Improved Algorithms for Multi-period Multi-class Packing Problems with Bandit Feedback

Wonyoung Kim [1] **Garud Iyengar** [1] **Assaf Zeevi** [1]

Abstract

We consider the linear contextual multi-class multi-period packing problem (LMMP) where the goal is to pack items such that the total vector of consumption is below a given budget vector and the total value is as large as possible. We consider the setting where the reward and the consumption vector associated with each action is a class-dependent linear function of the context, and the decision-maker receives bandit feedback. LMMP includes linear contextual bandits with knapsacks and online revenue management as special cases. We establish a new estimator which guarantees a faster convergence rate, and consequently, a lower regret in LMMP. We propose a bandit policy that is a closed-form function of said estimated parameters. When the contexts are non-degenerate, the regret of the proposed policy is sublinear in the context dimension, the number of classes, and the time horizon T when the budget grows at least as $\sqrt{T}$. We also resolve an open problem posed in Agrawal & Devanur (2016), and extend the result to a multi-class setting. Our numerical experiments clearly demonstrate that the performance of our policy is superior to other benchmarks in the literature.

1. Introduction

In the multi-period packing problem (MPP) the decision-maker "packs" the arrivals so that the total consumption across a set of resources is below a given budget vector and the reward is maximized. A variant of the packing problem, where items consume multiple resources and the decisions must be made sequentially with bandit feedback for a fixed time horizon, is known as bandits with knapsacks (Agrawal & Devanur, 2014a; Badanidiyuru et al., 2018; Immorlica et al., 2019). MPPs also arise in online revenue management (Besbes & Zeevi, 2012; Ferreira et al., 2018). MPPs in the literature assume that all arrivals belong to a single class. However, in several application domains (e.g., operations, healthcare, and e-commerce), the arrivals are heterogeneous, and personalizing decisions to each distinct population or class is of paramount importance. In this paper, we consider a class of linear *multi-class* multi-period packing problems (LMMP). At each round, there is a single arrival that belongs to one of J classes, and the decision-maker observes the d-dimensional context and the cost for K different available actions. The outcome of selecting an action is a random sample of the reward and a consumption vector for m resources with an expected value that is a class-dependent linear function of the d-dimensional contexts. The goal of the problem is to minimize the cumulative regret over a time horizon T while ensuring that the total resource consumed is at most B.

The LMMP problem is a generalization of several problems including linear contextual bandits with knapsacks (LinCBwK) introduced by Agrawal & Devanur (2016). They proposed an online mirror descent-based algorithm that achieves $\tilde{O}(OPT/B \cdot d\sqrt{T})$ regret when the budget B for each of the m resources is $\Omega(\sqrt{d}T^{3/4})$, where OPT is the reward obtained by the oracle policy. Although the regret bound is meaningful for $B = \Omega(d\sqrt{T})$, establishing the regret bound for smaller budget values was left as an open problem. Chu et al. (2011) established a regret bound sublinear in d for the linear contextual bandit setting, which is a special case of LinCBwK with no budget constraints. Thus, the following question remained open: *"Is there an algorithm for LinCBwK that achieves sublinear dependence on d with budget $B = \Omega(\sqrt{T})$?"*

We propose a novel algorithm and an improved estimation strategy that settles this open problem and generalizes the result to the more general class of LMMP. The proposed algorithm achieves $\tilde{O}(OPT/B\sqrt{JdT})$ regret with budget $B = \Omega(\sqrt{JdT})$ under non-degenerate contexts. While regret of the existing algorithms grows linearly in the number of classes J, our estimator is able to pool data from different classes and avoids linear dependence on J. To reiterate, the improved regret bound results from the novel estimator which has faster convergence rates.

[1]Columbia University, New York, NY, USA. Correspondence to: Wonyoung Kim <wk2389@columbia.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Our main contributions are summarized as follows:

- We propose a new problem class – linear multi-class multi-period packing problems (LMMP). This problem generalizes a variety of problems including LinCBwK and online revenue management problems to the multi-class setting.
- We propose a novel estimator that uses contexts for *all* actions (including the contexts in skipped rounds) and yields $O(\sqrt{Jd/n})$ convergence rate for J classes, context dimension d, and n admitted arrivals (Theorem 4.2).
- We propose a novel `AMF` (Allocate to the Maximum First) algorithm which achieves $\tilde{O}(OPT/B\sqrt{JdT})$ regret with budget $B = \Omega(\sqrt{JdT})$ where OPT is the reward obtained by oracle policy (Theorem 5.1). For the single class setting with $J = 1$, we improve the existing bound by $\sqrt{d}$ and show that the bound is valid when $B = \Omega(\sqrt{dT})$, thus resolving an open problem posed in Agrawal & Devanur (2016) regarding LinCBwK.
- We evaluate our proposed algorithm on a suite of synthetic experiments and demonstrate its superior performances.

All proofs omitted from the front matter can be found in the Appendix.

2. Related Works

There are two streams of work that are relevant for LMMP. In online revenue management literature, Gallego & Van Ryzin (1994) introduced the dynamic pricing problem where the demand is a known function of price (action). Besbes & Zeevi (2009) and Besbes & Zeevi (2012) extended the problem under unknown demands with multiple resource constraints. Ferreira et al. (2018) proposed a Thompson sampling-based algorithm and extended it to contextual bandits with knapsacks. When the expected demand is a linear function of the price vector, the dynamic pricing problem is a special case of linear contextual bandits with knapsack (LinCBwK) proposed by Agrawal & Devanur (2016).

The LinCBwk is a common generalization of bandits with knapsacks (Badanidiyuru et al., 2018; Immorlica et al., 2019; Li et al., 2021) and online stochastic packing problems (Feldman et al., 2010; Agrawal & Devanur, 2014b; Devanur et al., 2011). Recently, Sankararaman & Slivkins (2021) proved a logarithmic regret bound for LinCBwK when there exists a problem-dependent gap between the reward of the optimal action and the other actions. Instead of the gap assumption, we require non-degeneracy of the stochastic contexts (see Assumption 3 for a precise definition) to obtain a regret bound sublinear in d and extends to the case when the contexts are generated from J different class.

Amani et al. (2019) proposed a variant of LinCBwK where the selected action must satisfy a single constraint with high probability in all rounds, i.e., LinCBwK with anytime constraints. Moradipari et al. (2021) and Pacchiano et al. (2021) proposed a Thompson sampling-based algorithm and an upper confidence bound-based algorithm, respectively, for LinCBwK with a single anytime constraint. Liu et al. (2021) highlighted the difference between global and anytime constraints and proposed a pessimistic-optimistic algorithm for the anytime constraints. We focus on the global constraints; however, we note that the extension to the anytime constraints is straightforward with minor modifications.

2.1. Notation

Let $\mathbb{R}_+$ denote the set of positive real numbers. For two real numbers $a, b \in \mathbb{R}$, we write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For a natural number $N \in \mathbb{N}$, let $[N] := \{1, \dots, N\}$.

3. Linear Multi-period Packing Problem

Let $[J]$ denote the set of classes with arrival probabilities $p = \{p_j\}_{j\in[J]}$, where $p_{\min} := \min_{j\in[J]} p_j > 0$. For simplicity, we assume that the class arrival probabilities p are known while the same theoretical results can be obtained when the probabilities are unknown to the decision-maker (See Section B.7 for details). In each round $t \in [T]$, the covariates $\{\mathbf{x}_{k,t}^{(j)} \in [0,1]^d : k \in [K]\}$ are drawn from an unknown class-specific distribution $\mathbb{F}_j$ and the decision-maker observes an arrival of the form $(j_t, \{\mathbf{x}_{k,t}^{(j_t)} : k \in [K]\})$, where $j_t \in [J]$ is the arrived class. Upon observing the arrival, the decision-maker can either take one of K different actions or skip the arrival. When the arrival is skipped, the decision-maker does not obtain any rewards or consume any resources. When the decision-maker chooses an action $a_t \in [K]$, the reward and consumption of the resource are given by

$$\mathbb{E}\left[r_{a_t,t}^{(j_t)} \middle| \mathcal{H}_t\right] = \left\{\theta_\star^{(j_t)}\right\}^\top \mathbf{x}_{a_t,t}^{(j_t)} \in [-1,1],$$
$$\mathbb{E}\left[\mathbf{b}_{a_t,t}^{(j_t)} \middle| \mathcal{H}_t\right] = \left\{W_\star^{(j_t)}\right\}^\top \mathbf{x}_{a_t,t}^{(j_t)} \in [0,1]^m,$$

for some unknown class-specific parameters $\theta_\star^{(j)} \in [0,1]^d$ and $W_\star^{(j)} \in [0,1]^{d\times m}$. The sigma algebra $\mathcal{H}_t$ is generated by the class-specific variables $\{j_s, \mathbf{x}_{k,s}^{(j_s)} : s \in [t], k \in [K]\}$, actions $\{a_s : s \in \mathcal{A}_t\}$, consumption vectors $\{\mathbf{b}_{a_s,s}^{(j_s)} : s \in \mathcal{A}_{t-1}\}$ and rewards $\{r_{a_s,t}^{(j_s)} : s \in \mathcal{A}_{t-1}\}$, where $\mathcal{A}_t$ is the rounds admitted by the decision-maker until round t. The process terminates at the horizon T or runs out of budget

$B \in \mathbb{R}_+^m$ for some resources $r \in [m]$. The problem reduces to LinCBwK when the number of class is $J = 1$.

LMMP allows each class to have a different set of contexts and parameters, which is required in many applications such as e-commerce, clinical trials, and dynamic pricing. For example, consider an e-commerce setting with J classes of customers with J different preferences. At each decision point t, the decision-maker must make one of K different d-dimensional offers: the k-th offer will result in a random consumption of m resources with mean consumption vector $(W_\star^{(j_t)})^\top \mathbf{x}_{k,t}^{(j_t)}$ and results in random reward with mean $(\theta_\star^{(j_t)})^\top \mathbf{x}_{k,t}^{(j_t)}$. Note that the context $\mathbf{x}_{k,t}^{(j_t)}$ can include the price charged to the class j_t customers as one of the components. This feature of LMMP is novel to the literature and allows for a personalized decision for each class.

Let $\rho \in \mathbb{R}_+^m$ denote per-period budget vector for m resources. Without loss of generality, one can assume that $\rho = (B/T)\mathbf{1}_m$, by rescaling $W_\star^{(j)}$. We assume that ρ is known to the decision-maker, which is essential to target the correct optimal policy. Unlike the unconstrained finite horizon bandit problems, the optimal policy of LMMP depends on ρ. Without knowledge of ρ, the bandit policy cannot converge to the correct optimal policy, which leads to a cumulative regret linear in T. As a result, many related problems, such as LinCBwk (Agrawal & Devanur, 2016) and online revenue management (Ferreira et al., 2018), commonly assume knowledge of both B and T.

In our work, we assume that B is possibly unknown at first but known at the end of the round. Specifically, the decision-maker only observes the initial inventory B_1, which will increase to B before the end of the horizon T. This scenario is relevant to online inventory management, where a product's inventory is supplied at different time points. This assumption is more practical than in Agrawal & Devanur (2016) where B and OPT must be known to the decision-maker. When OPT is unknown, Agrawal & Devanur (2016) proposed to estimate OPT with $\sqrt{T}$ number of rounds, which requires the knowledge of T and budget $B = \Omega(\sqrt{d}T^{\frac{3}{4}})$. Instead of estimating OPT, we use ρ to avoid the required budget $B = \Omega(\sqrt{d}T^{\frac{3}{4}})$.

We benchmark the performance of the decision-maker's policy relative to that of an oracle who knows the distributions $\{\mathbb{F}_j : j \in [J]\}$ and the parameters $\{\theta_\star^{(j)}, W_\star^{(j)} : j \in [J]\}$, but does not know the arrivals $\{(j_t, \mathbf{x}_{k,t}^{(j)}) : t \in [T]\}$ a-priori. In this case, the optimal static policy for the oracle $\{\pi_k^{\star(j)} : j \in [J], k \in [K]\}$ is the solution to the following optimization problem:

$$\begin{aligned}
\max_{\pi_k^{(j)}} & \sum_{j=1}^{J}\sum_{k=1}^{K} p_j \pi_k^{(j)} \mathbb{E}_{\mathbf{x}_k \sim \mathbb{F}_j}\left[\left\{\theta_\star^{(j)}\right\}^\top \mathbf{x}_k\right] \\
\text{s.t.} & \sum_{j=1}^{j}\sum_{k=1}^{K} p_j \pi_k^{(j)} \mathbb{E}_{\mathbf{x}_k \sim \mathbb{F}_j}\left[\left\{W_\star^{(j)}\right\}^\top \mathbf{x}_k\right] \le \rho, \\
& \sum_{k=1}^{K} \pi_k^{(j)} \le 1,\ \forall j \in [J], \\
& \pi_k^{(j)} \ge 0,\ \forall j \in [J], \forall k \in [K],
\end{aligned} \tag{1}$$

Note that the oracle policy depends on ρ and thus both B and T. Then the expected reward obtained by the oracle is

$$OPT := T\sum_{j=1}^{J}\sum_{k=1}^{K} p_j \pi_k^{\star(j)} \mathbb{E}_{\mathbf{x}_k \sim \mathbb{F}_j}\left[\left\{\theta_\star^{(j)}\right\}^\top \mathbf{x}_k\right].$$

Let $\pi := \{\pi_{k,t}^{(j)} : j \in [J], k \in [K], t \in [T]\}$ denote the adapted (randomized) control policy of the decision-maker, i.e. she chooses action $k \in [K]$ when the arrival at time $t \in [T]$ belongs to class $j \in [J]$. Note that $\sum_{k=1}^K \pi_{k,t}^{(j)} \le 1$ in order to allow the decision-maker to skip an arrival and save the inventory for later use. Our goal is to compute a policy that minimizes the cumulative regret $\mathcal{R}_T^\pi$ defined as

$$\mathcal{R}_T^\pi := OPT - \mathbb{E}\left[\sum_{t=1}^{T} R_t^\pi\right],$$

where $R_t^\pi := \sum_{k=1}^K \pi_{k,t}^{(j_t)} \mathbb{E}\left[\left\{\theta_\star^{(j_t)}\right\}^\top \mathbf{x}_{k,t}^{(j_t)}\right]$ is the expected reward obtained by policy π at time t.

For the LMMP problem, we assume the following regularity conditions on the stochastic processes.

Assumption 1. (Sub-Gaussian errors) For each $t \in [T]$, the error of the reward $\eta_{k,t} = r_{k,t}^{(j_t)} - \left\{\theta_\star^{(j_t)}\right\}^\top \mathbf{x}_{k,t}^{(j_t)}$ is conditionally zero-mean σ_r-sub-Gaussian for a fixed constant $\sigma_r \ge 0$. In other words, $\mathbb{E}\left[\exp\left(v\eta_{k,t}\right) | \mathcal{H}_t\right] \le \exp\left(\frac{v^2\sigma_r^2}{2}\right)$ for all $v \in \mathbb{R}$. For the consumption vectors, $\mathbb{E}\left[\mathbf{v}^\top\{\mathbf{b}_{k,t}^{(j_t)} - (W_\star^{(j_t)})^\top \mathbf{x}_{k,t}^{(j_t)}\}\middle| \mathcal{H}_t\right] \le \exp(\frac{\|\mathbf{v}\|_2^2\sigma_b^2}{2})$ for all $\mathbf{v} \in \mathbb{R}^m$.

Assumption 2. (Independently distributed contexts) The set of contexts $\{\mathbf{x}_{k,t}^{(j)} : k \in [K]\}$ are generated independently over $t \in [T]$. The contexts and cost in the same round and class can be correlated with each other.

Assumption 3. (Positive definiteness of average covariances) For each $t \in [T]$ and $j \in [J]$, there exists $\alpha > 0$, such that

$$\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_{k,t}^{(j)}\left\{\mathbf{x}_{k,t}^{(j)}\right\}^\top\right]\right) \ge \alpha.$$

Assumptions 1 and 2 are standard in stochastic contextual bandits with knapsacks literature(Agrawal & Devanur, 2016; Sankararaman & Slivkins, 2021; Sivakumar et al., 2022). In the multi-class scenario, Assumption 2 implies that all the contexts are drawn independently over time steps, but their distribution may vary depending on the class. The independence of contexts is supported by Bastani & Bayati (2020) and Kim et al. (2021), as this assumption is practical in real-world applications such as clinical trials where patient health covariates are independent of those of other patients. Assumption 3 implies that the density of the covariate distribution is non-degenerate. This assumption is necessary to estimate all entries of the parameters through linear regression in the statistics literature. In our work, we use the estimator $\check{\Theta}_n$ for $\Theta_\star$ in pseudo-rewards as defined in equation (5). Therefore, the convergence of $\|\check{\Theta}_n - \Theta_\star\|_2$ and the accuracy of the pseudo-rewards are heavily dependent on Assumption 3. Recent literature on contextual bandits (without constraints) has utilized Assumption 3 to improve the dependency of d on the regret bound (Bastani & Bayati, 2020; Kim et al., 2021; Bastani et al., 2021; Oh et al., 2021). As noted by Kannan et al. (2018) and Sivakumar et al. (2020; 2022), contexts with measurement errors provide enough variability to satisfy Assumption 3.

