{\mathbb{P}(a^\star(t) = i)}{2}\left(1 - \sum_{j\neq i}\left(\exp\left(-\frac{n_i(t)\lambda_m \nu_{im+}\kappa^2}{32\nu_{iM}v^2}\right) + \exp\left(-\frac{n_j(t)\lambda_m\nu_{jm+}\kappa^2}{32\nu_{jM}v^2}\right)\right)\right), \end{aligned}$$

where κ is the positive constant defined in (17) *and $\mathscr{F}_{t-1}$ is the filtration defined in Lemma 1.*

The results above can be applied to all partially observable contextual bandits in any case of weight matrices.

E Results for the model with a shared parameter

For the model with a shared parameter, the weight matrices $\{J_i\}_{i=1}^N$ satisfy the condition introduced in (18). For the model with a single parameter, $n_i(t) = t$ for all $i \in [N]$. This means that a decision-maker can learn the shared parameter regardless of the chosen arm. The proof of the following theorems are in Appendix I and J.

Theorem 3. *For partially observable contextual bandits with a shared parameter, let η_i and $\widehat{\eta}_i(t)$ be the transformed true parameter in* (6) *and the estimate in* (12)*, respectively. Then, given $T > 0$, with probability at least $1-\delta$, for all $N_i^{(1)}(\delta, T) < t \leq T$, Algorithm 1 is guaranteed to have*

$$\|\widehat{\eta}_i(t) - \eta_i\| \leq R\sqrt{\frac{2\nu_{iM}}{\lambda_m \nu_{im+}}}\left(\sqrt{d_\mu \log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + v^{-1}h\right) t^{-1/2},$$

where $N_i^{(1)}(\delta, T) = 32\lambda_M d_y^2 \nu_{iM}^2 \log^2(Td_y/\delta)\log(T/\delta)/(\lambda_m^2 \nu_{im+}^2)$ *is the minimum samples for the ith arm selections;* $\lambda_m = \lambda_{\min}(\Sigma_y)$ *and* $\lambda_M = \lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$*;* ν_{im+} *and* ν_{iM} *be the non-zero minimum and maximum eigenvalues of* $J_i^\top DD^\top J_i$*, respectively.*

The next theorem provides a high probability regret upper bound of Thompson sampling for partially observable contextual bandits with a shared parameter.

Theorem 4. *Assume that Algorithm 1 is used in partially observable contextual bandits with a shared parameter. Then, with probability at least* $1-\delta$*,* $\mathrm{Regret}(T)$ *is of the order*

$$\mathrm{Regret}(T) = \mathcal{O}\left(Nd_\mu d_y^3 \log^4\left(\frac{TNd_y}{\delta}\right)\right).$$

The regret bound scales at most $\log^4 T$ with respect to the time horizon and linearly with N. Similarly to Theorem 2, $\sqrt{d_y \log(TNd_y/\delta)}$, $d_\mu \log(TNd_y/\delta)$ and $d_y^2 \log^2(TNd_y/\delta)$ are incurred by the truncation of observations, estimation error, and the minimum sample size, respectively. Note that a high probability upper regret bound under the normality assumption has been found for the model with a shared parameter by [16]. As compared to the setting in the work of [16], 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.

F Proof of Lemma 1

Lemma 1 provides a sub-Gaussian tail property of the reward estimation error w_t given μ and shows a self-normalized bound for vector-valued martingale by using the sub-Gaussian property. The reward estimation error w_t can be decomposed into two parts. The one is the reward error $\varepsilon_i(t)$ given (1) due to the randomness of rewards. This error is created even if the context $x(t)$ is known. The other is the context estimation error $(x(t)-\widehat{x}(t))^\top J_i\mu$ caused by unknown contexts. To show the sub-Gaussian property of reward estimation error, the next lemma provides a sub-Gaussian property of context estimation errors.

Lemma 4. *The estimate* $\widehat{x}(t)^\top J_i\mu_\star$ *in* (4) *has the mean* $x(t)^\top J_i\mu_\star$ *and a sub-Gaussian tail property such as*

$$\mathbb{E}\left[e^{\nu(\widehat{x}(t)-x(t))^\top J_i\mu}\middle| y(t)\right] \le e^{\frac{\nu^2 R_2^2}{2}},$$

for any $\nu > 0$ *and some* $R_2 > 0$*.*

Proof. Since $\widehat{x}(t)$ is the BLUP of $x(t)$, we have $\mathbb{E}[(\widehat{x}(t)-x(t))^\top J_i\mu] = 0$ and

$$\mathrm{Var}((\widehat{x}(t)-x(t))^\top J_i\mu | y(t)) = (J_i\mu)^\top (A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1} J_i\mu$$

based on the results of the work of [33]. Because $\|J_i\mu\| \le 1$, we can find $R_2 > 0$ such that

$$(J_i\mu)^\top (A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1} J_i\mu \le \lambda_{\max}((A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1}) = R_2^2, \tag{19}$$

for any $i = 1, \ldots, N$. Therefore, since $\zeta_i(t)$ has a sub-Gaussian density, we get

$$\mathbb{E}\left[e^{\nu(\widehat{x}(t)-x(t))^\top J_i\mu_\star}\middle| y(t)\right] \le e^{\frac{\nu^2 R_2^2}{2}}.$$

□

Lemma 5. *For any* $\nu > 0$*, we have*

$$\mathbb{E}\left[e^{\nu(r_i(t)-\widehat{x}(t)^\top J_i\mu_\star)}\middle| y(t)\right] \le e^{\frac{\nu^2 R^2}{2}}.$$

where $R^2 = R_1^2 + R_2^2$*, for* R_1 *and* R_2 *in* (2) *and* (19)*, respectively.*

Proof. By (7),

$$r_i(t) - \widehat{x}(t)^\top J_i\mu_\star = (x(t)^\top J_i\mu_\star - y(t)^\top \eta_i) + \varepsilon_i(t),$$

which implies $\mathbb{E}[r_i(t) - \widehat{x}(t)^\top J_i\mu_\star|y(t), a(t)] = 0$, since $y(t)^\top\eta_i$ is the BLUP of $x(t)^\top J_i\mu_\star$. Due to $\mathrm{Var}(x(t)^\top J_i\mu_\star - y(t)^\top\eta_i|y(t)) \le R_2^2$ by (19), we can find $R > 0$ such that

$\mathrm{Var}(r_i(t) - \widehat{x}(t)^\top J_i\mu_\star|y(t)) = \mathrm{Var}(\varepsilon_i(t)) + \mathrm{Var}(x(t)^\top J_i\mu_\star - y(t)^\top\eta_i|y(t)) \le R_1^2 + R_2^2 = R^2$

Since $\varepsilon_i(t)$ and $x(t)^\top J_i\mu_\star - y(t)^\top\eta_i$ have a sub-Gaussian distribution, $r_i(t) - \widehat{x}(t)^\top J_i\mu$ has a sub-Gaussian distribution as well. Thus,

$$\mathbb{E}[e^{\nu(r_i(t) - \widehat{x}(t)^\top J_i\mu)}|y(t)] = \mathbb{E}[e^{\nu\zeta_i(t)}|y(t)] \le e^{\frac{\nu^2R^2}{2}}.$$

□

Lemma 6. *For J_i such that $\mathbb{E}[r_i(t)|x(t)] = x(t)^\top J_i\mu_\star$, let*

$$D_t^\mu = \exp\left(\frac{(r_{a(t)}(t) - \widehat{x}(t)^\top J_{a(t)}\mu_\star)\widehat{x}(t)^\top J_{a(t)}\mu_\star}{R} - \frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right),$$

and $M_t^\mu = \prod_{\tau=1}^t D_\tau^\mu$. Then, $\mathbb{E}[M_\tau^\mu] \le 1$.

