A Regret Bound for Greedy Partially Observed Stochastic Contextual Bandits

Hongju Park [1] **Mohamad Kazem Shirani Faradonbeh** [1]

Abstract

Contextual bandits are widely-used models in reinforcement learning for incorporating both generic and idiosyncratic factors in reward functions. The existing approaches rely on full observation of the context vectors, while the problem of learning optimal arms from partially observed contexts remains immature. We show that in the latter setting, decisions can be made more guarded to *minimize* the risk of pulling sub-optimal arms. More precisely, efficiency is established for *Greedy* policies that treat the estimates of the unknown parameter and of the unobserved contexts as their true values. That includes non-asymptotic high probability regret bounds that grow logarithmically with the time horizon and linearly with the number of arms. Numerical results that showcase the efficacy of avoiding exploration are provided as well.

1. Introduction

Contextual bandits are ubiquitous models for sequential decision-making in environments with finite action spaces. The range of applications is extensive and includes different problems for which time-varying and action-dependent information are important, such as personalized recommendation of news articles, healthcare interventions, advertisements, and clinical trials (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 many applications, consequential variables for decision-making are not perfectly observed. Technically speaking, in the bandit problem, the context vectors are often observed in a partial, transformed, or noisy manner (Bensoussan, 2004; Bouneffouf et al., 2017; Tennenholtz et al., 2021). Furthermore, since perfect observation can be considered as a special case of imperfectly observed contexts, sequential decision-making algorithms for the latter family of problems provide a richer class of settings compared to the former. Accordingly, partial observation models are commonly used in different problems, including space-state models, robot control, and filtering (Nagrath, 2006; Lin et al., 2012; Kang et al., 2012).

We study contextual bandits with imperfectly observed context vectors. The probabilistic structure of the problem under study, as time proceeds, is as follows. At every time step, there are N available arms, each of which has the unobserved context that is denoted by $x_i(t) \in \mathbb{R}^{d_x}$ for arm i at time t. The context vectors are generated independent of the previous contexts and independent of the other arms, according to a multivariate normal distribution $\mathcal{N}(0_{d_x}, \Sigma_x)$. Moreover, the corresponding observation of $x_i(t)$ is $y_i(t) \in \mathbb{R}^{d_y}$, while the stochastic reward $r_i(t)$ of arm i is determined by the context vector and the unknown parameter μ_*:

$$y_i(t) = Ax_i(t) + \zeta_i(t), \tag{1}$$

$$r_i(t) = x_i(t)^\top \mu_* + \psi_i(t), \tag{2}$$

Above, $\zeta_i(t)$ and $\psi_i(t)$ are the noises of observation and reward, which are identically distributed and independent following the distributions $\mathcal{N}(0_{d_y}, \Sigma_y)$ and $\mathcal{N}(0, \gamma_r^2)$, respectively. Further, the known $d_y \times d_x$ sensing matrix A captures the relationship between $x_i(t)$ and the noiseless portion of $y_i(t)$. The above structure holds for all arms and at all time t. For this model, as compared to the classic contextual bandits, there is uncertainty in contexts as well as in the environment associated with the parameter μ_*.

At each time, the goal is to learn to choose the optimal arm $a^*(t)$ (which can change at every time step), by utilizing the available information by time t. That is, the agent chooses an arm based on the previously collected data from the model in (1); $\{a(\tau)\}_{1\leq\tau\leq t-1}, \{y_{a(\tau)}(\tau)\}_{1\leq\tau\leq t-1}$, $\{r_{a(\tau)}\}_{1\leq\tau\leq t-1}$, as well as the observations at the time; $\{y_i(t)\}_{1\leq i\leq N}$. Then, once the action is taken, the resulting reward of the chosen arm will be provided to the agent according to the equation in (2), while rewards of the other arms are not observed. Clearly, to choose high-reward arms, the agent needs accurate estimates of the unknown

[*] Equal contribution [1] Department of Statistics, University of Georgia, 310 Herty Dr, Athens, GA 30602, USA. Correspondence to: Hongju Park <hp97161@uga.edu>, Mohamad Kazem Shirani Faradonbeh <mohamadksf@uga.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

parameter μ_*, as well as those of the contexts $x_i(t)$, for $i = 1, \cdots, N$. However, because $x_i(t)$ is not observed, estimation of μ_* is feasible only through the observation $y_i(t)$. Thereby, design of efficient reinforcement learning policies with guaranteed performance is challenging.

Learning strategies for contextual bandits are investigated in the literature, assuming that the context vectors are fully observed. Early papers focus on reinforcement learning policies that utilize Upper-Confident-Bounds (UCB) for addressing the exploitation-exploration trade-off (Auer, 2002; Abbasi-Yadkori et al., 2011; Chu et al., 2011). Another popular and efficient family of policies use randomized exploration, usually in the (Bayesian) form of Thompson sampling (Chapelle & Li, 2011; Agrawal & Goyal, 2013; Faradonbeh et al., 2020; Modi & Tewari, 2020). Recently, for contexts generated under certain conditions, it is shown that exploration-free greedy policies have efficient performances (Bastani et al., 2021).

