Sequentially Adaptive Experimentation for Learning Optimal Options subject to Unobserved Contexts

Hongju Park HP97161@UGA.EDU
Department of Statistics, University of Georgia
Athens, GA 30602, USA

Mohamad Kazem Shirani Faradonbeh MOHAMADKSF@SMU.EDU
Department of Mathematics, Southern Methodist University
Dallas, TX 75205, USA

Abstract

Contextual bandits constitute a classical framework for interactive learning of best decisions subject to context information. In this setting, the goal is to sequentially learn arms of highest reward subject to the contextual information, while the unknown reward parameters of each arm need to be learned by experimenting it. Accordingly, a fundamental problem is that of balancing such experimentation (i.e., pulling different arms to learn the parameters), versus sticking with the best arm learned so far, in order to maximize rewards. 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 arms based on observations, which are noisy linear functions of the unobserved context vectors. Our theoretical analysis shows that adaptive experiments based on samples from the posterior distribution efficiently learn optimal arms. Specifically, we establish regret bounds that grow logarithmically with time. Extensive simulations for real-world data are presented as well to illustrate this efficacy.

Keywords: Contextual Bandits, Partial Observability, Posterior Sampling, Regret Bounds

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. However, comprehensive studies are currently missing for imperfectly observed contexts, which 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. 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 (Gottesman et al., 2019). Accordingly, partial observation models are studied in canonical settings such as linear systems

(Kargin et al., 2023), bandit monitoring (Kirschner et al., 2020; Lattimore, 2022; Tsuchiya et al., 2023), and Markov decision processes (Bensoussan, 2004; Krishnamurthy and Wahlberg, 2009). The above have recently motivated some work on contextual bandit policies with partially observed contexts (Park and Faradonbeh, 2021, 2022a,b).

In this paper, we study algorithms to balance exploration and exploitation for contextual bandits with partially observable contexts via Thompson sampling. We consider a linear reward, which is the common bandit setting, 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* (Agrawal and Goyal, 2013; Bastani and Bayati, 2020), or *shared* across all arms (Dani et al., 2008; Abeille and Lazaric, 2017). We focus more on challenging one of the former and establish a worst-case (poly-)logarithmic regret bounds for both settings with the supplement using numerical experiments.

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

$$r_i(t) = x_i(t)^\top \mu_i + \varepsilon_i(t), \tag{1}$$

where $x_i(t)$ is the **unobserved** d_x dimensional stochastic context of the ith arm generated independently at time t with $\mathbb{E}[x_i(t)] = \mathbf{0}_{d_x}$ and $\mathrm{Cov}(x_i(t)) = \Sigma_x$, μ_i is the arm-specific reward parameter of the ith arm in $\mathbb{R}^{d_y}$, and $\varepsilon_i(t)$ is the reward R_1-subgaussian noise.

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

$$y_i(t) = Ax_i(t) + \xi_i(t), \tag{2}$$

where A is a $d_y \times d_x$ sensing matrix, and $\xi_i(t)$ is the sensing (or measurement) noise, its covariance matrix being denoted by Σ_y. We assume that each element of $\xi_i(t)$ is subgaussian as well. 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 $\{y_{a(\tau)}(\tau)\}_{1\leq\tau\leq t-1}$, as well as the current one $\{y_i(t)\}_{i\in[N]}$. 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.

Now, we look into the optimal arm identification. Note that the optimal arm i maximizes $x_i(t)^\top \mu_i$ based on (1). Under the linear structure (2) with unknown stochastic contexts, the Best Linear Unbiased Prediction (BLUP) (Harville, 1976; Robinson, 1991) is the best approximate value of context $x(t)$ in terms of the unbiasedness and minimal variance, denoted by $\widehat{x}_i(t) := Dy_i(t)$, where $D = (A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1} A^\top \Sigma_y^{-1}$. Accordingly, the prediction of $x_i(t)^\top \mu_i$ is $\widehat{x}_i(t)^\top \mu_i$.

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

$$r_i(t) = y_i(t)^\top D^\top \mu_i + \zeta_i(t), \tag{3}$$

where $\zeta_i(t) = (x_i(t)^\top \mu_i - y_i(t)^\top D^\top \mu_i) + \varepsilon_i(t)$ is a noise centered at 0. $\zeta_i(t)$ is independent of others because of the independence of the prediction error, $x_i(t)^\top \mu_i - y_i(t)^\top D^\top \mu_i$. Here, μ_i is not estimable based on the equation (3), since the space spanned by $\{Dy_i(\tau)\}_{\tau=1:a(\tau)=i}^{t}$ does not

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

Set $B_i(1) = I_{d_y}$, $\widehat{\eta}_i(1) = \mathbf{0}_{d_y}$ for $i = 1, 2, \ldots, N$
for $t = 1, 2, \ldots,$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Sample $\widetilde{\eta}_i(t)$ from $\mathcal{N}(\widehat{\eta}_i(t), v^2 B_i^{-1}(t))$
  end for
  Select arm $a(t) = \text{argmax}_{i \in [N]} y_i(t)^\top \widetilde{\eta}_i(t)$
  Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_{a(t)} + \varepsilon_{a(t)}(t)$
  Update $B_i(t+1)$ and $\widehat{\eta}_i(t+1)$ by (7), (8) and (9) for $i = 1, 2, \ldots, N$
end for

generally include μ_i, if $d_y < d_x$. Thus, instead of μ_i, we estimate $D^\top \mu_i$ defined as the transformed parameter of the arm i, denoted by

$$\eta_i := D^\top \mu_i. \tag{4}$$

Thus, using (3) and (4), we get

$$r_i(t) = y_i(t)^\top \eta_i + \zeta_i(t). \tag{5}$$

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

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 above, regret is expressed as $\text{Regret}(T) = \sum_{t=1}^{T} \left(y_{a^\star(t)}(t)^\top \eta_{a^\star(t)} - y_{a(t)}(t)^\top \eta_{a(t)} \right)$, where $a(t)$ is the chosen arm by the decision maker at time t.

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: $r_i(t) = y_i(t)^\top \eta_i + \psi_i(t)$, where $\psi_i(t)$ has the normal distribution with the mean 0 and variance $v^2 = R_1^2$. In the beginning, the decision-maker starts with the initial value $\widehat{\eta}_i(1) = \mathbf{0}_{d_y}$ and $B_i(1) = I_{d_y}$ for all $i \in [N]$, which are the mean and (unscaled) covariance matrix of a prior distribution of η_i, respectively. The posterior distribution of η_i at time t is given as $\mathcal{N}(\widehat{\eta}_i(t), v^2 B_i(t)^{-1})$.

