Balancing exploration and exploitation in Partially Observed Linear Contextual Bandits via Thompson Sampling

Anonymous Authors[1]

Abstract

Contextual bandits constitute a popular framework for studying the exploration-exploitation trade-off under finitely many options with side information. In the majority of the existing works, contexts are assumed perfectly observed, while in practice it is more reasonable to assume that they are observed *partially*. In this work, we study reinforcement learning algorithms for contextual bandits with partial observations. First, we consider different structures for partial observability and their corresponding optimal policies. Subsequently, we present and analyze reinforcement learning algorithms for partially observed contextual bandits with noisy linear observation structures. For these algorithms that utilize Thompson sampling, we establish estimation accuracy and regret bounds under different structural assumptions.

1. Introduction

Contextual bandits provide the framework for sequential decision-making given the available information. In general, for contextual bandits, finite options can be taken given fully observed contexts. Contexts refer to information about available options, often representing individual characteristics in many applications (Li et al., 2010; Bouneffouf et al., 2012; Tewari & Murphy, 2017; Nahum-Shani et al., 2018; Durand et al., 2018; Varatharajah et al., 2018; Ren & Zhou, 2020). In contextual bandits, similarly to other reinforcement learning problems, the exploration-exploitation trade-off needs to be addressed to get satisfactory performances. There are two methods to address the trade-off in the main stream: OFU and Thompson Sampling.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML). Do not distribute.

The origin of Thompson sampling goes back to the literature (Thompson, 1933). Recently, Thompson sampling has become more popular for addressing the trade-off of exploration and exploitation because of its simplicity as well as good performance. As compared to methods with Optimism in the Face of Uncertainty (OFU), Thompson sampling has been known to have easier installment and heuristically better performance (Chapelle & Li, 2011; Agrawal & Goyal, 2013).

Meanwhile, stochastic contextual bandits have various assumptions about their features such as reward functions, context space, and action space. For reward functions, a popular one is a linear reward function (Dani et al., 2008; Hamidi & Bayati, 2020; Agrawal & Goyal, 2013), while more general models assume non-linearity for reward functions (Dumitrascu et al., 2018; Modi & Tewari, 2020). Next, for action space, a common action set is a pre-fixed finite set representing finite arms, which does not change over time (Agrawal & Goyal, 2013). On the contrary, the other general models have an infinite action set, which consists of d-dimensional context vectors (Abbasi-Yadkori et al., 2011). For linear contextual bandits with finite arms, a reward for each arm is generated based on a linear function of a given context and parameter with a noise. Reward functions can take various forms of inputs, contexts and parameters. For clarity, we define the terms *private* and *public* for contexts and parameters. Here, a public one is a common input for reward functions for all arms, while a private one is associated only with the reward function of the corresponding arm. Generally, the linear function can have three structures: private contexts and a public parameter (Agrawal & Goyal, 2013); a public context and private parameters; private contexts and private parameters. For example, for N-armed contextual bandits with a public context and private parameters, all the arms share a public context, but each arm has its own private parameter so there are N private parameters (Agrawal & Goyal, 2013). In this paper, we analyze all three cases, especially focusing on the one with private contexts and private parameters, which can be the general case of the other two.

The reinforcement learning community has paid sufficient attention to decision-making algorithms in the absence of information uncertainty. However, frameworks with imperfect information and decision-making algorithms for them have not drawn sufficient interest, even though the information for decision-making is often observed in a partial, transformed, or noisy manner in practice (Bensoussan, 2004). Imperfect observations are the problems of interest in various areas such as state-space models, robot control, image processing and filtering, which are associated with decision-making problems (Nise, 2020; Nagrath, 2006; Lin et al., 2012; Dougherty, 2020; Kang et al., 2012). The imperfect observations in contexts can be caused by many reasons: privacy regulations, measurement errors, and missing data (Lin et al., 2012; Kang et al., 2012; Sbeity & Younes, 2015; Azimi et al., 2019). Ignorance of the imperfectness of observations can cause imprecise decisions in many applications such as health care, advertisements, and clinical trials (Dyczkowski, 2018; Nahum-Shani et al., 2018; Li et al., 2010; Bouneffouf et al., 2012). For example, for sick septic patients, if missing information is not properly adjusted for clinical context, clinicians' decision-making may result in worse outcomes (Gottesman et al., 2019). To this end, we suggest decision-making algorithms for contextual bandits in the presence of imperfectly observed contexts.

Imperfect or partial observations in decision-making get more interest in the reinforcement learning community. A Partially Observable Markov Decision Process (POMDP), which is a generalization of a Markov decision process (MDP), was introduced to address imperfect observations in decision making (Åström, 1965; Kaelbling et al., 1998). Recently, some contextual bandits models have started to take the imperfectness of contexts into account as well. However, the existing studies consider some particular cases under certain assumptions. In cases where some elements of contexts are missing and the others are fully observed, UCB-type algorithms have been employed based on the correlations between these two types of elements have been used to minimize the regret (Tennenholtz et al., 2021). In addition, under the presence of only a public parameter, analyses about UCB-type algorithms and Thompson sampling have been done for contextual bandit with invertible linear observation function (Yun et al., 2017; Park & Faradonbeh, 2021) and greedy algorithms are shown to have logarithmic regret with respect to the time horizon for the general linear observation function under normality assumption (Park & Faradonbeh, 2022). But, analyses for the case with private parameters and the general linear observation function have not been studied yet. In this paper, we analyze Thompson sampling for partially observed contextual bandits relaxing the assumptions in the existing literature. We perform the finite-time worst-case analysis under the sub-gaussian assumption for observations, which is more general than the normality assumption. In addition, we construct the model with a general linear observation structure, which can include various cases.

The remainder of this paper is organized as follows. In Section 2, we formulate the model and discuss the relevant preliminary materials. Next, Thompson sampling for contextual bandits with partially observed contexts is presented in Section 3. In Section 4, we provide theoretical performance guarantees for the proposed algorithm. Finally, we conclude the paper and discuss future directions.

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, $C(A)$ is employed to denote the column space of the matrix A. Further, polylog(xy/z) is a polynomial of $\log x$, $\log y$ and $\log z^{-1}$. Finally, $P_{C(A)}$ is the projection operator onto $C(A)$, and $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) denotes the minimum (maximum) eigenvalue of A.

2. Problem Formulation

In this section, we discuss stochastic contextual bandits with unobserved contexts, where the reward of the ith arm is generated based on the following probabilistic assumption

$$r_i(t) = f(x(t), i) + \varepsilon_i(t), \tag{1}$$

where $x(t)$ is an unknown d_x-dimensional stochastic context at time t with the mean $\mathbf{0}_{d_x}$ and a covariance matrix Σ_x, f is a deterministic unknown linear function from $\mathbb{R}^{dim(x(t))+1}$ to $\mathbb{R}^1$ and $\varepsilon_i(t)$ is a sub-Gaussian noise generated independently such that

$$\mathbb{E}\left[e^{\lambda\varepsilon_i(t)}\right] \le e^{\frac{\lambda^2 R_1^2}{2}},$$

for some $R_1 > 0$. Instead of the context $x(t)$, a transformed noisy context, denoted as $y(t)$, can be observed based on the following observation model

$$y(t) = Ax(t) + \xi(t), \tag{2}$$

where A is a matrix in $\mathbb{R}^{d_y\times d_x}$; $\xi(t)$ is a sub-Gaussian noise vector centered at 0 with the positive definite covariance Σ_Y. A learner is aware of the probabilistic assumption of rewards (1), but does not know the function f. At each time t, the learner tries to choose the optimal arm given the history of actions $\{a(\tau)\}_{1\le\tau\le t-1}$, rewards $\{r_{a(\tau)}(\tau)\}_{1\le\tau\le t-1}$, and observations $\{y(\tau)\}_{1\le\tau\le t-1}$ as well as the current observation $y(t)$. f has a linearity assumption such that

$$f(x(t), i) = x(t)^\top J_i \mu_*, \tag{3}$$

where μ_* is the parameter of interest and J_i is a known matrix in $\mathbb{R}^{dim_x \times dim_\mu}$. Since the optimal policy does not know the value of $x(t)$ as well, $f(x(t), i)$ is not available for it. Thus, the optimal policy also needs to estimate $f(x(t), i)$ based on the observation $y(t)$.