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

$$\begin{aligned}\mathbb{E}[D_t^\mu|\mathscr{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\frac{(r_{a(t)}(t) - \widehat{x}(t)^\top J_{a(t)}\mu_\star)\widehat{x}(t)^\top J_{a(t)}\mu_\star}{R} - \frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right)\middle| y(t), a(t)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\zeta_{a(t)}(t)\widehat{x}(t)^\top J_{a(t)}\mu_\star}{R}\right)\middle| y(t), a(t)\right]\exp\left(-\frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right).\end{aligned}$$

Then, by Lemma 5, we have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\frac{\zeta_{a(t)}(t)\widehat{x}(t)^\top J_{a(t)}\mu_\star}{R}\right)\middle| y(t), a(t)\right]\exp\left(-\frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right) \\ &\quad\le \exp\left(\frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right)\exp\left(-\frac{1}{2}(\widehat{x}(t)^\top J_{a(t)}\mu_\star)^2\right) = 1.\end{aligned}$$

Therefore,

$$\mathbb{E}[M_t^\mu|\mathscr{F}_{t-1}] = \mathbb{E}[M_1^\mu\cdots D_{t-1}^\mu D_t^\mu|\mathscr{F}_{t-1}] = D_1^\mu\cdots D_{t-1}^\mu\mathbb{E}[D_t^\mu|\mathscr{F}_{t-1}] \le M_{t-1}^\mu.$$

□

Now, we continue the proof of Lemma 1. Let ϕ_μ be the probability density function of multivariate Gaussian distribution of $\mu_\star$ with the mean $\mathbf{0}_{d_\mu}$ and the positive covariance matrix $\lambda^{-1}I$, where $\lambda = v^{-2}$. By Lemma 9 of the work of [21], for $M_t = \mathbb{E}[M_t^\mu|\mathscr{F}_\infty]$, we have

$$\mathbb{P}_{\phi_\mu}\left(\|S_t\|^2_{B(t)^{-1}} > 2R^2\log\left(\frac{det(B(t))^{1/2}}{\delta det(\lambda I)^{1/2}}\right)\right) \le \mathbb{E}[M_t] \le \delta, \tag{20}$$

where $\mathbb{P}_{\phi_\mu}$ is the probability measure based on ϕ_μ and, $S_t = \sum_{\tau=1}^t J_{a(\tau)}^\top Dy(\tau)w_\tau$. Lemma 5 , Lemma 6 and (20) are sufficient conditions for the following inequality

$$\mathbb{P}_{\phi_\mu}\left(\|S_t\|_{B(t)^{-1}} > 2R^2\log\left(\frac{\mathrm{det}(B(t))^{1/2}}{\delta\mathrm{det}(\lambda I)^{1/2}}\right),\ \forall\, t > 0\right) \le \delta,$$

by Theorem 1 of the work of [21]. By Lemma 10 of the work of [21], we have

$$\mathrm{det}(B(t)) \le (\lambda + tL^2/d_x)^{d_x}.$$

Therefore, with probability of at least $1-\delta$, we have

$$\|\widehat{\mu}(t) - \mu_\star\|_{B(t)} = \|S_t\|_{B(t)^{-1}} \le R\sqrt{d_x\log\left(\frac{1 + L^2t/\lambda}{\delta}\right)} + vh,$$

which is a similar result to Theorem 2 of the work of [21].

G Proof of Lemma 2

First, to find the bound for $\|y(t)\|$, for $\delta > 0$, we define W_T such that

$$W_T = \left\{ \max_{\{1 \le \tau \le T\}} ||y(\tau)||_\infty \le v_T(\delta) \right\}, \tag{21}$$

where $v_T(\delta) = (2\lambda_M \log(2d_y T/\delta))^{1/2}$ and $\lambda_M = \lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$.

Lemma 7. *For the event W_T defined in* (21)*, we have $\mathbb{P}(W_T) \ge 1 - \delta$.*

Proof. Note that $y(t)$ has the mean $\mathbf{0}_{d_y}$ and the covariance $A\Sigma_x A^\top + \Sigma_y$ without knowing $x(t)$. Using the sub-Gaussian tail property, we have

$$\mathbb{P}\left(\|(A\Sigma_x A^\top + \Sigma_y)^{-1/2} \Sigma_y^{-1/2} y(t)\|_\infty \ge \varepsilon \right) \le 2d_y \cdot e^{-\frac{\varepsilon^2}{2}}.$$

Accordingly, we have

$$\mathbb{P}\left(\|y(t)\|_\infty \ge \lambda_M^{\frac{1}{2}} \varepsilon \right) \le 2d_y \cdot e^{-\frac{\varepsilon^2}{2}}.$$

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

$$\mathbb{P}\left(\max_{1 \le t \le T} \|y(t)\|_\infty \ge \lambda_M^{\frac{1}{2}} \varepsilon \right) \le 2Td_y \cdot e^{-\frac{\varepsilon^2}{2}}$$

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

$$\mathbb{P}\left(\max_{1 \le t \le T} \|y(t)\|_\infty \ge (2\lambda_M \log(2Td_y/\delta))^{1/2} \right) \le 2Td_y \cdot \exp\left(-\frac{2\log(2Td_y/\delta)}{2} \right) = \delta.$$

Thus,

$$\mathbb{P}(W_T) \ge 1 - \mathbb{P}\left(\max_{1 \le t \le T} \|y(t)\| \ge v_T(\delta) \right) \ge 1 - \delta.$$

□

Then, by Lemma 7, we have

$$\|y(t)\| \le \sqrt{d_y} v_T(\delta) := L = \mathcal{O}\left(\sqrt{d_y \log(Td_y/\delta)} \right), \tag{22}$$

for all $1 \le t \le T$ with probability at least $1 - \delta$.