Currently, study of efficient algorithms with theoretical performance guarantees for imperfect context observations is rather incomplete. Notably, imperfectness of observation frequently appears in different areas of applications. The causes are various, including privacy regulations (Sbeity & Younes, 2015), measurement errors (Lin et al., 2012; Kang et al., 2012), and missingness (Azimi et al., 2019). The existing analyses study some special cases, such as those with invertible sensing matrices. Asymptotic results are shown for UCB-type algorithms (Yun et al., 2017) and Thompson sampling (Park & Faradonbeh, 2021), as well as in presence of additional information (Tennenholtz et al., 2021), and in adversarial settings with partial monitoring (Lattimore & Szepesvári, 2019; Lattimore, 2022). For Greedy algorithms, numerical analyses indicate that they outperform Thompson sampling in partially observed contextual bandits with stochastic contexts (Park & Faradonbeh, 2022). Importantly, Greedy policies are of special interests in settings that exploration is not (ethically) permitted, such as precision medicine (Bastani et al., 2021). However, comprehensive theoretical performance guarantees are currently unavailable for Greedy policies.

We perform the *finite-time worst-case* analysis of Greedy reinforcement learning policies for imperfectly observed contextual bandits. We provide high probability regret bounds that consists of poly-logarithmic factors of the time horizon and of the failure probability. Furthermore, the effects of other quantities such as the number of arms, dimension, and properties of the noise processes, are fully characterized. Illustrative numerical experiments showcasing the efficiency of Greedy policies are also provided.

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

A highlight of this paper is as follows. In Section 2, we formulate the problem and discuss the preliminary materials. Next, a Greedy policy for contextual bandits with imperfect context observations is presented in Section 3. In Section 4, we provide theoretical performance guarantees, followed by numerical experiments in Section 5.

Notation $A^\top$ is the transpose of A and the ℓ_2 norm is $\|v\| = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$. Moreover, $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) denote the minimum (maximum) eigenvalues of A, respectively. In addition, $O(\cdot)$ is the order of magnitude such that $f(n) = O(g(n))$ denotes $\limsup_{n\to\infty} |f(n)|/g(n) < \infty$ for a real valued function f and a strictly positive valued function g. Finally, $\{X_i\}_{i\in E} = \{X_i : i \in E\}$, and $I(\cdot)$ is the indicator function.

2. Problem Formulation

First, we formally discuss the problem of contextual bandits with imperfect context observations. The bandit machine under consideration has N arms, each of which has its own unobserved context $x_i(t)$, for $i \in \{1, \cdots, N\}$. Equation (1) presents the observation model, where the observations $\{y_i(t)\}_{1\le i\le N}$ are linearly transformed functions of the contexts, perturbed by additive noise vectors $\{\zeta_i(t)\}_{1\le i\le N}$. Equation (2) describes the process of reward generation for different arms, depicting that *if* the agent selects arm i, then the resulting reward is an *unknown* linear function of the unobserved context vector, subject to some additional randomness due to the reward noise $\psi_i(t)$.

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

$$\mathcal{N}(y_i(t)^\top D^\top \mu_*, \gamma_{ry}^2), \tag{3}$$

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

Based on the conditional distribution in (3), in order to maximize the expected reward given the observations, one can consider the conditional expectation of the reward given the observation; $y_i(t)^\top D^\top \mu_*$. So, letting $\eta_* = D^\top \mu_*$ be the transformed parameter, we focus on the estimation of η_*. The rationale is twofold. First, the conditional expected reward given the observation can be inferred with only learning η_*, and one does not need to learn μ_*. Second, μ_* is not estimable when the rank of the sensing matrix A in the observation model is less than d_x. Therefore, the quality of learning the environment is reflected by the estimation accuracy of η_*. Note that consistent estimation of μ_* needs further assumptions of non-singularity of A and $d_y \geq d_x$.

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

$$a^*(t) = \arg\max_i y_i(t)^\top \eta_*. \tag{4}$$

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

$$\text{Regret}(T) = \sum_{t=1}^{T} \left(y_{a^*(t)}(t) - y_{a(t)}(t) \right)^\top \eta_*. \tag{5}$$

In other words, the regret at time T is the total difference in the rewards obtained up to time T, between the optimal arms $a^*(t)$ and the arm $a(t)$ chosen by the reinforcement learning policy based on the observations by the time t. Note that this difference does not directly depend on the unknown contexts $\{x_i(t)\}_{1 \leq i \leq N}$ since the optimal policy that selects $a^*(t)$ does *not* observe the context vectors, and decides merely based on the observations $y_i(t)$, for $i = 1, \cdots, N$.

3. Reinforcement Learning Policy

In this section, we explain the details of the Greedy algorithm for contextual bandits with imperfect context observations. Although inefficient in some reinforcement learning problems, Greedy algorithms have a logarithmic regret bound for contextual bandits with fully observed contexts under certain conditions such as covariate diversity (Bastani et al., 2021). Intuitively speaking, the latter condition expresses that the context vectors provide information along all dimensions in $\mathbb{R}^{d_x}$ with positive probability, so that additional exploration is not necessary.