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

$$\widetilde{\eta}_i(t) \sim \mathcal{N}(\widehat{\eta}_i(t), v^2 B_i(t)^{-1}), \tag{6}$$

for $i = 1, 2 \ldots, N$. Accordingly, the decision-maker pulls the arm $a(t)$ such that $a(t) = \underset{1 \le i \le N}{\text{argmax}}\ y_i(t)^\top \widetilde{\eta}_i(t)$. Then, once the decision-maker gains the reward of the chosen arm $a(t)$, it

can update $\widehat{\eta}_i(t)$ and $B_i(t)$ based on the recursions below:

$$B_i(t+1) = B_i(t) + y_i(t)^\top y_i(t), \tag{7}$$
$$\widehat{\eta}_i(t+1) = B_i(t+1)^{-1}\left(B_i(t)\widehat{\eta}_i(t) + y_i(t)r_i(t)\right), \tag{8}$$

if $i = a(t)$,

$$B_i(t+1) = B_i(t), \;\; \widehat{\eta}_i(t+1) = \widehat{\eta}_i(t), \tag{9}$$

otherwise. The pseudocode for the algorithm is provided in Algorithm 1.

4. Theoretical Performance Analyses

In this section, we establish the theoretical result of Algorithm 1 for partially observable contextual bandits with arm-specific parameters. The following result provides a high probability regret upper bound for Algorithm 1.

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

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

The above theorem demonstrates that the regret upper bound scales at most $\log^4 T$ with respect to the time. A poly-logarithmic regret bounds are unprecedented to the best of our knowledge. Specifically, 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 (Agrawal and Goyal, 2013) and the greedy first algorithm for the stochastic contextual bandits (Bastani et al., 2021).

5. Numerical Experiments

5.1 Simulation Experiments

In this 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 presents the regret, normalized by $(\log t)^2$, which represents the actual regret growth because the $(\log t)^2$ term in the regret bound of Theorem 1 is attributed to the minimum sample size.

Moving on, Figure 2 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. In Figure 2, 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.

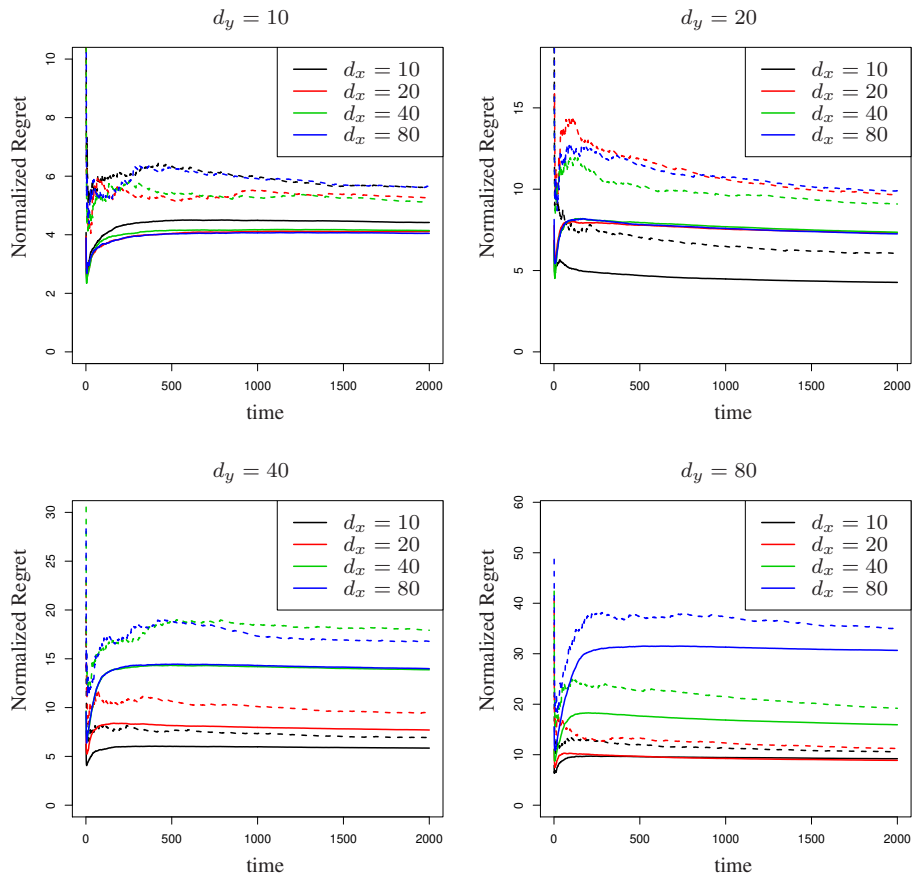

Figure 1: Plots of $\mathrm{Regret}(t)/(\log t)^2$ over time for the different dimensions of context at $N = 5$ and $d_y = 10, 20, 40, 80$. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

5.2 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 analyzed in previous studies by Bastani and Bayati (2020); Bietti et al. (2021) via contextual bandits with arm-specific parameters and shared context. These datasets involve 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 (2) with a sensing matrix A consisting of 0 and 1 only. We

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

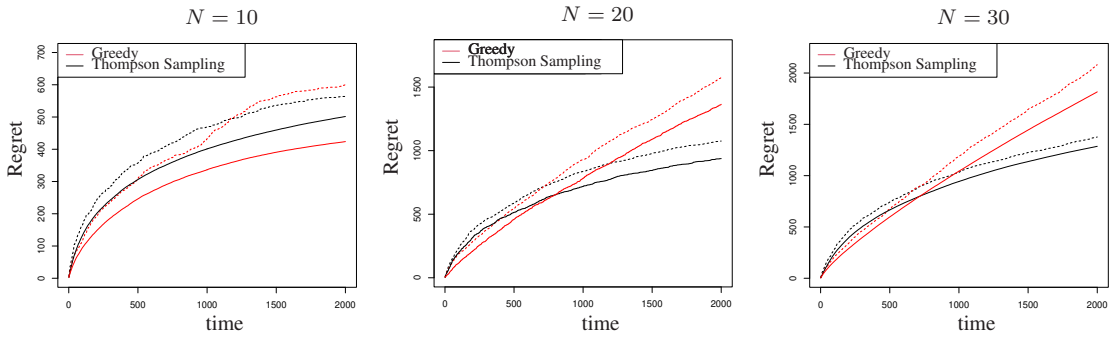

Figure 2: Plots of regrets over time with the different number of arms $N = 10,\ 20,\ 30$ for Thomson sampling versus the Greedy algorithm. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

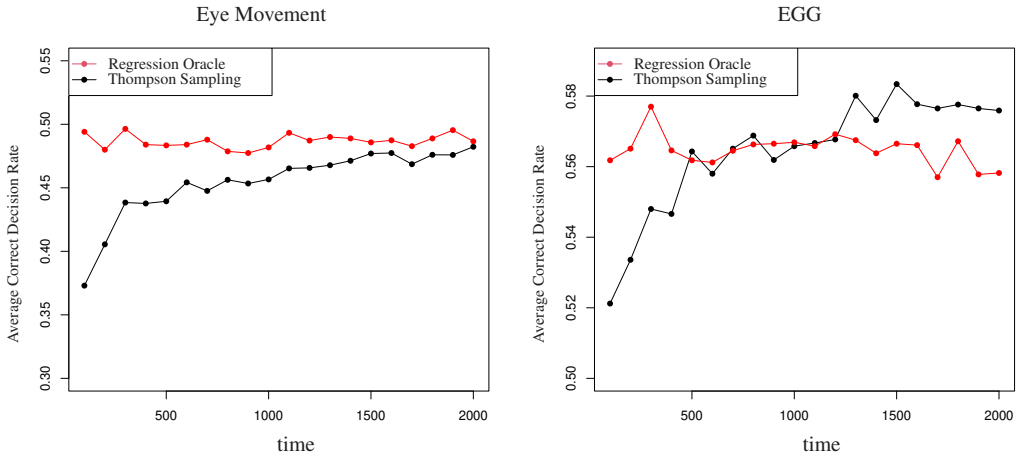

Figure 3: Plots of average correction decision rates of the regression oracle and Thompson sampling for Eye movement (left) and EGG dataset (right).