Lemma 8. *(Azuma Inequality) Consider the sequence $\{X_t\}_{1 \le t \le T}$ random variables adapted to some filtration $\{\mathcal{G}_t\}_{1 \le t \le T}$, such that $\mathbb{E}[X_t | \mathcal{G}_{t-1}] = 0$. Assume that there is a deterministic sequence $\{c_t\}_{1 \le t \le T}$ that satisfies $X_t^2 \le c_t^2$, almost surely. Let $\sigma^2 = \sum_{1 \le t \le T} c_t^2$. Then, for all $\varepsilon \ge 0$, it holds that*

$$\mathbb{P}\left(\sum_{t=1}^{T} M_t \ge \varepsilon \right) \le e^{-\varepsilon^2 / 2\sigma^2}.$$

The proof of Lemma 8 is provided in the work of [36]. Now, we construct a martingale and its different sequence to find an upper bound of a sum of random variables with Lemma 8. Let the sigma field generated by the contexts and chosen arms by time t

$$\mathcal{G}_{t-1} = \sigma\{x(1), a(1), x(2), a(2), \ldots, x(t), a(t)\}.$$

Consider $V_t = D^\top J_{a(t)} y(t) y(t)^\top J_{a(t)}^\top D$ in order to study the behavior of $B(t)$. Note that

$$\begin{aligned}\mathbb{E}[V_t|\mathscr{G}_{t-1}] &= J_{a(t)}^\top D\mathrm{Var}(y(t)|\mathscr{G}_{t-1})D^\top J_{a(t)} + J_{a(t)}^\top DAx(t)x(t)^\top A^\top D^\top J_{a(t)} \\ &\succeq \lambda_m J_{a(t)}^\top DD^\top J_{a(t)}, \end{aligned} \tag{23}$$

where $\lambda_m = \lambda_{\min}(\Sigma_y)$. For all $t > 0$ and $z \in C(J_i^\top D)$ such that $\|z\| = 1$, it holds that

$$z^\top \left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}]\right) z \geq z^\top \left(\sum_{\tau=1:a(\tau)=i}^{t-1} \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}]\right) z \geq \lambda_m \nu_{im+} n_i(t). \tag{24}$$

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

$$X_t^i = (V_t - \mathbb{E}[V_t|\mathscr{G}_{\tau-1}])I(a(\tau) = i), \tag{25}$$

$$Y_t^i = \sum_{\tau=1}^{t} (V_\tau - \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}])\, I(a(\tau) = i). \tag{26}$$

Then, $X_\tau^i = Y_\tau^i - Y_{\tau-1}^i$ and $\mathbb{E}\left[X_\tau^i|\mathscr{G}_{\tau-1}\right] = 0$. Thus, $z^\top X_\tau^i z$ is also a martingale difference sequence. Here, we are interested in the minimum eigenvalue of $\sum_{\tau=1}^{t-1} V_\tau I(a(\tau) = i)$, whose corresponding eigenvector is not orthogonal to $C(J_i^\top D)$. Because $(z^\top X_\tau^i z)^2 \leq d_y^2 \nu_{iM}^2 v_T(\delta)^4$ on the event W_T defined in (21) and thereby $\sum_{\tau=1}^{t-1} \left(z^\top X_\tau^i z\right)^2 \leq n_i(t) d_y^2 \nu_{iM}^2 v_T(\delta)^4$, using Lemma 8, we get the following inequality

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

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

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

for $\varepsilon \leq 0$. Now, using (23), we have the following inequality

$$\begin{aligned}&\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} (V(\tau) - \mathbb{E}[V_t|\mathscr{G}_{\tau-1}])\, I(a(\tau) = i)\right) z \leq n_i(t)\varepsilon\right) \\ \geq\ &\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} \left(V(\tau) - \lambda_m J_{a(\tau)}^\top DD^\top J_{a(\tau)}\right) I(a(\tau) = i)\right) z \leq n_i(t)\varepsilon\right). \end{aligned} \tag{27}$$

Putting together (24), (25), (26) and (27), we obtain

$$\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} V(\tau)I(a(\tau) = i)\right) z \leq n_i(t)(\lambda_m\nu_{im+} + \varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2d_y^2\nu_{iM}^2 v_T^4(\delta)}\right), \tag{28}$$

where $-\lambda_m\nu_{im+} \leq \varepsilon \leq 0$ is arbitrary. Indeed, using $B(t) \succeq \sum_{\tau=1}^{t-1} V(\tau)I(a(\tau) = i)$, for $-\lambda_m\nu_{im+} \leq \varepsilon \leq 0$, we have

$$\mathbb{P}\left(z^\top B(t)z \leq n_i(t)(\lambda_m\nu_{im+} + \varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2d_y^2\nu_{iM}^2 v_T^4(\delta)}\right). \tag{29}$$

In other words, by putting $\exp\left(-n_i(t)\varepsilon^2/(2d_y^2\nu_{iM}^2 v_T(\delta)^4)\right) = \delta/T$, (29) can be written as

$$z^\top B(t)z \geq n_i(t)\left(\lambda_m\nu_{im+} - \sqrt{\frac{2d_y^2\nu_{iM}^2 v_T(\delta)^4}{n_i(t)}\log\frac{T}{\delta}}\right),$$

for all $1 \leq t \leq T$ with the probability at least $1 - \delta$. If $n_i(t) \geq N_i^{(1)}(\delta, T) := 8d_y^2\nu_{iM}^2 v_T(\delta)^4 \log(T/\delta)/(\lambda_m^2\nu_{im+}^2) = \mathcal{O}(d_y^2 \log^3(d_y T/\delta))$, we have

$$\lambda_{\max}\left(D^\top J_i B(t)^{-1} J_i^\top D\right) \leq \frac{2\nu_{iM}}{\lambda_m\nu_{im+}} n_i(t)^{-1}.$$

H Proof of Lemma 3

For simplicity, let the event of the ith arm of being optimal at time t $A_{it} = \{y(t) \in A_i\}$. Then, we aim to have a lower bound of the probability $\mathbb{P}(a(t) = i|\mathscr{F}_{t-1})$ to find a lower bound of $n_i(t)$ with

$$\begin{aligned}\mathbb{P}(a(t) = i|\mathscr{F}_{t-1}) &\geq \mathbb{P}(a(t) = i|A_{it}, \mathscr{F}_{t-1})\mathbb{P}(A_{it}) \\ &\geq \left(1 - \sum_{j\neq i} \mathbb{P}(y(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)|A_{it}, \mathscr{F}_{t-1})\right)\mathbb{P}(A_{it}).\end{aligned}$$

Using the relationship below,

$$\begin{aligned}&\{y(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)\} \\ &\subset \left\{y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > \frac{1}{2}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) + y(t)^\top(\eta_i - \eta_j))\right\} \\ &\quad \bigcup \left\{y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) < -\frac{1}{2}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) + y(t)^\top(\eta_i - \eta_j))\right\}, \quad (30)\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{P}\left(y(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)|A_{it}, \mathscr{F}_{t-1}\right) \\ \leq\ &\mathbb{P}\left(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t) > \frac{1}{2}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) + y(t)^\top(\eta_i - \eta_j))\middle| A_{it}, \mathscr{F}_{t-1}\right) \\ +\ &\mathbb{P}\left(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > \frac{1}{2}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) + y(t)^\top(\eta_i - \eta_j))\middle| A_{it}, \mathscr{F}_{t-1}\right).\end{aligned}$$