4. Proposed Method

In this section, we present our proposed estimator for the parameters $\{\theta_\star^{(j)}, W_\star^{(j)} : j \in [J]\}$ and the proposed closed form bandit policy.

4.1. Proposed Estimator

In sequential decision-making problems with contexts, the decision-maker observes the contexts for all actions, *but* the reward for only selected actions, i.e. the rewards for unselected actions remain missing. A statistical missing data technique called the doubly robust (DR) method is employed to handle the missing rewards for linear contextual bandit problem (Kim & Paik, 2019; Dimakopoulou et al., 2019; Kim et al., 2021; 2023; 2022). However, extensions to LinCBwK or LMMP problems have not been explored yet.

To apply the DR method to the LMMP problem, we modify the randomization technique proposed by Kim et al. (2023). For each $n \in \mathbb{N}$, let $\tau(n)$ be the round when the n-th admission happens (recall that the bandit policy allows for skipping some arrivals). Clearly, $n \le \tau(n) < \tau(n+1)$ holds. Let

$$\Theta_\star := \begin{pmatrix} \theta_\star^{(1)} \\ \vdots \\ \theta_\star^{(J)} \end{pmatrix}, \mathbf{W}_\star := \begin{pmatrix} W_\star^{(1)} \\ \vdots \\ W_\star^{(J)} \end{pmatrix}, \tilde{X}_{k,n} := \begin{pmatrix} \mathbf{0}_d \\ \vdots \\ \mathbf{x}_{k,\tau(n)}^{(j_{\tau(n)})} \\ \mathbf{0}_d \end{pmatrix}$$

denote the stacked parameter vectors, and zero padded contexts where $\mathbf{x}_{k,\tau(n)}^{(j_{\tau(n)})}$ is located after the $j_{\tau(n)} - 1$ of $\mathbf{0}_d$ vectors. Then the score for the ridge estimator for $\Theta^\star$ at round $\tau(n)$ is:

$$\sum_{\nu=1}^{n} \left(r_{a_{\tau(\nu)},\tau(\nu)}^{(j_{\tau(\nu)})} - \Theta^\top \tilde{X}_{a_{\tau(\nu)},\nu} \right) \tilde{X}_{a_{\tau(\nu)},\nu}$$
$$= \sum_{\nu=1}^{n} \sum_{k=1}^{K} \mathbb{I}\left(a_{\tau(\nu)} = k\right) \left(r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} - \Theta^\top \tilde{X}_{k,\nu} \right) \tilde{X}_{k,\nu},$$

where $\Theta \in \mathbb{R}^{J\cdot d}$. Dividing the score by the probability $\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}$ gives the inverse probability weighted (IPW) score,

$$\sum_{\nu=1}^{n} \sum_{k=1}^{K} \frac{\mathbb{I}\left(a_{\tau(\nu)} = k\right)}{\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}} \left(r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} - \Theta^\top \tilde{X}_{k,\nu} \right) \tilde{X}_{k,\nu}.$$

To obtain the DR score, Bang & Robins (2005); Kim et al. (2021) proposed to subtract the nuisance tangent space generated by an imputed estimator $\check{\Theta}$:

$$\sum_{\nu=1}^{n} \sum_{k=1}^{K} \frac{\mathbb{I}\left(a_{\tau(\nu)} = k\right)}{\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}} \left(\tilde{X}_{k,\nu}^\top \check{\Theta} - \tilde{X}_{k,\nu}^\top \Theta \right) \tilde{X}_{k,\nu},$$

from the IPW score. Then the following DR score

$$\sum_{\nu=1}^{n} \sum_{k=1}^{K} \left\{ r_{k,\tau(\nu)}^{DR(\check{\Theta})} - \tilde{X}_{k,\nu}^\top \Theta \right\} \tilde{X}_{k,\nu}, \tag{2}$$

is obtained where

$$r_{k,\nu}^{DR(\check{\Theta})} := \frac{\mathbb{I}\left(a_{\tau(\nu)}=k\right)}{\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}} r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} + \left\{ 1 - \frac{\mathbb{I}\left(a_{\tau(\nu)}=k\right)}{\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}} \right\} \tilde{X}_{k,\nu}^\top \check{\Theta}. \tag{3}$$

The score (2) has a similar form with the score equation for the ridge estimator. The difference with the ridge estimator is that it uses contexts for all actions $k \in [K]$ with the pseudo-reward $r_{k,\nu}^{DR(\check{\Theta})}$ which is unbiased, i.e., $\mathbb{E}[r_{k,\nu}^{DR(\check{\Theta})}] = \mathbb{E}[r_{k,\tau(\nu)}^{(j_{\tau(\nu)})}]$, for any given $\check{\Theta} \in \mathbb{R}^{J\cdot d}$. Adding the ℓ_2 regularization norm and solving (2) leads to the DR estimator:

$$\left(\sum_{\nu=1}^{n} \sum_{k=1}^{K} \tilde{X}_{k,\nu} \tilde{X}_{k,\nu}^\top + I_{J\cdot d} \right)^{-1} \left(\sum_{\nu=1}^{n} \sum_{k=1}^{K} \tilde{X}_{k,\nu} r_{k,\tau(\nu)}^{DR(\check{\Theta})} \right).$$

The main advantage of the DR estimator is that it uses contexts from *all* K actions. However, in our policy, some $\pi_{k,\tau(\nu)}^{(j_{\tau(\nu)})}$ can be zero, and therefore, the pseudo-reward (3) is not defined. To handle this problem, we propose to introduce a random variable. After taking an action at round $\tau(\nu)$

and observing the selected action $a_{\tau(\nu)}$, the decision-maker samples h_ν from the distribution:

$$\phi_{k,\nu} := \mathbb{P}\left(h_\nu = k \middle| \mathcal{H}_{\tau(n)}\right) = \begin{cases} 1 - \frac{16(K-1)\log\left(\frac{dJ}{\delta}\right)}{\lambda_{\min}(F_\nu)} & k = a_{\tau(\nu)} \\ \frac{16\log\left(\frac{dJ}{\delta}\right)}{\lambda_{\min}(F_\nu)} & k \neq a_{\tau(\nu)} \end{cases} \tag{4}$$

where $F_\nu := \sum_{i,k=1}^{\nu,K} \tilde{X}_{k,i}\tilde{X}_{k,i}^\top + 16d(K-1)\log\left(\frac{dJ}{\delta}\right) I_{J\cdot d}$ is the Gram matrix of contexts from ν admitted rounds and $\delta \in (0,1)$ is the confidence level. We would like to emphasize that h_ν is sampled after observing the actions $a_{\tau(\nu)}$ and does not affect the policies until round $\tau(\nu)$.

Sampling the random variables h_ν after choosing actions is motivated by Kim et al. (2023) which uses bootstrap methods (Efron & Tibshirani, 1994) and resampling methods (Good, 2006). To obtain the unbiased pseudo-rewards similar to (3), we resample the action from another distribution with non-zero probabilities. The probabilities $\{\phi_{k,\nu} : k \in [K]\}$ are designed to control the level of exploration and exploitation for future rounds based on the ratio of confidence level to the number of admitted rounds. When the minimum eigenvalue of F_ν is small compared to $\log(1/\delta)$, the distribution of h_ν is less concentrated on $a_{\tau(\nu)}$ and tends to explore other actions. As ν increases, the probabilities $\{\phi_{k,\nu} : k \in [K]\}$ concentrates on $a_{\tau(\nu)}$, and the decision-maker tends to exploit.

Since we obtain non-zero probabilities $\{\phi_{k,\nu} : k \in [K], \nu \in [n]\}$, we define novel unbiased pseudo-rewards:

$$\tilde{r}_{k,\nu} := \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} + \left\{1 - \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}}\right\} \tilde{X}_{k,\nu}^\top \check{\Theta}_n, \tag{5}$$

where the imputation estimator $\check{\Theta}_n$ is an IPW estimator with new probabilities:

$$\begin{aligned} \check{\Theta}_t &:= A_n^{-1}\left\{\sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \tilde{X}_{k,\nu} r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu} r_{a_{\tau(\nu)},\tau(\nu)}^{(j_{\tau(\nu)})}\right\}, \\ A_n &:= \sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu}^\top + I_{J\cdot d}, \\ \Psi_n &:= \left\{\nu \in [n] : h_\nu = a_{\tau(\nu)}\right\}. \end{aligned}$$

The set Ψ_n is introduced because we cannot observe $\frac{\mathbb{I}(h_\nu=k)}{\phi_{k,\nu}} r_{k,\tau(\nu)}^{(j_{\tau(\nu)})}$ in case of $h_\nu \neq a_{\tau(\nu)}$. In other words, we use the pseudo-rewards in (5) only at the rounds that satisfy $h_\nu = a_{\tau(\nu)}$. Then our estimator with n admitted samples is defined as

$$\begin{aligned} \widehat{\Theta}_n &:= V_n^{-1}\left\{\sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \tilde{X}_{k,\nu}\tilde{r}_{k,\nu} + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu} r_{a_{\tau(\nu)},\tau(\nu)}^{(j_{\tau(\nu)})}\right\} \\ V_n &:= \sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu}^\top + I_{J\cdot d}. \end{aligned} \tag{6}$$

Analogous to the construction of (6), we can also define the estimator for the resource consumption parameters $\{W_\star^{(j)} : j \in [J]\}$,

$$\widehat{\mathbf{W}}_n := V_n^{-1}\left[\sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \tilde{X}_{k,\nu}\tilde{\mathbf{b}}_{k,\nu}^\top + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu}\mathbf{b}_{a_{\tau(\nu)}\tau(\nu)}^{(j_{\tau(\nu)})\top}\right], \tag{7}$$

where the pseudo-consumption vectors and the imputation estimator are

$$\begin{aligned} \tilde{\mathbf{b}}_{k,\nu} &:= \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \mathbf{b}_{a_{\tau(\nu)},\tau(\nu)}^{(j_{\tau(\nu)})} + \left\{1 - \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}}\right\} \check{\mathbf{W}}_n^\top \tilde{X}_{k,\nu}, \\ \check{\mathbf{W}}_n &:= A_n^{-1}\left[\sum_{\nu\in\Psi_n}\sum_{k=1}^{K} \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \tilde{X}_{k,\nu}\left\{\mathbf{b}_{k,\nu}^{(j_{\tau(\nu)})}\right\}^\top + \sum_{\nu\notin\Psi_n} \tilde{X}_{a_{\tau(\nu)},\nu}\left\{\mathbf{b}_{a_{\tau(\nu)},\tau(\nu)}^{(j_{\tau(\nu)})}\right\}^\top\right]. \end{aligned}$$

The two estimators use the novel Gram matrix V_n defined in (6) consisting of contexts from *all* K actions. Now, we present estimation error bounds normalized by the novel Gram matrix V_n.

Theorem 4.1. *(Self-normalized bound for the estimator) Suppose Assumptions 1 and 2 hold. For each $t \in [T]$, let n_t denote the number of admitted arrivals until round t and $\Psi_{n_t} := \{\nu \in [n_t] : h_\nu = a_{\tau(\nu)}\}$, where h_ν is defined in (4). Suppose $F_{n_t} := \sum_{\nu=1}^{n_t}\sum_{k=1}^{K} \tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top + 16d(K-1)\log\frac{Jd}{\delta} I_{J\cdot d}$ satisfies*

$$\lambda_{\min}(F_{n_t}) \geq 4Kd\left\{\sum_{\nu=1}^{n_t} \frac{144\,(K-1)\log\left(\frac{Jd}{\delta}\right)}{\lambda_{\min}(F_\nu)} + 35\log\frac{Jd}{\delta}\right\}, \tag{8}$$

for $\delta \in (0,1)$. For each $r \in [m]$, let $\widehat{\mathbf{W}}_{n_t,r}$ and $\mathbf{W}_{\star,r}$ be the r-th column of $\widehat{\mathbf{W}}_{n_t}$ and $\mathbf{W}_\star$, respectively. Denote $\beta_\sigma(\delta) := 8\sqrt{Jd} + 96\sigma\sqrt{Jd\log\frac{4}{\delta}}$. Then with probability at least $1 - 4(m+1)\delta$,

$$\begin{aligned} \left\|\widehat{\Theta}_{n_t} - \Theta^*\right\|_{V_{n_t}} &\leq \beta_{\sigma_r}(\delta), \\ \max_{r\in[m]} \left\|\widehat{\mathbf{W}}_{n_t,r} - \mathbf{W}_{\star,r}\right\|_{V_{n_t}} &\leq \beta_{\sigma_b}(\delta). \end{aligned} \tag{9}$$

Compared to the self-normalized bound in Abbasi-Yadkori et al. (2011) uses the Gram matrix consisting of selected contexts only, our bounds are normalized by V_{n_t} This change in the Gram matrix enables us to develop a fast convergence rate. The condition (8) is required for the eigenvalues of the Gram matrix F_{n_t} to be large so that the probability $\phi_{a_{\tau(\nu)},\nu}$ is large and the estimators use the pseudo rewards and pseudo consumption vectors for most of the rounds. We show in Lemma 5.3 that the condition (8) requires at most rounds logarithmic in T, and does not affect the main order of the regret bound.

Using the novel estimators, we define the estimates for utility and resource consumption. Denote $\mathcal{C}_t^{(j)} := \{s \in [t] : j_s = j\}$ and

$$\begin{aligned}\widehat{u}_{k,t}^{(j)} :=& \left|\mathcal{C}_t^{(j)}\right|^{-1} \sum_{s\in\mathcal{C}_t^{(j)}} \left\{\widehat{\theta}_{t-1}^{(j)}\right\}^\top \mathbf{x}_{k,s}^{(j)},\\ \widehat{\mathbf{b}}_{k,t}^{(j)} :=& \left|\mathcal{C}_t^{(j)}\right|^{-1} \sum_{s\in\mathcal{C}_t^{(j)}} \left\{\widehat{W}_{t-1}^{(j)}\right\}^\top \mathbf{x}_{k,s}^{(j)}.\end{aligned} \tag{10}$$

The estimates (10) use the average of contexts in the same class to estimate the expected value over the context distribution. In this way, the decision-maker effectively uses previous contexts in all rounds including the *skipped rounds*. Next, we establish a convergence rate for the estimators $\widehat{u}_{k,t}^{(j)}$ and $\widehat{\mathbf{b}}_{k,t}^{(j)}$.

Theorem 4.2. *(Convergence rate for the estimates) Suppose Assumptions 1-3 hold. Denote the expected utility* $u_k^{\star(j)} := \mathbb{E}_{\mathbf{x}_k\sim\mathbb{F}_j}\left[\left\{\theta_\star^{(j)}\right\}^\top \mathbf{x}_k\right]$ *and consumption* $\mathbf{b}_k^{\star(j)} := \mathbb{E}_{\mathbf{x}_k\sim\mathbb{F}_j}\left[\left\{W_\star^{(j)}\right\}^\top \mathbf{x}_k\right]$. *Set* $\gamma_{t,\sigma}(\delta) := \frac{16\sqrt{J\log(JKT)}}{\sqrt{t}} + \frac{6\beta_\sigma(\delta)}{\sqrt{n_t}}$, *where* n_t *is the number of admitted arrivals until round* t *and* $\beta_\sigma(\delta)$ *is defined in Theorem 4.1. Suppose* $t \geq 8d\alpha^{-1}p_{\min}^{-1}\log JT$, $\delta \in (0, T^{-1})$ *and* F_{n_t} *satisfies* (8). *Then with probability at least* $1 - 4(m+1)\delta - 7T^{-1}$,

$$\begin{aligned}&\sqrt{\sum_{j=1}^J p_j \max_{k\in[K]}\left|u_k^{\star(j)} - \widehat{u}_{k,t+1}^{(j)}\right|^2} \leq \gamma_{t,\sigma_r}(\delta),\\ &\sqrt{\sum_{j=1}^J p_j \max_{k\in[K]}\left\|\mathbf{b}_k^{\star(j)} - \widehat{\mathbf{b}}_{k,t+1}^{(j)}\right\|_\infty^2} \leq \gamma_{t,\sigma_b}(\delta).\end{aligned} \tag{11}$$

The convergence rate of the estimates is $\tilde{O}(\sqrt{Jd}n_t^{-1/2})$. In deriving the fast rate, the novel Gram matrix V_{n_t} plays a significant role. To prove Theorem 4.2, we bound the sum of squared maximum prediction error as follows:

$$\begin{aligned}&\frac{1}{n_t}\sum_{s\in\Psi_{n_t}} \max_{k\in[K]}\left\{\left(\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right\}^2\\ &= \frac{1}{n_t}\sum_{s\in\Psi_{n_t}} \max_{k\in[K]}\left(\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right)\left(\mathbf{x}_{k,s}^{(j)}\mathbf{x}_{k,s}^{(j)}\right)^\top\left(\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right)\\ &\leq \frac{1}{n_t}\sum_{s\in\Psi_{n_t}}\left(\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right)\left(\sum_{k=1}^K \mathbf{x}_{k,s}^{(j)}\mathbf{x}_{k,s}^{(j)}\right)^\top\left(\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right)\\ &\leq \frac{1}{n_t}\left\|\theta_\star^{(j)} - \widehat{\theta}_t^{(j)}\right\|_{V_{n_t}}^2.\end{aligned}$$

Such a bound is not available if the Gram matrix is constructed using only contexts corresponding to selected actions. In this way, we obtain a faster convergence rate for the estimates for utility and consumption vectors.