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 3 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 biased estimation with complex reasons such as non-linearity in the data and arm sampling bias incurred by actions with higher optimal probabilities.

6. Concluding Remarks

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. Lastly, we showed that Thompson sampling guarantees regret bounds scaling poly-logarithmically.

References

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

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

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

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

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

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

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

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

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

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

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.

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

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.

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

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

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

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.

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

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, 2022b.

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

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.

Contents

Appendices

The appendices are organized as follows. First, Appendix A provides the notations used in this paper. Second, Appendix B and C explain the necessary assumptions for the theoretical analysis in Section 4 and the relationship between the suggested framework and original contextual bandits, respectively. Then, Appendix D presents the theoretical results for the general model, with the comprehensive proofs found in Appendix H, I, and J. Following this, Appendix E provides the worst-case regret upper bounds for the model with a shared parameter, accompanied by its proof detailed in Appendix L. Next, Appendix G illustrates the estimation accuracy of transformed parameters. Lastly, Appendix establishes the square-root estimation accuracy of parameters supplemented by the proof in Appendix M followed by the complete proof for Theorem1 in Appendix N.

Appendix 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 use the notation $\|v\| = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$ for the ℓ_2 norm. Finally, $P_{C(M)}$ is projection on $C(M)$, and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues.

Appendix 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 (Dani et al., 2008; Goldenshluger and Zeevi, 2013; Bastani et al., 2021; Kargin et al., 2023). The first assumption is about the boundedness of the parameter space.

Assumption 1 (Parameter Set) *For transformed parameter η_i, there exists a positive constant c_η such that $\|\eta_i\| \le c_\eta$, for all $i = 1, \dots, N$.*

To proceed, we define exhaustive and exclusive sets in the observation space to represent the event of each arm being optimal.

Definition 1 *Let $y(t) = (y_1(t)^\top, y_2(t)^\top, \dots, y_N(t)^\top)^\top$ and $A_i^\star \subset \mathbb{R}^{Nd_y}$ be the region in the space of $y(t)$ that makes arm i optimal: $a^\star(t) = i$. Then, denote the optimality probability of arm i by*

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

The definition holds for a normalized observation because the norm of an observation does not affect optimality. The next assumption is the margin condition of observations, which is slightly modified based on Definition 2 and Assumption 2 in the work of Bastani et al. (2021).

Assumption 2 (Margin Condition) *Consider the observation $y(t) = (y_1(t)^\top, y_2(t)^\top, \dots, y_N(t)^\top)^\top$ and the transformed parameters $\{\eta_i\}_{i\in[N]}$ as defined in (4). Then, given the event $\{y(t) \in A_i^\star\}$, we assume that there is $C' > 0$, such that for all $u > 0$;*

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

As a result of Assumption 2, for all $i,\ j \in [N]$, there exist a subset $A_i \subseteq A_i^\star$ and $\kappa > 0$ such that

$$\mathbb{P}(y(t) \in A_i) > \frac{1}{2}\mathbb{P}(y(t) \in A_i^\star) \qquad \text{and} \qquad \mathbb{P}(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j > \kappa | y(t) \in A_i) = 1. \quad (10)$$

We refer κ to a suboptimality gap with a positive probability, which is dependent on problems. Using this gap, we can analyze the stochastic contextual bandit problem in a similar way to analyses of the conventional multi-armed bandit problem with the suboptimality gap κ.

Appendix C. Relation to Contextual Bandits

The framework suggested in this section is generalized contextual bandits considering the uncertainty of contexts. An observation $y(t)$ presented in this framework can be considered a perfect observed context, if the context is observed without noise and transformation. That is, the model with the identity sensing matrix $A = I_{d_x}$ and the covariance of observation $\Sigma_y = \mathbf{0}_{d_y \times d_y}$ is reduced to the conventional contextual bandits, which are commonly discussed in the literature. Consequently, the following algorithm and theoretical results suggested in the remaining sections are valid for the conventional contextual bandits.

Appendix 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 (8). This result came from Theorem 1 of Abbasi-Yadkori et al. (2011) with some modifications.

Lemma 1 *Let $w_t = r_{a(t)}(t) - \widehat{x}_{a(t)}(t)^\top \eta_{a(t)}$ 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

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

where $\lambda_M = \lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$.

The next lemma guarantees the linear growth of eigenvalues of covariance matrices $\{B_i(t)\}_{i\in[N]}$ defined in (7). 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 $B_i(t)^{-1}$ in (7), with probability at least $1 - \delta$, if $N^{(1)}(\delta, T) \leq n_i(t) \leq T$ for given $T > 0$, we have*

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

where $N^{(1)}(\delta, T) = 8\log(T/\delta)/\lambda_m^2$.

Lemma 3 *Let $\widehat{\eta}_i(t)$ be the estimate in (8). Then, if $N^{(1)}(\delta, T) < n_i(t) \leq T$, with probability at least $1-\delta$, for all $i \in [N]$, we have*

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

Lemma 4 *Let $\widetilde{\eta}_i(t)$ be a sample in (6). Then, if $N^{(1)}(\delta, T) < n_i(t) \leq T$, with probability at least $1-\delta$, for all $i \in [N]$, we have*

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

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 5 *Let the minimum sample size be*

$$N^{(2)}(\delta, T, \kappa) = \max \left(N^{(1)}(\delta, T),\ 16L^2\lambda_m^{-1} \left(R\sqrt{d_y \log\left(1 + L^2T/\delta\right)} + c_\eta \right)^2 \kappa^{-2} \right).$$

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

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

where κ is the positive constant defined in (10) and $\mathscr{F}_{t-1}$ is the filtration defined in Lemma 1.

The results above can be applied to all partially observable contextual bandits with any type of parameter setup.

Appendix E. Results for the model with a shared parameter

In this sub-section, we present the theoretical results of the model with a shared parameter. For the model with a single shared parameter, $\eta_i = \eta_\star$ and $n_i(t) = t$ 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 is in L. The next theorem provides a high probability regret upper bound of Thompson sampling for partially observable contextual bandits with a shared parameter.

Theorem 2 *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*

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

The regret bound scales at most $\log^4 T$ with respect to the time horizon and linearly with N. $\sqrt{d_y \log(T/\delta)}$ and $\sqrt{d_y} \log(TNd_y/\delta)$ are incurred by the estimation errors 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 Park and Faradonbeh (2022b). As compared to the setting in the work of Park and Faradonbeh (2022b), 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.