First, assuming the function f to be known, we investigate how to find the estimate of $f(x(t), i)$. To find an estimate of $f(x(t), i)$, we first find an estimate of $x(t)$. To proceed, based on (2), we aim to find an estimate of context $x(t)$. Since $x(t)$ is an unobserved random variable, the minimizer of the expected norm of the difference between $x(t)$ and a linear unbiased predictor $Dy(t)$ such that

$$Dy(t) = \underset{Dy(t), D \in \mathbb{R}^{d_y \times d_x}}{\arg\min} \mathbb{E}[(x(t) - Dy(t))^\top (x(t) - Dy(t))]. \quad (4)$$

can be a predictor of $x(t)$. A solution of (4) is the best linear unbiased prediction (BLUP) of $x(t)$, denoted as $\widehat{x}(t)$,

$$\widehat{x}(t) := (A^\top \Sigma_Y^{-1} A + \Sigma_X^{-1})^{-1} A^\top \Sigma_Y^{-1} y(t) = Dy(t), \quad (5)$$

where $D = (A^\top \Sigma_Y^{-1} A + \Sigma_X^{-1})^{-1} A^\top \Sigma_Y^{-1}$ (Robinson, 1991). Because f is a linear function, $f(x(t), i)$ can be represented as $x(t)^\top \mu$ for a $\mu \in \mathbb{R}^{dim(x(t))}$. Then, by the extension of Gauss-Markov theorem, we have a BLUP of $x(t)^\top \mu$, $\widehat{x}(t)^\top \mu = f(\widehat{x}(t), i)$. Since $\widehat{x}(t)$ is a function of $y(t)$, $f(\widehat{x}(t), i)$ also can be written as $f_*(y(t), i)$ for a function f_*. That is,

$$f_*(y(t), i) := f(\widehat{x}(t), i).$$

Specifically, for the ith arm, $f(x(t), i) = x(t)^\top J_i \mu_*$ is predictable with $y(t)$ given $\mu_i := J_i \mu_*$, where the estimate of $f(x(t), i) = x(t)^\top \mu_i$ is

$$f_*(y(t), i) = y(t)^\top D^\top J_i \mu_*. \quad (6)$$

Now, we investigate the estimation of $f_*(y(t), i)$ given $y(t)$. Define

$$\eta_i := D^\top J_i \mu_*. \quad (7)$$

Thus, using (1), (2), (6) and (7), we get

$$r_i(t) = y(t)^\top \eta_i + \zeta_i(t) \quad (8)$$

where $\zeta_i(t) = (x(t)^\top J_i \mu_* - y(t)^\top \eta_i) + \varepsilon_i(t)$ is a noise independent from the others. η_i is always guaranteed to be estimable thanks to the full rank Σ_Y. In fact, given the observation $y(t)$, the estimation of η_i is necessary and sufficient to estimate $f_*(y(t), i)$, while $J_i \mu_*$ and μ_* are not estimable because of rank deficiencies. For these reasons, instead of $J_i \mu_*$, we estimate η_i.

The optimal arm is the arm maximizing the expected reward given the observations. Thus, the optimal arm at time t can be presented as

$$a^*(t) = \underset{1 \le i \le N}{\arg\max} f_*(y(t), i) = \underset{1 \le i \le N}{\arg\max}\, y(t)^\top \eta_i.$$

The framework described is a general observational structure for partially observed contextual bandits. The following two settings are the most common structures for contextual bandits.

1. A single parameter and multiple contexts (SPMC)

   $f(x(t), i) = x_i(t)^\top \mu_*$ and $y_i(t) = A_0 x_i(t) + \xi_i(t)$

   $x_i(t)$ represents the context of the ith arm at time t and $A = diag(A_0, \dots, A_0)$. The context $x(t)$ at time t is a concatenation of the contexts of all arms such that $x(t) = [x_1(t)^\top, x_2(t)^\top, \dots, x_N(t)^\top]^\top$. $J_i = [\mathbf{0}_{d_x \times d_x} \ \cdots \ \underbrace{I_{d_x}}_{ith} \ \cdots \ \mathbf{0}_{d_x \times d_x}]^\top$. In this case, the optimal arm can be represented as

   $$a^*(t) = \arg\max_i y(t)^\top D^\top J_i \mu_* = \arg\max_i y_i(t)^\top \eta_*,$$

   where $\eta_* = D_0^\top \mu_*$. Note that the column space of J_i is the same for all i under this assumption. That is, regardless of which arm has been chosen, the decision maker can learn the parameter η_*.

2. Multiple parameters and multiple contexts (general case)

   $f(x(t), i) = x_i(t)^\top \mu_{i*}$ and $y_i(t) = A x_i(t) + \xi_i(t)$

   $x_i(t)$ represents the context of the ith arm at time t. The context $x(t)$ at time t is a concatenation of the contexts of all arms such that $x(t) = [x_1(t)^\top, x_2(t)^\top, \dots, x_N(t)^\top]^\top$. μ_{i*} denotes the parameter of the ith arm, which is associated only with the reward of the ith arm. μ_* is written as $\mu_* = [\mu_{*1}, \mu_{*2}, \dots, \mu_{*N}]$. $J_i = diag(\mathbf{0}_{d_x \times d_x}, \ \cdots \ \underbrace{I_{d_x}}_{ith}, \ \cdots \ \mathbf{0}_{d_x \times d_x})$.

   $$a^*(t) = \arg\max_i y(t)^\top D^\top J_i \mu_* = \arg\max_i y_i(t)^\top \eta_{*i},$$

   where $\eta_{*i} = D_0 \mu_{*i}$.

We consider the second case as the general case because it includes all the other cases.

Regret is a performance measure, which can be written as the cumulative sum of expected reward differences between the optimal and chosen arms over time

$$\text{Regret}(T) = \sum_{t=1}^{T} y(t)^\top (\eta_{a^*(t)} - \eta_{a(t)}), \quad (9)$$

where $a(t)$ is the chosen arm at time t. The learner eventually aims to minimize the regret by trying to choose the optimal arm at each time. Accordingly, the goals of this paper are to find algorithms minimizing the regret and regret bounds of the algorithms, which are attracting attention in the reinforcement learning community. Here, f_* is the function of interest because it is the best information about the reward given the observation $y(t)$.

3. Reinforcement Learning Policy

In this section, we describe Thompson sampling algorithm for contextual bandits with partial observations. The algorithm assumes the probabilistic structure of the reward generation of the arm i given the observation

$$r_i(t) = y(t)^\top D^\top J_i \mu_* + \varepsilon_i(t),$$

where $\varepsilon_i(t) \sim \mathcal{N}(\mathbf{0}, v^2)$. With a prior distribution of μ_*, $\mathcal{N}(0, v^2\lambda^{-1}I)$, the posterior distribution at time t can be given as $\mathcal{N}(\widehat{\mu}(t), v^2 B(t)^{-1})$, where

$$\widehat{\mu}(t) = B(t)^{-1} \sum_{\tau=1}^{t-1} r_{a(\tau)}(\tau) J_{a(t)}^\top D y(\tau), \tag{10}$$

$$B(t) = \lambda I + \sum_{\tau=1}^{t-1} J_{a(t)}^\top D y(t) y(t)^\top D^\top J_{a(t)}. \tag{11}$$

At time t, with

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

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

by generating a sample

$$\widetilde{\eta}_i(t) \sim \mathcal{N}(\widehat{\eta}_i(t), v^2 B_i(t)^{-1}) \tag{14}$$

which is the posterior distribution, the optimal arm estimation can be done by

$$a(t) = \underset{1 \le i \le N}{\arg\max}\, y(t)^\top \widetilde{\eta}_i(t). \tag{15}$$

Here, $D^\top J_a \widetilde{\mu}(t)$ can be an estimate of η_i. We can update the $\widehat{\mu}(t)$ based on the recursions below:

$$B(t+1) = B(t) + J_{a(t)}^\top D y(t) y(t)^\top D^\top J_{a(t)}, \tag{16}$$

$$\widehat{\mu}(t+1) = B(t+1)^{-1}\left(B(t)\widehat{\mu}(t) + J_{a(t)}^\top D y(t) r_{a(t)}(t)\right), \tag{17}$$

where $B(1) = \lambda I$ and $\widehat{\mu}(1) = \mathbf{0}_{d_\mu}$.

The pseudo-code of Thompson sampling for contextual bandit with partial observation is given in Algorithm 1. Algorithm starts with initial values $B(1) = \lambda I$ and $\widehat{\mu}(1) = \mathbf{0}_{d_\mu}$. Then, at each time, based on the posterior, generate samples and select an estimate of the optimal arm maximizing the quantity in (15). With the reward gained from the chosen arm, update the posterior mean and covariance.

Algorithm 1 : Thompson sampling algorithm for contextual bandits with partial observations

Set $B(1) = \lambda I_{d_\mu}$, $\widehat{\mu}(1) = \mathbf{0}_{d_\mu}$ for $i = 1, \ldots, N$
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 $a(t) = \arg\max_i y(t)^\top \widetilde{\eta}_i(t)$
  Gain reward $r_{a(t)}(t) = f(x(t), a(t)) + \varepsilon_{a(t)}(t)$
  Update $B(t+1)$ and $\widehat{\mu}(t+1)$ by (16) and (17)
end for

4. Results

Next, we establish theoretical results for Algorithm 1 suggested in the previous section. The results provide a high probability regret bound for Algorithm 1 and estimation error bounds of the estimators defined in (12). Without loss of generality, we assume that $\|J_i\mu_*\| \le 1$ for all $i \in \{1, 2, \ldots, N\}$. We first show the results for the general setting encompassing the first (SPMC) and second settings (MPMC) introduced in Section 2. The complete proof of the following results is provided in Appendix.