4.2. Proposed Algorithm

Let $(K+1)$-th action denote skipping the arrival and $\pi_{K+1,t}^{(j)} := \mathbb{P}\left(\text{Skip the round } t \middle| \mathcal{H}_t\right)$ denote the probability of skipping the arrival. Since the decision-maker must choose an action or skip the round, we have $\sum_{k=1}^{K+1}\pi_{k,t}^{(j)} = 1$. When the decision-maker skips round t, we set $\mathbf{x}_{K+1,t}^{(j)} := 0$ and $\mathbf{b}_{K+1,t}^{(j)} := 0$. In round t, the randomized bandit policy is given by the optimal solution of the following optimization problem:

$$\begin{aligned}\max_{\pi_{k,t}^{(j_t)}} &\sum_{k=1}^{K+1}\pi_{k,t}^{(j_t)}\left(\widehat{u}_{k,t}^{(j_t)} + \frac{\gamma_{t-1,\sigma_r}(\delta)}{\sqrt{p_{j_t}}}\mathbb{I}\left(k\in[K]\right)\right),\\ \text{s.t.} &\sum_{k=1}^{K+1}\pi_{k,t}^{(j_t)}\left(\widehat{\mathbf{b}}_{k,t}^{(j_t)} - \frac{\gamma_{t-1,\sigma_b}(\delta)}{\sqrt{p_{j_t}}}\mathbf{1}_m\right) \leq \rho_t \vee 0,\\ &\sum_{k=1}^{K+1}\pi_{k,t}^{(j_t)} = 1,\\ &\pi_{k,t}^{(j_t)} \geq 0, \quad \forall k\in[K+1],\end{aligned} \tag{12}$$

where $\rho_t := t\rho - \sum_{s=1}^{t-1}\mathbf{b}_{a_s,s}^{(j_s)}$ is the difference between the used resources and planned budget until round t. The algorithm is optimistic in that it uses upper confidence bound (UCB) in rewards and lower confidence bound (LCB) in consumption while it regulates the consumption to be less than $t\rho$ with ρ_t. In this way, the problem (12) balances between admitting the arrivals and saving the resources for later use. Next, we show that the optimal solution (12) is available in a closed form.

Lemma 4.3. *(Optimal policy for bandit) Let* $\tilde{u}_{k,t}^{(j_t)} := \widehat{u}_{k,t}^{(j_t)} + p_{j_t}^{-1/2}\gamma_{t-1,\sigma_r}(\delta)\mathbb{I}\left(k\in[K]\right)$ *and* $\tilde{b}_{k,t}^{(j_t)}(r) := \widehat{b}_{k,t}^{(j_t)}(r) - p_{j_t}^{-1/2}\gamma_{t-1,\sigma_b}(\delta)$, *for* $r\in[m]$. *For* $i\in[K+1]$, *let* $\tilde{u}_{k\langle i\rangle,t}^{(j_t)}$

Algorithm 1 Allocate to the Maximum First algorithm (AMF)

INPUT: confidence lengths $\gamma_\theta, \gamma_b > 0$, confidence level $\delta \in (0,1)$.
Initialize $F_0 := 16d(K-1)\log\frac{Jd}{\delta} I_{J\cdot d}$, $\rho_1 := \rho$, $\widehat{\Theta}_0 := \mathbf{0}_{J\cdot d}$, $\widehat{\mathbf{W}}_0 := \mathbf{0}_{J\cdot d\times m}$
for $t = 1$ **to** T **do**
  Observe arrival $(j_t, \{\mathbf{x}_{k,t}^{(j_t)}\}_{k\in[K]})$.
  if F_{t-1} does not satisfy (8) **then**
    Take action $a_t = \arg\max_{k\in[K]} \rho \|\widehat{\mathbf{b}}_{k,t}^{(j_t)}\|_\infty^{-1}$.
  else
    Compute $\widehat{u}_{k,t}^{(j_t)}$ and $\widehat{\mathbf{b}}_{k,t}^{(j_t)}$ with $\widehat{\theta}_{t-1}^{(j_t)}$ and $\widehat{\mathbf{W}}_{t-1}^{(j_t)}$.
    Compute $\tilde{u}_{k,t}^{(j_t)} := \widehat{u}_{k,t}^{(j_t)} + \frac{\gamma_\theta}{\sqrt{p_{j_t} n_{t-1}}}$ and $\tilde{\mathbf{b}}_{k,t}^{(j_t)} := \widehat{\mathbf{b}}_{k,t}^{(j_t)} - \frac{\gamma_b}{\sqrt{p_{j_t} n_{t-1}}}\mathbf{1}_m$.
    Take action a_t with the policy $\widehat{\pi}_{1,t}^{(j_t)}, \dots, \widehat{\pi}_{K+1,t}^{(j_t)}$ defined in (13).
  end if
  if $a_t \in [K]$ **then**
    Observe $r_{a_t,t}^{(j_t)}$ and $\mathbf{b}_{a_t,t}^{(j_t)}$, then estimate $\widehat{\Theta}_t$ and $\widehat{\mathbf{W}}_t$ as in (6) and (7), respectively.
    Update $F_t = F_{t-1} + \sum_{k=1}^K \tilde{X}_{k,t}\tilde{X}_{k,t}^\top$.
  end if
  Update available resource $\rho_{t+1} = \rho_t + \rho - \mathbf{b}_{a_t,t}^{(j_t)}$.
  if $\sum_{s=1}^t \mathbf{b}_{a_s,s}^{(j_s)} \geq T\rho$ **then**
    Exit
  end if
end for

be a sequence of ordered variables of $\tilde{u}_{k,t}^{(j_t)}$ in decreasing order, i.e. $\tilde{u}_{k\langle 1\rangle,t}^{(j_t)} \geq \tilde{u}_{k\langle 2\rangle,t}^{(j_t)} \geq \cdots \geq \tilde{u}_{k\langle K+1\rangle t}^{(j_t)}$. When there is a tie between $\tilde{u}_{k\langle i\rangle,t}^{(j_t)}$ and $\tilde{u}_{k\langle i+1\rangle,t}^{(j_t)}$, the index $k\langle i\rangle$ with the higher value for

$$\left(\min_{r\in[m]} \frac{\rho_t(r) \vee 0 - \sum_{h=1}^{i-1} \widehat{\pi}_{k\langle h\rangle,t}^{(j_t)} \tilde{b}_{k\langle h\rangle,t}^{(j_t)}(r)}{\tilde{b}_{k\langle h\rangle,t}^{(j_t)}(r)}\right)$$

goes first. Then the policy defined as,

$$\begin{aligned}
\widehat{\pi}_{k\langle 1\rangle,t}^{(j_t)} &= \left(\min_{r\in[m]} \frac{\rho_t(r)\vee 0}{\tilde{b}_{k\langle 1\rangle,t}^{(j_t)}(r)}\right) \wedge 1, \\
\widehat{\pi}_{k\langle i\rangle,t}^{(j_t)} &= \left(\min_{r\in[m]} \frac{\rho_t(r)\vee 0 - \sum_{h=1}^{i-1} \widehat{\pi}_{k\langle h\rangle,t}^{(j_t)} \tilde{b}_{k\langle h\rangle,t}^{(j_t)}(r)}{\tilde{b}_{k\langle i\rangle,t}^{(j_t)}(r)}\right) \\
&\quad \wedge \left(1 - \sum_{h=i}^{i-1} \widehat{\pi}_{k\langle h\rangle,t}^{(j_t)}\right), \forall i \in [2, K+1],
\end{aligned} \tag{13}$$

is the optimal solution to (12).

Since the objective function of (12) is linear, we can obtain the maximum value by permuting the objective coefficients in decreasing order and allocating the greatest possible probability value in decreasing order of the objective coefficients. Note that $\widehat{\pi}_{k,t}^{(j_t)}$ is automatically set to zero when the utility is negative. This is because of the probability of skipping the arrival, $\widehat{\pi}_{K+1,t}^{(j_t)} = 1 - \sum_{h=1}^{l-1} \widehat{\pi}_{k\langle h\rangle,t}^{(j_t)}$, when $\tilde{u}_{K+1,t}^{(j_t)}$ is the l-th largest weighted utility function and all the remaining probability is allocated to $\widehat{\pi}_{K+1,t}^{(j_t)}$. Therefor, the probabilities for actions k with $\tilde{u}_{k,t}^{(j_t)} < \tilde{u}_{K+1,t}^{(j_t)} := 0$ are all zero.

Our proposed algorithm, Allocate to the Maximum First (AMF) is presented in Algorithm 1. The algorithm first explores with the least consumption action until the eigenvalue condition for the estimator (8) holds. In each round of exploration, the Gram matrix of all actions is added to F_{n_t}, and any choice of action increases the eigenvalue of F_{n_t}. Once the condition (8) holds, the algorithm solves the problem (12) by computing the closed-form policy (13). The computational complexity of our algorithm is $\tilde{O}(d^3mKT + Jd^3T)$ where the main order occurs from updating the estimators and computing the eigenvalues of J symmetric positive-definite matrix F_{n_t}. Note that computing estimators does not depend on J because the algorithm updates only j_t-th variables for each $t \in [T]$.

5. Regret Analysis

In this section, we present our regret bound and regret analysis for the proposed AMF algorithm.

Theorem 5.1. *(Regret bound of AMF) Suppose Assumptions 1-3 hold. Let $M_{\alpha,p,T} := 2304\alpha^{-2}p_{\min}^{-2}\log T + 280\alpha^{-1}p_{\min}^{-1}$ and $C_\sigma(\delta) := 96 + 1152\sigma\sqrt{\log\frac{4}{\delta}}$. Suppose T and ρ satisfies $T \geq 8d\alpha^{-1}p_{\min}^{-1}\log JdT$, and $\rho \geq \sqrt{Jd/T}$. Setting $\gamma_\theta = 16\sqrt{J\log JKT} + 6\beta_{\sigma_r}(\delta)$ and $\gamma_b = 16\sqrt{J\log JKT} + 6\beta_{\sigma_b}(\delta)$, the regret bound of AMF is*

$$\begin{aligned}
\mathcal{R}_T^{\widehat{\pi}} \leq \left(2 + \frac{OPT}{\rho T}\right)\Bigg\{ &\frac{4d\log JdT}{\alpha p_{\min}} + 2dM_{\alpha,p,T}\log\frac{Jd}{\delta} + 15 \\
&+ \left(96\sqrt{\log JKT} + 3C_{\sigma_r\vee\sigma_b}(\delta)\right)\sqrt{JdT\log T} + 10mT^3\delta\Bigg\}.
\end{aligned}$$

For $\delta \in (0, m^{-1}T^{-3})$, the regret bound is

$$\mathcal{R}_T^{\widehat{\pi}} = O\left(\frac{OPT}{T\rho}\sqrt{JdT\log mJKT}\log T\right). \tag{14}$$

The regret bound (14) holds when the hyperparameter $\delta = m^{-1}T^{-3}$, which requires the knowledge of T. However, in practice, selecting another value of δ does not affect the performance of the algorithm. We provide the discussion on the sensitivity to the hyperparameter choice in Section 6.3.

Setting $B = T\rho$, the main term of the regret bound is $\tilde{O}(OPT/B\sqrt{JdT})$ for $B = \Omega(\sqrt{JdT})$. The sublinear

dependence of the regret bound on J, d, and T is a direct consequence of the improved $\tilde{O}(\sqrt{Jd/n_t})$ convergence rate for the parameter estimates. Agrawal & Devanur (2016) establish a regret bound $\mathcal{R}_T^\pi = \tilde{O}(OPT/B \cdot d\sqrt{T})$ for the LinCBwK when $B = \Omega(\sqrt{d}T^{3/4})$. Our bound for LMMP (which subsumes LinCBwK as a special case) is improved by a $\sqrt{d}$ factor and is valid under budget constraints that relaxed from $\Omega(\sqrt{d}T^{\frac{3}{4}})$ to $\Omega(\sqrt{d}T^{\frac{1}{2}})$.

For the proof of the regret bound, we first present the lower bound of the reward obtained by our algorithm.

Lemma 5.2. *Let $\tilde{u}_{k,t}^{(j)}$ and $\widehat{\mathbf{b}}_{k,t}^{(j)}$ be the estimates defined in (10). Denote $\widehat{\pi}$ the policy of AMF. Define the good events,*

$$\begin{aligned} \mathcal{E}_t &:= \left\{ \tilde{u}_{k,t}^{(j)} \text{ and } \widehat{\mathbf{b}}_{k,t}^{(j)} \text{ satisfies (11).} \right\}, \\ \mathcal{M}_t &:= \left\{ F_{n_t} \text{ satisfies (8).} \right\}, \end{aligned} \tag{15}$$

and $\mathcal{G}_t := \mathcal{E}_t \cap \mathcal{M}_{t-1}$. Let τ be the stopping time for the algorithm and $\xi := \inf_{t\in[T]} \{\mathcal{M}_{t-1} \cap \{\rho_t > 0\}\}$ be the starting time after the exploration for condition (8). Then, the total reward

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t^{\widehat{\pi}}\right] \geq \frac{OPT}{T}\mathbb{E}\left[\tau - \xi\right] - \left(2 + \frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}(\mathcal{G}_t^c)$$
$$-2\left(1+\frac{OPT}{\rho T}\right)\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r\vee\sigma_b}(\delta)^2\mathbb{I}\left(a_t \in [K]\right)\right]}.$$

The lower bound consists of three main terms. The first term $\frac{OPT}{T}\mathbb{E}\left[\tau - \xi\right]$ relates to the period for which the algorithm uses the optimal policy (13). The second term $(2+\frac{OPT}{\rho T})\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right)$ is the sum of the probability of bad events $\mathcal{M}_{t-1}^c$ over which the minimum eigenvalue of the Gram matrix F_{n_t} is not large enough for the fast convergence rate, and the event $\mathcal{E}_t^c$ over which the estimator goes out of the confidence interval. And, the third term consists of the sum of confidence lengths for the reward and consumption.

The following result bounds τ, ξ and the sum of bad events $\{\mathcal{M}_t^c : t \in [T]\}$.

Lemma 5.3. *Suppose Assumptions 1-3 holds and $\rho > \sqrt{Jd/T}$. Let $M_{\alpha,p,T}$ and $\gamma_{t,\sigma}(\delta)$ denote the variables defined in Theorem 5.1 and Theorem 4.2, respectively. Then, for any $\delta \in (0, 1/T^2)$, the starting time $\xi := \inf_{t\in[T]}\{\mathcal{M}_{t-1} \cap \{\rho_t > 0\}\}$ and the stopping time τ of the AMF algorithm is bounded as*

$$\mathbb{E}\left[\xi\right] \leq \frac{1 + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right) + T^2\delta}{\rho} + 1,$$
$$\mathbb{E}[T - \tau] \leq \frac{4(m+1)T\delta + 7 + 2\gamma_{1,\sigma_b}(\delta)}{\rho},$$

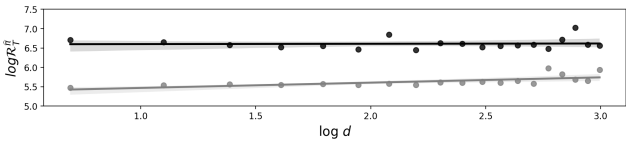
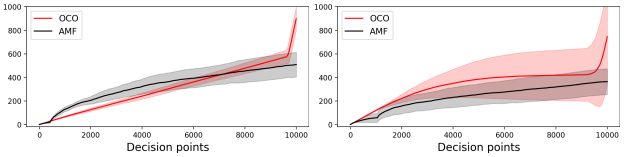
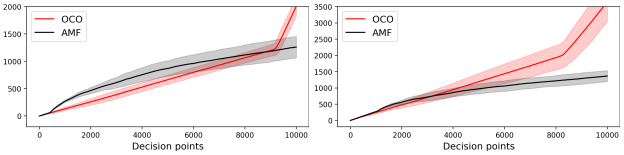


Figure 1. Logarithm of cumulative regret of the proposed AMF algorithm on various dimension d when the per-period budget is $\rho = \sqrt{d/T}$. The gray (resp. black) line is the best fit line on the points when $T = 5000$ (resp. $T = 20000$).