Appendix F. Results for the model with arm-specific parameters

The following results provide estimation error bounds of the estimators defined in (9) 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 3 *Let η_i and $\widehat{\eta}_i(t)$ be the transformed true parameter in (4) and its estimate in (8), respectively. Then, with probability at least $1-\delta$, Algorithm 1 guarantees*

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

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

Appendix G. Numerical illustration of Estimation Accuracy

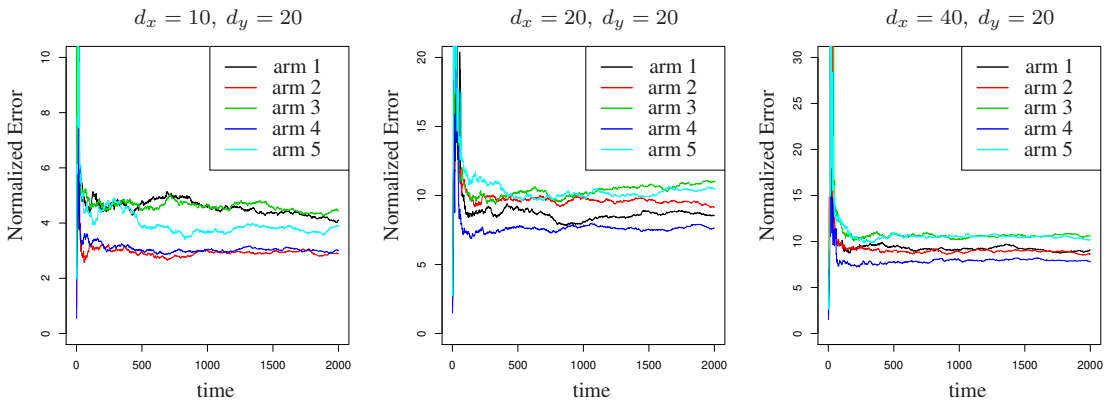

Figure 4: Plots of normalized estimation errors $\sqrt{t}\|\widehat{\eta}_i(t) - \eta_i\|$ of Algorithm 1 over time for partially observable stochastic contextual bandits with five arm-specific parameters and dimensions of observations and contexts $d_y = 20$, $d_x = 10,\ 20,\ 40$.

Figure 4 showcases the average estimation errors of the estimates in (9) for five different arm-specific parameters defined in (4), changing dimensions of observations and contexts. These errors are normalized by $t^{-0.5}$ based on Theorem 3. 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.

Appendix H. 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 $x_i(t)$ is known. The other is the context estimation error $(x_i(t)-\widehat{x}_i(t))^\top \mu_i$ 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 6 *The context estimate $\widehat{x}_i(t)^\top \mu_i$ has the mean $x_i(t)^\top \mu_i$ and a sub-Gaussian tail property such as*

$$\mathbb{E}\left[e^{\nu(\widehat{x}_i(t)-x_i(t))^\top \mu_i} \middle| y(t) \right] \leq e^{\frac{\nu^2 R_2^2}{2}},$$

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

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

$$\mathrm{Var}((\widehat{x}_i(t) - x_i(t))^\top \mu_i | y_i(t)) = \mu_i^\top (A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1} \mu_i$$

based on the results of the work of Robinson (1991). Because $\|\mu_i\| \leq 1$ for all $i \in [N]$, we can find $R_2 > 0$ such that

$$\mu_i^\top (A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1} \mu_i \leq \lambda_{\max}((A^\top \Sigma_y^{-1} A + \Sigma_x^{-1})^{-1}) = R_2^2, \tag{11}$$

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

$$\mathbb{E}\left[e^{\nu(\widehat{x}_i(t)-x_i(t))^\top \mu_i} \middle| y(t) \right] \leq e^{\frac{\nu^2 R_2^2}{2}}.$$

■

Lemma 7 *For any $\nu > 0$, we have*

$$\mathbb{E}\left[e^{\nu(r_i(t)-y_i(t)^\top \eta_i)} \middle| y(t) \right] \leq e^{\frac{\nu^2 R^2}{2}}.$$

where $R^2 = R_1^2 + R_2^2$ for R_1 from R_1-subgaussian property of reward errors and R_2 in (11).

Proof By (5),

$$r_i(t) - y_i(t)^\top \eta_i = (x_i(t)^\top \mu_i - y_i(t)^\top \eta_i) + \varepsilon_i(t),$$

which implies $\mathbb{E}[r_i(t) - \widehat{x}_i(t)^\top \mu_i | y_i(t), a(t)] = 0$, since $y_i(t)^\top \eta_i$ is the BLUP of $x_i(t)^\top \mu_i$. Due to $\mathrm{Var}(x_i(t)^\top \mu_i - y(t)^\top \eta_i | y(t)) \leq R_2^2$ by (11), we can find $R > 0$ such that

$$\mathrm{Var}(r_i(t) - \widehat{x}_i(t)^\top \mu_i | y_i(t)) = \mathrm{Var}(\varepsilon_i(t)) + \mathrm{Var}(x_i(t)^\top \mu_i - y_i(t)^\top \eta_i | y_i(t)) \leq R_1^2 + R_2^2 = R^2$$

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

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

■

Lemma 8 *Let*

$$D_{it}^\eta = \exp\left(\frac{(r_i(t) - y_i(t)^\top \eta) y_i(t)^\top \eta}{R} - \frac{1}{2}(y_i(t)^\top \eta)^2\right) \mathbb{I}(a(t) = i),$$

and $M_{it}^\eta = \prod_{\tau=1}^t D_{i\tau}^\eta$. *Then,* $\mathbb{E}[M_{it}^\eta] \leq 1$.

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

$$\begin{aligned}
\mathbb{E}[D_{it}^\eta | \mathscr{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\frac{(r_i(t) - y_i(t)^\top \eta) y_i(t)^\top \eta}{R} - \frac{1}{2}(y_i(t)^\top \eta)^2\right) \middle| y(t), a(t)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\zeta_i(t) y_i(t)^\top \eta}{R}\right) \middle| y(t), a(t)\right] \exp\left(-\frac{1}{2}(y_i(t)^\top \eta)^2\right).
\end{aligned}$$

Then, by Lemma 7, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\zeta_i(t) y_i(t)^\top \eta}{R}\right) \middle| y(t), a(t)\right] \exp\left(-\frac{1}{2}(y_i(t)^\top \eta)^2\right) \\
&\quad \leq \exp\left(\frac{1}{2}(y_i(t)^\top \eta)^2\right) \exp\left(-\frac{1}{2}(y_i(t)^\top \eta)^2\right) = 1.
\end{aligned}$$

Therefore,

$$\mathbb{E}[M_{it}^\eta | \mathscr{F}_{t-1}] = \mathbb{E}[M_{i1}^\eta D_{i2}^\eta \cdots D_{i(t-1)}^\eta D_{it}^\eta | \mathscr{F}_{t-1}] = D_1^\eta \cdots D_{i(t-1)}^\eta \mathbb{E}[D_{it}^\eta | \mathscr{F}_{t-1}] \leq M_{t-1}^\eta.$$