Since $y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) \leq \|y(t)\|\left(\lambda_{\max}(B_i^+(t))^{1/2} + \lambda_{\max}(B_j^+(t))^{1/2}\right)\|\widehat{\mu}(t) - \mu_\star\|_{B(t)}$, by Lemma 1 and 2, if $n_i(t) \geq N_i^{(1)}(\delta, T)$ and $n_j(t) \geq N_j^{(1)}(\delta, T)$, we have

$$\begin{aligned}&|\dot{y}(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j)| \\ \leq\ &\left(R\sqrt{d_\mu \log\left(1 + \frac{L^2T}{\delta}\right)} + v^{-1}h\right)\left(\sqrt{\frac{2\nu_{iM}}{\lambda_m \nu_{im+}} n_i(t)^{-1}} + \sqrt{\frac{2\nu_{jM}}{\lambda_m \nu_{jm+}} n_j(t)^{-1}}\right).\end{aligned}$$

To lower the value on the RHS of (31) less than $\kappa/2$, we need the minimum samples $n_i(t) > 16\lambda_m^{-1}\nu_{iM}\nu_{im+}^{-1}\left(R\sqrt{d_\mu \log(1 + L^2T/\delta)} + v^{-1}h\right)^2 \kappa^{-2}$ and $n_j(t) > 16\lambda_m^{-1}\nu_{jM}\nu_{jm+}^{-1}\left(R\sqrt{d_\mu \log(1 + L^2T/\delta)} + v^{-1}h\right)^2 \kappa^{-2}$, for the arm i and j, respectively. Then, we have

$$|\dot{y}(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j)| \leq \frac{\kappa}{2},$$

and thereby

$$\frac{1}{2}(\dot{y}(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) + y(t)^\top(\eta_i - \eta_j)) \geq \frac{\kappa}{4},$$

because $\dot{y}(t)^\top(\eta_i - \eta_j) \geq \kappa$ given A_{it} by (17). Accordingly, we have

$$\begin{aligned}&\mathbb{P}(y(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)|A_{it}, \mathscr{F}_{t-1}) \\ &\leq \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > \kappa/4|A_{it}, \mathscr{F}_{t-1}) + \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > \kappa/4|A_{it}, \mathscr{F}_{t-1}).\end{aligned}$$

Based on (14), by Lemma 2, we have

$$\begin{aligned}\mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t))>c|A_{it},\mathscr{F}_{t-1}) &\leq \mathbb{E}\left[\exp\left(-\frac{c^2}{2v^2\dot{y}(t)^\top B_i^+(t)\dot{y}(t)}\right)\right]\\ &\leq \exp\left(-\frac{n_i(t)\lambda_m\nu_{im+}c^2}{2\nu_{iM}v^2}\right)\end{aligned}$$

for any $c\geq 0$. Thus, if $n_i(t)>N_i^{(2)}(\delta,T,\kappa)$ and $n_j(t)>N_j^{(2)}(\delta,T,\kappa)$ for $j\neq i$, we have

$$\mathbb{P}(\dot{y}(t)^\top\widetilde{\eta}_i(t)<\dot{y}(t)^\top\widetilde{\eta}_j(t)|A_{it},\mathscr{F}_{t-1})\leq\exp\left(-\frac{n_i(t)\lambda_m\nu_{im+}\kappa^2}{32\nu_{iM}v^2}\right)+\exp\left(-\frac{n_j(t)\lambda_m\nu_{jm+}\kappa^2}{32\nu_{jM}v^2}\right),$$

and thereby

$$\mathbb{P}(a(t)=i|A_{it},\mathscr{F}_{t-1})\geq 1-\sum_{j\neq i}\left(\exp\left(-\frac{n_i(t)\lambda_m\nu_{im+}\kappa^2}{32\nu_{iM}v^2}\right)+\exp\left(-\frac{n_j(t)\lambda_m\nu_{jm+}\kappa^2}{32\nu_{jM}v^2}\right)\right).$$

Therefore, if $n_i(t)>N_i^{(2)}(\delta,T,\kappa)$ and $n_j(t)>N_j^{(2)}(\delta,T,\kappa)$,

$$\begin{aligned}&\mathbb{P}(a(t)=i|\mathscr{F}_{t-1})\geq\mathbb{P}(a(t)=i|A_{it},\mathscr{F}_{t-1})\mathbb{P}(A_{it})\\ \geq\ &\frac{\mathbb{P}(a^\star(t)=i)}{2}\left(1-\sum_{j\neq i}\left(\exp\left(-\frac{n_i(t)\lambda_m\nu_{im+}\kappa^2}{32\nu_{iM}v^2}\right)+\exp\left(-\frac{n_j(t)\lambda_m\nu_{jm+}\kappa^2}{32\nu_{jM}v^2}\right)\right)\right).\end{aligned}$$

I Proof of Theorem 3

By Lemma 1, for all $1\leq t\leq T$, with probability of at least $1-\delta$, we have

$$\|B(t)^{\frac{1}{2}}(\widehat{\mu}(t)-\mu_\star)\|\leq R\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h.$$

Suppose that $D^\top J_i$ has the singular value decomposition $U_i\Sigma_iV_i^\top$. Using $(V_i\Sigma_i^-U_i^\top)D^\top J_i\preceq I$, where Σ_i^- is the pseudo-inverse matrix of Σ_i, we get

$$\|B(t)^{\frac{1}{2}}(V_i\Sigma_i^-U_i^\top)D^\top J_i(\widehat{\mu}(t)-\mu_\star)\|\leq\|B(t)^{\frac{1}{2}}(\widehat{\mu}(t)-\mu_\star)\|. \tag{31}$$

Accordingly, we have

$$\gamma_{\min}((V_i\Sigma_i^-U_i^\top)^\top B(t)(V_i\Sigma_i^-U_i^\top))^{\frac{1}{2}}\|D^\top J_i(\widehat{\mu}(t)-\mu_\star)\|\leq\|B(t)^{\frac{1}{2}}(V_i\Sigma_i^-U_i^\top)D^\top J_i(\widehat{\mu}(t)-\mu_\star)\|, \tag{32}$$

where $\gamma_{\min}(M)$ is the smallest non-zero eigenvalue of M for a square matrix M. Finally, by putting together (31), (32) and Lemma 2, if $n_i(t)>N_i^{(1)}(\delta,T)$, we have

$$\begin{aligned}\|\widehat{\eta}_i(t)-\eta_i\| &\leq \lambda_{\max}(D^\top J_iB(t)^{-1}J_i^\top D)^{\frac{1}{2}}R\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)\\ &\leq R\sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)n_i(t)^{-\frac{1}{2}}.\end{aligned} \tag{33}$$

Because $n_i(t)=t$ for all i, for all $N_i^{(1)}(\delta,T)<t\leq T$, we have

$$\|\widehat{\eta}_i(t)-\eta_i\|\leq R\sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)t^{-1/2}.$$

Corollary 1. *Let $\widetilde{\eta}_i(t)$ be a sample in* (14)*. Then, if $N_i^{(1)}(\delta,T)<t\leq T$, with probability at least $1-\delta$, for all $i\in[N]$, we have*

$$\|\widetilde{\eta}_i(t)-\eta_i\|\leq\sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\left(v\sqrt{2d_y\log\frac{2TN}{\delta}}+R\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)\right)t^{-1/2}.$$

Proof. Using $\mathbb{P}\left(\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| > \epsilon\right) \leq \mathbb{P}\left(\sqrt{d_y}Z > \epsilon\right)$, where $Z \sim \mathcal{N}\left(0, v^2\lambda_{\max}(B_i^+(t))\right)$, we have