4.1. Results for the general setting

Theorem 4.1. *Let $w_t = r_{a(t)}(t) - \widehat{x}(t)^\top J_{a(t)}\mu$ and $\mathscr{F}_t = \sigma\{\{y(\tau)\}_{\tau=1}^{t+1}, \{a(\tau)\}_{\tau=1}^{t+1}\}$. Then, w_t is $\mathscr{F}_{t-1}$-measurable and conditionally R-sub-Gaussian for some $R > 0$ such that*

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

For any $\delta > 0$, assuming that $\|\mu_*\| \le h$ and $B(1) = \lambda I$, $\lambda > 0$, with probability at least $1 - \delta$, we have

$$\|\widehat{\mu}(t) - \mu_*\|_{B(t)} = \left\| \sum_{\tau=1}^{t-1} J_{a(\tau)}^\top D y(\tau) w_\tau \right\|_{B(t)}$$

$$\le R\sqrt{d_\mu \log\left(\frac{1 + L^2 t/\lambda}{\delta}\right)} + \lambda^{1/2} h,$$

where $L = \sqrt{d_y} v_T(\delta)$, $v_T(\delta) = (2\lambda_M \log(2d_y T/\delta))^{1/2}$, $\lambda_M = \lambda_{\max}(A\Sigma_X A^\top + \Sigma_Y)$, $d_y = dim(y(t))$ and $d_\mu = dim(\mu_*)$.

Theorem 4.1 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.

The next theorem provides the lower bound of the smallest eigenvalue of sample covariance matrix $B_i(t)$, which is associated with the error of estimation η_i. We denote $n_i(t)$ as the count of the i arm chosen up to the time t.

Theorem 4.2. *Let $\ell_i(t)=\sum_{j:C(J_i)=C(J_j)} n_j(t)$. For $B(t)$ in (16), on the event W_T defined in (20), with probability at least $1-\delta$, if $\ell_i(t)\geq v_T(\delta)^4/(2\lambda_m^2\nu_{im}^2)\log(T/\delta)$, we have*

$$\lambda_{\min}\left(D^\top J_i B(t) J_i^\top D\right)\geq \frac{\nu_{iM}\lambda_m\nu_{im}}{2}\ell_i(t)$$

and

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

Definition 4.3. $A_i^*\in\mathbb{R}^{d_y}$ is the set such that $a^*(t)=i$, if and only if $y(t)\in A_i^*$.

Proposition 4.4. *For any arm i, there exist a set $A_i\subseteq A_i^*$ and $\epsilon_i>0$ such that $P(y(t)\in A_i)>\frac{1}{2}P(y(t)\in A_i^*)$ and $y(t)^\top(\eta_i-\eta_j)>\epsilon_i$, if $y(t)\in A_i$.*

The proposition above helps to find a lower bound of the probability $\mathbb{P}(a(t)=i|\mathscr{F}_{t-1})$ in the next theorem, which can provide a lower bound of the number of each arm being chosen.

Theorem 4.5. *Let*

$$m_{ij}(T)=\max\left(v_T(\delta)^4\frac{\log(T/\delta)}{2\lambda_m^2\nu_{jm}^2},\ \nu_{jM}\lambda_m\nu_{jm}q(T)\epsilon_i^{-1}\right),$$

where $q(T)=R\sqrt{d_\mu\log\left(1+\frac{L^2T}{\delta}\right)}+\lambda^{\frac{1}{2}}h$ and ϵ_i is defined in Proposition 4.4. Then, if $\ell_i(t)>m_{ii}(T)$ and $\ell_j(t)>m_{ij}(T)$,

$$\mathbb{P}(a(t)=i|\mathscr{F}_{t-1})\geq \frac{\mathbb{P}(a^*(t)=i)}{2}\left(1-\sum_{j\neq i}\left(e^{-\frac{\ell_i(t)\epsilon_i^2}{8v^2}}+e^{-\frac{\ell_j(t)\epsilon_i^2}{8v^2}}\right)\right).$$

The results above can be applied to both the two common cases defined in Section 2. Now, we focus on regret analysis. We investigate regret bounds for two settings discussed in Section 2. First, we consider setting 1, where all arms share the parameter.

4.2. Regret upper bound under the SPMC assumption

Under the SPMC assumption, the column spaces of J_i for different arms are identical. Thus, $\ell_i(t)=t$ for all $i\in[N]$. The next theorem guarantees the estimation accuracy under the SPMC assumption, which is proportional to $t^{-0.5}$. This implies that the parameter of interest η_i can be learned regardless of which arm is chosen.

Theorem 4.6. *Let η_i and $\widehat{\eta}_i(t)$ be the transformed true parameter in (7) and the estimate in (12), respectively. Then, under the SPMC assumption, if $t>8(v_T(\delta)^4/(\lambda_m^2\nu_{im}^2))\log(T/\delta)$, with probability at least $1-\delta$, for all $0<t\leq T$, we have*

$$\|\widehat{\eta}_i(t)-\eta_i\|\leq\frac{R\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m\nu_{im}}}{2t^{\frac{1}{2}}}q(T).$$

where ν_{iM} and ν_{im} are the maximum and the non-zero minimum eigenvalue of $J_i^\top DD^\top J_i$, respectively; $\lambda_{\min}(\Sigma_Y)=\lambda_m$; $q(T)$ is defined in Theorem 4.5.

The next theorem shows a poly-logarithmic upper bound with respect to the time horizon under the SPMC assumption.

Theorem 4.7. *Assume that Algorithm 1 is used in a bandit under the SPMC assumption. Then, with probability at least $1-\delta$, $\mathrm{Regret}(T)$ is of the order*

$$\mathrm{Regret}(T)=\mathcal{O}\left(N(d_\mu+\sqrt{d_\mu d_y})polylog\left(\frac{TNd_y}{\delta}\right)\right).$$

4.3. Regret upper bound for the general assumption

Under the general assumption, note that $\ell_i(t)=n_i(t)$, since all the column spaces of J_i do not overlap each other. The next theorem presents the estimation error of $\widehat{\eta}_i$ and a lower bound of $n_i(t)$. The estimation error is proportional to the inverse of the square root of $h_i(t)$, which is a lower bound of $n_i(t)$.

Theorem 4.8. *Let η_i and $\widehat{\eta}_i(t)$ be the transformed true parameter in (7) and the estimate in (12), respectively. Then, under the general assumption, if $t>max(8(v_T(\delta)^4/(\lambda_m^2\nu_{im}^2))\log(T/\delta),123)$, with probability at least $1-\delta$, for all $0<t\leq T$, we have*

$$\|\widehat{\eta}_i(t)-\eta_i\|_2\leq\frac{R\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m\nu_{im}}}{\sqrt{p_i t}}\cdot\left(\sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)}+\lambda^{\frac{1}{2}}h\right). \tag{18}$$

From the theorem above, we can find the frequency $n_i(t)$ increases linearly with the time horizon. Accordingly, in the next theorem, the regret upper bound also grows with at most poly-logarithmic rate thanks to the linear growth of $n_i(t)$ even under the general assumption.

Theorem 4.9. *Assume that Algorithm 1 is used in a bandit under the general assumption. Then, with probability at least* $1-\delta$, $\mathrm{Regret}(T)$ *is of the order*

$$\mathrm{Regret}(T) = \mathcal{O}\left(\max_{i,j} p_i^{-0.5} N(d_\mu + \sqrt{d_y d_\mu}) polylog\left(\frac{TNd_y}{\delta}\right)\right).$$

where $p_m = \min_i \mathbb{P}(a^*(t) = i)$.

5. Numerical Experiments

In this section, we show the results in Section 4 based on numerical simulation. First, to see the relationships between the regret and dimension of observations and contexts, we simulate various cases under the general assumption for $N = 5$ 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 shows normalized regret over time for different dimensions of observations and contexts. Because the regret grows poly-logarithmically with respect to t, we normalize the regret by $(\log t)^2$. Next, Figure 2 shows the normalized errors for different cases of dimensions of observations and contexts at $N = 5$. Since the estimation errors decrease with $t^{-0.5}$ in Theorem 4.8, we describe $\sqrt{t}\|\widehat{\eta}_i(t) - \eta_i\|_2$ over time. We evaluate the average estimation errors of η_i for 5 different arms over time. Since the errors decrease rate $t^{-0.5}$ and $\sqrt{t}$ cancel out each other, the normalized errors for all the arms are flattened over time. This shows that the estimations of η_i are available regardless of whether the dimension of observations is greater or less than that of contexts.