(a) Regret comparison with budget $B = \sqrt{d}T^{3/4}$



(b) Regret comparison with budget $B = \sqrt{dT}$

Figure 2. Regret of AMF and OCO algorithms for $K = 20$ and $m = 20$. The line and shade represent the average and standard deviation based on 20 independent experiments. Additional results on different K and m are in Section A.3.

and for $\mathcal{M}_t$ defined as in (15),

$$\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{M}_{t-1}^c\right) \leq T^2\delta + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right).$$

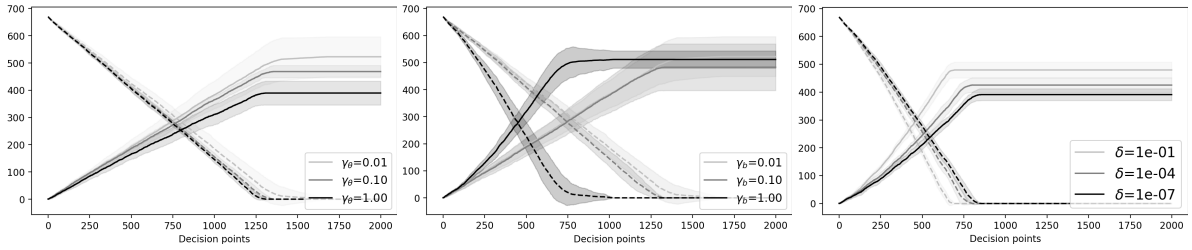
The regret bound follows from bounding the probability of $\mathcal{E}_t^c$ with Theorem 4.2 and showing that the sum of square of $\gamma_{t,\sigma}(\delta)$ is $O(Jd\log T)$. The bound holds because the summation of $\gamma_{t,\sigma}(\delta)^2 = \tilde{O}(\frac{Jd}{n_t})$ over the rounds that $a_t \in [K]$ happens is $\sum_{n=1}^{n_T} O(Jd/n) = O(Jd\log T)$.

6. Numerical Results

We demonstrate the cumulative regrets with given budgets (Section 6.1 and 6.2) and the sensitivity of our proposed AMF to the hyperparameter choice (Section 6.3). For the computation of the regret, we use a setting where the instantaneous regret is computable for each round. (For the details of the setting, see Appendix A.1.)

6.1. Regret $\mathcal{R}_T^{\widehat{\pi}}$ as a function of d

Figure 1 plots $\log(\mathcal{R}_T^{\widehat{\pi}})$ vs. $\log(d)$ for a single-class ($J = 1$) LMMP for $T = \{5000, 20000\}$ and the budget $B = \sqrt{dT}$, where our $\tilde{O}(\frac{OPT}{B}\sqrt{JdT})$ regret bound implies that $\log(\mathcal{R}_T^{\widehat{\pi}})$ is constant over d. The regression line on the plot is nearly flat and the slope of the best-fit line is 0.136



(a) On various γ_θ (b) On various γ_b (c) On various δ

Figure 3. The reward and inventory of AMF on various hyperparameters γ_θ, γ_b and δ. The solid (resp. dashed) line represents the reward (resp. inventory). The line and shade represent the average and standard deviation based on 10 repeated experiments, respectively.

(resp. 0.008) for $T = 5000$ (resp. $T = 20000$). The weak increase in $T = 5000$ is captured by the $O(d \log JdmT)$ term in our bound, which diminishes for large T.

6.2. Comparison of AMF with OCO

In order to compare AMF with OCO (Agrawal & Devanur, 2016), we set $J = 1$. The hyperparameters for AMF were set to $\gamma_\theta = 1$, $\gamma_b = 1$ and $\delta = 0.01$.

Figure 2(a) (resp. (b)) plots the cumulative regret of the two algorithms with budget $B = \sqrt{d}T^{\frac{3}{4}}$ (resp. $B = \sqrt{dT}$). Note that OCO requires a minimum budget $B = \sqrt{d}T^{\frac{3}{4}}$ whereas AMF requires a lower minimum budget of $B = \sqrt{dT}$. The regret lines cross because AMF is allowed to skip arrivals whereas OCO does not skip arrivals. The sudden bend points at the end of the round in OCO show that it runs out of budget and has regret = 1. In all cases, our algorithm performs better and the performance gap increases as d increases. Note that the regret plot for OCO never flattens out for most cases, where the regret of AMF flattens as t increases. This is because our new estimator which uses contexts from *all* actions with unbiased pseudo-rewards (5) for unselected actions and has a significantly faster convergence rate as compared with the estimator used in OCO.

6.3. Sensitivity Analysis

We demonstrate the sensitivity of the proposed AMF algorithm to its three hyperparameters: γ_θ, γ_b, and δ. Figure 3(a) and 3(b) show the reward and inventory of our algorithm on various $\gamma_\theta \in \{0.01, 0.1, 1\}$ and $\gamma_b \in \{0.01, 0.1, 1\}$. We present on these sets since the variability of the reward and the inventory of the algorithm are hardly visible outside the sets. Figure 3(c) shows the reward and inventory of AMF on various $\delta \in \{10^{-1}, 10^{-4}, 10^{-7}\}$. When $\delta \geq 10^{-1}$ (resp. $\delta \leq 10^{-7}$) the reward and inventories are same with $\delta = 10^{-1}$ (resp. $\delta = 10^{-7}$). The change in the reward and the consumption of the proposed AMF is visible only when the hyperparameters change drastically. This shows that choice of hyperparameters is not sensitive. The effect of γ_θ and γ_b diminishes fast by $n_t^{-1/2}$ term and our policy finds the order of the utilities rather than their absolute values. For δ, which controls the sampling probabilities (4) in estimators and the exploration rounds in (8), it also has a small effect. This is because the minimum eigenvalue of F_{n_t} increases in $\Omega(n_t)$-rate and reduces the effect of $\log \frac{1}{\delta}$ terms in (4) and (8). Therefore, our algorithm guarantees a similar performance for other hyperparameters than specified in Theorem 5.1. For details of the experimental settings and recommendation of the specific hyperparameter choices, see Appendix A.2.

7. Conclusion

We introduce a new problem class LMMP that extends upon LinCBwK and online revenue management to a multi-class setting. To address this problem, we propose a novel estimator that utilizes unbiased pseudo-rewards and contexts of *all* actions to learn class-specific parameters of all classes. We use this transfer-learning-based estimator to propose an algorithm in which the policy is available in closed form and the worst-case regret is $\tilde{O}(\frac{OPT}{B}\sqrt{JdT})$ when the budget $B(T) = \Omega(\sqrt{JdT})$. This result improves both the regret bound and the minimum budget required, resolving an open problem in LinCBwK. Numerical experiments demonstrate superior performances over benchmarks for the single-class case, and robustness to hyperparameters changes for multiple-class data.

Acknowledgements

We thank the anonymous referees for offering many helpful comments and valuable feedback. Wonyoung Kim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00240142) and Garud Iyengar was supported by NSF EFMA-2132142, ARPA-E PERFORM Program, and ONR N000142312374.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014a.

Agrawal, S. and Devanur, N. R. Fast algorithms for online stochastic convex programming. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1405–1424. SIAM, 2014b.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pp. 9252–9262, 2019.

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

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

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

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

Besbes, O. and Zeevi, A. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Besbes, O. and Zeevi, A. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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, 2011.

Devanur, N. R., Jain, K., Sivan, B., and Wilkens, C. A. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 29–38, 2011.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3445–3453, 2019.

Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V. S., and Stein, C. Online stochastic packing applied to display ad allocation. In *European Symposium on Algorithms*, pp. 182–194. Springer, 2010.

Ferreira, K. J., Simchi-Levi, D., and Wang, H. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

Gallego, G. and Van Ryzin, G. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

Good, P. I. *Resampling methods*. Springer, 2006.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219. IEEE, 2019.

Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.

Kim, G. and Paik, M. C. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pp. 5869–5879, 2019.

Kim, W., Kim, G.-S., and Paik, M. C. Doubly robust thompson sampling with linear payoffs. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Kim, W., Lee, K., and Paik, M. C. Double doubly robust thompson sampling for generalized linear contextual bandits. *arXiv preprint arXiv:2209.06983*, 2022.

Kim, W., Paik, M. C., and Oh, M.-H. Squeeze all: Novel estimator and self-normalized bound for linear contextual bandits. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3098–3124. PMLR, 25–27 Apr

2023. URL https://proceedings.mlr.press/v206/kim23d.html.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, J. R., Peres, Y., and Smart, C. K. A gaussian upper bound for martingale small-ball probabilities. *Ann. Probab.*, 44(6):4184–4197, 11 2016. doi: 10.1214/15-AOP1073.

Li, X., Sun, C., and Ye, Y. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6483–6492. PMLR, 18–24 Jul 2021.

Liu, X., Li, B., Shi, P., and Ying, L. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021. doi: 10.1109/TSP.2021.3089822.

Oh, M.-h., Iyengar, G., and Zeevi, A. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pp. 8271–8280. PMLR, 2021.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2827–2835. PMLR, 13–15 Apr 2021.

Sankararaman, K. A. and Slivkins, A. Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34:23191–23204, 2021.

Sivakumar, V., Wu, S., and Banerjee, A. Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *International Conference on Machine Learning*, pp. 9026–9035. PMLR, 2020.

Sivakumar, V., Zuo, S., and Banerjee, A. Smoothed adversarial linear contextual bandits with knapsacks. In *International Conference on Machine Learning*, pp. 20253–20277. PMLR, 2022.

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

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

A. Supplementary for Experiments

A.1. Settings of Parameters and Contexts for Regret Computation

For numerical experiments, we devise a setting where explicit regret computation is available. We set $J = 1$ for OCO to be compatible with the setting. For $x \in \mathbb{R}_+$, let $\lceil x \rceil$ be the smallest integer greater than equal to x. For parameters, we set $\theta_\star = (-1, \cdots, -1, \lceil d/2 \rceil^{-1}, \cdots, \lceil d/2 \rceil^{-1})$ and

$$W_\star = \begin{pmatrix} \rho\lceil d/2 \rceil^{-1} & \cdots & \rho\lceil d/2 \rceil^{-1} \\ \vdots & \cdots & \vdots \\ \rho\lceil d/2 \rceil^{-1} & \cdots & \rho\lceil d/2 \rceil^{-1} \\ \rho & \cdots & \rho \\ \vdots & \vdots & \vdots \\ \rho & \cdots & \rho \end{pmatrix},$$

where the $\lceil d/2 \rceil^{-1}$ and $\rho\lceil d/2 \rceil^{-1}$ terms are in the first $\lceil d/2 \rceil$ entries. For contexts, we set the optimal action by $(0, \cdots, 0, 1, \cdots, 1)$, and for other actions, we set $(U_{0,0.05}, \cdots, U_{0,0.05}, U_{-0.05,0}, \cdots, U_{-0.05,0})$,where $U_{a,b}$ is the Uniform random variable supported on $[a, b]$. Then we have the optimal arm with reward 1 and consumption ρ, while other arms have reward less than 1 and consumption more than ρ.

A.2. Experiment Settings for Sensitivity Analysis

The settings of the experiment in Section 6.3 is described as follows. The number of classes is $J = 3$ with a uniform prior $p = (1/3, 1/3, 1/3)^\top$ and every $d = 5$ elements of $K = 10$ contexts are generated from the uniform distribution on $[\frac{kj}{KJ} - 1, \frac{kj}{KJ} + 1]$ for $k \in [K]$ and $j \in [J]$. The costs are generated from the uniform distribution on $[\frac{k(J-j+1)-1}{KJ}, \frac{k(J-j+1)+1}{KJ}]$ for $k \in [K]$ and $j \in [J]$. Each element of $\theta_\star^{(j)}$ and $W_\star^{(j)}$ is generated from $U_{0,1}$ and fixed throughout the experiment. The generated rewards and consumption vectors are not truncated to one to impose greater variability, as our algorithm does not show apparent sensitivity on bounded rewards and consumption vectors. The algorithm consumes the budget faster than in previous experiments because the consumption vector is not bounded to 1.

Based on the experiments, we recommend using grid search on $\gamma_\theta \times \gamma_b \in [0, 1]^2$ to maximize the reward.

However, we recommend using $\delta = 0.1$, which is greater than the specified value in Theorem 5.1 for the algorithm to start using its policy in earlier rounds.

A.3. Additional Results on Regret Comparison.

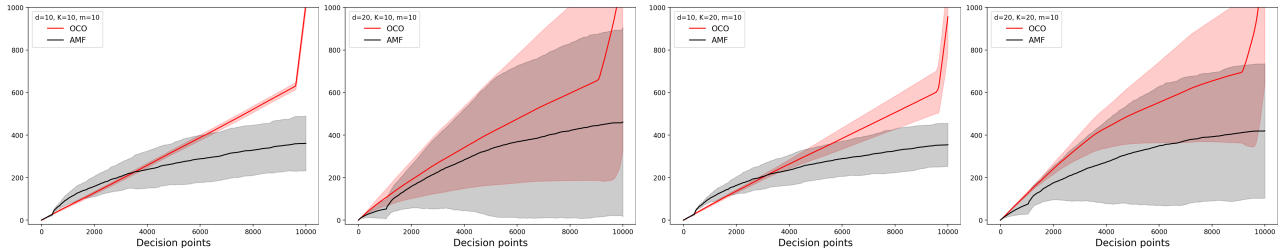
Figure 4 (a)-(d) show the regret comparison of AMF and OCO on different terms of $K = 10, 20$, $m = 10, 20$, and $B = dT^{3/4}$. Similar to the results in Figure 2(a), our algorithm has less regret than OCO in all cases, especially at the end of the rounds. The crossing line occurs when our algorithm skips in the middle round when $\rho_t < 0$ while OCO does not skip until the inventory runs out.

Figure 5 (a)-(d) show the regret of AMF and OCO algorithm on various $K = 10, 20$ and $m = 10, 20$ with budget $B = \sqrt{dT}$. Even in the smaller budget, our algorithm AMF does not run out the inventory and gains more reward than OCO. The gap of the performance tends to be larger than $B = \sqrt{d}T^{\frac{3}{4}}$ case.

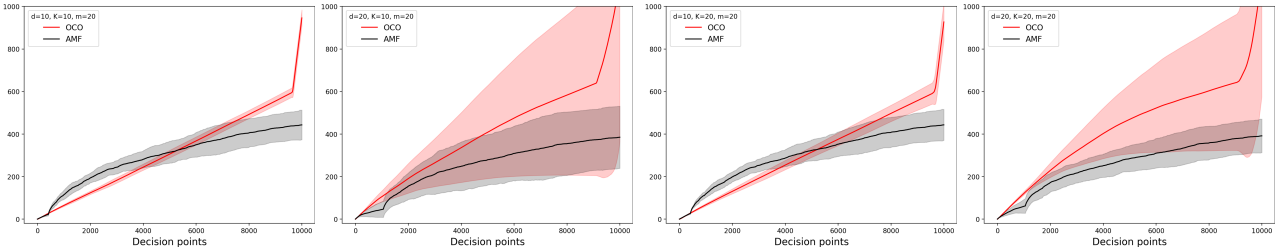
B. Missing Proofs

B.1. Proof of Theorem 4.1

Proof. Because the construction of $\widehat{\Theta}_t$ and $\widehat{\mathbf{W}}_t$ is the same, the bound for the $\widehat{\mathbf{W}}_t$ follows immediately from the bound for $\widehat{\Theta}_t$ by replacing $\{r^{(j_{\tau(\nu)})}_{a_{\tau(\nu)},\tau(\nu)} : \nu \in [n_t]\}$ with m entries of $\{\mathbf{b}^{(j_{\tau(\nu)})}_{a_{\tau(\nu)},\tau(\nu)} : \nu \in [n_t]\}$. Thus, it is sufficient to prove the bound for $\widehat{\Theta}_t$.



(a) Regret comparison under $K = 10$ and $m = 10$ (b) Regret comparison under $K = 20$ and $m = 10$



(c) Regret comparison under $K = 10$ and $m = 20$ (d) Regret comparison under $K = 20$ and $m = 20$

Figure 4. Regret comparison of `AMF` and `OCO` algorithms under $B = dT^{3/4}$. The line and shade represent the average and standard deviation based on 20 repeated experiments.