■

Now, we continue the proof of Lemma 1. Let ϕ_η be the probability density function of multivariate Gaussian distribution of η with the mean $\mathbf{0}_{d_y}$ and the covariance matrix I_{d_y}. By Lemma 9 of the work of Abbasi-Yadkori et al. (2011), for $M_t = \mathbb{E}[M_{it}^\eta | \mathscr{F}_\infty]$, we have

$$\mathbb{P}_{\phi_\mu}\left(\|S_{it}\|^2_{B_i(t)^{-1}} > 2R^2 \log\left(\frac{det(B_i(t))^{1/2}}{\delta}\right)\right) \leq \mathbb{E}[M_{it}] \leq \delta, \tag{12}$$

where $\mathbb{P}_{\phi_\eta}$ is the probability measure based on ϕ_η, and $S_{it} = \sum_{\tau=1}^{t} y(\tau) w_\tau \mathbb{I}(a(\tau) = i)$. Lemma 7 , Lemma 8 and (12) are sufficient conditions for the following inequality

$$\mathbb{P}_{\phi_\eta}\left(\|S_{it}\|^2_{B_i(t)^{-1}} > 2R^2 \log\left(\frac{\det(B_i(t))^{1/2}}{\delta} \right), \ \forall\, t > 0 \right) \leq \delta,$$

by Theorem 1 of the work of Abbasi-Yadkori et al. (2011). By Lemma 10 of the work of Abbasi-Yadkori et al. (2011), we have

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

and subsequently we have

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

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

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

which is a similar result to Theorem 2 of the work of Abbasi-Yadkori et al. (2011).

Appendix I. Proof of Lemma 2

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

$$W_T = \left\{ \max_{\{i \in [N], \tau \in [T]\}} ||y_i(\tau)||_\infty \leq v_T(\delta) \right\}, \tag{13}$$

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

Lemma 9 *For the event W_T defined in (13), we have $\mathbb{P}(W_T) \geq 1 - \delta$.*

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

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

Accordingly, we have

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

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

$$\mathbb{P}\left(\max_{i \in [N], \tau \in [T]} \|y(t)\|_\infty \geq \lambda_M^{\frac{1}{2}} \varepsilon \right) \leq 2TNd_y \cdot e^{-\frac{\varepsilon^2}{2}}$$

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

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

Thus,

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

■

Then, by Lemma 9, we have

$$\|y(t)\| \leq \sqrt{d_y} v_T(\delta) := L = \mathcal{O}\left(\sqrt{d_y \log(TNd_y/\delta)}\right), \tag{14}$$

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

Lemma 10 *(Azuma Inequality) Consider the sequence $\{X_t\}_{1\leq t\leq T}$ random variables adapted to some filtration $\{\mathscr{G}_t\}_{1\leq t\leq T}$, such that $\mathbb{E}[X_t|\mathscr{G}_{t-1}] = 0$. Assume that there is a deterministic sequence $\{c_t\}_{1\leq t\leq T}$ that satisfies $X_t^2 \leq c_t^2$, almost surely. Let $\sigma^2 = \sum_{1\leq t\leq T} c_t^2$. Then, for all $\varepsilon \geq 0$, it holds that*

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

The proof of Lemma 10 is provided in the work of Azuma (1967). Now, we construct a martingale and its different sequence to find an upper bound of a sum of random variables with Lemma 10. Let the sigma field generated by the contexts and chosen arms by time t

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

Consider $V_t = y_{a(t)}(t) y_{a(t)}(t)^\top$ in order to study the behavior of $B_i(t)$. Note that

$$\begin{aligned}\mathbb{E}[V_t|\mathscr{G}_{t-1}] &= \mathrm{Var}(y_i(t)|\mathscr{G}_{t-1}) + A x_i(t) x_i(t)^\top A^\top \\ &\succeq \lambda_m I_{d_y},\end{aligned} \tag{15}$$

where $\lambda_m = \lambda_{\min}(\Sigma_y)$. For all $t > 0$ and $\|z\| = 1$, it holds that

$$z^\top \left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}]\right) z \geq z^\top \left(\sum_{\tau=1:a(\tau)=i}^{t-1} \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}]\right) z \geq \lambda_m n_i(t). \tag{16}$$

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

$$X_t^i = (V_t - \mathbb{E}[V_t|\mathscr{G}_{\tau-1}])\mathbb{I}(a(\tau) = i), \tag{17}$$

$$Y_t^i = \sum_{\tau=1}^{t} (V_\tau - \mathbb{E}[V_\tau|\mathscr{G}_{\tau-1}])\, \mathbb{I}(a(\tau) = i). \tag{18}$$

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 $\sum_{\tau=1}^{t-1} V_\tau \mathbb{I}(a(\tau) = i)$. Because $(z^\top X_\tau^i z)^2 \leq \|y_i(t)\|^4 \leq L^4$ and thereby $\sum_{\tau=1}^{t-1} \left(z^\top X_\tau^i z\right)^2 \leq n_i(t) L^4$, using Lemma 10, we get the following inequality

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

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

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

for $\varepsilon \leq 0$. Now, using (15), we have the following inequality

$$\begin{aligned}
&\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} \left(V(\tau) - \mathbb{E}[V_t | \mathscr{G}_{\tau-1}]\right) \mathbb{I}(a(\tau) = i)\right) z \leq n_i(t)\varepsilon\right) \\
\geq\ &\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} \left(V(\tau) - \lambda_m I_{d_y}\right) \mathbb{I}(a(\tau) = i)\right) z \leq n_i(t)\varepsilon\right). \qquad (19)
\end{aligned}$$

Putting (16), (17), (18) and (19) together, we obtain

$$\mathbb{P}\left(z^\top \left(\sum_{\tau=1}^{t-1} V(\tau)\mathbb{I}(a(\tau) = i)\right) z \leq n_i(t)(\lambda_m + \varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2L^4}\right), \qquad (20)$$

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

$$\mathbb{P}\left(z^\top B_i(t) z \leq n_i(t)(\lambda_m + \varepsilon)\right) \leq \exp\left(-\frac{n_i(t)\varepsilon^2}{2L^4}\right), \qquad (21)$$

for $-\lambda_m \leq \varepsilon \leq 0$. In other words, by putting $\exp\left(-n_i(t)\varepsilon^2/(2L^4)\right) = \delta/T$, (21) can be written as

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

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

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

Appendix J. Proof of Lemma 5

For simplicity, let the event of the ith arm of being optimal at time t $A_{it} = \{a^\star(t) = i\}$. 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

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

Using the relationship below,

$$\begin{aligned}&\{y_i(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)\}\\ &\subset \left\{y_j(t)^\top(\widetilde{\eta}_j(t)-\eta_j) > \frac{1}{2}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)\right\}\bigcup\left\{y_i(t)^\top(\widetilde{\eta}_i(t)-\eta_i) < -\frac{1}{2}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)\right\},\end{aligned} \tag{22}$$

we have

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

Since $y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i) \leq L\|\widehat{\eta}_i(t)-\eta_i\|$, by Lemma 1 and 2, if $n_i(t),\ n_j(t) \geq N^{(1)}(\delta,T)$, we have

$$|y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i)| \leq L\sqrt{\frac{2}{\lambda_m}}\left(R\sqrt{d_y\log\left(1+\frac{L^2T}{\delta}\right)}+c_\eta\right)\frac{1}{n_i(t)^{1/2}}. \tag{23}$$