Algorithm 1 : Greedy policy for contextual bandits with imperfect context observations

```
Set $B(1) = \Sigma^{-1}$, $\widehat{\eta}(1) = \eta$
for $t = 1, 2, \ldots,$ do
    Observe observations $y_i(t)$ for $i = 1, \ldots, N$
    Select arm $a(t) = \arg\max_{1 \leq i \leq N} y_i(t)^\top \widehat{\eta}(t)$
    Gain reward $r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_* + \psi_{a(t)}(t)$
    Update $B(t+1)$ and $\widehat{\eta}(t+1)$ by (7) and (8)
end for
```

As discussed in Section 2, it suffices for the policy to learn the arm i that maximizes $\mathbb{E}[r_i(t)|y_i(t)] = y_i(t)^\top \eta_*$. To that end, we estimate η_* using the least-squares estimator

$$\widehat{\eta}(t) = \arg\max_{\eta} \sum_{\tau=1}^{t} (r_{a(\tau)}(\tau) - y_{a(\tau)}(\tau)^\top \eta)^2. \tag{6}$$

Then, the Greedy policy works with $\widehat{\eta}(t)$ in lieu of the truth η_*. So, the algorithm selects the arm $a(t)$ at time t, such that $a(t) = \arg\max_{1 \leq i \leq N} y_i(t)^\top \widehat{\eta}(t)$. Thanks to the structure of the parameter estimates in (6), once can update $\widehat{\eta}(t)$ in a recursive fashion. The recursions relies on updating the empirical inverse covariance matrix $B(t)$ as well, according to

$$B(t+1) = B(t) + y_{a(t)}(t) y_{a(t)}(t)^\top, \tag{7}$$

and

$$\widehat{\eta}(t+1) = B(t+1)^{-1} \left(B(t)\widehat{\eta}(t) + y_{a(t)}(t) r_{a(t)}(t) \right), \tag{8}$$

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

4. Theoretical Performance Guarantees

In this section, we present a theoretical result for Algorithm 1 presented in the previous section. The result provides a worst-case analysis and establishes a high probability upper-bound for the regret as defined in (5).

Theorem 4.1. *Assume that Algorithm 1 is used in a bandit with N arms and the observation dimension d_y. Then, with probability at least $1 - 4\delta$, $\text{Regret}(T)$ is of the order*

$$\text{cond}(A\Sigma_x A^\top + \Sigma_y) \gamma_{ry} N d_y^{3/2} \left(\log \frac{N d_y T}{\delta} \right)^{5/2} \log \frac{d_y T}{\delta},$$

where cond *denotes the condition number of a matrix and γ_{ry}^2 is the conditional variance of reward defined in (3).*

The proof of Theorem 4.1 is provided in the appendix. Next, we discuss the intuitions of different terms in the regret bound presented in the above theorem.

The regret bound above scales linearly with the number of arms N, and with $d_y^{3/2}$ for the dimension of the observations d_y, while it grows poly-logarithmically with time horizon T. The dimension of the unobserved context vectors does not appear in the above regret bound because the optimal policy in (4) does not observe the exact values of the context vectors. So, similar to Algorithm 1, the optimal policy needs to estimate the context vectors as well.

The rationale of the linear growth of the regret bound in Theorem 4.1 with the number of arms N is that for larger N, the policy is more likely to choose one of the sub-optimal arms, which makes the larger growths of the regret more likely. In addition, the terms $d_y^{3/2}$ and $\lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)$ constituting $\mathrm{cond}(A\Sigma_x A^\top + \Sigma_y)$ are generated by the ℓ_2 norm of the stochastic observation vectors, for which the following high probability upper-bounds hold; $\|y_i(t)\|_2^2 = O(\lambda_{\max}(A\Sigma_x A^\top + \Sigma_y) d_y \log(NTd_y/\delta))$. Similarly, the poly-logarithmic scaling of the regret bound in terms of T and δ, originates from the magnitude of the random vectors ($\upsilon_T(\delta) = O((\log(NTd_y/\delta))^{1/2})$) that are used in the analysis of Algorithm 1. On the other hand, γ_{ry} indicates the role of the reward noise. Simply, if the reward observations are noisier, it will be harder to learn the optimal arms. Finally, the factor $\lambda_{\min}(A\Sigma_x A^\top + \Sigma_y)$ consisting of $\mathrm{cond}(A\Sigma_x A^\top + \Sigma_y)$ in the regret bound is the intrinsic decrease in the cumulative reward caused by uncertainties about the unknown parameter μ_*.

The proof process is divided into two steps. The first step consists of establishing a high probability upper-bound for $(y_{a^*(t)}(t) - y_{a(t)}(t))^\top \eta_*$. Then, in the second step we find upper-bounds on the frequency of pulling sub-optimal arms. Consequently, by combining the above-mentioned bounds, we prove Theorem 4.1.

5. Numerical Illustrations

In this section, we perform numerical analyses for the theoretical result in the previous section. We simulate cases with $N = 10, 20, 50$ arms and different dimensions of the observations $d_y = 5, 20, 50$, while the context dimensions are fixed to $d_x = 20$. Each case is repeated 100 times, and the average values of the quantities of interest across 100 scenarios are reported, as well as the worst among all the repetitions.

In Figure 1, the left plot depicts the average (solid) and worst-case (dashed) regret among all scenarios, normalized by $\log t$. The number of arms N varies as shown in the graph, while the observation dimension is fixed to $d_y = 10$. Next, the graph on the right hand side illustrates that the normalized regrets increase over time for different d_y, for the fixed number of arms $N = 5$. In Figure 1, the worst-case regret curves are well above the average ones as expected, but curves for both average-case and worst-case become flat as time goes on, implying that the worst-case regret grows logarithmically in terms of t.