6. Conclusion

We studied Thompson sampling for contextual bandits with partial observations under relaxed assumptions. Indeed, the suggested model formulation covers various possible cases for observation structures and provides estimation processes for contexts. Further, we show that the parameter estimates converge to the truth, and that as time goes by, the presented algorithm learns the unknown true parameter accurately. Finally, we proved that Thompson sampling has upper bounds with a poly-logarithmic rate for the most common two cases.

A problem of future interest is the modeling, estimation and algorithms for the unknown observation structure, where the sensing matrix A is unknown. Further, relaxing the linear observation structure to non-linear can be a problem of interest.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

Azimi, I., Pahikkala, T., Rahmani, A. M., Niela-Vilén, H., Axelin, A., and Liljeberg, P. Missing data resilient decision-making for healthcare iot through personalization: A case study on maternal health. *Future Generation Computer Systems*, 96:297–308, 2019.

Bensoussan, A. *Stochastic control of partially observable systems*. Cambridge University Press, 2004.

Bouneffouf, D., Bouzeghoub, A., and Gançarski, A. L. A contextual-bandit algorithm for mobile context-aware recommender system. In *International conference on neural information processing*, pp. 324–331. Springer, 2012.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Dougherty, E. R. *Digital image processing methods*. CRC Press, 2020.

Dumitrascu, B., Feng, K., and Engelhardt, B. Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, 31, 2018.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pp. 67–82. PMLR, 2018.

Dyczkowski, K. Intelligent medical decision support system based on imperfect information. *Studies in Computational Intelligence. Springer, Cham, Switzerland. doi*, 10:978–3, 2018.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

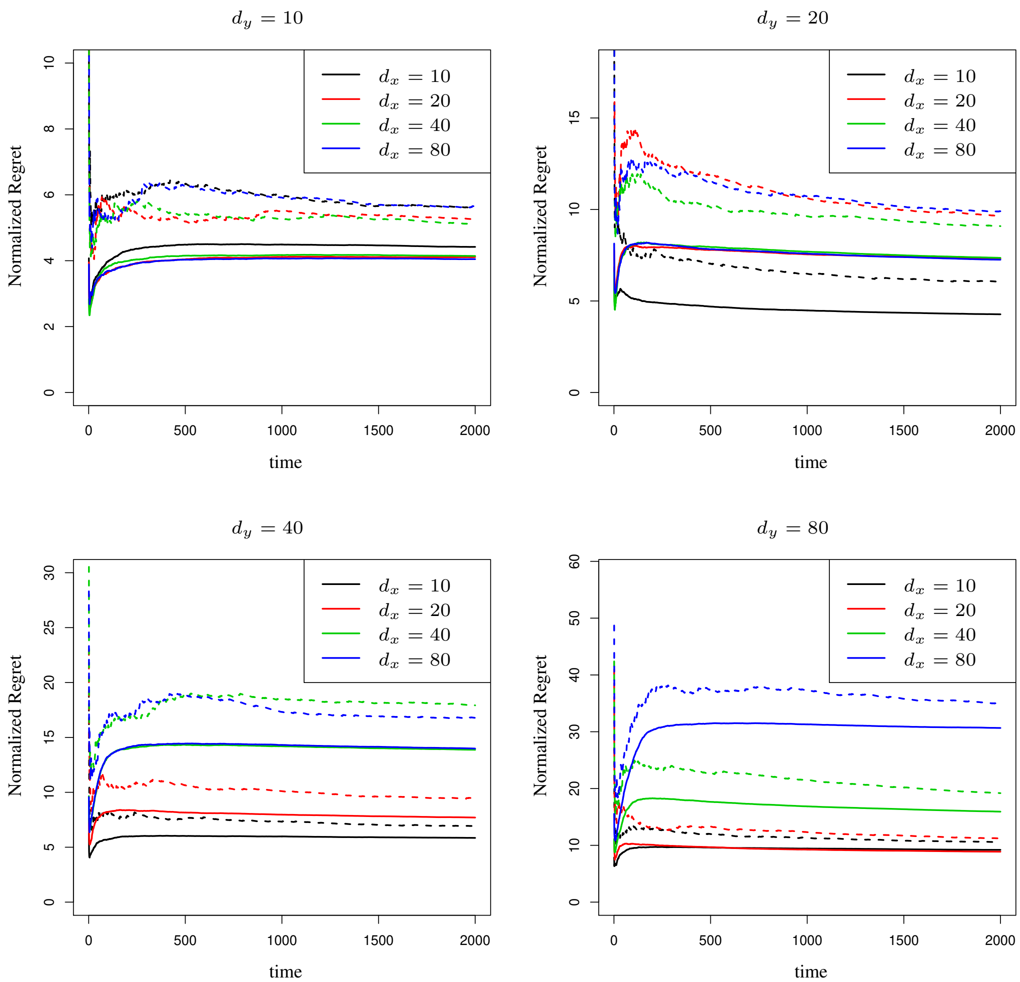

Figure 1. Plots of $\mathrm{Regret}(t)/(\log t)^2$ over time for the different dimensions of context at $N = 5$ and $d_y = 10, 20, 40, 80$. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kang, Y., Roh, C., Suh, S.-B., and Song, B. A lidar-based decision-making method for road boundary detection using multiple kalman filters. *IEEE Transactions on Industrial Electronics*, 59(11):4360–4368, 2012.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Lin, J.-W., Chen, C.-W., and Peng, C.-Y. Kalman filter decision systems for debris flow hazard assessment. *Natural hazards*, 60(3):1255–1266, 2012.

Modi, A. and Tewari, A. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 829–838. PMLR, 2020.

Nagrath, I. *Control systems engineering*. New Age International, 2006.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. 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, 2018.

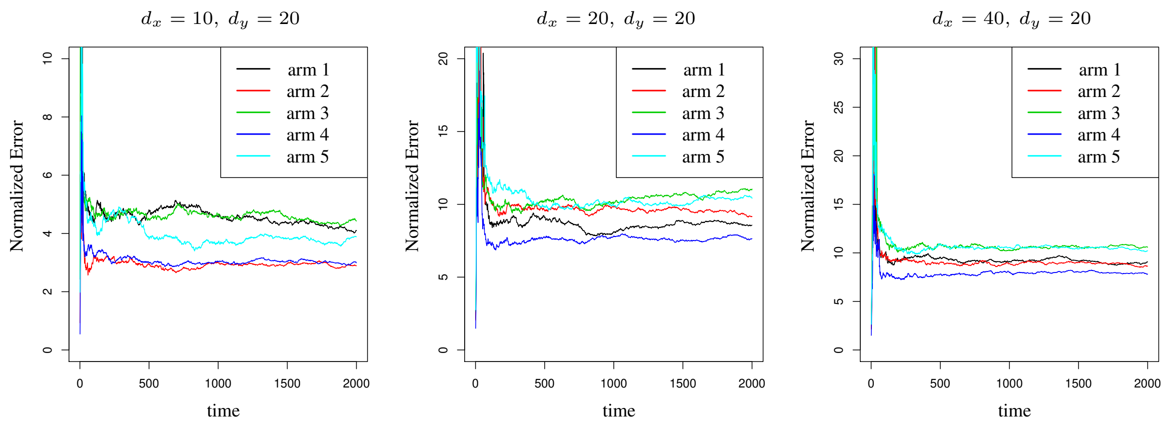

Figure 2. Plots of average normalized errors $\sqrt{t}\|\widehat{\eta}_i(t) - \eta_i\|_2$ over time at $N = 5$ and $d_y = 20$ for $d_x = 10,\ 20,\ 40$.

Nise, N. S. *Control systems engineering*. John Wiley & Sons, 2020.

Park, H. and Faradonbeh, M. K. S. Analysis of thompson sampling for partially observable contextual multi-armed bandits. *IEEE Control Systems Letters*, 6:2150–2155, 2021.

Park, H. and Faradonbeh, M. K. S. Worst-case performance of greedy policies in bandits with imperfect context observations. *arXiv preprint arXiv:2204.04773*, 2022.

Ren, Z. and Zhou, Z. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*, 2020.

Robinson, G. K. That blup is a good thing: the estimation of random effects. *Statistical science*, pp. 15–32, 1991.

Sbeity, H. and Younes, R. Review of optimization methods for cancer chemotherapy treatment planning. *Journal of Computer Science & Systems Biology*, 8(2):74, 2015.

Tennenholtz, G., Shalit, U., Mannor, S., and Efroni, Y. Bandits with partially observable confounded data. In *Conference on Uncertainty in Artificial Intelligence. PMLR*, 2021.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pp. 495–517. Springer, 2017.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Varatharajah, Y., Berry, B., Koyejo, S., and Iyer, R. A contextual-bandit-based approach for informed decision-making in clinical trials. *arXiv preprint arXiv:1809.00258*, 2018.

Yun, S.-Y., Nam, J. H., Mo, S., and Shin, J. Contextual multi-armed bandits under feature uncertainty. *arXiv preprint arXiv:1703.01347*, 2017.