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

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

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

If $t > N_i^{(1)}(\delta, T)$, by Lemma 2, we have

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

Therefore, by Theorem 4, for $t > N_i^{(1)}(\delta, T)$, we have

$$\|\widetilde{\eta}_i(t) - \eta_i\| \leq \sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\left(v\sqrt{2d_y\log\frac{2TN}{\delta}} + R\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + v^{-1}h\right)\right)t^{-1/2}.$$

□

J Proof of Theorem 4

We decompose the regret as follows:

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{T} y(t)^\top(\eta_{a^\star(t)}(t) - \eta_{a(t)}(t)) \\
&\leq \sum_{t=1}^{T} y(t)^\top(\eta_{a^\star(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t) + \widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}(t))I(a^\star(t) \neq a(t)) \\
&\leq \sqrt{d_y}v_T(\delta)\sum_{t=1}^{T}(\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\|)I(a^\star(t) \neq a(t)),
\end{aligned}$$

since $\|y(t)\| \leq \sqrt{d_y}v_T(\delta)$ for all $t \in [T]$ by (22). By Corollary 1, if $t > N_i^{(1)}(\delta, T)$, with probability at least $1 - \delta$, we have

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

where

$$\begin{aligned}
g(\delta) &= 2\max_{i\in[N]}\left(\sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\right)\left(v\sqrt{2d_y\log\frac{2TN}{\delta}} + R\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + v^{-1}h\right) \\
&= \mathcal{O}\left((d_y^{1/2} + d_\mu^{1/2})\sqrt{\log(TN/\delta)}\right).
\end{aligned}$$

Now, we construct a martingale sequence with respect to the filtration $\{\mathscr{F}_{t-1}\}_{t=1}^T$ defined in Lemma 3. To that end, let $G_1 = H_1 = 0$,

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

and $H_t = \sum_{\tau=1}^{t} G_\tau$. 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 $c_\tau = 2\tau^{-1/2}$. Since $\sum_{\tau=1}^{T}|G_\tau| \leq \sum_{\tau=2}^{T} c_\tau^2 \leq 4\log T$, by Lemma 8, we have

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

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

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}} I(a^\star(t) \neq a(t)) \leq \sqrt{8\log T \log \delta^{-1}} + \sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbb{P}(a^\star(\tau) \neq a(\tau)|\mathscr{F}_{\tau-1}). \tag{34}$$

Now, we proceed to the upper bound of the second term on the right side in (34).

Let $A_{it}^\star = \{y(t) \in A_i^\star\}$, where $A_i^\star$ is defined in Definition 1. By applying the same logic in (30), we get

$$\begin{aligned}
& \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\
+\ & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star).
\end{aligned}$$

By Theorem 3, with probability of at least $1-\delta$, we have

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

for all $t \in [T]$ and $i \in [N]$, where

$$h_i(\delta, T) = R\sqrt{\frac{2\nu_{iM}}{\lambda_m \nu_{im+}}}\left(\sqrt{d_\mu \log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + v^{-1}h\right) = \mathcal{O}\left(\sqrt{d_\mu \log(T/\delta)}\right).$$

Accordingly, we have

$$\begin{aligned}
& \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -h_i(t)t^{-1/2} + 0.5\dot{y}(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\
+\ & \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -h_j(t)t^{-1/2} + 0.5\dot{y}(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star).
\end{aligned} \tag{35}$$

Let $E_{ij1t} = \{h_i(\delta, T)t^{-1/2} < 0.25\dot{y}(t)^\top(\eta_i - \eta_j)\}$ and $E_{ij2t} = \{h_j(\delta, T)t^{-1/2} < 0.25\dot{y}(t)^\top(\eta_i - \eta_j)\}$. Then, we can decompose the first term on the RHS in (35) as follows:

$$\begin{aligned}
& \mathbb{P}\left(\dot{y}(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h_i(t)}{t^{1/2}} + 0.5\dot{y}(t)^\top(\eta_i - \eta_j)\middle|\mathscr{F}_{t-1}, A_{it}^\star\right) \\
&= \mathbb{P}\left(\dot{y}(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h_i(t)}{t^{1/2}} + 0.5\dot{y}(t)^\top(\eta_i - \eta_j)\middle|E_{ij1t}, \mathscr{F}_{t-1}, A_{it}^\star\right)\mathbb{P}(E_{ij1t}|\mathscr{F}_{t-1}, A_{it}^\star) \\
&+ \mathbb{P}\left(\dot{y}(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h_i(t)}{t^{1/2}} + 0.5\dot{y}(t)^\top(\eta_i - \eta_j)\middle|E_{ij1t}^c, \mathscr{F}_{t-1}, A_{it}^\star\right)\mathbb{P}(E_{ij1t}^c|\mathscr{F}_{t-1}, A_{it}^\star).
\end{aligned} \tag{36}$$

By Theorem 3 and Assumption 2, if $t > N_i^{(1)}(\delta, T)$, we have

$$\mathbb{P}(E_{ij1t}^c|\mathscr{F}_{t-1}, A_{it}^\star) = \mathbb{P}\left(4h_i(\delta, T)t^{-1/2} > \dot{y}(t)^\top(\eta_i - \eta_j)\middle|\mathscr{F}_{t-1}, A_{it}^\star\right) \leq \frac{4h_i(\delta, T)C'}{t^{1/2}}.$$

Thus, the probability in (36) can be written as

$$\begin{aligned}
& \mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -y(t)^\top(\widehat{\eta}_i(t) - \eta_i) + 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & \mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) + \frac{4h_i(\delta, T)C'}{t^{1/2}}.
\end{aligned}$$

Using $\dot{y}(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) \sim \mathcal{N}(0, v^2\dot{y}(t)^\top B_i^+(t)\dot{y}(t))$ given $\dot{y}(t)$, the first term on the RHS above can be written as

$$\begin{aligned}
&\mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > 0.25\dot{y}(t)^\top(\eta_i-\eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) + \frac{4h_i(\delta,T)C'}{t^{1/2}} \\
\leq\ & \int_0^\infty \mathbb{P}(\dot{y}(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > 0.25\dot{y}(t)^\top(\eta_i-\eta_j)|\dot{y}(t), \mathscr{F}_{t-1}, A_{it}^\star)\mathbb{P}(\dot{y}(t)^\top(\eta_i-\eta_j)=u)du \\
&\qquad + \frac{4h_i(\delta,T)C'}{t^{1/2}} \\
\leq\ & \int_0^\infty \exp\left(-\frac{t\lambda_m\nu_{im+}u^2}{32\nu_{iM}v^2}\right)\mathbb{P}(\dot{y}(t)^\top(\eta_i-\eta_j)=u|A_{it}^\star)du + \frac{4h_i(\delta,T)C'}{t^{1/2}}.
\end{aligned}$$