Similarly,

$$|y_j(t)^\top(\widehat{\eta}_j(t)-\eta_j)| \leq L\sqrt{\frac{2}{\lambda_m}}\left(R\sqrt{d_y\log\left(1+\frac{L^2T}{\delta}\right)}+c_\eta\right)\frac{1}{n_j(t)^{1/2}}. \tag{24}$$

To lower the value on the RHS of (23) less than $\kappa/4$, we need the minimum samples $n_i(t),\ n_j(t) > 32L^2\lambda_m^{-1}\left(R\sqrt{d_y\log(1+L^2T/\delta)}+c_\eta\right)^2\kappa^{-2}$, for the arm i and j. Then, we have

$$y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i) - \frac{1}{2}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j) \leq -\frac{\kappa}{4},$$

because $y_i(t)^\top\eta_i - y_j(t)^\top\eta_j \geq \kappa$ given A_{it} by (10). Similarly, we have

$$y_j(t)^\top(\widehat{\eta}_j(t)-\eta_j) + \frac{1}{2}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j) \geq \frac{\kappa}{4}.$$

Accordingly, we have

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

Based on (6), by Lemma 2, we have

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

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

$$\mathbb{P}(y_i(t)^\top \widetilde{\eta}_i(t) < y_j(t)^\top \widetilde{\eta}_j(t) | A_{it}, \mathscr{F}_{t-1}) \leq \exp\left(-\frac{n_i(t)\lambda_m \kappa^2}{32 v^2 L^2}\right) + \exp\left(-\frac{n_j(t)\lambda_m \kappa^2}{32 v^2 L^2}\right),$$

and thereby

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

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

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

Appendix K. Proof of Lemma 4

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

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

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

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

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

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

Therefore, by Theorem 4, for $N^{(1)}(\delta, T) < n_i(t) \leq T$, we have

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

Appendix L. Proof of Theorem 2

Note that $n_i(t) = t$ for all $i \in [N]$ for the shared parameter setup. We decompose the regret as follows:

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

since $\|y(t)\| \leq L$. By Lemma 4, if $t > N^{(1)}(\delta, T)$, 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

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

Now, we construct a martingale sequence with respect to the filtration $\{\mathscr{F}_{t-1}\}_{t=1}^{T}$ defined in Lemma 5. To that end, let $G_1 = H_1 = 0$,

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

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 $c_\tau = \tau^{-1/2}$. Since $\sum_{\tau=1}^{T} |G_\tau| \leq \sum_{\tau=2}^{T} c_\tau^2 \leq \log T$, by Lemma 10, we have

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

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

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{I}(a^\star(t) \neq a(t)) \leq \sqrt{2 \log T \log \delta^{-1}} + \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{P}(a^\star(\tau) \neq a(\tau) | \mathscr{F}_{\tau-1}). \tag{25}$$

Now, we proceed to the upper bound of the second term on the right side in (25).

Let $A_{it}^\star = \{y(t) \in A_i^\star\}$, where $A_i^\star$ is defined in Definition 1. By using

$$\begin{aligned}
&\{y_i(t)^\top \widetilde{\eta}_i(t) < y(t)^\top \widetilde{\eta}_j(t)\} \\
&\subset \left\{ y_j(t)^\top (\widetilde{\eta}_j(t) - \eta_j) > \frac{1}{2}(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \right\} \bigcup \left\{ y_i(t)^\top (\widetilde{\eta}_i(t) - \eta_i) < -\frac{1}{2}(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \right\},
\end{aligned} \tag{26}$$

we get

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

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

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

for all $N^{(1)}(\delta, T) < t \leq T$ and $i \in [N]$, where

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

Accordingly, we have

$$\begin{aligned}
&\mathbb{P}(y_i(t)^\top \widetilde{\eta}_j(t) - y_j(t)^\top \widetilde{\eta}_i(t) > 0 | \mathscr{F}_{t-1}, A^\star_{it}) \\
\leq\ &\mathbb{P}(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -h(\delta, T) L t^{-1/2} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | \mathscr{F}_{t-1}, A^\star_{it}) \\
+\ &\mathbb{P}(y_j(t)^\top (\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -h(\delta, T) L t^{-1/2} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) | \mathscr{F}_{t-1}, A^\star_{it}). \quad (27)
\end{aligned}$$

Let $E_{ijt} = \{h(\delta, T) L t^{-1/2} < 0.25(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j)\}$. Then, we can decompose the first term on the RHS in (27) as follows:

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T) L}{t^{1/2}} + (y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| \mathscr{F}_{t-1}, A^\star_{it} \right) \\
&= \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T) L}{t^{1/2}} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| E_{ijt}, \mathscr{F}_{t-1}, A^\star_{it} \right) \mathbb{P}(E_{ijt} | \mathscr{F}_{t-1}, A^\star_{it}) \\
&+ \mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T) L}{t^{1/2}} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| E^c_{ijt}, \mathscr{F}_{t-1}, A^\star_{it} \right) \mathbb{P}(E^c_{ijt} | \mathscr{F}_{t-1}, A^\star_{it}). \quad (28)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > -\frac{h(\delta, T) L}{t^{1/2}} + 0.5(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| E_{ijt}, \mathscr{F}_{t-1}, A^\star_{it} \right) \\
\leq\ &\mathbb{P}\left(y_i(t)^\top (\widetilde{\eta}_i(t) - \widehat{\eta}_i(t)) > 0.25(y_i(t)^\top \eta_i - y_j(t)^\top \eta_j) \middle| \mathscr{F}_{t-1}, A^\star_{it} \right). \quad (29)
\end{aligned}$$

By Assumption 2, if $t > N^{(1)}(\delta, T)$, we have

$$\mathbb{P}(E^c_{ijt} | \mathscr{F}_{t-1}, A^\star_{it}) = \mathbb{P}\left(4h(\delta, T) L t^{-1/2} > y_i(t)^\top \eta_i - y_j(t)^\top \eta_j \middle| \mathscr{F}_{t-1}, A^\star_{it} \right) \leq \frac{4h(\delta, T) L C'}{t^{1/2}}. \quad (30)$$

Thus, by (29) and (30), the probability in on the LHS of (28) can be written as

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

Using $y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) \sim \mathcal{N}(0, v^2 y_i(t)^\top B_i(t)^{-1} y_i(t))$ given $y(t)$, the first term on the RHS above can be written as

$$\begin{aligned}
&\mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > 0.25(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)|\mathscr{F}_{t-1}, A^\star_{it}) + \frac{4h(\delta,T)LC'}{t^{1/2}} \\
\leq\ & \int_0^\infty \mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > 0.25u|y(t), \mathscr{F}_{t-1}, A^\star_{it})\mathbb{P}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j = u)du \\
&\qquad + \frac{4h(\delta,T)LC'}{t^{1/2}} \\
\leq\ & \int_0^\infty \exp\left(-\frac{t\lambda_m u^2}{32v^2L^2}\right)\mathbb{P}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j = u|A^\star_{it})du + \frac{4h(\delta,T)LC'}{t^{1/2}}.
\end{aligned}$$