Figure 2 presents the average and worst-case regret (not normalized by $\log T$) at time $T = 2000$ for different values of $N = 10, 20, 50$ and $d_y = 5, 20, 50$. The plot shows that the regret at $T = 2000$ increases as N and d_y become larger. In addition, it shows that the dimension of observations d_y has a greater effect on the regret than that of the number of arms N, which is consistent with the result of Theorem 4.1.

6. Conclusion

This work investigates reinforcement learning algorithms for contextual bandits where the contexts are observed imperfectly. We focus on establishing theoretical results for regret analysis, and establish a high probability regret bound for Greedy algorithms. The presented regret bound grows poly-logarithmically with the time horizon T.

There exist multiple interesting future directions introduced in this paper. First, it will be of interest to study reinforcement learning policies for settings that each arm has its own parameter. Further, regret analysis for contextual bandits under imperfect context observations where the covariance matrices of context vectors and the sensing matrix are unknown, is another problem for future work.

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.

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

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.

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

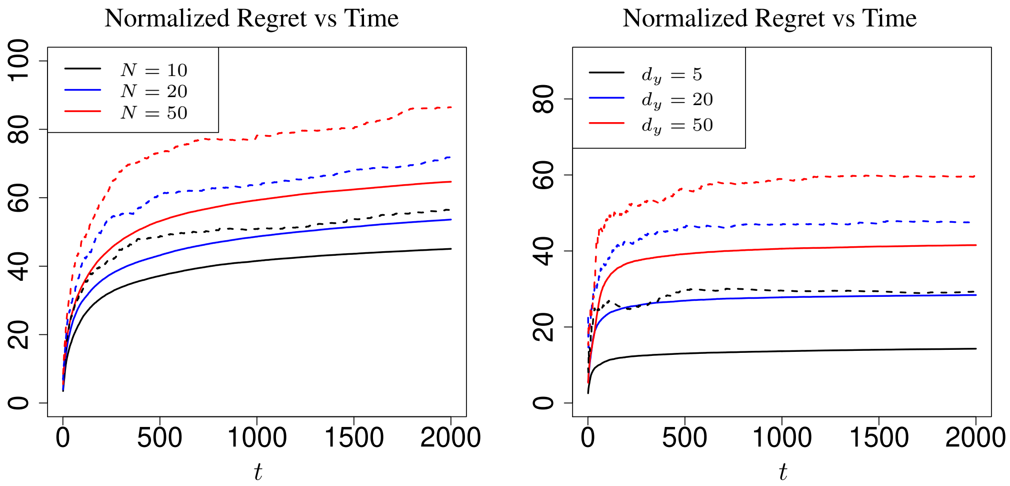

Figure 1. Plots of $\text{Regret}(t)/\log t$ over time for the different number of arms $N = 10, 20, 100$ and $d_y = 5, 20, 50$. The solid and dashed lines represent average and worst regret curves, respectively.

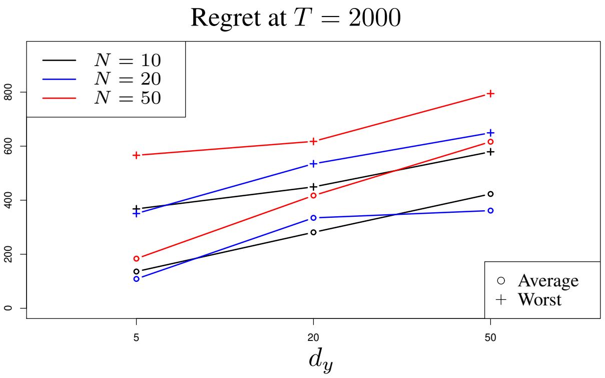

Figure 2. Plot of average and worst-case $\text{Regret}(T)$ at $T = 2000$ for different number of arms $N = 10, 20, 50$ and dimension of observations $d_y = 5, 20, 50$.

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.

Bouneffouf, D., Rish, I., Cecchi, G. A., and Féraud, R. Context attentive bandits: Contextual bandit with restricted context. *arXiv preprint arXiv:1705.03821*, 2017.

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

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

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.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. On adaptive linear–quadratic regulators. *Automatica*, 117: 108982, 2020.

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.

Lattimore, T. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *arXiv preprint arXiv:2202.10997*, 2022.

Lattimore, T. and Szepesvári, C. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pp. 2111–2139. PMLR, 2019.

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.

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. Efficient algorithms for learning to control bandits with unobserved contexts. *arXiv preprint arXiv:2202.00867*, 2022.

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

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.

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. Proofs of the Technical Results

In the sequel, $C(A)$ and $C(A)^\perp$ are employed to denote the column-space of the matrix A and its orthogonal subspace, respectively. In addition, $P_{C(A)}$ is the projection operator onto $C(A)$.

A.1. Proof of Theorem 4.1

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

$$W_T = \left\{ \max_{\{1 \leq \tau \leq t \text{ } and \text{ } 1 \leq i \leq N\}} ||S_y^{-1/2} y_i(\tau)||_\infty \leq v_T(\delta) \right\}, \tag{9}$$

where $v_T(\delta) = (2\log(N d_y T/\delta))^{1/2}$.