Step 1. Estimation error decomposition: Let us fix $t \in [T]$ throughout the proof. For each $\nu \in [n_t]$ and $k \in [K]$, denote $\mathbf{X}_{k,\nu} := \tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top$. Then we can write

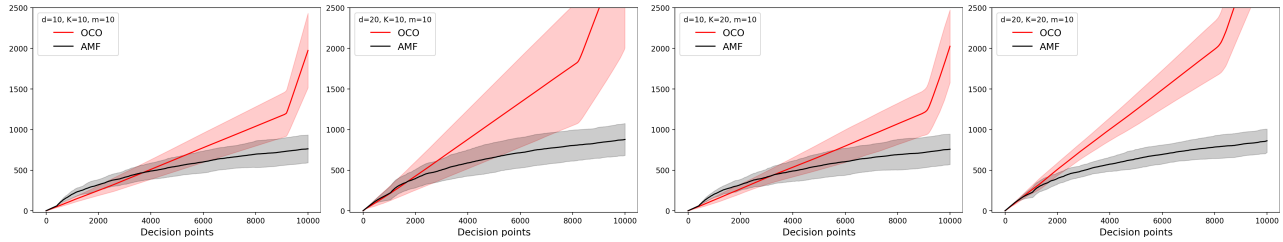
$$V_{n_t} := \sum_{\nu \in \Psi_{n_t}} \sum_{k=1}^{K} \mathbf{X}_{k,\nu} + \sum_{\nu \notin \Psi_{n_t}} \mathbf{X}_{a_{\tau(\nu)},\nu} + I_{J\cdot d},$$
$$A_{n_t} := \sum_{\nu \in \Psi_{n_t}} \sum_{k=1}^{K} \frac{\mathbb{I}(h_\nu = k)}{\phi_{k,\nu}} \mathbf{X}_{k,\nu} + \sum_{\nu \notin \Psi_{n_t}} \mathbf{X}_{k,\nu} + I_{J\cdot d}.$$

Denote the errors $\tilde{\eta}_{k,\nu} := \tilde{r}_{k,\nu} - \tilde{X}_{k,\nu}^\top \Theta_\star$ and $\eta_{k,\nu} := r_{k,\tau(\nu)}^{(j_{\tau(\nu)})} - \tilde{X}_{k,\nu}^\top \Theta_\star$. By the definition of the estimator $\widehat{\Theta}_{n_t}$,

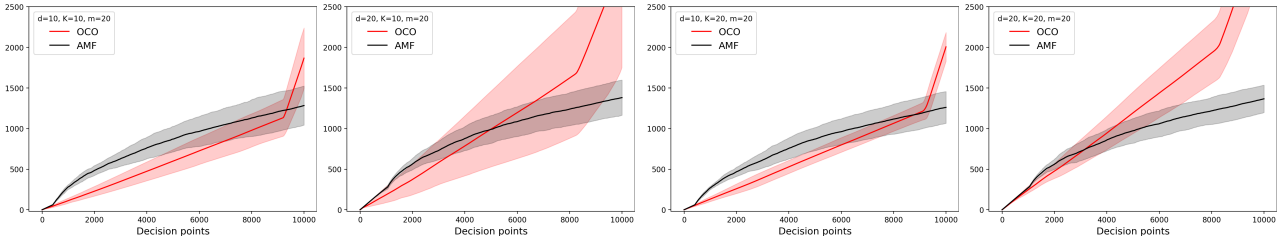
$$\begin{aligned}
&\left\|\widehat{\Theta}_{n_t} - \Theta^*\right\|_{V_{n_t}} \\
&= \left\| V_{n_t}^{-1/2} \left\{ -\Theta^* + \sum_{\nu \in \Psi_{n_t}} \sum_{k=1}^{K} \tilde{\eta}_{k,\nu} \tilde{X}_{k,\nu} + \sum_{\nu \notin \Psi_{n_t}} \eta_{k,\nu} \tilde{X}_{a_{\tau(\nu)},\nu} \right\} \right\|_2 \\
&\le \lambda_{\max}\left(V_{n_t}^{-1/2}\right) \left\|\Theta^*\right\|_2 + \left\| V_{n_t}^{-1/2} \left\{ \sum_{\nu \in \Psi_{n_t}} \sum_{k=1}^{K} \tilde{\eta}_{k,\nu} \tilde{X}_{k,\nu} + \sum_{\nu \notin \Psi_{n_t}} \eta_{k,\nu} \tilde{X}_{a_{\tau(\nu)},\nu} \right\} \right\|_2 \\
&\le \sqrt{Jd} + \left\| V_{n_t}^{-1/2} \left\{ \sum_{\nu \in \Psi_{n_t}} \sum_{k=1}^{K} \tilde{\eta}_{k,\nu} \tilde{X}_{k,\nu} + \sum_{\nu \notin \Psi_{n_t}} \eta_{k,\nu} \tilde{X}_{a_{\tau(\nu)},\nu} \right\} \right\|_2,
\end{aligned} \tag{16}$$

where and the last inequality holds because $\left\|\theta_\star^{(j)}\right\|_2 \le \sqrt{d}$. Plugging in $\tilde{r}_{k,\nu}$ defined in (5),

$$\tilde{\eta}_{k,\nu}\tilde{X}_{k,\nu} = \left(1 - \frac{\mathbb{I}(h_\nu = k)}{\phi_{k,\nu}}\right) \tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top \left(\check{\Theta}_t - \Theta^*\right) + \frac{\mathbb{I}(h_\nu = k)}{\phi_{k,\nu}} \eta_{k,\nu}\tilde{X}_{k,\nu},$$



(a) Regret comparison under $K = 10$ and $m = 10$ (b) Regret comparison under $K = 20$ and $m = 10$



(c) Regret comparison under $K = 10$ and $m = 20$ (d) Regret comparison under $K = 20$ and $m = 20$

Figure 5. Regret comparison of `AMF` and `OCO` algorithms under $B = \sqrt{dT}$. The line and shade represent the average and standard deviation based on 20 repeated experiments.

and the term $\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{\eta}_{k,\nu}\tilde{X}_{k,\nu}$ is decomposed as,

$$\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{\eta}_{k,\nu}\tilde{X}_{k,\nu}=\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\left\{\left(1-\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\right)\mathbf{X}_{k,\nu}\left(\check{\Theta}_t-\Theta^*\right)+\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\eta_{k,\nu}\tilde{X}_{k,\nu}\right\}. \tag{17}$$

By definition of the IPW estimator $\check{\Theta}_t$,

$$\begin{aligned}
&\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\left(1-\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\right)\mathbf{X}_{k,\nu}\left(\check{\Theta}_t-\Theta^*\right)\\
&=\left\{\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\left(1-\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\right)\mathbf{X}_{k,\nu}\right\}A_{n_t}^{-1}\left(-\Theta^*+\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\eta_{k,\nu}\tilde{X}_{k,\nu}+\sum_{\nu\notin\Psi_{n_t}}\eta_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu}\right)\\
&=\left(V_{n_t}-A_{n_t}\right)A_{n_t}^{-1}\left(-\Theta^*+\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\eta_{k,\nu}\tilde{X}_{k,\nu}+\sum_{\nu\notin\Psi_{n_t}}\eta_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu}\right)\\
&:=\left(V_{n_t}-A_{n_t}\right)A_{n_t}^{-1}\left(-\Theta^*+S_{n_t}\right),
\end{aligned} \tag{18}$$

where

$$S_{n_t}:=\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\eta_{k,\nu}\tilde{X}_{k,\nu}+\sum_{\nu\notin\Psi_{n_t}}\eta_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu},$$

then,

$$\begin{aligned}
\left\|\widehat{\Theta}_{n_t}-\Theta^*\right\|_{V_{n_t}}&\underset{(16)}{\leq}\sqrt{Jd}+\left\|V_{n_t}^{-1/2}\left\{\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{\eta}_{k,\nu}\tilde{X}_{k,\nu}+\sum_{\nu\notin\Psi_{n_t}}\eta_{k,\nu}\tilde{X}_{a_{\tau(\nu)},\nu}\right\}\right\|_2\\
&\underset{(17),(18)}{=}\sqrt{Jd}+\left\|V_{n_t}^{-1/2}\left\{\left(V_{n_t}-A_{n_t}\right)A_{n_t}^{-1}\left(-\Theta^*+S_{n_t}\right)+S_{n_t}\right\}\right\|_2.
\end{aligned}$$

By triangular inequality,

$$\begin{aligned}
\left\|\widehat{\Theta}_t-\Theta^*\right\|_{V_{n_t}} \leq & \sqrt{Jd}+\left\|V_{n_t}^{-1/2}\left\{\left(V_{n_t}-A_{n_t}\right)A_{n_t}^{-1}\left(-\Theta^*+S_{n_t}\right)+S_{n_t}\right\}\right\|_2 \\
\leq & \sqrt{Jd}+\left\|V_{n_t}^{-1/2}\left(V_{n_t}-A_{n_t}\right)A_{n_t}^{-1}\left(-\Theta^*+S_{n_t}\right)\right\|_2+\|S_{n_t}\|_{V_{n_t}^{-1}} \\
\leq & \sqrt{Jd}+\left\|\left(V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}-I_{J\cdot d}\right)\left(-V_{n_t}^{-1/2}\Theta^*+V_t^{-1/2}S_{n_t}\right)\right\|_2+\|S_{n_t}\|_{V_{n_t}^{-1}} \\
\leq & \sqrt{Jd}+\left\|V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}-I_{J\cdot d}\right\|_2\left\|-V_{n_t}^{-1/2}\Theta^*+V_t^{-1/2}S_{n_t}\right\|_2+\|S_{n_t}\|_{V_{n_t}^{-1}} \\
\leq & \sqrt{Jd}+\left\|V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}-I_{J\cdot d}\right\|_2\left(\sqrt{Jd}+\|S_{n_t}\|_{V_{n_t}^{-1}}\right)+\|S_{n_t}\|_{V_{n_t}^{-1}} \\
= & \left(\left\|V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}-I_{J\cdot d}\right\|_2+1\right)\left(\sqrt{Jd}+\|S_{n_t}\|_{V_{n_t}^{-1}}\right).
\end{aligned} \tag{19}$$

Step 2. Bounding the $\|\cdot\|_2$ of the matrix in (19) We claim that

$$V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}\succeq\frac{1}{8}I_{J\cdot d} \tag{20}$$

Define $F_{n_t}:=\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\mathbf{X}_{k,\nu}+16Kd\log(\frac{Jd}{\delta})I_{J\cdot d}$. Then we have $V_{n_t}\preceq F_{n_t}$ and $V_{n_t}^{1/2}A_{n_t}^{-1}V_{n_t}^{1/2}\succeq F_{n_t}^{-1/2}A_{n_t}F_{n_t}^{-1/2}$. Now we decompose the matrix A_{n_t} as

$$\begin{aligned}
F_{n_t}^{-1/2}A_{n_t}F_{n_t}^{-1/2}=&F_{n_t}^{-1/2}\left\{\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}+I_{J\cdot d}\right\}F_{n_t}^{-1/2} \\
&+F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\left\{\mathbf{X}_{a_{\tau(\nu)},\nu}-\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right\}\right]F_{n_t}^{-1/2} \\
=&F_{n_t}^{-1/2}\left\{\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}+16Kd\log\frac{Jd}{\delta}I_{J\cdot d}\right\}F_{n_t}^{-1/2} \\
&+F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\left\{\mathbf{X}_{a_{\tau(\nu)},\nu}-\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right\}+\left(1-16Kd\log\frac{Jd}{\delta}\right)I_{J\cdot d}\right]F_{n_t}^{-1/2}
\end{aligned} \tag{21}$$

For each $\nu\in[n_t]$, the matrix $\sum_{k=1}^{K}\frac{\mathbb{I}(h_\nu=k)}{\phi_{k,\nu}}F_{n_t}^{-1/2}\mathbf{X}_{k,\nu}F_{n_t}^{-1/2}$ symmetric positive definite and

$$\begin{aligned}
\lambda_{\max}\left(8\log\frac{Jd}{\delta}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}F_{n_t}^{-1/2}\mathbf{X}_{k,\nu}F_{n_t}^{-1/2}\right)\leq&8\log\frac{Jd}{\delta}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\lambda_{\max}\left(F_{n_t}^{-1}\right) \\
\leq&8\log\frac{Jd}{\delta}\frac{\lambda_{\min}(F_\nu)}{16\log\left(\frac{Jd}{\delta}\right)}\lambda_{\max}\left(F_{n_t}^{-1}\right) \\
\leq&\frac{1}{2}\frac{\lambda_{\min}(F_\nu)}{\lambda_{\min}(F_{n_t})} \\
\leq&\frac{1}{2}.
\end{aligned} \tag{22}$$

With the filtration $\mathcal{F}_0 := \mathcal{H}_t$ and $\mathcal{F}_n := \mathcal{F}_0 \cup \{h_\nu : \nu \in [n]\}$, we use Lemma C.3 to have with probability at least $1-\delta$,

$$\begin{aligned}
&8\log\frac{Jd}{\delta} F_{n_t}^{-1/2}\left\{\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu} + 16Kd\log\frac{Jd}{\delta}I_{J\cdot d}\right\}F_{n_t}^{-1/2}\\
&\succeq 8\log\frac{Jd}{\delta} F_{n_t}^{-1/2}\left\{\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\mathbf{X}_{k,\nu} + I_{J\cdot d}\right\}F_{n_t}^{-1/2}\\
&= 4\log\frac{Jd}{\delta}I_{J\cdot d} - \log\frac{Jd}{\delta}I_{J\cdot d}\\
&= 3\log\frac{Jd}{\delta}I_{J\cdot d},
\end{aligned}$$

which implies

$$F_{n_t}^{-1/2}\left\{\sum_{\nu=1}^{n_t}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu} + 16Kd\log\frac{Jd}{\delta}I_{J\cdot d}\right\}F_{n_t}^{-1/2} \succeq \frac{3}{8}I_{J\cdot d}, \tag{23}$$

and the left hand side of (21) is bounded as

$$F_{n_t}^{-1/2}A_{n_t}F_{n_t}^{-1/2} \succeq \frac{3}{8}I_{J\cdot d} + F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\left\{\mathbf{X}_{a_{\tau(\nu)},\nu} - \sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right\} + \left(1-16Kd\log\frac{Jd}{\delta}\right)I_{J\cdot d}\right]F_{n_t}^{-1/2}. \tag{24}$$

To bound the other term, observe that for $\nu\notin\Psi_{n_t}$,

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\middle|\mathcal{H}_t\right] = \sum_{i\neq a_{\tau(\nu)}}\sum_{k=1}^{K}\phi_{i,\nu}\frac{\mathbb{I}\left(i=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu} = \sum_{k\neq a_{\tau(\nu)}}\mathbf{X}_{k,\nu}.$$

Because (22) holds for $\nu\notin\Psi_{n_t}$, we can use Lemma C.3, to have with probability at least $1-\delta$

$$8\log\frac{Jd}{\delta}F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right]F_{n_t}^{-1/2} \preceq 12\log\frac{Jd}{\delta}F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\sum_{k\neq a_{\tau(\nu)}}\mathbf{X}_{k,\nu}\right]F_{n_t}^{-1/2} + \log\frac{Jd}{\delta}I_{J\cdot d}.$$

Rearranging the terms,

$$F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right]F_{n_t}^{-1/2} \preceq \frac{3}{2}F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\sum_{k\neq a_{\tau(\nu)}}\mathbf{X}_{k,\nu}\right]F_{n_t}^{-1/2} + \frac{1}{8}I_{J\cdot d}.$$

Thus the second term in (24) is bounded as,

$$\begin{aligned}
&F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\left\{\mathbf{X}_{a_{\tau(\nu)},\nu} - \sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu=k\right)}{\phi_{k,\nu}}\mathbf{X}_{k,\nu}\right\}\right]F_{n_t}^{-1/2}\\
&\succeq F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\left\{\mathbf{X}_{a_{\tau(\nu)},\nu} - \frac{3}{2}\sum_{\nu\notin\Psi_t}\sum_{k\neq a_{\tau(\nu)}}\mathbf{X}_{k,\nu}\right\}\right]F_{n_t}^{-1/2} - \frac{1}{8}I_{J\cdot d}\\
&\succeq -\frac{3}{2}F_{n_t}^{-1/2}\left[\sum_{\nu\notin\Psi_{n_t}}\sum_{k=1}^{K}\mathbf{X}_{k,\nu}\right]F_{n_t}^{-1/2} - \frac{1}{8}I_{J\cdot d}\\
&\succeq -\left\{\frac{3dK}{2}\left|\Psi_{n_t}^c\right|\lambda_{\max}\left(F_{n_t}^{-1}\right) + \frac{1}{8}\right\}I_{J\cdot d},
\end{aligned} \tag{25}$$

where the last inequality holds by $\lambda_{\max}(\mathbf{X}_{k,\nu}) \leq d$. Plugging in (24),

$$\begin{aligned} F_{n_t}^{-1/2} A_{n_t} F_{n_t}^{-1/2} \succeq & \frac{1}{4} I_{J\cdot d} - \left\{ \frac{3dK}{2} \left|\Psi_{n_t}^c\right| \lambda_{\max}\left(F_{n_t}^{-1}\right) \right\} I_{J\cdot d} - \left(16Kd \log \frac{Jd}{\delta} - 1 \right) F_{n_t}^{-1} \\ \succeq & \frac{1}{4} I_{J\cdot d} - \left(\frac{3dK}{2} \left|\Psi_{n_t}^c\right| + 16Kd \log \frac{Jd}{\delta} \right) \lambda_{\max}\left(F_{n_t}^{-1}\right) I_{J\cdot d}. \end{aligned} \tag{26}$$