A. Appendix

Proof of Theorem 4.1

Lemma A.1. *Given $y(t)$, the estimate $\widehat{x}(t)^\top J_i\mu$ has the mean $x(t)^\top J_i\mu$ 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 a BLUP, $\mathbb{E}[(\widehat{x}(t) - x(t))^\top J_i\mu] = 0$. In addition, using $\widehat{x}(t) = Dy(t) = D(Ax(t) + \xi(t))$,

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

Because $\|J_i\mu\| \le 1$, we can find $R_2 > 0$ such that

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

for any $J_i\mu \in \mathbb{R}^{dim(x(t))}$. Therefore, since $\xi(t)$ has a sub-Gaussian density, we get

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

□

Lemma A.2. *For any $\nu > 0$, we have*

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

where $R = R_1 + R_2$.

Proof. By (8),

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

which implies $\mathbb{E}[r_i(t) - \widehat{x}(t)^\top J_i\mu | y(t), a(t)] = 0$ because $\widehat{x}(t)^\top J_i\mu$ is a unbaised predictor of $x(t)^\top J_i\mu$. Due to $\mathrm{Var}(\xi(t)^\top \eta_i | y(t)) \le R_2^2$ by (19), we have

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

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

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

□

Lemma A.3. *For $J_i\mu$ such that $\mathbb{E}[r_i(t)|x(t)] = x(t)^\top J_i\mu$, let*

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

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

Proof.

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

Then,

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

□

Let f_μ be the normal density of μ with the mean zero and the positive covariance matrix $\lambda^{-1}I$. By Lemma 9 in (Abbasi-Yadkori et al., 2011), for $M_t = \mathbb{E}[M_t^\mu|\mathscr{F}_\infty]$, we have

$$P\left(\|S_\tau\|^2_{B(\tau)^{-1}} > 2\log\left(\frac{det(B(\tau))^{1/2}}{\delta det(\lambda I)^{1/2}}\right)\right) \le \mathbb{E}[M_\tau] \le \delta,$$

where $S_t = \sum_{\tau=1}^t J_{a(\tau)}^\top Dy(\tau)w_\tau$. By Theorem 1 in (Abbasi-Yadkori et al., 2011), we have

$$P\left(\|S_\tau\|_{B(\tau)^{-1}} > 2\log\left(\frac{det(B(\tau))}{\delta det(\lambda I)}\right),\ \forall\,\tau>0\right) \le \delta.$$

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

$$W_T = \left\{\max_{\{1\le\tau\le T\}}||y(\tau)||_\infty \le v_T(\delta)\right\}, \tag{20}$$

where $v_T(\delta) = (2\lambda_M\log(2d_yT/\delta))^{1/2} = O(\lambda_M^{\frac{1}{2}}\log(d_yT/\delta))$ and $\lambda_M = \lambda_{\max}(A\Sigma_XA^\top + \Sigma_Y)$.

Lemma A.4. *For the event W_T defined in (20), we have $\mathbb{P}(W_T) \ge 1-\delta$.*

Proof. Note that $y(t)$ has a sub-Gaussian density with the mean $Ax(t)$ and the covariance Σ_Y. Then, using the sub-Gaussian tail property, we have $\mathbb{P}\left(\|(A\Sigma_XA^\top+\Sigma_Y)^{-1/2}y(t)\|_\infty \ge \varepsilon\right) \le 2d_y\cdot e^{-\frac{\varepsilon^2}{2}}$. By simple calculations, we have

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

By plugging $(2\log(2d_yT/\delta))^{1/2}$ into ε, we have

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

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 A.4, we have

$$\|y(t)\| \leq \sqrt{d_y} v_T(\delta) := L = \mathcal{O}(\sqrt{\lambda_M d_y} \log(d_y T/\delta))$$

for all $1 \leq t \leq T$ with the at least probability $1-\delta$. Therefore, by Theorem 2 in (Abbasi-Yadkori et al., 2011), we have

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

Lemma A.5. *(Azuma Inequality, (Tropp, 2012)) Consider the sequence $\{X_k\}_{1\leq k\leq K}$ random variables adapted to some filtration $\{\mathscr{G}_k\}_{1\leq k\leq K}$, such that $\mathbb{E}[X_k|\mathscr{G}_{k-1}] = 0$. Assume that there is a deterministic sequence $\{c_k\}_{1\leq k\leq K}$ that satisfy $X_k^2 \leq c_k^2$, almost surely. Let $\sigma^2 = \sum_{1\leq k\leq K} c_k^2$. Then, for all $\varepsilon \geq 0$, it holds that*

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

Proof of Theorem 4.2

Proof. Let $\mathscr{F}_t = \sigma\{x(1), a(1), x(2), a(2), \ldots, x(t), a(t)\}$. Consider $V_t = D^\top J_{a(t)} y(t) y(t)^\top J_{a(t)}^\top D$ to identify the behavior of $B(t)$. Note that

$$\begin{aligned}
\mathbb{E}[J_{a(t)}^\top D y(t) y(t)^\top D^\top J_{a(t)} | \mathscr{F}_t] &= J_{a(t)}^\top D \mathrm{Var}(y(t)|\mathscr{F}_t) D^\top J_{a(t)} + J_{a(t)}^\top D A x(t) x(t)^\top A^\top D^\top J_{a(t)} \\
&\succeq \lambda_m J_{a(t)}^\top D D^\top J_{a(t)}
\end{aligned}$$

where $\lambda_{\min}(\Sigma_Y) = \lambda_m$. Let ν_{im} be the non-zero minimum eigenvalue of $J_i^\top D D^\top J_i$. Then, for all $t > 0$ and $z \in C(J_i^\top D)$ such that $\|z\| = 1$, it holds that

$$z^\top \left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau|\mathscr{F}_\tau]\right) z \geq z^\top \left(\sum_{\tau=1:a(\tau)=i}^{t-1} \mathbb{E}[V_\tau|\mathscr{F}_\tau]\right) z \geq \lambda_m \nu_{im} n_i(t). \tag{21}$$

Now, we focus on a high probability lower-bound for the smallest eigenvalue of $B(t)$. Let

$$X_\tau^i = (V_\tau - \mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}]) I(a(\tau) = i), \tag{22}$$

$$Y_\tau^i = \sum_{j=1}^{\tau} (V_j - \mathbb{E}[V_j|\mathscr{F}_{j-1}]) I(a(j) = i). \tag{23}$$

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

$$\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} X_\tau^i\right) z \leq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{8\ell_i(t) v_T^4(\delta)}\right).$$

By plugging $\ell_i(t)\varepsilon$ into ε, we have

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

for $\varepsilon \leq 0$. Now, using (21) and (22), we obtain

$$P\left(z^\top\left(\sum_{\tau=1}^{t-1}V(\tau)I(a(\tau)=i)\right)z\leq\ell_i(t)(\lambda_m\nu_{im}+\varepsilon)\right)\leq\exp\left(-\frac{\ell_i(t)\varepsilon^2}{8v_T^4(\delta)}\right),\tag{24}$$

where $-\lambda_m\nu_{im}\leq\varepsilon\leq 0$ is arbitrary. Indeed, using $B(t)\succeq\sum_{\tau=1}^{t-1}V(\tau)I(a(\tau)=i)$, on the event W_T defined in (20), for $-\lambda_m\nu_{im}\leq\varepsilon\leq 0$ we have

$$\mathbb{P}\left(z^\top B(t)z\leq\ell_i(t)(\lambda_m\nu_{im}+\varepsilon)\right)\leq\exp\left(-\frac{\ell_i(t)\varepsilon^2}{2v_T^4(\delta)}\right).\tag{25}$$

In other words, by equating $\exp\left(-\ell_i(t)\varepsilon^2/(2v_T(\delta)^4\right)$ to δ/T, (25) can be written as

$$z^\top B(t)z\geq\ell_i(t)\left(\lambda_m\nu_{im}-\sqrt{\frac{2v_T(\delta)^4}{\ell_i(t)}\log\frac{T}{\delta}}\right),\tag{26}$$

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

$$\lambda_{\min}\left(D^\top J_iB(t)J_i^\top D\right)\leq\nu_{iM}\ell_i(t)\left(\lambda_m\nu_{im}-\sqrt{\frac{2v_T(\delta)^4}{\ell_i(t)}\log\frac{T}{\delta}}\right).$$

Accordingly, we have

$$\lambda_{\max}\left(D^\top J_iB(t)^{-1}J_i^\top D\right)\leq\nu_{iM}\ell_i(t)^{-1}\left(\lambda_m\nu_{im}-\sqrt{\frac{2v_T(\delta)^4}{\ell_i(t)}\log\frac{T}{\delta}}\right)^{-1}.$$