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

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

Lemma A.2. *Let $\sigma\{X_1, \ldots, X_n\}$ be the sigma-field generated by random vectors $X_1, \ldots, X_n$. For the observation of chosen arm $y_{a(t)}(t)$ at time t, the estimator $\widehat{\eta}(t)$ defined in (8), and the filtration $\{\mathscr{F}_t\}_{1 \leq t \leq T}$ defined according to*

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

we have

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

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

Lemma A.2 sets the stage for analysis of the (unnormalized) empirical inverse covariance $B(t)$ in (7)

Lemma A.3. *(Matrix Azuma Inequality (Tropp, 2012)) Consider the sequence $\{M_k\}_{1 \leq k \leq K}$ of symmetric $d \times d$ random matrices adapted to some filtration $\{\mathscr{G}_k\}_{1 \leq k \leq K}$, such that $\mathbb{E}[M_k|\mathscr{G}_{k-1}] = 0$. Assume that there is a deterministic sequence of symmetric matrices $\{A_k\}_{1 \leq k \leq K}$ that satisfy $M_k^2 \preceq A_k^2$, almost surely. Let $\sigma^2 = \| \sum_{1 \leq k \leq K} A_k^2 \|$. Then, for all $\varepsilon \geq 0$, it holds that*

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

Lemma A.4 provides a high probability lower bound for the minimum eigenvalue of $B(t)$. Then, Lemma A.5 bounds the estimation error.

Lemma A.4. *For $B(t)$ in (7) and $t \leq T$, on the event W_T defined in (9), by Lemma A.2 and A.3, with the probability at least $1 - \delta$, we have*

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

Lemma A.5. *In Algorithm 1, let $\widehat{\eta}(t)$ be the parameter estimate, as defined in (8). Then, for $t \leq T$, on the event W_T defined (9), we have*

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

Next, Lemma A.6 gives an upper bound for the probability that Algorithm 1 does not choose the optimal arm at time t. Finally, Lemma A.7 studies the weighted sum of indicator functions $I(a^*(t) \neq a(t))$ that count the effective number of times that the algorithm chooses sub-optimal arms.

Lemma A.6. *Given $B(t)$, an upper bound of probability of choosing a sub-optimal arm is bounded as follows:*

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

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

Lemma A.7. *For $I(a^*(t) \neq a(t))$, on the event W_T, with the probability at least $1-\delta$, we have*

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

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

Note that $\mathrm{Regret}(T)$ is the sum of the conditional expected reward difference $\left(y_{a^*(t)}(t) - y_{a(t)}(t)\right)^\top \eta_*$ for $1 \leq t \leq T$. The difference $\left(y_{a^*(t)}(t) - y_{a(t)}(t)\right)^\top \eta_*$ at time t is greater than 0, only when $a^*(t) \neq a(t)$. Thus, the regret can be rewritten as $\mathrm{Regret}(T) = \sum_{t=1}^T \left(y_{a^*(t)}(t) - y_{a(t)}(t)\right)^\top \eta_* I(a^*(t) \neq a(t))$. To find an upper bound of the regret, we find high probability upper bounds for $\left(y_{a^*(t)}(t) - y_{a(t)}(t)\right)^\top \eta_*$ and $I(a^*(t) \neq a(t))$, respectively. For both upper bounds, the inverse of the (unnormalized) empirical covariance matrix $B(t)$ in (7) matters in that the matrix determines the size of estimation error $\|\widehat{\eta}(t) - \eta_*\|$.

By, Lemma A.4, we have

$$\lambda_{\min}(B(t)) \geq \lambda_{\min}(S_y)(t-1)\left(1 - \sqrt{\frac{32\upsilon_T(\delta)^4}{t-1}\log\frac{d_yT}{\delta}}\right), \tag{10}$$

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

$$(y_{a^*(t)}(t) - y_{a(t)}(t))^\top \eta_* \leq \lambda_{\max}(S_y)^{1/2}\upsilon_T(\delta)\|\widehat{\eta}(t) - \eta_*\|, \tag{11}$$

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

$$\mathrm{Regret}(T) \leq \sum_{t=1}^T \lambda_{\max}(S_y)^{1/2}\upsilon_T(\delta)\|\widehat{\eta}(t) - \eta_*\| I(a^*(t) \neq a(t)). \tag{12}$$

Here, we denote $\lambda_t = \lambda_{\max}(B(t)^{-1}) = (\lambda_{\min}(B(t)))^{-1}$. By (10), we can find $t^* = 128\upsilon_T(\delta)^4 \log \frac{d_yT}{\delta} + 1$, such that

$$\lambda_t \leq \frac{2}{\lambda_{\min}(S_y)(t-1)}, \tag{13}$$

with the probability at least $1-\delta$, for all $t^* < t \leq T$. By Lemma A.5 and (13), for all $t^* < t \leq T$, with the probability at least $1-3\delta$, we have

$$\lambda_{\max}(S_y)^{1/2}\upsilon_T(\delta)\|\widehat{\eta}(t) - \eta_*\| \leq a_1(t-1)^{-1/2}, \tag{14}$$

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

$$\mathrm{Regret}(T) \leq \sum_{t<t^*} 2\lambda_{\max}(S_y)^{1/2}\upsilon_T(\delta)\|\eta_*\| + \sum_{t^* \leq t \leq T} a_1(t-1)^{-1/2} I(a^*(t) \neq a(t)), \tag{15}$$

with the probability at least $1-3\delta$. Now, we consider the probability to choose the optimal arm at time t. By Lemma A.6, we have

$$\sum_{t^* \leq t \leq T} \frac{\mathbb{P}(a^*(t) \neq a(t) | B(t))}{\sqrt{t-1}} \leq \frac{2^{3/2} N \lambda_{\max}(S_y)^{1/2} d_y v_T(\delta) \gamma_{ry}}{\|\eta_*\| \lambda_{\min}(S_y)^{1/2}} \log T. \tag{16}$$