Since $\mathbb{P}(\dot{y}(t)^\top(\eta_i-\eta_j)=u|A_{it}^\star) < C'$ by Assumption 2, we have

$$\int_0^\infty \exp\left(-\frac{t\lambda_m\nu_{im+}u^2}{32\nu_{iM}v^2}\right)\mathbb{P}(\dot{y}(t)^\top(\eta_i-\eta_j)=u|A_{it}^\star)du \leq C'v\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}t}}.$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > -y(t)^\top(\widehat{\eta}_i(t)-\eta_i) + 0.5y(t)^\top(\eta_i-\eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & C't^{-1/2}\left(v\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}}} + 4h_i(\delta,T)\right). \qquad (37)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t)-\widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t)-\eta_j) + 0.5y(t)^\top(\eta_i-\eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & C't^{-1/2}\left(v\sqrt{\frac{32\nu_{jM}}{\lambda_m\nu_{jm+}}} + 4h_j(\delta,T)\right). \qquad (38)
\end{aligned}$$

Using (35), we have

$$\begin{aligned}
&\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t)-\widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star) \\
\leq\ & C't^{-1/2}\left(v\left(\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}}} + \sqrt{\frac{32\nu_{jM}}{\lambda_m\nu_{jm+}}}\right) + 4h_i(\delta,T) + 4h_j(\delta,T)\right).
\end{aligned}$$

By summing the probabilities in (39) over $i, j \in [N]$, we have

$$\begin{aligned}
&\sum_{i=1}^N\sum_{j=1}^N \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t)-\widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star)\mathbb{P}(A_{it}^\star) \\
\leq\ & \frac{C'}{\sqrt{t}}\sum_{i=1}^N\sum_{j=1}^N \mathbb{P}(A_{it}^\star)\left(v\left(\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}}} + \sqrt{\frac{32\nu_{jM}}{\lambda_m\nu_{jm+}}}\right) + 4h_i(\delta,T) + 4h_j(\delta,T)\right) \\
\leq\ & \frac{2c_MC'N}{\sqrt{t}}, \qquad (39)
\end{aligned}$$

where $c_M = \max_{i\in[N]}\left(v\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}}} + 4h_i(\delta,T)\right) = \mathcal{O}(\sqrt{d_\mu\log(T/\delta)})$. Note that

$$\mathbb{P}(a^\star(t) \neq a(t)|\mathscr{F}_{t-1}) \leq \sum_{i=1}^N\sum_{j=1}^N \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t)-\widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star)\mathbb{P}(A_{it}^\star), \qquad (40)$$

by the inclusion-exclusion formula. Putting (39), (40) and the minimal sample size $\max_{i\in[N]} N_i^{(1)}(\delta, T)$ together, we have

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(t) \neq a(t) | \mathscr{F}_{t-1}) &\leq \max_{i\in[N]} N_i^{(1)}(\delta, T) + 2c_M C' N \sum_{t=2}^{T} \frac{1}{t} \\ &\leq \max_{i\in[N]} N_i^{(1)}(\delta, T) + 2c_M C' N \log T.\end{aligned}$$

By (34), with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} I(a^\star(t) \neq a(t)) \leq \max_{i\in[N]} N_i^{(1)}(\delta, T) + \sqrt{8 \log T \log \delta^{-1}} + 2c_M C' N \log T.$$

Therefore, since $N_i^{(1)}(\delta, T) = \mathcal{O}(\log^3(TNd_y/\delta))$,

$$\begin{aligned}\text{Regret}(T) &\leq \sqrt{d_y} v_T(\delta) g(\delta) \left(\max_{i\in[N]} N_i^{(1)}(\delta, T) + \sqrt{8 \log T \log \delta^{-1}} + 2c_M C' N \log T \right) \\ &= \mathcal{O}\left(N d_\mu d_y^3 \log^{3.5}\left(\frac{TNd_y}{\delta} \right) \right).\end{aligned}$$

K Proof of Theorem 1

Before starting the proof, we specify the constants described in the statement in Theorem 1. L is the bound of the ℓ_2-norm of observation $L = \sqrt{2d_y \lambda_M \log(2d_y T/\delta)}$ such that $\|y(t)\| \leq L$. $\lambda_m = \lambda_{\min}(\Sigma_y)$ and $\lambda_M = \lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$. p_i is the probability of optimality of the ith arm, as defined in Definition 1. κ is the suboptimality gap defined in (17). Furthermore, ν_{im+} and ν_{iM} be the non-zero minimum and maximum eigenvalues of $J_i^\top D D^\top J_i$, respectively.

First, we show that the number of selections of each arm scales linearly with a high probability.

Lemma 9. *For partially observable stochastic contextual bandits, with probability at least $1-\delta$, if $t > s_i^{(3)}(\delta, T, \kappa)$, Algorithm 1 guarantees*

$$n_i(t) > \frac{p_i t}{4},$$

where $N_i^{(4)}(\delta, T, \kappa) := \max(2(a_{i1} + (4/p_i)a_{i2}^2) + 2\sqrt{(a_{i1} + (4/p_i)a_{i2}^2)^2 - a_{i1}^2}, N_i^{(3)}(\delta, T, \kappa)) = \mathcal{O}(d_\mu d_y^2 \log^3(TNd_y/\delta))$, $a_{i1} = \sum_{j\neq i}(s_i''(\delta) + s_j''(\delta)) + 2N/T$, $a_{i2} = \sqrt{2\log(2/\delta)}$, $N_i^{(3)}(\delta, T, \kappa) = \max(N_i^{(2)}(\delta, T, \kappa),\ 64(\nu_{iM} v^2/(\lambda_m \nu_{im+} \kappa^2)) \log T)$ and $N_i^{(2)}(\delta, T, \kappa)$ is defined in Theorem 3.

Proof. By Lemma 3, if $n_i(t) > N_i^{(2)}(\delta, T, \kappa)$ and $n_j(t) > N_j^{(2)}(\delta, T, \kappa)$,

$$\begin{aligned}&\mathbb{P}(a(t) = i | \mathscr{F}_{t-1}) \\ \geq\ & \frac{\mathbb{P}(a^\star(t) = i)}{2} \left(1 - \sum_{j\neq i} \left(\exp\left(-\frac{n_i(t) \lambda_m \nu_{im+} \kappa^2}{32 \nu_{iM} v^2} \right) + \exp\left(-\frac{n_j(t) \lambda_m \nu_{jm+} \kappa^2}{32 \nu_{jM} v^2} \right) \right) \right).\end{aligned}$$

If $n_i(t) \geq 64(\nu_{iM} v^2/(\lambda_m \nu_{im+} \kappa^2)) \log T$, we have $\exp\left(-(n_i(t)\lambda_m \nu_{im+} \kappa^2)/(32 \nu_{iM} v^2)\right) \leq T^{-2}$. Now, we assume $n_i(t) > N_i^{(3)}(\delta, T, \kappa) = \max(N_i^{(2)}(\delta, T, \kappa),\ 64(\nu_{iM} v^2/(\lambda_m \nu_{im+} \kappa^2)) \log T)$ for all $i \in [N]$. Since $I(a(t) = i) - (1/2)\mathbb{P}(a^\star(t) = i)\left(1 - \sum_{j\neq i} \mathbb{P}(a(t) = j | A_{it}, \mathscr{F}_{t-1})\right)$ is a

submartingale difference,

$$\begin{aligned}
&\sum_{\tau=1}^{t}\mathbb{P}(a(\tau)=i|\mathscr{F}_{\tau-1})\\
\geq\ & \frac{\mathbb{P}(a^\star(t)=i)}{2}\left(t-\sum_{\tau=1}^{t}\sum_{j\neq i}\mathbb{P}\left(\dot{y}(\tau)^\top(\widetilde{\eta}_j(\tau)-\widetilde{\eta}_i(\tau))>\kappa\middle|\, A_{i\tau},\mathscr{F}_{\tau-1}\right)\right)\\
\geq\ & \frac{\mathbb{P}(a^\star(t)=i)}{2}\left(t-\sum_{j\neq i}(N_i^{(3)}(\delta,T,\kappa)+N_j^{(3)}(\delta,T,\kappa))-\frac{2N}{T}\right).
\end{aligned}$$