If $\ell_i(t)\geq v_T(\delta)^4\log(T/\delta)/(2\lambda_m^2\nu_{im}^2)$, we have

$$\lambda_{\min}\left(D^\top J_iB(t)J_i^\top D\right)\geq\frac{\nu_{iM}\lambda_m\nu_{im}}{2}\ell_i(t),$$

and

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

□

A.1. Proof of Proposition 4.4

Proof. We assume that each arm has a positive probability of being the optimal arm at each time, and the event of being the optimal arm does not depend on the history. Let $A_i^*\subset\mathbb{R}^{d_y}$ be the event such that $\arg\max_j y(t)^\top\eta_j=i$, if $y(t)\in A_i^*$. The probability of being the optimal arm for the arm i is denoted as

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

and does not change over time. Note that, for $c>0$, $cy(t)\in A_i^*$, if $y(t)\in A_i^*$. A_i^* is a convex set, because $(sy_1+(1-s)y_2)^\top\eta_i=\max_j(sy_1+(1-s)y_2)^\top\eta_j$ for $y_1,\ y_2\in A_i$ and $c>0$. Thus, we take a subset $A_i\subseteq A_i^*$ and $\epsilon_i>0$ such that $\mathbb{P}(y(t)\in A_i)\geq p_i/2$ and $(y(t)/\|y(t)\|)^\top(\eta_i-\eta_j)>\epsilon_i$ for any j, if $y(t)\in A_i$. □

A.2. Proof of Theorem 4.5

Denote $A_{it}=\{y(t)\in A_i\}$. Then, we want to have a lower bound of the probability $\mathbb{P}(a(t)=i)$ to find a lower bound of $n_i(t)$ using

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

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

Since $y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) \leq \|y(t)\| \left(\lambda_{\max}(B_i(t)^{-1}) + \lambda_{\max}(B_j(t)^{-1})\right) \|\widehat{\mu}(t) - \mu_*\|_{B(t)}$, using Theorem 1 and 2, if $\ell_i(t) \geq v_T(\delta)^4 \log(T/\delta)/(2\lambda_m^2 \nu_{im}^2)$ and $\ell_j(t) \geq v_T(\delta)^4 \log(T/\delta)/(2\lambda_m^2 \nu_{jm}^2)$, we have

$$y(t)^\top(\widehat{\eta}_i(t) - \eta_i - \widehat{\eta}_j(t) + \eta_j) \leq \|y(t)\| \left(R\sqrt{d_\mu \log\left(1 + \frac{L^2 t}{\delta}\right)} + \lambda^{\frac{1}{2}} h\right) \left(\frac{\nu_{iM}\lambda_m \nu_{im}}{2} \ell_i(t)^{-1} + \frac{\nu_{jM}\lambda_m \nu_{jm}}{2} \ell_j(t)^{-1}\right).$$

Assume $\ell_i(t) > \frac{\nu_{iM}\lambda_m\nu_{im}\left(R\sqrt{d_\mu \log\left(1+\frac{L^2T}{\delta}\right)}+\lambda^{\frac{1}{2}}h\right)}{\epsilon_i}$ and $\ell_j(t) > \frac{\nu_{jM}\lambda_m\nu_{jm}\left(R\sqrt{d_\mu \log\left(1+\frac{L^2T}{\delta}\right)}+\lambda^{\frac{1}{2}}h\right)}{\epsilon_i}$, then we have

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

Accordingly, we have

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

If

$$\ell_i(t) > \max\left(v_T(\delta)^4 \log(T/\delta)/(2\lambda_m^2 \nu_{im}^2), \nu_{iM}\lambda_m\nu_{im}\left(R\sqrt{d_\mu \log\left(1 + \frac{L^2 T}{\delta}\right)} + \lambda^{\frac{1}{2}} h\right)\epsilon_i^{-1}\right) := m_{ii}(T) \tag{27}$$

and

$$\ell_j(t) > \max\left(v_T(\delta)^4 \log(T/\delta)/(2\lambda_m^2 \nu_{jm}^2), \nu_{jM}\lambda_m\nu_{jm}\left(R\sqrt{d_\mu \log\left(1 + \frac{L^2 T}{\delta}\right)} + \lambda^{\frac{1}{2}} h\right)\epsilon_i^{-1}\right) := m_{ij}(T), \tag{28}$$

we have

$$\mathbb{P}(y(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)|A_{it}, \mathscr{F}_{t-1}) \leq e^{-\frac{\ell_i(t)\epsilon_i^2}{8v^2}} + e^{-\frac{\ell_j(t)\epsilon_i^2}{8v^2}}.$$

Thus, if $\ell_i(t)$ and $\ell_j(t)$ satisfy (27) and (28), respectively, we have

$$\mathbb{P}(a(t) = i|A_{it}, \mathscr{F}_{t-1}) \geq 1 - \sum_{j\neq i}\left(e^{-\frac{\ell_i(t)\epsilon_i^2}{8v^2}} + e^{-\frac{\ell_j(t)\epsilon_i^2}{8v^2}}\right).$$

Therefore,

$$\mathbb{P}(a(t) = i|\mathscr{F}_{t-1}) \geq \mathbb{P}(a(t) = i|A_{it}, \mathscr{F}_{t-1})\mathbb{P}(A_{it}) \geq \frac{p_i}{2}\left(1 - \sum_{j\neq i}\left(e^{-\frac{\ell_i(t)\epsilon_i^2}{8v^2}} + e^{-\frac{\ell_j(t)\epsilon_i^2}{8v^2}}\right)\right).$$

The results above can be applied to all two cases defined in Section 2. Now, we focus on regret analysis. We investigate regret bounds for two settings discussed in Section 2. First, we consider setting 1, where all arms share the parameter.

Proof of Theorem 4.6

By Theorem 4.1, for all $1 \le t \le T$, we have

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

Suppose that $D^\top J_i$ has the singular value decomposition $U_i \Sigma_i V_i^\top$. Using $(V_i \Sigma_i^- U_i^\top) D^\top J_i \preceq I$, we get

$$\|B(t)^{\frac{1}{2}}(V_i \Sigma_i^- U_i^\top) D^\top J_i(\widehat{\mu}(t) - \mu_*)\| \le \|B(t)^{\frac{1}{2}}(\widehat{\mu}(t) - \mu_*)\|. \tag{29}$$

Accordingly,

$$\lambda_{mnz}(((V_i \Sigma_i^- U_i^\top)^\top B(t)(V_i \Sigma_i^- U_i^\top))^{\frac{1}{2}} \|D^\top J_i(\widehat{\mu}(t) - \mu_*)\| \le \|B(t)^{\frac{1}{2}}(V_i \Sigma_i^- U_i^\top) D^\top J_i(\widehat{\mu}(t) - \mu_*)\|, \tag{30}$$

where $\lambda_{mnz}(M)$ is the smallest non-zero eigenvalue of M for a square matrix M. Finally, by putting together (29), (30) and Theorem 4.2, we have

$$\begin{aligned}
\|\widehat{\eta}_i(t) - \eta_i\| &\le \lambda_{\max}(D^\top J_i B(t)^{-1} J_i^\top D)^{\frac{1}{2}} R\left(\sqrt{d_\mu \log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} h\right) \\
&\le \nu_{iM}^{\frac{1}{2}} \ell_i(t)^{-\frac{1}{2}} R\left(\lambda_m \nu_{im} - \sqrt{\frac{2v_T(\delta)^4}{\ell_i(t)} \log\frac{T}{\delta}}\right)^{-\frac{1}{2}} \left(\sqrt{d_\mu \log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} h\right)
\end{aligned}$$

If $\ell_i(t) > 8(v_T(\delta)^4/(\lambda_m^2 \nu_{im}^2)) \log(T/\delta)$, with $\ell_i(t) = t$ for all i under the SPMC assumption, we have

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

Lemma A.6. *Let $\widetilde{\eta}_i(t)$ be a sample in (14). Then, if $t > 8(v_T(\delta)^4/(\lambda_m^2 \nu_{im}^2)) \log(T/\delta)$, with probability at least $1 - \delta$, for all $i \in [N]$ and $0 < t \le T$, we have*

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

Proof. Using $\mathbb{P}(\|\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)\| > \epsilon) \le \mathbb{P}\left(\sqrt{d_y} Z > \epsilon\right)$, where $Z \sim \mathcal{N}(0, v^2 \max(B_i(t))^{-1})$, we have

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

By putting $2 \cdot e^{-\frac{\epsilon^2}{2v^2 \max(B_i(t)^{-1})}} = \frac{\delta}{TN}$, we have

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

If $t > 8(v_T(\delta)^4/(\lambda_m^2 \nu_{im}^2)) \log(T/\delta)$, we have

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

Therefore, by Theorem 4.8,

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

$\square$

A.3. Proof of Theorem 4.7

Let $reg(t) = (y(t)/\|y(t)\|)^\top(\eta_{a^*(t)}(t) - \eta_{a(t)}(t))$. Then,

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

since $\|y(t)\| \le v_T(\delta)$ for all $t \in [T]$. By Lemma **??**, if $t > 8(v_T(\delta)^4/(\lambda_m^2\nu_{im}^2))\log(T/\delta)$, with a probability at least $1-\delta$, we have

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

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

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

and $H_t = \sum_{\tau=1}^{t} G_\tau$. Since $\mathbb{E}[G_\tau|\mathscr{F}_{\tau-1}] = 0$, the above sequences $\{G_\tau\}_{\tau\ge0}$ and $\{H_\tau\}_{\tau\ge0}$ are a martingale difference sequence and a martingale with respect to the filtration $\{\mathscr{F}_\tau\}_{1\le\tau\le T}$, respectively. Let $c_\tau = 2\tau^{-1/2}$. Since $\sum_{\tau=1}^{T}|G_\tau| \le \sum_{\tau=2}^{T} c_\tau^2 \le 4\log T$, by Lemma A.5, we have

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

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

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}}I(a^*(t) \neq a(t)) \le \sqrt{32\log T\log\delta^{-1}} + \sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbb{P}(a^*(\tau) \neq a(\tau)|\mathscr{F}_{\tau-1}^*). \tag{31}$$

Now, we proceed to the upper bound of the second term on the right side in (31).