Since $\mathbb{P}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j = u|A^\star_{it}) < C'$ by Assumption 2, we have

$$\int_0^\infty \exp\left(-\frac{t\lambda_m u^2}{32v^2L^2}\right)\mathbb{P}(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j = u|A^\star_{it})du \leq vLC'\sqrt{\frac{32}{\lambda_m t}}.$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t)) > -y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i) + 0.5(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)|\mathscr{F}_{t-1}, A^\star_{it}) \\
\leq\ & LC't^{-1/2}\left(v\sqrt{\frac{32}{\lambda_m}} + 4h(\delta,T)\right). \qquad (31)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}(y_j(t)^\top(\widetilde{\eta}_j(t)-\widehat{\eta}_j(t)) > -y_j(t)^\top(\widehat{\eta}_j(t)-\eta_j) + 0.5(y_i(t)^\top\eta_i - y_j(t)^\top\eta_j)|\mathscr{F}_{t-1}, A^\star_{it}) \\
\leq\ & LC't^{-1/2}\left(v\sqrt{\frac{32}{\lambda_m}} + 4h(\delta,T)\right). \qquad (32)
\end{aligned}$$

Using (27), we have

$$\begin{aligned}
&\mathbb{P}(y_j(t)^\top\widetilde{\eta}_j(t) - y_i(t)^\top\widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A^\star_{it}) \\
\leq\ & LC't^{-1/2}\left(v\left(\sqrt{\frac{32}{\lambda_m}} + \sqrt{\frac{32}{\lambda_m}}\right) + 4h(\delta,T) + 4h(\delta,T)\right).
\end{aligned}$$

By summing the probabilities in (33) over $i, j \in [N]$, we have

$$\begin{aligned}
&\sum_{i=1}^N\sum_{j=1}^N \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A^\star_{it})\mathbb{P}(A^\star_{it}) \\
\leq\ & \frac{LC'}{\sqrt{t}}\sum_{i=1}^N\sum_{j=1}^N \mathbb{P}(A^\star_{it})\left(v\left(\sqrt{\frac{32}{\lambda_m}} + \sqrt{\frac{32}{\lambda_m}}\right) + 4h(\delta,T) + 4h(\delta,T)\right) \\
\leq\ & \frac{2c_M(\delta,T)LC'N}{\sqrt{t}}, \qquad (33)
\end{aligned}$$

where $c_M(\delta, T) = v\sqrt{\frac{32}{\lambda_m}} + 4h(\delta, T) = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$. Note that

$$\mathbb{P}(a^\star(t) \neq a(t)|\mathscr{F}_{t-1}) \leq \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A^\star_{it})\mathbb{P}(A^\star_{it}), \tag{34}$$

by the inclusion-exclusion formula. Putting (33), (34) and the minimal sample size $N^{(1)}(\delta, T)$ together, we have

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

By (25), with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t) \neq a(t)) \leq N^{(1)}(\delta, T) + \sqrt{2\log T \log \delta^{-1}} + 2c_M(\delta, T)LC'N\log T.$$

Therefore, since $N^{(1)}(\delta, T) = 8L^4\log(T/\delta)/\lambda_m^2$ and $L = \sqrt{2\lambda_M d_y \log(TNd_y/\delta)}$,

$$\begin{aligned}\text{Regret}(T) &\leq Lg(\delta)\left(N^{(1)}(\delta, T) + \sqrt{2\log T\log\delta^{-1}} + 2c_M(\delta, T)C'N\log T\right) \\ &= \mathcal{O}\left(Nd_y^3\log^4\left(\frac{T}{\delta}\right)\right).\end{aligned}$$

Appendix M. Proof of Theorem 3

Before starting the proof, we specify the constants described in the statement in Theorem 3. L is the bound of the ℓ_2-norm of observation 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 (10). First, we show that the number of selections of each arm scales linearly with a high probability.

Lemma 11 *For partially observable stochastic contextual bandits, with probability at least* $1-\delta$, *if* $\tau^{(4)}(\delta, T, \kappa) < t \leq T$, *Algorithm 1 guarantees*

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

where $\tau^{(4)}(\delta, T, \kappa) := \max(2(a_1 + (4/p_i)a_2^2) + 2\sqrt{(a_1 + (4/p_i)a_2^2)^2 - a_1^2}, N^{(3)}(\delta, T, \kappa)) = \mathcal{O}(d_y^{1/2}\log(T/\delta)\kappa^{-2})$, $a_1 = N^{(3)}(\delta, T, \kappa) + 2N/T$, $a_2 = \sqrt{2\log(2/\delta)}$, $N^{(3)}(\delta, T, \kappa) = \max(N^{(2)}(\delta, T, \kappa),\ 64(v^2L^2/(\lambda_m\kappa^2))\log T)$ *and* $N^{(2)}(\delta, T, \kappa)$ *is defined in Theorem 5.*

Proof By Lemma 5, if $n_i(t),\ n_j(t) > N^{(2)}(\delta, T, \kappa)$,

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

If $n_i(t) \geq 64(v^2L^2/(\lambda_m\kappa^2))\log T$, we have $\exp\left(-(n_i(t)\lambda_m\kappa^2)/(32v^2L^2)\right) \leq T^{-2}$. Now, we assume $n_i(t) > N^{(3)}(\delta,T,\kappa) = \max(N^{(2)}(\delta,T,\kappa),\ 64(v^2L^2/(\lambda_m\kappa^2))\log T)$ for all $i \in [N]$.

Note that

$$\mathbb{P}(a(t)=i|\mathscr{F}_{t-1}) \geq \mathbb{P}(a(t)=i|A_{it},\mathscr{F}_{t-1})\mathbb{P}(A_{it}) \geq \frac{\mathbb{P}(a^\star(t)=i)}{2}\left(1-\sum_{j\neq i}\mathbb{P}(a(t)=j|A_{it},\mathscr{F}_{t-1})\right)$$

by (10). Thus, $\mathbb{I}(a(t)=i)-(1/2)\mathbb{P}(a^\star(t)=i)\left(1-\sum_{j\neq i}\mathbb{P}(a(t)=j|A_{it},\mathscr{F}_{t-1})\right)$ is a submartingale difference. Accordingly, we have

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

Using Lemma 10, 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 $1-\delta$,

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

The following inequality

$$\frac{p_i}{2}\left(t-N^{(3)}(\delta,T,\kappa)-\frac{2N}{T}\right)-\sqrt{2t\log(2/\delta)} > \frac{p_i}{4}t,$$

is satisfied, if $t > 2(a_1+(4/p_i)a_2^2)+2\sqrt{(a_1+(4/p_i)a_2^2)^2-a_1^2}$, which is defined as $\tau_i^{(4)}(\delta,T,\kappa)$, where $a_1 = N^{(3)}(\delta,T,\kappa)+2N/T$ and $a_2=\sqrt{2\log(2/\delta)}$ based on the quadratic formula. With probability at least $1-\delta$, we have

$$n_i(t) > \frac{p_i t}{4}, \tag{35}$$

if $t > \tau_i^{(4)}(\delta,T,\kappa) = \mathcal{O}(p_i^{-1}d_y^{1/2}\log(T/\delta)\kappa^{-2})$. ■