By Lemma C.3, with probability at least $1-\delta$,

$$\frac{1}{2}\left|\Psi_{n_t}^c\right| = \frac{1}{2} \sum_{\nu=1}^{n_t} \mathbb{I}\left(h_\nu \neq a_{\tau(\nu)}\right) \leq \frac{3}{2} \sum_{\nu=1}^{n_t} \sum_{k \neq a_{\tau(\nu)}} \phi_{k,\nu} + \log \frac{1}{\delta},$$

which implies

$$\begin{aligned} & \left(\frac{3dK}{2} \left|\Psi_{n_t}^c\right| + 16Kd \log \frac{Jd}{\delta} \right) \lambda_{\max}\left(F_{n_t}^{-1}\right) \\ & \leq \frac{dK}{2\lambda_{\min}(F_{n_t})} \left(9 \sum_{\nu=1}^{n_t} \sum_{k \neq a_{\tau(\nu)}} \phi_{k,\nu} + 3 \log \frac{1}{\delta} + 32 \log \frac{Jd}{\delta} \right) \\ & \leq \frac{dK}{2\lambda_{\min}(F_{n_t})} \left(9 \sum_{\nu=1}^{n_t} \sum_{k \neq a_{\tau(\nu)}} \phi_{k,\nu} + 35 \log \frac{Jd}{\delta} \right) \\ & = \frac{dK}{2\lambda_{\min}(F_{n_t})} \left(\sum_{\nu=1}^{n_t} \frac{144\,(K-1) \log \frac{Jd}{\delta}}{\lambda_{\min}(F_\nu)} + 35 \log \frac{Jd}{\delta} \right) \\ & \leq \frac{1}{8}, \end{aligned} \tag{27}$$

where the last inequality holds by the assumption (8). Plugging in (26), with probability at least $1-2\delta$,

$$F_{n_t}^{-1/2} \left[\sum_{\nu \notin \Psi_{n_t}} \left\{ \mathbf{X}_{a_{\tau(\nu)},\nu} - \sum_{k=1}^{K} \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \mathbf{X}_{k,\nu} \right\} \right] F_{n_t}^{-1/2} \succeq -\frac{1}{4} I_{J\cdot d}.$$

With (24),

$$F_{n_t}^{-1/2} A_{n_t} F_{n_t}^{-1/2} \succeq \frac{1}{8} I_{J\cdot d},$$

which proves (20) and the claim implies

$$\left\| V_{n_t}^{1/2} A_{n_t}^{-1} V_{n_t}^{1/2} - I_{J\cdot d} \right\|_2 \leq 7.$$

Step 3. Bounding the self-normalized vector-valued martingale S_{n_t} Let $\mathcal{F}_0$ be a sigma algebra generated by contexts $\{\mathbf{x}_{k,s}^{(j_s)} : k \in [K], s \in [t]\}$, and Ψ_t. Define filtration as $\mathcal{F}_\nu := \sigma(\mathcal{F}_0 \cup \mathcal{H}_{\tau(\nu+1)})$. Then S_ν is a $\mathbb{R}^{J\cdot d}$-valued martingale because

$$\begin{aligned} \mathbb{E}\left[S_\nu - S_{\nu-1} \middle| \mathcal{F}_{\nu-1} \right] &= \mathbb{E}\left[\mathbb{I}\left(\nu \in \Psi_{n_t}\right) \sum_{k=1}^{K} \frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}} \eta_{k,\nu} \tilde{X}_{k,\nu} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right) \eta_{a_{\tau(\nu)},\tau(\nu)} \tilde{X}_{k,\nu} \middle| \mathcal{F}_{\nu-1} \right] \\ &= \mathbb{E}\left[\mathbb{I}\left(\nu \in \Psi_{n_t}\right) \sum_{k=1}^{K} \frac{\mathbb{I}\left(a_{\tau(\nu)} = k\right)}{\phi_{k,\nu}} \eta_{k,\nu} \tilde{X}_{k,\nu} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right) \eta_{a_{\tau(\nu)},\tau(\nu)} \tilde{X}_{k,\nu} \middle| \mathcal{F}_{\nu-1} \right] \\ &= \mathbb{E}\left[\left\{ \frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right) \right\} \eta_{a_{\tau(\nu)},\nu} \tilde{X}_{k,\nu} \middle| \mathcal{F}_{\nu-1} \right] \\ &= \mathbb{E}\left[\left\{ \frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right) \right\} \eta_{a_{\tau(\nu)},\nu} \tilde{X}_{k,\nu} \middle| \mathcal{H}_{\tau(\nu)} \right] \\ &= 0, \end{aligned}$$

where the second equality holds by definition of Ψ_{n_t} and the fourth inequality holds because the distribution of $\{\mathbf{x}_{k,s}^{(j_s)} : k \in [K], s \in (\tau(\nu), t]\}$ is independent of $\mathcal{H}_{\tau(\nu)}$ by Assumption 2. By Assumption 1, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left[\lambda\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right)\right\}\eta_{k,a_{\tau(\nu)}}\right]\middle|\mathcal{F}_{\nu-1}\right] \leq \mathbb{E}\left[\exp\left[\frac{\lambda^2\sigma^2}{2}\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right)\right\}^2\right]\middle|\mathcal{F}_{\nu-1}\right] \leq \exp\left[2\lambda^2\sigma_r^2\right],$$

Thus, $\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right)\right\}\eta_{k,a_{\tau(\nu)}}$ is $2\sigma_r$-sub-Gaussian. Because

$$\begin{aligned}
\|S_{n_t}\|_{V_t^{-1}} &= \left\|\sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\frac{\mathbb{I}\left(h_\nu = k\right)}{\phi_{k,\nu}}\eta_{k,\nu}\tilde{X}_{k,\nu} + \sum_{\nu\notin\Psi_{n_t}}\eta_{a_{\tau(\nu)},\nu}\tilde{X}_{a_{\tau(\nu)},\nu}\right\|_{V_{n_t}^{-1}} \\
&= \left\|\sum_{\nu=1}^{n_t}\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right)\right\}\eta_{a_{\tau(\nu)},\nu}\tilde{X}_{k,\nu}\right\|_{V_{n_t}^{-1}} \\
&= \left\|\sum_{\nu=1}^{n_t}\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_{n_t}\right)\right\}\eta_{a_{\tau(\nu)},\nu}V_{n_t}^{-1/2}\tilde{X}_{k,\nu}\right\|_2,
\end{aligned}$$

by Lemma C.6, with probability at least $1-\delta$,

$$\begin{aligned}
\|S_t\|_{V_t^{-1}} \leq& \left\|\sum_{\nu=1}^{n_t}\left\{\frac{\mathbb{I}\left(\nu \in \Psi_{n_t}\right)}{\phi_{a_{\tau(\nu)},\nu}} + \mathbb{I}\left(\nu \notin \Psi_t\right)\right\}\eta_{a_{\tau(\nu)},\nu}V_{n_t}^{-1/2}\tilde{X}_{k,\nu}\right\|_2, \\
\leq& 12\sigma_r\sqrt{\sum_{\nu=1}^{n_t}\left\|V_{n_t}^{-1/2}\tilde{X}_{k,\nu}\right\|_2^2\log\frac{4}{\delta}} \\
\leq& 12\sigma_r\sqrt{Jd\log\frac{4}{\delta}},
\end{aligned}$$

where the last inequality holds because

$$\begin{aligned}
\sum_{\nu=1}^{n_t}\left\|V_{n_t}^{-1/2}\tilde{X}_{k,\nu}\right\|_2^2 &= \sum_{\nu=1}^{n_t}\tilde{X}_{k,\nu}^\top V_{n_t}^{-1}\tilde{X}_{k,\nu} = \mathrm{Tr}\left(\sum_{\nu=1}^{n_t}\tilde{X}_{k,\nu}^\top V_{n_t}^{-1}\tilde{X}_{k,\nu}\right) \\
&= \mathrm{Tr}\left(\sum_{\nu=1}^{n_t}\tilde{X}_{k,\nu}\tilde{X}_{k,\nu}^\top V_{n_t}^{-1}\right) \leq \mathrm{Tr}\left(V_{n_t}V_{n_t}^{-1}\right) = Jd.
\end{aligned}$$

With (19), the proof is completed □

B.2. Proof of Theorem 4.2

Proof. Similar to the proof of Theorem 4.1, the bound for consumption vector immediately follows from the bound for the utilities. Therefore we provide the proof for the utility bound.

Step 1. Decomposition: For each $k \in [K]$ and $j \in [J]$,

$$\begin{aligned}
\left|u_k^{\star(j)} - \tilde{u}_{k,t+1}^{(j)}\right| &\le \left|\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left(\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \widehat{\theta}_t^{(j)}\right)\right| \\
&\le \left|\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left(\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| \\
&\quad + \left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)} - \theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right| \\
&\quad \left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| \\
&\quad + \left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)} - \theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|.
\end{aligned}$$

Taking maximum over $k \in [K]$ gives the decomposition,

$$\begin{aligned}
\max_{k\in[K]}\left|u_k^{\star(j)} - \tilde{u}_{k,t+1}^{(j)}\right| &\le \max_{k\in[K]}\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| \\
&\quad + \max_{k\in[K]}\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)} - \theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|.
\end{aligned} \tag{28}$$

Step 2. Bounding the difference between expectation and empirical distribution: The random variables $\left\{\left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)} : s \in [t]\right\}$ are IID by Assumption 2. Using Lemma C.1,

$$\begin{aligned}
&\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| \\
&= \frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\left|\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| \\
&\le \frac{4}{\sqrt{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}}\sqrt{\log JKT}.
\end{aligned}$$

with probability at least $1 - 3(JKT)^{-1}$. By Lemma C.3, with probability at least $1 - (JT)^{-1}$,

$$\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right) \ge \frac{1}{2}p_j\left(t+1\right) - 2\log JT \ge \frac{1}{4}p_j\left(t+1\right), \tag{29}$$

where the last inequality holds by the assumption $t \ge 8d\alpha^{-1}p_{\min}^{-1}\log JT$. Summing up the probability bounds, with probability at least $1 - 5T^{-1}$,

$$\begin{aligned}
\max_{k\in[K]}\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\mathbb{E}\left[\mathbf{x}_k^{(j)}\right]^\top \theta_\star^{(j)} - \left\{\mathbf{x}_{k,s}^{(j)}\right\}^\top \theta_\star^{(j)}\right)\right| &\le \frac{4}{\sqrt{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)}}\sqrt{\log JKT} \\
&\le \frac{8}{\sqrt{p_j\left(t+1\right)}}\sqrt{\log JKT}.
\end{aligned}$$

Plugging in the decomposition (28), for each $j \in [J]$,

$$\max_{k\in[K]}\left|u_k^{\star(j)}-\tilde{u}_{k,t+1}^{(j)}\right| \leq \frac{16}{\sqrt{p_j\,(t+1)}}\sqrt{\log JKT} + \max_{k\in[K]}\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)}\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\left(\widehat{\theta}_{t-1}^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|.$$

Taking square and summing up over $j \in [J]$ gives

$$\sum_{j=1}^{J} p_j \max_{k\in[K]}\left|u_k^{\star(j)}-\tilde{u}_{k,t+1}^{(j)}\right|^2 \leq \frac{16J\log JKT}{t+1} + \sum_{j=1}^{J} p_j \max_{k\in[K]}\left|\frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)}\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\left(\widehat{\theta}_{t}^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|^2, \tag{30}$$

Step 3. Bounding the prediction error: By Cauchy-Schwartz inequality and (29),

$$\begin{aligned}
&\sum_{j=1}^{J}\max_{k\in[K]} p_j \frac{1}{\left\{\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\right\}^2}\left|\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|^2\\
&\leq \sum_{j=1}^{J}\max_{k\in[K]} p_j \frac{1}{\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)}\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\left\{\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right\}^2\\
&\leq \sum_{j=1}^{J}\max_{k\in[K]} \frac{4p_j}{p_j\,(t+1)}\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\left\{\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right\}^2\\
&= \frac{4}{(t+1)}\sum_{j=1}^{J}\max_{k\in[K]}\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top\left\{\sum_{s=1}^{t+1}\mathbb{I}\,(j_s=j)\,\mathbf{x}_{k,s}^{(j)}\left(\mathbf{x}_{k,s}^{(j)}\right)^\top\right\}\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)\\
&\leq \frac{4}{(t+1)}\sum_{j=1}^{J}\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top\left\{\sum_{s=1}^{t+1}\sum_{k=1}^{K}\mathbb{I}\,(j_s=j)\,\mathbf{x}_{k,s}^{(j)}\left(\mathbf{x}_{k,s}^{(j)}\right)^\top\right\}\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)\\
&= \frac{4}{(t+1)}\left(\widehat{\Theta}_t-\Theta^\star\right)^\top\left\{\sum_{s=1}^{t+1}\sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top\right\}\left(\widehat{\Theta}_t-\Theta^\star\right),
\end{aligned} \tag{31}$$

where $\Theta^\star := (\theta_\star^{(1)},\ldots,\theta_\star^{(J)})^T \in \mathbb{R}^{J\cdot d}$ and

$$\tilde{X}_{k,s} := \begin{pmatrix} \mathbf{0}_d \\ \vdots \\ \mathbf{x}_{k,s}^{(j_s)} \\ \mathbf{0}_d \end{pmatrix} \in \mathbb{R}^{J\cdot d},$$

where the context $\mathbf{x}_{k,s}^{(j_s)}$ is located after j_s-1 of $\mathbf{0}_d$ vectors. We claim that

$$\frac{1}{t+1}\sum_{s=1}^{t+1}\sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top \preceq 2\mathbb{E}\left[\tilde{X}_{k,1}\tilde{X}_{k,1}^\top\right] \preceq \frac{8}{|\Psi_{n_t}|}\sum_{s\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top, \tag{32}$$

with probability at least $1-2T^{-1}$. The matrix $\mathbf{X}_s := \sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top$is symmetric nonnegative definite which satisfies

$$\lambda_{\max}\left(\frac{1}{2dK}\mathbf{X}_s\right) \leq \frac{1}{2}.$$

By Lemma C.3, with probability at least $1 - T^{-1}$,

$$\frac{1}{2Kd}\sum_{s=1}^{t+1}\mathbf{X}_s \preceq \frac{3}{4Kd}\sum_{s=1}^{t+1}\mathbb{E}\left[\mathbf{X}_s\right] + \left(\log JdT\right) I_{J\cdot d},$$

which implies

$$\frac{1}{t+1}\sum_{s=1}^{t+1}\mathbf{X}_s \preceq \frac{3}{2\left(t+1\right)}\sum_{s=1}^{t+1}\mathbb{E}\left[\mathbf{X}_s\right] + \frac{2dK}{t+1}\left(\log JdT\right) I_{J\cdot d}. \tag{33}$$

By Assumption 3, for $s \in [t+1]$,

$$\begin{aligned}
\lambda_{\min}(\mathbb{E}\left[\mathbf{X}_s\right]) =& \lambda_{\min}\left(\begin{pmatrix} p_1\mathbb{E}_{\mathbf{x}_k\sim\mathbb{F}_1}\left[\sum_{k=1}^K \mathbf{x}_k\mathbf{x}_k^\top\right] & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_J\mathbb{E}_{\mathbf{x}_k\sim\mathbb{F}_J}\left[\sum_{k=1}^K \mathbf{x}_k\mathbf{x}_k^\top\right] \end{pmatrix}\right) \\
\geq& \lambda_{\min}\left(\begin{pmatrix} p_1 K\alpha I_d & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_J K\alpha I_d \end{pmatrix}\right) \\
\geq& K p_{\min}\alpha.
\end{aligned}$$

For $t \geq 8d\alpha^{-1}p_{\min}^{-1}\log JdT$,

$$\sum_{s=1}^{t+1}\mathbb{E}\left[\mathbf{X}_s\right] \succeq \sum_{s=1}^{t+1}\lambda_{\min}(\mathbb{E}\left[\mathbf{X}_s\right]) I_{J\cdot d} \succeq \left(t+1\right) K p_{\min}\alpha I_{J\cdot d} \succeq 4dK\left(\log\frac{Jd}{\delta}\right)$$

Plugging in (33) proves the first inequality of (32),

$$\frac{1}{t+1}\sum_{s=1}^{t+1}\mathbf{X}_s \preceq \frac{2}{\left(t+1\right)}\sum_{s=1}^{t+1}\mathbb{E}\left[\mathbf{X}_s\right] = 2\mathbb{E}\left[\mathbf{X}_1\right],$$

where the equality holds because $\mathbb{E}\mathbf{X}_s = \mathbb{E}\mathbf{X}_1$ for all $s \in [T]$. To prove the second inequality,

$$\mathbb{E}\left[\mathbf{X}_1\right] = \left|\Psi_{n_t}\right|^{-1}\sum_{\nu\in\Psi_{n_t}}\mathbb{E}\left[\mathbf{X}_{\tau(\nu)}\right],$$

and by Lemma C.3, with probability at least $1 - T^{-1}$,

$$\frac{1}{2Kd}\sum_{\nu\in\Psi_{n_t}}\mathbf{X}_{\tau(\nu)} \succeq \frac{1}{4Kd}\sum_{\nu\in\Psi_{n_t}}\mathbb{E}\left[\mathbf{X}_{\tau(\nu)}\right] - \left(\log JdT\right) I_{J\cdot d}.$$