Using Lemma 8, we have

$$\mathbb{P}\left(n_i(t)-\sum_{\tau=1}^{t}\mathbb{P}(a(\tau)=i|\mathscr{F}_{\tau-1})<-\epsilon\right)\leq e^{-\frac{\epsilon^2}{t}},$$

for any $\epsilon>0$. Accordingly, with probability at least $1-\delta$,

$$n_i(t)>\frac{\mathbb{P}(a^\star(t)=i)}{2}\left(t-\sum_{j\neq i}(N_i^{(3)}(\delta,T,\kappa)+N_j^{(3)}(\delta,T,\kappa))-\frac{2N}{T}\right)-\sqrt{2t\log(2/\delta)}.$$

The following inequality

$$\frac{p_i}{2}\left(t-\sum_{j\neq i}(N_i^{(3)}(\delta,T,\kappa)+N_j^{(3)}(\delta,T,\kappa))-\frac{2N}{T}\right)-\sqrt{2t\log(2/\delta)}>\frac{p_i}{4}t,$$

is satisfied, if $t>2(a_{i1}+(4/p_i)a_{i2}^2)+2\sqrt{(a_{i1}+(4/p_i)a_{i2}^2)^2-a_{i1}^2}$, where $a_{i1}=\sum_{j\neq i}(N_i^{(3)}(\delta,T,\kappa)+N_j^{(3)}(\delta,T,\kappa))+2N/T$ and $a_{i2}=\sqrt{2\log(2/\delta)}$ based on the quadratic formula. By (41), with probability at least $1-\delta$, we have

$$n_i(t)>\frac{p_i t}{4}, \tag{41}$$

if $t>N_i^{(4)}(\delta,T,\kappa):=\max(2(a_{i1}+(4/p_i)a_{i2}^2)+2\sqrt{(a_{i1}+(4/p_i)a_{i2}^2)^2-a_{i1}^2},N_i^{(3)}(\delta,T,\kappa))=\mathcal{O}(Nd_\mu d_y^2\log^3(TNd_y/\delta))$. □

Now, we are ready to prove Theorem 1. From (33), we have

$$\|\widehat{\eta}_i(t)-\eta_i\|\leq R\sqrt{\frac{2\nu_{iM}}{\lambda_m\nu_{im+}}}\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)n_i(t)^{-1/2}, \tag{42}$$

if $n_i(t)>N_i^{(1)}(\delta,T)$. Thus, putting (41) and (42) together, if $t>\tau_i(\delta,T,\kappa):=\max(N_i^{(4)}(\delta,T,\kappa),4p_i^{-1}N_i^{(1)}(\delta,T))=\mathcal{O}(p_i^{-1}\kappa^{-2}Nd_\mu d_y^2\log^3(TNd_y/\delta))$, with probability at least $1-\delta$, we have the following estimation accuracy

$$\|\widehat{\eta}_i(t)-\eta_i\|\leq R\sqrt{\frac{8\nu_{iM}}{\lambda_m\nu_{im+}p_i}}\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+v^{-1}h\right)t^{-1/2}.$$

L Proof of Theorem 2

The regret can be written as

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{t=1}^{T}y(t)^\top(\eta_{a^\star(t)}(t)-\eta_{a(t)}(t))I(a^\star(t)\neq a(t))\\
&\leq \sum_{t=1}^{T}y(t)^\top(\eta_{a^\star(t)}(t)-\widetilde{\eta}_{a^\star(t)}(t)+\widetilde{\eta}_{a(t)}(t)-\eta_{a(t)}(t))I(a^\star(t)\neq a(t)),
\end{aligned}$$

because $y(t)^\top(\widetilde{\eta}_{a(t)} - \widetilde{\eta}_{a^\star(t)}(t)) \geq 0$. Since $\|y(t)\| \leq \sqrt{d_y} v_T(\delta)$ for all $t \in [T]$, we have

$$\begin{aligned} & \sum_{t=1}^{T} y(t)^\top(\eta_{a^\star(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t) + \widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}(t)) I(a^\star(t) \neq a(t)) \\ \leq \quad & \sqrt{d_y} v_T(\delta) \sum_{t=1}^{T} (\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\|) I(a^\star(t) \neq a(t)). \end{aligned}$$

By Theorem 1 and the same logic as the proof of Corollary 1, if $t > \max_{i\in[N]} \tau_i(\delta, T, \kappa)$, we have

$$\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\| \leq R \max_{i\in[N]} \left(\sqrt{\frac{8\nu_{iM}}{\lambda_m \nu_{im+} p_i}} \right) g'(\delta) t^{-1/2},$$

where

$$\begin{aligned} g'(\delta) &= 2 \max_{i\in[N]} \left(\sqrt{\frac{8\nu_{iM}}{p_i \lambda_m \nu_{im+}}} \right) \left(v \sqrt{2 d_y \log \frac{2TN}{\delta}} + R \sqrt{d_\mu \log \left(\frac{1 + TL^2/\lambda}{\delta} \right)} + v^{-1} h \right) \\ &= \mathcal{O}\left((d_y^{1/2} + d_\mu^{1/2}) \sqrt{\log(TN d_y/\delta)} \right). \end{aligned}$$

To proceed, with the probability of at least $1 - \delta$, we utilize the martingale constructed in Theorem 4 with the intermediate result

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} I(a^\star(t) \neq a(t)) \leq \sqrt{8 \log T \log \delta^{-1}} + \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(\tau) \neq a(\tau) | \mathscr{F}_{\tau-1}). \tag{43}$$

To find a bound $\mathbb{P}(a^\star(\tau) \neq a(\tau)|\mathscr{F}_{\tau-1})$, we decompose the following probability using the inclusion-exclusion formula as follows:

$$\begin{aligned} & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0 | \mathscr{F}_{t-1}, A_{it}^\star) \\ \leq \quad & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5 y(t)^\top(\eta_i - \eta_j) | \mathscr{F}_{t-1}, A_{it}^\star) \\ + \quad & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5 y(t)^\top(\eta_i - \eta_j) | \mathscr{F}_{t-1}, A_{it}^\star). \end{aligned} \tag{44}$$