Assumption A.7. The support of standardized observation $y(t)/\|y(t)\|$ is a subset of a unit sphere with the dimension d_y. The density of $y(t)/\|y(t)\|$ is bounded by a constant C,

$$P(y(t)/\|y(t)\| = y) < C.$$

Accordingly, $d_{ij}(t) = (y(t)/\|y(t)\|)^\top(\eta_i - \eta_j)|(a^*(t) = i)$ has a density f_{ij} bounded by a constant, $c_{ij} > 0$.

Let $A_{it}^* = \{y(t) \in A_i\}$.

$$\begin{aligned}&\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^*) = \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \eta_j - \widetilde{\eta}_i(t) - \eta_i) > y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A_{it}^*)\\ \le\ &\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \eta_j) > 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \eta_i) > 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*)\\ \le\ &\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A_{it}^*)\\ +\ &\mathbb{P}(y(t)^\top(\widehat{\eta}_j(t) - \eta_j) > 0.25y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i) > 0.25y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*)\end{aligned}$$

By Theorem 4.6 and Assumption 1, if $t > 8(v_T(\delta)^4/(\lambda_m^2\nu_{im}^2))\log(T/\delta)$, we have

$$\mathbb{P}(y(t)^\top(\widehat{\eta}_i(t) - \eta_i) > 0.25y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*) \le \mathbb{P}\left(\frac{2h_i(T)}{\sqrt{t}} > y(t)^\top/\|y(t)\|(\eta_i - \eta_j)\,\middle|\,\mathscr{F}_{t-1}, A_{it}^*\right) \le \frac{2h_i(T)c_{ij}}{\sqrt{t}}$$

$$\mathbb{P}(y(t)^\top(\widehat{\eta}_j(t) - \eta_j) > 0.25y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^*) \le \mathbb{P}\left(\frac{2h_j(T)}{\sqrt{t}} > y(t)^\top/\|y(t)\|(\eta_i - \eta_j)\,\middle|\,\mathscr{F}_{t-1}, A_{it}^*\right) \le \frac{2h_j(T)c_{ij}}{\sqrt{t}}$$

where

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

Because

$$\mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A^*_{it}, y(t)) \leq e^{-\frac{t(y(t)^\top(\eta_i-\eta_j))^2}{32\|y(t)\|^2 v^2}}$$
$$\mathbb{P}\left(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A^*_{it}, y(t)\right) \leq e^{-\frac{t(y(t)^\top(\eta_i-\eta_j))^2}{32\|y(t)\|^2 v^2}},$$

based on Assumption 1, we have

$$\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |\mathscr{F}_{t-1}, A^*_{it}) \leq E[e^{-\frac{t(y(t)^\top(\eta_i-\eta_j))^2}{8\|y(t)\|^2 v^2}}|\mathscr{F}_{t-1}, A^*_{it}]$$
$$= \int_0^{\|\eta_i-\eta_j\|} e^{-\frac{tz^2}{8v^2}} f_{ij}(z)dz \leq \frac{2c_{ij}v}{\sqrt{t}}$$

Accordingly, we have

$$\begin{aligned}\mathbb{P}(a^*(t) \neq a(t)|\mathscr{F}^*_{t-1}) &\leq \sum_{i=1}^N\sum_{j=1}^N P(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A^*_{it})p_i \\ \leq \quad & \sum_{i=1}^N\sum_{j=1}^N p_i\left(\frac{4c_{ij}v}{\sqrt{t}} + \frac{4h_j(T)c_{ij}}{\sqrt{t}}\right) = \frac{4}{\sqrt{t}}\sum_{i=1}^N\sum_{j=1}^N p_i c_{ij}\left(v + h_j(T)\right) = \frac{4Nc_M}{\sqrt{t}}.\end{aligned}$$

where $c_M = \max_{ij} c_{ij}\left(v + h_j(T)\right) = \mathcal{O}(\sqrt{d_\mu}\log T)$. Thus, we have

$$\sum_{t=1}^T \frac{1}{\sqrt{t}}\mathbb{P}(a^*(t) \neq a(t)|\mathscr{F}^*_{t-1}) \leq 4Nc_M\sum_{t=1}^T\frac{1}{t} \leq 4Nc_M \log T.$$

By (31), with a probability at least $1-\delta$, we have

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} I(a^*(t) \neq a(t)) \leq \sqrt{32\log T \log\delta^{-1}} + 4Nc_M\log T.$$

Therefore,

$$\begin{aligned}\mathrm{Regret}(T) \quad &\leq \quad R\max_i(\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m\nu_{im}})\left(v\sqrt{2d_y\log\frac{2TN}{\delta}} + \sqrt{d_\mu\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}}h\right)\left(\sqrt{32\log T\log\delta^{-1}} + 4Nc_M\log T\right) \\ &= \quad \mathcal{O}\left(N(d_\mu + \sqrt{d_\mu d_y})\mathrm{polylog}\left(\frac{TNd_y}{\delta}\right)\right).\end{aligned}$$

A.4. Proof of Theorem 4.8

Lemma A.8. *Under the general assumption, with a probability at least $1-\delta$, the algorithm 1 guarantees*

$$n_i(t) > \frac{p_i}{2}\left(t - \sum_{j\neq i}(m'_{ii}(T) + m'_{ij}(T)) - (N-1)/T\right) - \sqrt{2t\log(2/\delta)},$$

where $m'_{ii}(T) = \max(m_{ii}(T),\ 16(v^2/\epsilon_i^2)\log T)$ and $m'_{ij}(T) = \max(m_{ij}(T),\ 16(v^2/\epsilon_i^2)\log T)$.

Proof. By Theorem 3, if $\ell_i(t) = n_i(t) > m_{ii}(T)$ and $\ell_j(t) = n_j(t) > m_{ij}(T)$,

$$\mathbb{P}(a(t) = i|F_{t-1}) \geq \mathbb{P}(a(t) = i|F_{t-1}, A_{it})\mathbb{P}(A_{it}) \geq \frac{p_i}{2}\left(1 - \sum_{j\neq i}\left(e^{-\frac{\ell_i(t)\epsilon_i^2}{8v^2}} + e^{-\frac{\ell_j(t)\epsilon_i^2}{8v^2}}\right)\right),$$

where $p_i = \mathbb{P}(a^*(t) = i)$. If $\ell_i(t) \geq m'_{ii}(T) := \max(m_{ii}(T),\ 16(v^2/\epsilon_i^2)\log T)$, we have $\exp\left(-(\ell_i(t)\epsilon_i^2)/(8v^2)\right) \leq T^{-2}$. Similarly, if $\ell_j(t) \geq m'_{ij}(T) := \max(m_{ij}(T),\ 16(v^2/\epsilon_i^2)\log T)$, we have $\exp\left(-(\ell_j(t)\epsilon_i^2)/(8v^2)\right) \leq T^{-2}$. Since $I(a(t) = i) - (p_i/2)\left(1 - \sum_{j\neq i}\mathbb{P}(a(t) = j|A_{it})\right)$ is a submartingale difference,

$$\begin{aligned}\sum_{\tau=1}^{t} P(a(\tau) = i|F_{\tau-1}) &\geq \frac{p_i}{2}\left(t - \sum_{\tau=1}^{t}\sum_{j\neq i} P\left(y(\tau)^\top(\widetilde{\eta}_j(\tau) - \widetilde{\eta}_i(\tau)) > \epsilon_i \middle| A_{i\tau}, F_{\tau-1}\right)\right) \\ &\geq \frac{p_i}{2}\left(t - \sum_{j\neq i}(m'_{ii}(T) + m'_{ij}(T)) - (N-1)/T\right).\end{aligned}$$

$$P\left(n_i(t) - \sum_{\tau=1}^{t} P(a(\tau) = i|F_{\tau-1}) < -\epsilon\right) \leq e^{-\frac{\epsilon^2}{T}}.$$

With a probability of at least $1 - \delta$,

$$n_i(t) > \frac{p_i}{2}\left(t - \sum_{j\neq i}(m'_{ii}(T) + m'_{ij}(T)) - (N-1)/T\right) - \sqrt{2t\log(2/\delta)}.$$

□

Now we are ready to prove Theorem 6.