Now, we are ready to prove Theorem 3. From Lemma (1), we have

$$\|\widehat{\eta}_i(t)-\eta_i\| \leq R\sqrt{\frac{2}{\lambda_m}}\left(\sqrt{d_y\log\left(\frac{1+TL^2}{\delta}\right)}+c_\eta\right)n_i(t)^{-1/2}, \tag{36}$$

if $n_i(t) > N^{(1)}(\delta, T)$. Thus, putting (35) and (36) together, if $t > \tau_i := \max(\tau^{(4)}(\delta, T, \kappa), 4p_i^{-1}N^{(1)}(\delta, T)) = \mathcal{O}(p_i^{-1}\kappa^{-2}d_y^2 \log^3(TNd_y/\delta))$, with probability at least $1-\delta$, we have the following estimation accuracy

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

Appendix N. Proof of Theorem 1

The regret can be written as

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

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

$$\begin{aligned}&\sum_{t=1}^{T}(y_{a^\star(t)}(t)^\top (\eta_{a^\star(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)) - y_{a(t)}(t)^\top (\eta_{a(t)}(t) - \widetilde{\eta}_{a^\star(t)}(t)))\mathbb{I}(a^\star(t) \neq a(t)) \\ \leq\ & L\sum_{t=1}^{T}(\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\|)\mathbb{I}(a^\star(t) \neq a(t)).\end{aligned}$$

By Theorem 3 and Lemma 4 with Lemma 11, if $t > \max_{i\in[N]} \tau_i$, we have

$$\|\widetilde{\eta}_{a^\star(t)}(t) - \eta_{a^\star(t)}\| + \|\widetilde{\eta}_{a(t)}(t) - \eta_{a(t)}\| \leq Rg'(\delta)t^{-1/2},$$

where

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

To proceed, with the probability of at least $1-\delta$, we utilize the martingale constructed in Theorem 2 with the intermediate result

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t) \neq a(t)) \leq \sqrt{2\log T \log \delta^{-1}} + \sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbb{P}(a^\star(\tau) \neq a(\tau)|\mathscr{F}_{\tau-1}). \tag{37}$$

To find a bound $\mathbb{P}(a^\star(\tau) \neq a(\tau)|\mathscr{F}_{\tau-1})$, using the same logic as (22) and (26), we decompose the following probability as follows:

$$\begin{aligned}&\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widetilde{\eta}_i(t)) > 0|\mathscr{F}_{t-1}, A_{it}^\star) \\ \leq\ &\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star) \\ +\ &\mathbb{P}(y(t)^\top(\widetilde{\eta}_j(t) - \widehat{\eta}_j(t)) > -y(t)^\top(\widehat{\eta}_j(t) - \eta_j) + 0.5y(t)^\top(\eta_i - \eta_j)|\mathscr{F}_{t-1}, A_{it}^\star).\end{aligned} \tag{38}$$

Similarly to (31) and (32), we have

$$\begin{aligned}&\mathbb{P}(y_i(t)^\top(\widetilde{\eta}_i(t)-\widehat{\eta}_i(t))>-y_i(t)^\top(\widehat{\eta}_i(t)-\eta_i)+0.5(y_i(t)^\top\eta_i-y_j(t)^\top\eta_j)|\mathscr{F}_{t-1},A_{it}^\star)\\ \leq\ & LC'\sqrt{\frac{4}{p_it}}\left(v\sqrt{\frac{32}{\lambda_m}}+4h(\delta,T)\right),\end{aligned}\tag{39}$$

if $t>\tau_i$, and,

$$\begin{aligned}&\mathbb{P}(y_j(t)^\top(\widetilde{\eta}_j(t)-\widehat{\eta}_j(t))>-y_j(t)^\top(\widehat{\eta}_j(t)-\eta_j)+0.5(y_i(t)^\top\eta_i-y_j(t)^\top\eta_j)|\mathscr{F}_{t-1},A_{it}^\star)\\ \leq\ & LC'\sqrt{\frac{4}{p_jt}}\left(v\sqrt{\frac{32}{\lambda_m}}+4h(\delta,T)\right),\end{aligned}\tag{40}$$

if $t>\tau_j$. Accordingly, based on (38), (39), and (40), we obtain the following bounds for the probabilities

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

By summing the probabilities up over $i,j\in[N]$, if $t>\tau_M:=\max_{i\in[N]}\tau_i=\mathcal{O}((p_{\min}^+)^{-1}\kappa^{-2}d_y^2\log^3(TNd_y/\delta))$, we have the following upper bound for the probability of choosing a suboptimal arm

$$\begin{aligned}&\mathbb{P}(a^\star(t)\neq a(t)|\mathscr{F}_{t-1})\\ \leq\ & \sum_{i=1}^N\sum_{j=1}^N\mathbb{P}(y_j(t)^\top\widetilde{\eta}_j(t)-y_i(t)\widetilde{\eta}_i(t)>0|\mathscr{F}_{t-1},A_{it}^\star)\mathbb{P}(A_{it}^\star)\\ \leq\ & \frac{2LC'}{\sqrt{p_{\min}^+t}}\sum_{i=1}^N\sum_{j=1}^N\mathbb{P}(A_{it}^\star)\left(v\left(\sqrt{\frac{32}{\lambda_m}}+\sqrt{\frac{32}{\lambda_m}}\right)+4h(\delta,T)+4h(\delta,T)\right)\\ \leq\ & \frac{4c_M(\delta,T)LC'N}{\sqrt{p_{\min}^+t}},\end{aligned}\tag{41}$$

where $c_M(\delta,T)=\max_{i\in[N]}\left(v\sqrt{\frac{32}{\lambda_m}}+4h(\delta,T)\right)=\mathcal{O}(\sqrt{d_y\log(T/\delta)})$. Putting (41) and the minimum sample size τ_M together, we have

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

where $\lceil\cdot\rceil$ is the ceiling function. By (25), with probability at least $1-\delta$, we have

$$\sum_{t=1}^T\frac{1}{\sqrt{t}}\mathbb{I}(a^\star(t)\neq a(t))\leq\tau_M+\sqrt{2\log T\log\delta^{-1}}+\frac{8c_M(\delta,T)C'N}{\sqrt{p_{\min}^+}}\log T.$$

Therefore, putting together $L = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$, $g'(\delta) = \mathcal{O}\left((p^+_{\min})^{-1/2} d_y^{1/2} \sqrt{\log(TNd_y/\delta)}\right)$, $c_M(\delta, T) = \mathcal{O}(\sqrt{d_y \log(TNd_y/\delta)})$ $\tau_M = \mathcal{O}((p^+_{\min})^{-1} \kappa^{-2} d_y^2 \log^3(TNd_y/\delta))$,

$$\begin{aligned}
\text{Regret}(T) &\leq Lg'(\delta) \left(\tau_M + \sqrt{2 \log T \log \delta^{-1}} + \frac{2c_M(\delta, T) C'}{\sqrt{p^+_{\min}}} \log T \right) \\
&= \mathcal{O}\left(\frac{d_y^3}{(p^+_{\min})^{3/2} \kappa^2} \log^4 \left(\frac{TNd_y}{\delta} \right) \right).
\end{aligned}$$