Rearranging the terms,

$$\sum_{\nu\in\Psi_{n_t}}\mathbb{E}\left[\mathbf{X}_{\tau(\nu)}\right] \preceq 2\sum_{\nu\in\Psi_{n_t}}\mathbf{X}_{\tau(\nu)} + 4Kd\left(\log JdT\right) I_{J\cdot d} \tag{34}$$

By definition of F_{n_t},

$$\begin{aligned}
\sum_{\nu\in\Psi_{n_t}}\mathbf{X}_{\tau(\nu)} =& F_{n_t} - \sum_{\nu\notin\Psi_{n_t}}\mathbf{X}_{\tau(\nu)} - 16d(K-1)\log\frac{Jd}{\delta} I_{J\cdot d} \\
\succeq& F_{n_t} - \left(Kd\left|\Psi_{n_t}^c\right| + 16d(K-1)\log\frac{Jd}{\delta}\right) I_{J\cdot d}.
\end{aligned} \tag{35}$$

Because the condition (8) holds, we can use (27),

$$\left(\frac{3dK}{2}\left|\Psi_{n_t}^c\right| + 16Kd\log\frac{Jd}{\delta}\right)\lambda_{\max}\left(F_{n_t}^{-1}\right) \leq \frac{1}{8},$$

to obtain

$$\begin{aligned}
Kd\left|\Psi_{n_t}^c\right| + 16d(K-1)\log\frac{Jd}{\delta} \leq & \frac{2}{3}\left(\frac{3dK}{2}\left|\Psi_{n_t}^c\right| + 16Kd\log\frac{Jd}{\delta}\right) + \frac{16}{3}Kd\log\frac{Jd}{\delta} \\
\leq & \frac{1}{12\lambda_{\max}(F_{n_t})} + \frac{16}{3}Kd\log\frac{Jd}{\delta} \\
= & \frac{\lambda_{\min}(F_{n_t})}{12} + \frac{16}{3}Kd\log\frac{Jd}{\delta}.
\end{aligned}$$

Plugging in (35),

$$\begin{aligned}
\sum_{\nu\in\Psi_{n_t}} \mathbf{X}_{\tau(\nu)} \succeq & F_{n_t} - \left\{\frac{1}{12}\lambda_{\min}(F_{n_t}) + \frac{16}{3}Kd\log\frac{Jd}{\delta}\right\} I_{J\cdot d} \\
\succeq & \left\{\frac{11}{12}\lambda_{\min}(F_{n_t}) - \frac{16}{3}Kd\log\frac{Jd}{\delta}\right\} I_{J\cdot d} \\
\succeq & \left\{\frac{44}{3}(K-1)d\log\frac{Jd}{\delta} - \frac{16}{3}Kd\log\frac{Jd}{\delta}\right\} I_{J\cdot d} \\
\succeq & \left\{\frac{22}{3}Kd\log\frac{Jd}{\delta} - \frac{16}{3}Kd\log\frac{Jd}{\delta}\right\} I_{J\cdot d} \\
\succeq & 2Kd\log\frac{Jd}{\delta} I_{J\cdot d} \\
\succeq & 2Kd\log JdT I_{J\cdot d}
\end{aligned} \tag{36}$$

where the third inequality holds by $F_{n_t} \succeq 16(K-1)\log\frac{Jd}{\delta}$ and the last inequality holds by $\delta < T^{-1}$. Plugging in (34),

$$\sum_{\nu\in\Psi_{n_t}} \mathbb{E}\left[\mathbf{X}_{\tau(\nu)}\right] \preceq 2\sum_{\nu\in\Psi_{n_t}} \mathbf{X}_{\tau(\nu)} + 4Kd\left(\log JdT\right) I_{J\cdot d} \preceq 4\sum_{\nu\in\Psi_{n_t}} \mathbf{X}_{\tau(\nu)},$$

proves the second inequality in claim (32).

From (31),

$$\begin{aligned}
& \sum_{j=1}^{J} \max_{k\in[K]} p_j \frac{1}{\left\{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\right\}^2} \left|\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)} - \theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|^2 \\
& \leq \frac{4}{(t+1)}\left(\widehat{\Theta}_t - \Theta^\star\right)^\top \left\{\sum_{s=1}^{t+1}\sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top\right\}\left(\widehat{\Theta}_t - \Theta^\star\right) \\
& \leq \frac{32}{|\Psi_{n_t}|}\left(\widehat{\Theta}_t - \Theta^\star\right)^\top \left\{\sum_{s\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{X}_{k,s}\tilde{X}_{k,s}^\top\right\}\left(\widehat{\Theta}_t - \Theta^\star\right) \\
& \leq \frac{32}{|\Psi_{n_t}|}\left(\widehat{\Theta}_t - \Theta^\star\right)^\top \{V_{n_t}\}\left(\widehat{\Theta}_t - \Theta^\star\right) \\
& = \frac{32}{|\Psi_{n_t}|}\left\|\widehat{\Theta}_t - \Theta^\star\right\|_{V_{n_t}}^2
\end{aligned} \tag{37}$$

On bounding the normalizing matrix, the novel Gram matrix V_{n_t} plays a crucial role. To obtain an upper bound for (37), we need a matrix whose eigenvalue is greater than that of:

$$\sum_{\nu\in\Psi_{n_t}} \mathbf{X}_{\tau(\nu)} = \sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K}\tilde{X}_{k,\tau(\nu)}\tilde{X}_{k,\tau(\nu)}^\top, \tag{38}$$

However, with $\sum_{\nu\in\Psi_{n_t}} \tilde{X}_{a_{\tau(\nu)},\tau(\nu)}\tilde{X}^\top_{a_{\tau(\nu)},\tau(\nu)}$, a Gram matrix consist of only selected contexts, we cannot bound the matrix (38). Instead, by using a Gram matrix V_t, we can bound (38) as,

$$\begin{aligned}\sum_{\nu\in\Psi_{n_t}} \mathbf{X}_{\tau(\nu)} =& \sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K} \tilde{X}_{k,\tau(\nu)}\tilde{X}^\top_{k,\tau(\nu)}\\ \preceq& \sum_{\nu\in\Psi_{n_t}}\sum_{k=1}^{K} \tilde{X}_{k,\tau(\nu)}\tilde{X}^\top_{k,\tau(\nu)} + \sum_{\nu\notin\Psi_{n_t}} \tilde{X}_{a_{\tau(\nu)},\tau(\nu)}\tilde{X}^\top_{a_{\tau(\nu)},\tau(\nu)}\\ \preceq& V_{n_t},\end{aligned}$$

and prove the bound (37) to relate the prediction error to the self-normalized bound. From (37), by Theorem 4.1

$$\begin{aligned}\sum_{j=1}^{J}\max_{k\in[K]} p_j \frac{1}{\left\{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\right\}^2}\left|\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|^2 \leq& \frac{32}{|\Psi_{n_t}|}\left\|\widehat{\Theta}_t-\Theta^\star\right\|_{V_{n_t}}^2\\ \leq& \frac{32}{|\Psi_{n_t}|}\beta_{\sigma_r}(\delta)^2,\end{aligned}$$

with probability at least $1-4(m+1)\delta$. Because $|\Psi_{n_t}|+\left|\Psi_{n_t}^c\right|=n_t$ and (8) implies

$$\frac{3Kd}{2}\left|\Psi_{n_t}^c\right|\lambda_{\max}\left(F_{n_t}^{-1}\right)\leq\left(\frac{3Kd}{2}\left|\Psi_{n_t}^c\right|+16Kd\log\frac{Jd}{\delta}\right)\lambda_{\max}\left(F_{n_t}^{-1}\right)\leq\frac{1}{8},$$

we obtain

$$|\Psi_{n_t}|\geq n_t-\left|\Psi_{n_t}^c\right|\geq n_t-\frac{\lambda_{\min}(F_{n_t})}{12Kd}\geq n_t-\frac{\lambda_{\max}(F_{n_t})}{12Kd}\geq n_t-\frac{n_tKd}{12Kd}=\frac{11}{12}n_t.$$

Thus,

$$\begin{aligned}\sum_{j=1}^{J}\max_{k\in[K]} p_j \frac{1}{\left\{\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\right\}^2}\left|\sum_{s=1}^{t+1}\mathbb{I}\left(j_s=j\right)\left(\widehat{\theta}_t^{(j)}-\theta_\star^{(j)}\right)^\top \mathbf{x}_{k,s}^{(j)}\right|^2 \leq& \frac{32}{|\Psi_{n_t}|}\beta_{\sigma_r}(\delta)^2\\ \leq& \frac{12}{11}\cdot\frac{32}{n_t}\beta_{\sigma_r}(\delta)^2\\ \leq& \frac{36}{n_t}\beta_{\sigma_r}(\delta)^2.\end{aligned}$$

From (30),

$$\sqrt{\sum_{j=1}^{J} p_j \max_{k\in[K]}\left|u_k^{\star(j)}-\tilde{u}_{k,t+1}^{(j)}\right|^2}\leq\frac{16\sqrt{J\log JKT}}{\sqrt{t}}+\frac{6\beta_{\sigma_r}(\delta)}{\sqrt{n_t}}$$

and the proof is completed. □

B.3. Proof of Lemma 4.3

Proof. Suppose a feasible policy $\tilde{\pi}_{k,t}^{(j)}$ for the optimization problem (1) satisfies

$$\sum_{k=1}^{K+1}\tilde{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}>\sum_{k=1}^{K+1}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)},$$

which is equivalent to

$$\sum_{l=1}^{K+1}\tilde{\pi}_{k\langle l\rangle,t}^{(j_t)}\tilde{u}_{k\langle l\rangle,t}^{(j_t)}>\sum_{l=1}^{K+1}\widehat{\pi}_{k\langle l\rangle,t}^{(j_t)}\tilde{u}_{k\langle l\rangle,t}^{(j_t)}. \tag{39}$$

Without loss of generality we assume $\widehat{u}^{(j_t)}_{k\langle l\rangle,t} \geq 0$ (Because $\sum_{l=1}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} = \sum_{l=1}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t} = 1$, we can subtract $\widehat{u}^{(j_t)}_{k\langle K+1\rangle,t}$ on both side of (39)). By the constraints on the resources,

$$\tilde{\pi}^{(j_t)}_{k\langle 1\rangle,t} \leq \left(\min_{r\in[m]} \frac{\rho_t(r)}{\mathbf{b}^{(j_t)}_{k\langle 1\rangle,t}(r)} \right) \wedge 1 = \widehat{\pi}^{(j_t)}_{k\langle 1\rangle,t}$$

Suppose $\tilde{\pi}^{(j_t)}_{k\langle 1\rangle,t} < \widehat{\pi}^{(j_t)}_{k\langle 1\rangle,t}$. Because $\sum_{l=1}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} = \sum_{l=1}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t} = 1$, by Lemma C.2,

$$\sum_{l=1}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t} \leq \sum_{l=1}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t},$$

which contradicts with (39). Thus we have $\tilde{\pi}^{(j_t)}_{k\langle 1\rangle,t} = \widehat{\pi}^{(j_t)}_{k\langle 1\rangle,t}$ and

$$\sum_{l=2}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t} > \sum_{l=2}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t}. \tag{40}$$

Again, by the constraints on the resources,$\tilde{\pi}^{(j_t)}_{k\langle 2\rangle,t} \leq \widehat{\pi}^{(j_t)}_{k\langle 2\rangle,t}$. Suppose $\tilde{\pi}^{(j_t)}_{k\langle 2\rangle,t} < \widehat{\pi}^{(j_t)}_{k\langle 2\rangle,t}$. Because $\sum_{l=2}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} = \sum_{l=2}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t}$, by Lemma C.2,

$$\sum_{l=2}^{K+1} \tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t} \leq \sum_{l=2}^{K+1} \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t} \tilde{u}^{(j_t)}_{k\langle l\rangle,t}.$$

which contradicts with (40). Thus we have $\tilde{\pi}^{(j_t)}_{k\langle 2\rangle,t} = \widehat{\pi}^{(j_t)}_{k\langle 2\rangle,t}$ Recursively, we have $\tilde{\pi}^{(j_t)}_{k\langle l\rangle,t} = \widehat{\pi}^{(j_t)}_{k\langle l\rangle,t}$ for all $l \in [K+1]$. Thus there exist no feasible solution $\tilde{\pi}^{(j)}_{k,t}$ such that (39) holds and the proof is completed. □

B.4. Proof of Lemma 5.2

Proof. For each $t \in [T]$, denote the good events $\mathcal{G}_t := \mathcal{E}_t \cap \mathcal{M}_{t-1}$.

Step 1. Bounds for the estimates $\tilde{u}^{(j_t)}_{k,t}$ and $\tilde{\mathbf{b}}^{(j_t)}_{k,t}$: For each $t \in [T]$ and $k \in [K]$,

$$\begin{aligned}
\tilde{u}^{(j_t)}_{k,t} = \tilde{u}^{(j_t)}_{k,t} - u^{\star(j_t)}_k + u^{\star(j_t)}_k =& \frac{\gamma_{t-1,\sigma_r}(\delta)}{\sqrt{p_{j_t}}} + \widehat{u}^{(j_t)}_{k,t} - u^{\star(j_t)}_k + u^{\star(j_t)}_k \\
\geq& \frac{\gamma_{t-1,\sigma_r}(\delta) - \sqrt{p_{j_t}} \max_{k\in[K]} \left| \widehat{u}^{(j_t)}_{k,t} - u^{\star(j_t)}_k \right|}{\sqrt{p_{j_t}}} + u^{\star(j_t)}_k.
\end{aligned}$$

Under the event $\mathcal{G}_t$,

$$\begin{aligned}
\sqrt{p_{j_t}} \max_{k\in[K]} \left| u^{\star(j_t)}_k - \widehat{u}^{(j_t)}_{k,t} \right| =& \sqrt{p_{j_t} \max_{k\in[K]} \left| u^{\star(j_t)}_k - \widehat{u}^{(j_t)}_{k,t} \right|^2} \\
\leq& \sqrt{\sum_{j=1}^{J} p_j \max_{k\in[K]} \left| u^{\star(j)}_k - \widehat{u}^{(j)}_{k,t} \right|^2} \\
\leq& \gamma_{t-1,\sigma_r}(\delta),
\end{aligned}$$

which implies

$$\tilde{u}^{(j_t)}_{k,t} \geq u^{\star(j_t)}_k. \tag{41}$$

Similarly,

$$\tilde{\mathbf{b}}^{(j_t)}_{k,t} \leq \mathbf{b}^{\star(j_t)}_k. \tag{42}$$

Another useful bound for $\tilde{u}_{k,t}^{(j_t)}$ is

$$\mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)}-u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right]\leq 2\gamma_{t-1,\sigma_r}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]}. \tag{43}$$

This bound is proved by the tower property of conditional expectation and Cauchy-Schwartz inequality,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)}-u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right]=&\mathbb{E}\left[\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j_t)}-u_k^{\star(j_t)}\right|\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right]\\
=&\mathbb{E}\left[\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j_t)}-u_k^{\star(j_t)}\right|\mathbb{I}\left(a_t\in[K]\right)\mathbb{I}\left(\mathcal{G}_t\right)\right]\\
=&\mathbb{E}\left[\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j)}-u_k^{\star(j)}\right|\mathbb{I}\left(a_t\in[K]\right)\mathbb{I}\left(\mathcal{G}_t\right)\right]\\
\leq&\mathbb{E}\left[\sqrt{\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j)}-u_k^{\star(j)}\right|^2}\sqrt{\sum_{j=1}^{J}p_j\mathbb{I}\left(a_t\in[K]\right)}\mathbb{I}\left(\mathcal{G}_t\right)\right]
\end{aligned}$$

By definition of $\tilde{u}_{k,t}^{(j)}$ and triangular inequality for ℓ_2-norm,

$$\begin{aligned}
\sqrt{\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j)}-u_k^{\star(j)}\right|^2}\mathbb{I}\left(\mathcal{G}_t\right)=&\sqrt{\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\widehat{u}_{k,t}^{(j)}-u_k^{\star(j)}+\frac{\gamma_{t-1,\sigma_r}(\delta)}{\sqrt{p_j}}\right|^2}\mathbb{I}\left(\mathcal{G}_t\right)\\
\leq&\left(\sqrt{\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\widehat{u}_{k,t}^{(j)}-u_k^{\star(j)}\right|^2}+\sqrt{\sum_{j=1}^{J}p_j\left(\frac{\gamma_{t-1,\sigma_r}(\delta)}{\sqrt{p_j}}\right)^2}\right)\mathbb{I}\left(\mathcal{G}_t\right)\\
\leq&2\gamma_{t-1,\sigma_r}(\delta)\mathbb{I}\left(\mathcal{G}_t\right)\\
\leq&2\gamma_{t-1,\sigma_r}(\delta).
\end{aligned}$$