Similarly to (37) and (38), if $t > \tau_i(\delta, T, \kappa)$ for all $i \in [N]$, we have

$$\begin{aligned} & \mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -y(t)^\top(\widehat{\eta}_i(t) - \eta_i) + 0.5 y(t)^\top(\eta_i - \eta_j) | \mathscr{F}_{t-1}, A_{it}^\star) \\ \leq \quad & C' \sqrt{\frac{4}{p_i t}} \left(v \sqrt{\frac{32\nu_{iM}}{\lambda_m \nu_{im+}}} + 4h_i(\delta, T) \right). \end{aligned} \tag{45}$$

Subsequently, if $t > N_j(\delta, T, \kappa)$, we have

$$\begin{aligned} & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5 y(t)^\top(\eta_i - \eta_j) | \mathscr{F}_{t-1}, A_{it}^\star) \\ \leq \quad & C' \sqrt{\frac{4}{p_j t}} \left(v \sqrt{\frac{32\nu_{jM}}{\lambda_m \nu_{jm+}}} + 4h_j(\delta, T) \right). \end{aligned} \tag{46}$$

Accordingly, based on (44), (45), and (46), we obtain the following bounds for the probabilities

$$\begin{aligned} & \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0 | \mathscr{F}_{t-1}, A_{it}^\star) \\ \leq \quad & \frac{2C'}{\sqrt{p_{\min}^+}} \left(v \left(\sqrt{\frac{32\nu_{iM}}{\lambda_m \nu_{im+}}} + \sqrt{\frac{32\nu_{jM}}{\lambda_m \nu_{jm+}}} \right) + 4h_i(\delta, T) + 4h_j(\delta, T) \right) t^{-1/2}, \end{aligned}$$

where $p_{\min}^{+} = \min_{i\in[N]:p_i>0} p_i$. By summing the probabilities up over $i, j \in [N]$, if $t > \tau(\delta, T, \kappa) := \max_{i\in[N]} \tau_i(\delta, T, \kappa)$, we have the following upper bound for the probability of choosing a suboptimal arm

$$\begin{aligned}
& \mathbb{P}(a^\star(t) \neq a(t)|\mathscr{F}_{t-1}) \\
\leq & \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star)\mathbb{P}(A_{it}^\star) \\
\leq & \frac{2C'}{\sqrt{p_{\min}^{+} t}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{P}(A_{it}^\star)\left(v\left(\sqrt{\frac{32\nu_{iM}}{\lambda_m \nu_{im+}}} + \sqrt{\frac{32\nu_{jM}}{\lambda_m \nu_{jm+}}}\right) + 4h_i(\delta, T) + 4h_j(\delta, T)\right) \\
\leq & \frac{4c_M C' N}{\sqrt{p_{\min}^{+} t}}, \qquad (47)
\end{aligned}$$

where $c_M = \max_{i\in[N]}\left(v\sqrt{\frac{32\nu_{iM}}{\lambda_m\nu_{im+}}} + 4h_i(\delta, T)\right) = \mathcal{O}(\sqrt{d_\mu \log(T/\delta)})$. Putting (47) and the minimum sample size $\tau(\delta, T, \kappa)$ together, we have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(t) \neq a(t)|\mathscr{F}_{t-1}) \leq \tau(\delta, T, \kappa) + \frac{8c_M C' N}{\sqrt{p_{\min}^{+}}} \sum_{t=\lceil\tau(\delta,T,\kappa)\rceil}^{T} \frac{1}{t} \leq \tau(\delta, T, \kappa) + \frac{8c_M C' N}{\sqrt{p_{\min}^{+}}} \log T,$$

where $\lceil\cdot\rceil$ is the ceiling function. By (34), with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} I(a^\star(t) \neq a(t)) \leq \tau(\delta, T, \kappa) + \sqrt{8 \log T \log \delta^{-1}} + \frac{8c_M C' N}{\sqrt{p_{\min}^{+}}} \log T.$$

Therefore, due to the order of the minimum sample size $\tau(\delta, T, \kappa) = \mathcal{O}((p_{\min}^{+})^{-1}\kappa^{-2} N d_\mu d_y^2 \log^3(TNd_y/\delta))$,

$$\begin{aligned}
\text{Regret}(T) & \leq \sqrt{d_y} v_T(\delta) g'(\delta) \left(\tau(\delta, T, \kappa) + \sqrt{8 \log T \log \delta^{-1}} + \frac{8c_M C' N}{\sqrt{p_{\min}^{+}}} \log T\right) \\
& = \mathcal{O}\left(\frac{N d_\mu d_y^3}{p_{\min}^{+} \kappa^2} \log^4\left(\frac{TNd_y}{\delta}\right)\right).
\end{aligned}$$

References

[1] Jeng-Wen Lin, Cheng-Wu Chen, and Cheng-Yi Peng. Kalman filter decision systems for debris flow hazard assessment. *Natural hazards*, 60(3):1255–1266, 2012.

[2] IJ Nagrath. *Control systems engineering*. New Age International, 2006.

[3] Edward R Dougherty. *Digital image processing methods*. CRC Press, 2020.

[4] Yeonsik Kang, Chiwon Roh, Seung-Beum Suh, and Bongsob Song. A lidar-based decision-making method for road boundary detection using multiple kalman filters. *IEEE Transactions on Industrial Electronics*, 59(11):4360–4368, 2012.

[5] Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

[6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[7] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

[8] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Thompson sampling for partially observable linear-quadratic control. In *2023 American Control Conference (ACC)*, pages 4561–4568. IEEE, 2023.

[9] Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.

[10] Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *Conference on Learning Theory*, pages 1547–1575. PMLR, 2022.

[11] Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 1484–1515. PMLR, 2023.

[12] Alain Bensoussan. *Stochastic control of partially observable systems*. Cambridge University Press, 2004.

[13] Vikram Krishnamurthy and Bo Wahlberg. Partially observed markov decision process multiarmed bandits—structural results. *Mathematics of Operations Research*, 34(2):287–302, 2009.

[14] Hongju Park and Mohamad Kazem Shirani Faradonbeh. Analysis of thompson sampling for partially observable contextual multi-armed bandits. *IEEE Control Systems Letters*, 6:2150–2155, 2021.

[15] Hongju Park and Mohamad Kazem Shirani Faradonbeh. Efficient algorithms for learning to control bandits with unobserved contexts. *IFAC-PapersOnLine*, 55(12):383–388, 2022.

[16] Hongju Park and Mohamad Kazem Shirani Faradonbeh. Worst-case performance of greedy policies in bandits with imperfect context observations. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 1374–1379. IEEE, 2022.

[17] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[18] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[19] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[20] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

[21] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[22] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[23] Bianca Dumitrascu, Karen Feng, and Barbara Engelhardt. Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, 31, 2018.

[24] Melody Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[25] Nirandika Wanigasekara and Christina Yu. Nonparametric contextual bandits in metric spaces with unknown metric. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[27] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

[28] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[29] Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. Greedy algorithm almost dominates in smoothed contextual bandits. *SIAM Journal on Computing*, 52(2):487–524, 2023.

[30] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

[31] Sunrit Chakraborty, Saptarshi Roy, and Ambuj Tewari. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR, 2023.

[32] David Harville. Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, 4(2):384–395, 1976.

[33] George K Robinson. That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[34] Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *The Journal of Machine Learning Research*, 22(1):5928–5976, 2021.

[35] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[36] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.