Proof. The following inequality

$$\frac{p_i}{2}\left(t - \sum_{j\neq i}(m'_{ii}(T) + m'_{ij}(T)) - (N-1)/T\right) - \sqrt{2t\log(2/\delta)} > \frac{p_i}{4}t,$$

is satisfied, if $t > m''_i(T) = 2(a_{i1} + (4/p_i)a_{i2}^2) + 2\sqrt{(a_{i1} + (4/p_i)a_{i2}^2)^2 - a_{i1}^2}$, where $a_{i1} = \sum_{j\neq i}(m'_{ii}(T) + m'_{ij}(T)) + (N-1)/T$ and $a_{i2} = \sqrt{2\log(2/\delta)}$ based on the quadratic formula. By Lemma A.8, with a probability at least $1-\delta$, $n_i(t) > (p_i t)/4$, if $t > m''_i(T)$. Similarly to Theorem 4, we have

$$\|\widehat{\eta}_i(t) - \eta_i\|_2 \leq \nu_{iM}^{\frac{1}{2}}(p_i t/4)^{-\frac{1}{2}} R\left(\lambda_m\nu_{im} - \sqrt{\frac{2v_T(\delta)^4}{p_i t/4}\log\frac{T}{\delta}}\right)^{-\frac{1}{2}}\left(\sqrt{d_\mu\log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}}h\right).$$

Thus, if $t \geq (32/p_i)(v_T(\delta)^4/(\lambda_m^2\nu_{im}^2))\log(T/\delta)$,

$$\|\widehat{\eta}_i(t) - \eta_i\|_2 \leq \frac{R\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m \nu_{im}}}{\sqrt{p_i t}} \left(\sqrt{d_\mu \log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} h \right).$$

□

Theorem A.9. *Assume that Algorithm 1 is used in a bandit the MPMC assumption. Then, with probability at least* $1-\delta$, $\mathrm{Regret}(T)$ *is of the order*

$$\mathrm{Regret}(T) = \mathcal{O}\left(\left(\max_i p_i^{-1}\right) N \sqrt{d_y d_\mu} poly\left(\log\left(\frac{TNd_y}{\delta}\right)\right)\right).$$

Proof. The regret can be decomposed as

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

since $\|y(t)\| \leq v_T(\delta)$ for all $t \in [T]$.

$$\begin{aligned} & (\|\widetilde{\eta}_{a^*(t)}(t) - \eta_{a^*(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\|) I(a^*(t) \neq a(t)) \\ = & \sum_{j=1}^N (\|\widetilde{\eta}_{a^*(t)}(t) - \eta_{a^*(t)}\| + \|\widetilde{\eta}_j(t) - \eta_j\|) I(a^*(t) \neq a(t), a(t) = j) \end{aligned}$$

By Theorem **??**, if $t > m_i''(T)$, we have

$$\|\widehat{\eta}_i(t) - \eta_i\|_2 \leq \frac{R\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m \nu_{im}}}{2\sqrt{p_i t}} \left(\sqrt{d_\mu \log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} h \right).$$

$$\begin{aligned} & \mathbb{P}(y(t)^\top (\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0 | F_{t-1}, A_{it}^*) = \mathbb{P}(y(t)^\top (\widetilde{\eta}_j(t) - \eta_j - \widetilde{\eta}_i(t) - \eta_i) > y(t)^\top (\eta_i - \eta_j) > | F_{t-1}, A_{it}^*) \\ \leq & \mathbb{P}(y(t)^\top (\widetilde{\eta}_j(t) - \eta_j) > 0.5 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top (\widetilde{\eta}_i(t) - \eta_i) > 0.5 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) \\ \leq & \mathbb{P}(y(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25 y(t)^\top (\eta_i - \eta_j) > | F_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25 y(t)^\top (\eta_i - \eta_j) > | F_{t-1}, A_{it}^*) \\ + & \mathbb{P}(y(t)^\top (\widehat{\eta}_j(t) - \eta_j) > 0.25 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) + \mathbb{P}(y(t)^\top (\widehat{\eta}_i(t) - \eta_i) > 0.25 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) \end{aligned}$$

By Theorem 4.6 and Assumption 1, if $t > 8(v_T(\delta)^4/(\lambda_m^2 \nu_{im}^2)) \log(T/\delta)$, we have

$$\mathbb{P}(y(t)^\top (\widehat{\eta}_i(t) - \eta_i) > 0.25 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) \leq \mathbb{P}\left(\frac{h_i(T)}{\sqrt{p_i t}} > y(t)^\top / \|y(t)\| (\eta_i - \eta_j) \middle| F_{t-1}, A_{it}^* \right) \leq \frac{h_i(T) c_{ij}}{\sqrt{p_i t}}$$

$$\mathbb{P}(y(t)^\top (\widehat{\eta}_j(t) - \eta_j) > 0.25 y(t)^\top (\eta_i - \eta_j) | F_{t-1}, A_{it}^*) \leq \mathbb{P}\left(\frac{h_j(T)}{\sqrt{p_j t}} > y(t)^\top / \|y(t)\| (\eta_i - \eta_j) \middle| F_{t-1}, A_{it}^* \right) \leq \frac{h_j(T) c_{ij}}{\sqrt{p_j t}}$$

where

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

Because

$$\mathbb{P}(y(t)^\top(\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |F_{t-1}, A^*_{it}, y(t)) \le e^{-\frac{tp_i(y(t)^\top(\eta_i-\eta_j))^2}{128\|y(t)\|^2 v^2}}$$

$$\mathbb{P}\left(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |F_{t-1}, A^*_{it}, y(t)\right) \le e^{-\frac{tp_j(y(t)^\top(\eta_i-\eta_j))^2}{128\|y(t)\|^2 v^2}},$$

based on Assumption 1, we have

$$\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > 0.25 y(t)^\top(\eta_i - \eta_j) > |F_{t-1}, A^*_{it}) \le E[e^{-\frac{tp_j(y(t)^\top(\eta_i-\eta_j))^2}{128\|y(t)\|^2 v^2}}|F_{t-1}, A^*_{it}]$$
$$= \int_0^{\|\eta_i-\eta_j\|} e^{-\frac{tp_j z^2}{128 v^2}} f_{ij}(z)dz \le \frac{16 c_{ij} v}{\sqrt{p_j t}}$$

$$\begin{aligned} P(a^*(t) \neq a(t)|F_{t-1}) &\le \sum_{i=1}^N p_i \sum_{j=1}^N P(a(t) = j|F_{t-1}, A^*_{it}) \\ &\le \sum_{i=1}^N p_i \sum_{j=1}^N \left(\frac{h_i(T)c_{ij}}{\sqrt{p_i t}} + \frac{h_j(T)c_{ij}}{\sqrt{p_j t}} + \frac{16c_{ij}v}{\sqrt{p_i t}} + \frac{16c_{ij}v}{\sqrt{p_j t}}\right) \le \frac{2Nc_M}{\sqrt{t}} \end{aligned}$$

where $c_M = \max_{i,j} p_i^{-0.5}(h_i(T) + 16v)c_{ij} = O(\max_i p_i^{-0.5}\sqrt{d_\mu}\log T)$.

Since $t^{-1/2} I(a^*(t) \neq a(t)) - t^{-1/2} P(a^*(t) \neq a(t)|F_{t-1})$ is a martingale difference w.r.t F_t, by Azuma, with a probability at least $1-\delta$, we have

$$\sum_{t=1}^T t^{-1/2} I(a^*(t) \neq a(t)) \le \sum_{t=1}^T t^{-1/2} P(a^*(t) \neq a(t)|F_{t-1}) + \sqrt{64 \log T \log \delta^{-1}}$$

Thus, we have

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} I(a^*(t) \neq a(t)) \le \sqrt{64 \log T \log \delta^{-1}} + 2Nc_M \log T.$$

Therefore,

$$\begin{aligned} \mathrm{Regret}(T) &\le R\max_i(\nu_{iM}^{\frac{1}{2}}\sqrt{\lambda_m\nu_{im}})\left(v\sqrt{2d_y \log\frac{2TN}{\delta}} + \sqrt{d_\mu \log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}}h\right)\left(\sqrt{64\log T \log \delta^{-1}} + 2Nc_M \log T\right) \\ &= \mathcal{O}\left(\max_{i,j} p_i^{-0.5} N(d_\mu + \sqrt{d_y d_\mu})\text{polylog}\left(\frac{TNd_y}{\delta}\right)\right). \end{aligned}$$

$\square$