Now, we construct an upper bound about the indicator function $I(a^*(t) \neq a(t))$ in (12), by Lemma A.7.

$$\begin{aligned} \sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} I(a^*(t) \neq a(t)) &\leq \sqrt{32 \log T \log(T\delta^{-1})} \\ &+ \sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t-1}} \mathbb{P}(a^*(\tau) \neq a(\tau) | B(\tau)), \end{aligned} \tag{17}$$

with the probability at least $1-\delta$. Therefore, by (16) and (17), with the probability at least $1-4\delta$, the following inequalities hold for the regret of the algorithm, which yield to the desired result:

$$\begin{aligned} & \mathrm{Regret}(T) \\ = \; & \sum_{t=1}^{T} (y_{a^*(t)}(t) - y_{a(t)}(t))^\top \eta_* I(a^*(t) \neq a(t)) \\ \leq \; & 2\lambda_{\max}(S_y)^{1/2} v_T(\delta) \|\eta_*\| t^* + \sum_{t^* \leq t \leq T} a_1 \frac{1}{\sqrt{t-1}} I(a^*(t) \neq a(t)) \\ = \; & O\left(\frac{\lambda_{\max}(S_y)}{\lambda_{\min}(S_y)} \gamma_{ry} N d_y^{2/3} \left(\log \frac{N d_y T}{\delta} \right)^{5/2} \log \frac{d_y T}{\delta} \right). \end{aligned} \tag{18}$$

Finally, using $S_y = A \Sigma_x A^\top + \Sigma_y$ and $\lambda_{\max}(S_y)/\lambda_{\min}(S_y) = O((\lambda_{\max}(\Sigma_A) + \lambda_{\max}(\Sigma_y))/(\lambda_{\min}(\Sigma_A) + \lambda_{\min}(\Sigma_y)))$, with the probability at least $1-4\delta$, we have

$$\begin{aligned} & \mathrm{Regret}(T) = \\ & O\left(\frac{(\lambda_{\max}(\Sigma_A) + \lambda_{\max}(\Sigma_y))\gamma_{ry}}{\lambda_{\min}(\Sigma_A) + \lambda_{\min}(\Sigma_y)} N d_y^{3/2} \left(\log \frac{N d_y T}{\delta} \right)^{5/2} \log \frac{d_y T}{\delta} \right). \end{aligned} \tag{19}$$

□

A.2. Proofs of Lemmas

A.2.1. Proof of Lemma A.1

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

$$\mathbb{P}\left(|\lambda_{\max}(S_y)^{-1/2} y_{ij}(t)| \geq \varepsilon \right) \leq 2 \cdot e^{-\frac{\varepsilon^2}{2}} \tag{20}$$

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

$$\mathbb{P}\left(|\lambda_{\max}(S_y)^{-1/2} y_{ij}(t)| \geq v_T(\delta) \right) \leq 2 \cdot e^{-\frac{v_T(\delta)^2}{2}} = 2 \cdot e^{-\log \frac{2 N d_y T}{\delta}} = \frac{\delta}{N d_y T}. \tag{21}$$

Thus,

$$\mathbb{P}(W_T) \geq 1 - \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d_y} \mathbb{P}\left(|\lambda_{\max}(S_y)^{-1/2} y_{ij}(t)| \geq v_T(\delta) \right) \geq 1 - \delta. \tag{22}$$

A.2.2. Proof of Lemma A.2

We use the following decomposition

$$S_y^{-0.5} y_{a(t)}(t) = P_{C(S_y^{0.5} \widehat{\eta}(t))} S_y^{-0.5} y_{a(t)}(t) + P_{C(S_y^{0.5} \widehat{\eta}(t))^\perp} S_y^{-0.5} y_{a(t)}(t). \tag{23}$$

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

$$Z(\nu,N)=\operatorname*{arg\,max}_{Z_i,1\leq i\leq N}\left\{Z_i^\top\nu\right\},\tag{24}$$

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

$$Z(\nu,N)\overset{d}{=}P_{C(\nu)}Z(\nu,N)+(I_d-P_{C(\nu)})Z_i,\tag{25}$$

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

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

Thus, $S_y^{-0.5}y_{a(t)}$ has the same distribution as $P_{C(S_y^{0.5}\widehat{\eta}(t))}S_y^{-0.5}y_{a(t)}+P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp}S_y^{-0.5}y_i(t)$, where $P_{C(S_y^{0.5}\widehat{\eta}(t))}S_y^{-0.5}y_{a(t)}$ and $P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp}S_y^{-0.5}y_i(t)$ are statistically independent. By the decomposition (23) and the independence,
$\mathbb{E}\left[S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}|\mathscr{F}_{t-1}\right]$ can be written as

$$\begin{aligned}
&\mathbb{E}\left[S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}|\mathscr{F}_{t-1}\right]\\
&=\mathbb{E}\left[(P_{C(S_y^{0.5}\widehat{\eta}(t))}+P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp})S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}(P_{S_y^{0.5}\widehat{\eta}(t)}+P_{S_y^{0.5}\widehat{\eta}(t)^\perp})|\mathscr{F}_{t-1}\right]\\
&=\mathbb{E}\left[P_{C(S_y^{0.5}\widehat{\eta}(t))}S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}P_{C(S_y^{0.5}\widehat{\eta}(t))}|\mathscr{F}_{t-1}\right]+P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp}.
\end{aligned}\tag{26}$$

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

$$\begin{aligned}
&P_{S_y^{0.5}\widehat{\eta}(t)}\mathbb{E}[S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}|\widehat{\eta}(t)]P_{S_y^{0.5}\widehat{\eta}(t)}\\
=\;&\frac{S_y^{0.5}\widehat{\eta}(t)\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y\widehat{\eta}(t)}\mathbb{E}[S_y^{-0.5}y_{a(t)}(t)y_{a(t)}(t)^\top S_y^{-0.5}|\widehat{\eta}(t)]\frac{S_y^{0.5}\widehat{\eta}(t)\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y\widehat{\eta}(t)}\\
=\;&\frac{S_y^{0.5}\widehat{\eta}(t)}{\widehat{\eta}(t)^\top S_y\widehat{\eta}(t)}\mathbb{E}[(\widehat{\eta}(t)^\top S_y^{0.5}S_y^{-0.5}y_{a(t)}(t))^2|\widehat{\eta}(t)]\frac{\widehat{\eta}(t)^\top S_y^{0.5}}{\widehat{\eta}(t)^\top S_y\widehat{\eta}(t)}\\
=\;&P_{S_y^{0.5}\widehat{\eta}(t)}\mathbb{E}\left[\left(\left(\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^\top S_y^{-0.5}y_{a(t)}(t)\right)^2\middle|\widehat{\eta}(t)\right],
\end{aligned}\tag{27}$$

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

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

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

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

we have

$$y_{a(t)}(t)^\top S_y^{-0.5}\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\overset{d}{=}\max_{1\leq i\leq N}\{V_i:V_i\sim N(0,1)\}.\tag{28}$$

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

$$\mathbb{E}\left[\left(y_{a(t)}(t)^\top S_y^{-0.5}\overrightarrow{S_y^{0.5}\widehat{\eta}(t)}\right)^2\middle|\widehat{\eta}(t)\right] = \mathbb{E}\left[\left(\max_{1\le i\le N}(\{V_i : V_i \sim N(0,1)\})\right)^2\right]. \tag{29}$$

We define the quantity in (29) as k_N,

$$k_N = \mathbb{E}\left[\left(\max_{1\le i\le N}(\{V_i : V_i \sim N(0,1)\})\right)^2\right], \tag{30}$$

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

$$\mathbb{E}[S_y^{-0.5} y_{a(t)}(t) y_{a(t)}(t)^\top S_y^{-0.5} | \widehat{\eta}(t)] = P_{C(S_y^{0.5}\widehat{\eta}(t))} k_N + P_{C(S_y^{0.5}\widehat{\eta}(t))^\perp} = P_{C(S_y^{0.5}\widehat{\eta}(t))}(k_N - 1) + I_{d_y}. \tag{31}$$

A.2.3. Proof of Lemma A.4

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

$$\lambda_{\min}\left(\sum_{\tau=1}^{t-1} \mathbb{E}[V_\tau | \widehat{\eta}(\tau)]\right) \geq t - 1. \tag{32}$$

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

$$\begin{aligned} X_\tau &= V_\tau - \mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}], \\ Y_\tau &= \sum_{j=1}^{\tau} \left(V_j - \mathbb{E}[V_j|\mathscr{F}_{j-1}]\right). \end{aligned} \tag{33}$$

Then, $X_\tau = Y_\tau - Y_{\tau-1}$ and $\mathbb{E}\left[X_\tau|\mathscr{F}_{\tau-1}\right] = 0$. Thus, X_τ is a martingale difference sequence. Because $v_T^2(\delta)I - V_t \succeq 0$ for all $t \leq T$, $4v_T^4(\delta)I - X_\tau^2 \succeq 0$, for all $\tau \leq T$, on the event W_T. By Lemma A.3, we get

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t-1} X_\tau\right) \leq (t-1)\varepsilon\right) \leq d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32 v_T^4(\delta)}\right), \tag{34}$$

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

$$\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau - \sum_{\tau=1}^{t-1}\mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}]\right) \leq \lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) - \lambda_{\min}\left(\sum_{\tau=1}^{t-1}\mathbb{E}[V_\tau|\mathscr{F}_{\tau-1}]\right) \tag{35}$$

and (32), we obtain

$$P\left(\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) \leq (t-1)(1+\varepsilon)\right) \leq d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32 v_T^4(\delta)}\right), \tag{36}$$

where $-1 \leq \varepsilon \leq 0$ is arbitrary, and we used the fact that $\lambda_{\min}\left(\sum_{\tau=1}^{t-1} V_\tau\right) \geq 0$. Indeed, using $\sum_{\tau=1}^{t-1} V_\tau = S_y^{-0.5} B(t) S_y^{-0.5}$, on the event W_T defined in (9), for $-1 \leq \varepsilon \leq 0$ we have

$$\mathbb{P}\left(\lambda_{\min}(B(t)) \leq \lambda_{\min}(S_y)(t-1)(1+\varepsilon)\right) \leq d_y \cdot \exp\left(-\frac{(t-1)\varepsilon^2}{32 v_T^4(\delta)}\right), \tag{37}$$

where $\lambda_{\min}(S_y) = \lambda_{\min}(S_y)$. In other words, by equating $d_y \cdot \exp\left(-(t-1)\varepsilon^2/(32v_T^4(\delta)\right)$ to δ/T, (37) can be written as

$$\lambda_{\min}(B(t)) \geq \lambda_{\min}(S_y)(t-1)\left(1 - \sqrt{\frac{32 v_T(\delta)^4}{t-1}\log\frac{d_y T}{\delta}}\right), \tag{38}$$

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

A.2.4. Proof of Lemma A.5

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

$$\begin{aligned} \mathbb{E}[\widehat{\eta}(t)|\mathscr{F}_{t-1}] &= B(t)^{-1}\left(\Sigma^{-1} + \sum_{\tau=1}^{t-1} y_{a(\tau)}(\tau) y_{a(\tau)}(\tau)^\top\right)\eta_* = \eta_* \\ \mathrm{Cov}(\widehat{\eta}(t)|\mathscr{F}_{t-1}) &= B(t)^{-1}\gamma_{ry}^2. \end{aligned} \tag{39}$$

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

$$\begin{aligned} \mathbb{P}\left(\|\widehat{\eta}(t) - \eta_*\| > \varepsilon | B(t)\right) &\leq \mathbb{P}\left(d_y Z^2 > \varepsilon^2\right) \\ &\leq 2 \cdot \exp\left(-\frac{\varepsilon^2}{2 d_y \lambda_{\max}(B(t)^{-1})\gamma_{ry}^2}\right), \end{aligned} \tag{40}$$

where $\varepsilon \geq 0$.

A.2.5. Proof of Lemma A.6

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

$$\begin{aligned} (y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_* &= (y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_{**} + (y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top(\eta_* - \eta_{**}) \\ &= \|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\| \, \|\eta_* - \eta_{**}\| \cos\theta(y_{a^*(t)}(t) - y_{a^{**}(t)}(t), \eta_* - \eta_{**}) \\ &= \|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\| \, \|\eta_* - \eta_{**}\|. \end{aligned} \tag{41}$$

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

$$\begin{aligned} \mathbb{P}(a^*(t) \neq a(t) | \{y_i(t)\}_{1 \leq i \leq N}, B(t)) &= \mathbb{P}\left(\|\widehat{\eta}(t) - \eta_*\| > \frac{(y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_*}{\|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\|} \middle| \{y_i(t)\}_{1 \leq i \leq N}, B(t)\right) \\ &\leq 2 \cdot \exp\left(-\frac{\left(\frac{(y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_*}{\|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\|}\right)^2}{2 d_y \lambda_{\max}(B(t)^{-1})\gamma_{ry}^2}\right). \end{aligned} \tag{42}$$

Using $\|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\|^2 \leq 2\lambda_{\max}(\Sigma_A) d_y v_T(\delta)^2$ on the event W_T, we have

$$2 \cdot \exp\left(-\frac{\left(\frac{(y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_*}{\|y_{a^*(t)}(t) - y_{a^{**}(t)}(t)\|}\right)^2}{2 d_y \lambda_t \sigma_{ry}^2}\right) \leq 2 \cdot \exp\left(-\frac{((y_{a^*(t)}(t) - y_{a^{**}(t)}(t))^\top \eta_*)^2}{2 d_y^2 v_T(\delta)^2 \lambda_{\max}(\Sigma_A) \lambda_t \sigma_{ry}^2}\right). \tag{43}$$

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

$$f_{X_{(N-1)}, X_{(N)}}(x_{N-1}, x_N) = N(N-1)\phi(x_N)\phi(x_{N-1})\Phi(x_{N-1})^{N-2}, \tag{44}$$

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

$$\begin{aligned} f_D(d) &= \int f_{D,X_{N-1}}(d, x_{N-1}) dx_{N-1} \\ &= \int N(N-1)\phi(x_{N-1}+d)\phi(x_{N-1})\Phi(x_{N-1})^{N-2} dx_{N-1} \\ &\leq N\phi(0). \end{aligned} \tag{45}$$

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

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

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

where the density of Δ_t, $f_{\Delta_t}(\Delta_t)$, is bounded by $N/\sqrt{2\pi\eta_*^\top S_y \eta_*}$ by (45).

A.2.6. PROOF OF LEMMA A.7

We construct a martingale difference sequence that satisfies the conditions in Lemma A.3. To that end, let $G_1 = H_1 = 0$,

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

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

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

Since $\mathbb{E}[G_\tau|\mathscr{F}_{\tau-1}^*] = 0$, the above sequences $\{G_\tau\}_{\tau\geq 0}$ and $\{H_\tau\}_{\tau\geq 0}$ are a martingale difference sequence and a martingale with respect to the filtration $\{\mathscr{F}_\tau^*\}_{1\leq\tau\leq T}$, respectively. Let $c_\tau = 2(\tau-1)^{-1/2}$. Since $\sum_{\tau=1}^T |G_\tau| \leq \sum_{\tau=2}^T c_\tau^2 \leq 4\log T$, by Lemma A.3, we have

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

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

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