Then by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)}-u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right]\leq&\mathbb{E}\left[\sqrt{\sum_{j=1}^{J}p_j\max_{k\in[K]}\left|\tilde{u}_{k,t}^{(j)}-u_k^{\star(j)}\right|^2}\sqrt{\sum_{j=1}^{J}p_j\mathbb{I}\left(a_t\in[K]\right)}\mathbb{I}\left(\mathcal{G}_t\right)\right]\\
\leq&2\gamma_{t-1,\sigma_r}(\delta)\mathbb{E}\left[\sqrt{\sum_{j=1}^{J}p_j\mathbb{I}\left(a_t\in[K]\right)}\right]\\
\leq&2\gamma_{t-1,\sigma_r}(\delta)\sqrt{\mathbb{E}\left[\sum_{j=1}^{J}p_j\mathbb{I}\left(a_t\in[K]\right)\right]}\\
=&2\gamma_{t-1,\sigma_r}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]},
\end{aligned}$$

which proves (43). Similarly,

$$\mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left\|\tilde{\mathbf{b}}_{k,t}^{(j_t)}-\mathbf{b}_k^{\star(j_t)}\right\|_\infty\mathbb{I}\left(\mathcal{G}_t\right)\right]\leq 2\gamma_{t-1,\sigma_b}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]} \tag{44}$$

Step 2. Reward decomposition: Let τ be the stopping time of the algorithm and let $\mathcal{U}:=\{t\in[\tau]:\rho_t>0\}$. Then for $t\notin\mathcal{U}$, the allocated resource is $\rho_t\vee 0=0$ and the algorithm skips the round. Thus,

$$\mathbb{E}\left[\sum_{t=1}^{T}R_t^{\widehat{\pi}}\right]=\mathbb{E}\left[\sum_{t\in\mathcal{U}}R_t^{\widehat{\pi}}\right].$$

Then, the reward is decomposed as

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in\mathcal{U}} R_t^{\widehat{\pi}}\right] =&\mathbb{E}\left[\sum_{t\in\mathcal{U}} R_t^{\widehat{\pi}}\mathbb{I}\left(\mathcal{G}_t\right)\right] + \mathbb{E}\left[\sum_{t\in\mathcal{U}} R_t^{\widehat{\pi}}\mathbb{I}\left(\mathcal{G}_t^c\right)\right]\\ \geq&\mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}u_k^{\star(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right] - \sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right)\\ \geq&\mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right] - \mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)} - u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right] - \sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right)\\ \geq&\mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right] - \sum_{t=1}^{T}\mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)} - u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right] - \sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right).\end{aligned}$$

By the bound (43),

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\left|\tilde{u}_{k,t}^{(j_t)} - u_k^{\star(j_t)}\right|\mathbb{I}\left(\mathcal{G}_t\right)\right] \leq&2\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]}\\ \leq&2\sqrt{T\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^2\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]}\\ =&2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]}\end{aligned}$$

where the last ineqaulity holds by Cauchy-Schwartz inequality. Thus, the reward is decomposed as

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} R_t^{\widehat{\pi}}\right] =&\mathbb{E}\left[\sum_{t\in\mathcal{U}} R_t^{\widehat{\pi}}\right]\\ \geq&\mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right] - 2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]} - \sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right)\end{aligned} \tag{45}$$

Step 3. A lower bound for ρ_t: Denote $u_1 < u_2 < \ldots < u_{|\mathcal{U}|}$ the indexes in $\mathcal{U}$. For $s \notin \mathcal{U}$, we have $\rho_s = \mathbf{0}_m$ and $\mathbf{b}_{a_s,s}^{(j_s)} = \mathbf{0}_m$. Thus for $\nu \in [|\mathcal{U}| - 1]$,

$$\rho_{u_{\nu+1}} = u_{\nu+1}\rho - \sum_{s=1}^{u_{\nu+1}-1}\mathbf{b}_{a_s,s}^{(j_s)} = u_{\nu+1}\rho - \sum_{s=1}^{u_\nu}\mathbf{b}_{a_s,s}^{(j_s)}. \tag{46}$$

By the resource constrain at round u_ν,

$$\begin{aligned}\sum_{k=1}^{K}\widehat{\pi}_{k,u_\nu}^{(j_{u_\nu})}\tilde{\mathbf{b}}_{k,u_\nu}^{(j_{u_\nu})} \leq&u_\nu\rho - \sum_{s=1}^{u_\nu-1}\mathbf{b}_{a_s,s}^{(j_s)}\\ =&u_\nu\rho + \mathbf{b}_{a_{u_\nu},u_\nu}^{(j_{u_\nu})} - \sum_{s=1}^{u_\nu}\mathbf{b}_{a_s,s}^{(j_s)}.\end{aligned}$$

Plugging in (46),

$$\begin{aligned}\rho_{u_{\nu+1}} \geq&\left(u_{\nu+1} - u_\nu\right)\rho - \mathbf{b}_{a_{u_\nu},u_\nu}^{(j_{u_\nu})} + \sum_{k=1}^{K}\widehat{\pi}_{k,u_\nu}^{(j_{u_\nu})}\tilde{\mathbf{b}}_{k,u_\nu}^{(j_{u_\nu})}\\ \geq&\left(u_{\nu+1} - u_\nu\right)\rho - \mathbf{b}_{a_{u_\nu},u_\nu}^{(j_{u_\nu})} + \sum_{k=1}^{K}\widehat{\pi}_{k,u_\nu}^{(j_{u_\nu})}\mathbf{b}_k^{\star(j_{u_\nu})} + \sum_{k=1}^{K}\widehat{\pi}_{k,u_\nu}^{(j_{u_\nu})}\left(\tilde{\mathbf{b}}_{k,u_\nu}^{(j_{u_\nu})} - \mathbf{b}_k^{\star(j_{u_\nu})}\right).\end{aligned}$$

Taking conditional expectation on both sides gives

$$\begin{aligned}\mathbb{E}\left[\rho_{u_{\nu+1}}\middle|\, j_{u_{\nu+1}}\right] \geq& \mathbb{E}\left[u_{\nu+1}-u_\nu\middle|\, j_{u_{\nu+1}}\right]\rho+\mathbb{E}\left[-\mathbf{b}^{(j_{u_\nu})}_{a_{u_\nu},u_\nu}+\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\mathbf{b}^{\star(j_{u_\nu})}_k\middle|\, j_{u_{\nu+1}}\right]\\ &+\mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\left(\tilde{\mathbf{b}}^{(j_{u_\nu})}_{k,u_\nu}-\mathbf{b}^{\star(j_{u_\nu})}_k\right)\middle|\, j_{u_{\nu+1}}\right]\\ =&\mathbb{E}\left[u_{\nu+1}-u_\nu\middle|\, j_{u_{\nu+1}}\right]\rho+\mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\left(\tilde{\mathbf{b}}^{(j_{u_\nu})}_{k,u_\nu}-\mathbf{b}^{\star(j_{u_\nu})}_k\right)\right]\\ \geq&\mathbb{E}\left[u_{\nu+1}-u_\nu\middle|\, j_{u_{\nu+1}}\right]\rho+\mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\left(\tilde{\mathbf{b}}^{(j_{u_\nu})}_{k,u_\nu}-\mathbf{b}^{\star(j_{u_\nu})}_k\right)\mathbb{I}\left(\mathcal{G}_{u_\nu}\right)\right]-\mathbb{P}\left(\mathcal{G}^c_{u_\nu}\right)\mathbf{1}_m,\end{aligned}$$

where the equality holds by Assumption 1 and

$$\begin{aligned}\mathbb{E}\left[\left\{-\mathbf{b}^{(j_{u_\nu})}_{a_{u_\nu},u_\nu}+\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\mathbf{b}^{\star(j_{u_\nu})}_k\right\}\right] =&\mathbb{E}\left[\left\{-\mathbf{b}^{\star(j_{u_\nu})}_{a_{u_\nu}}+\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\mathbf{b}^{\star(j_{u_\nu})}_k\right\}\right]\\ =&\mathbb{E}\left[\left\{-\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\mathbf{b}^{\star(j_{u_\nu})}_k+\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\mathbf{b}^{\star(j_{u_\nu})}_k\right\}\right]\\ =&0.\end{aligned}$$

For the last term, by the bound (44),

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\left(\tilde{\mathbf{b}}^{(j_{u_\nu})}_{k,u_\nu}-\mathbf{b}^{\star(j_{u_\nu})}_k\right)\mathbb{I}\left(\mathcal{G}_{u_\nu}\right)\right] \geq&-\mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu})}_{k,u_\nu}\left\|\tilde{\mathbf{b}}^{(j_{u_\nu})}_{k,u_\nu}-\mathbf{b}^{\star(j_{u_\nu})}_k\right\|_\infty\mathbb{I}\left(\mathcal{G}_{u_\nu}\right)\right]\mathbf{1}_m\\ \geq&-\mathbb{E}\left[2\gamma_{u_\nu-1,\sigma_b}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu}\in[K]\right)\middle|\, u_\nu\right]}\right]\mathbf{1}_m.\end{aligned}$$

Thus we obtain a lower bound,

$$\begin{aligned}\mathbb{E}\left[\rho_{u_{\nu+1}}\middle|\, j_{u_{\nu+1}}\right] \geq&\mathbb{E}\left[u_{\nu+1}-u_\nu\middle|\, j_{u_{\nu+1}}\right]\rho-\mathbb{P}\left(\mathcal{G}^c_{u_\nu}\right)\mathbf{1}_m\\ &-2\mathbb{E}\left[\gamma_{u_\nu-1,\sigma_b}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu}\in[K]\right)\middle|\, u_\nu\right]}\right]\mathbf{1}_m.\end{aligned}\tag{47}$$

Step 4. An upper bound for OPT In the optimization problem (1), all constraints are linear with respect to the variable and there exist a feasible solution. Thus the problem satisfies the Slater's condition and strong duality (Boyd et al., 2004). Then,

$$\frac{OPT}{T}=\max_{\pi^{(j)}_k}\min_{\lambda\in\mathbb{R}^m_+}\min_{\mu^{(j)}\geq 0}\min_{\nu^{(j)}_k\geq 0} L\left(\pi^{(j)}_k,\lambda,\mu^{(j)},\nu^{(j)}_k\right),$$

where L is the Lagrangian function:

$$\begin{aligned}L\left(\pi^{(j)}_k,\lambda,\mu^{(j)},\nu^{(j)}_k\right):=&\sum_{j=1}^J\sum_{k=1}^K p_j\pi^{(j)}_k u^{\star(j)}_k+\left(\rho-\sum_{j=1}^J\sum_{k=1}^K p_j\pi^{(j)}_k\mathbf{b}^{\star(j)}_k\right)^\top\lambda\\ &+\sum_{j=1}^J\mu^{(j)}\left(1-\sum_{k=1}^K\pi^{(j)}_{k,1}\right)+\sum_{j=1}^J\sum_{k=1}^K\nu^{(j)}_k\pi^{(j)}_{k,1}.\end{aligned}$$

Minimizing over $\mu^{(j)}$ and $\nu_k^{(j)}$ gives

$$\begin{aligned}
&\min_{\mu_t^{(j)}\geq 0}\min_{\nu_{k,t}^{(j)}\geq 0} L\left(\pi_k^{(j)},\lambda,\mu^{(j)},\nu_k^{(j)}\right)\\
&=\begin{cases}\sum_{j=1}^J\sum_{k=1}^K p_j\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{j=1}^J\sum_{k=1}^K p_j\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda & \sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0\\ -\infty & o.w.\end{cases},
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{OPT}{T}=&\max_{\pi_k^{(j)}}\min_{\lambda\in\mathbb{R}_+^m}\min_{\mu_t^{(j)}\geq 0}\min_{\nu_{k,t}^{(j)}\geq 0} L\left(\pi_k^{(j)},\lambda,\mu^{(j)},\nu_k^{(j)}\right)\\
\leq&\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0}\min_{\lambda\in\mathbb{R}_+^m}\sum_{j=1}^J\sum_{k=1}^K p_j\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{j=1}^J\sum_{k=1}^K p_j\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\\
=&\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0}\min_{\lambda\in\mathbb{R}_+^m}\sum_{j=1}^J p_j\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\right\}\\
\leq&\min_{\lambda\in\mathbb{R}_+^m}\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0}\sum_{j=1}^J p_j\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\right\}\\
\leq&\min_{\lambda\in\mathbb{R}_+^m}\sum_{j=1}^J p_j\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0}\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\right\}.
\end{aligned}$$

Let $\{\bar{\pi}_k^{(j)}: j\in[J],k\in[K]\}$ be the maximizer. If $\rho-\sum_{k=1}^K\bar{\pi}_k^{(j)}\mathbf{b}_k^{\star(j)}$ is negative for some element and $j\in[J]$, then the optimal value becomes $-\infty$. Thus

$$\begin{aligned}
\frac{OPT}{T}\leq&\min_{\lambda\in\mathbb{R}_+^m}\sum_{j=1}^J p_j\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0}\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\right\}\\
=&\min_{\lambda\in\mathbb{R}_+^m}\sum_{j=1}^J p_j\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0,\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\geq 0}\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}+\left(\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\right)^\top\lambda\right\}\\
=&\sum_{j=1}^J p_j\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0,\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\geq 0}\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}\right\}.
\end{aligned}$$

For each $j\in[J]$ and $\mathbf{v}\in\mathbb{R}_+^m$, let $\tilde{\pi}_{k,\mathbf{v}}^{(j)}$ be the solution to the optimization problem:

$$\begin{aligned}
&\max_{\pi_{k,\mathbf{v}}^{(j)}}\sum_{k=1}^K\pi_{k,\mathbf{v}}^{(j)}u_k^{\star(j)}\\
&\text{s.t.}\sum_{k=1}^K\pi_{k,\mathbf{v}}^{(j)}\mathbf{b}_k^{\star(j)}\leq\mathbf{v}.
\end{aligned}\tag{48}$$

Then,

$$\begin{aligned}
\frac{OPT}{T}\leq&\sum_{j=1}^J p_j\max_{\sum_{k=1}^K\pi_k^{(j)}\leq 1,\pi_k^{(j)}\geq 0,\rho-\sum_{k=1}^K\pi_k^{(j)}\mathbf{b}_k^{\star(j)}\geq 0}\left\{\sum_{k=1}^K\pi_k^{(j)}u_k^{\star(j)}\right\}\\
=&\sum_{j=1}^J p_j\sum_{k=1}^K\tilde{\pi}_{k,\rho}^{(j)}u_k^{\star(j)}.
\end{aligned}$$

For each $\nu \in [|\mathcal{U}| - 1]$,

$$\begin{aligned}\mathbb{E}\left[(u_{\nu+1} - u_\nu)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[(u_{\nu+1} - u_\nu)\sum_{j=1}^{J} p_j \sum_{k=1}^{K} \tilde{\pi}_{k,\rho}^{(j)} u_k^{\star(j)}\right] \\ = & \mathbb{E}\left[(u_{\nu+1} - u_\nu)\sum_{k=1}^{K} \tilde{\pi}_{k,\rho}^{(j_{u_{\nu+1}})} u_k^{\star(j_{u_{\nu+1}})}\right]\end{aligned}$$

In (48), all constraints are linear with respect to the variable and there exist a feasible solution. Thus the problem satisfies the Slater's condition and strong duality (Boyd et al., 2004). The dual problem of (48) is

$$\begin{aligned}\min_{\lambda_{\mathbf{v}}^{(j)} \in \mathbb{R}_+^m} \quad & \mathbf{v}^\top \lambda_{\mathbf{v}}^{(j)} \\ \text{s.t.} \quad & \left(\mathbf{b}_k^{\star(j)}\right)^\top \lambda_{\mathbf{v}}^{(j)} \geq u_k^{\star(j)}, \quad \forall k \in [K].\end{aligned} \tag{49}$$

Let $\tilde{\lambda}_{\mathbf{v}}^{(j)}$ be the solution to (49). By strong duality, for each $\nu \in [|\mathcal{U}| - 1]$,

$$\begin{aligned}& \mathbb{E}\left[(u_{\nu+1} - u_\nu)\sum_{k=1}^{K} \tilde{\pi}_{k,\rho}^{(j_{u_{\nu+1}})} u_k^{\star(j_{u_{\nu+1}})}\right] \\ & = \mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho^\top \tilde{\lambda}_\rho^{(j_{u_{\nu+1}})}\right] \\ & = \mathbb{E}\left[\mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho \middle| j_{u_{\nu+1}}\right]^\top \tilde{\lambda}_\rho^{(j_{u_{\nu+1}})}\right] \\ & = \mathbb{E}\left[\left(\mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right) + 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right]\right)\mathbf{1}_m^\top \tilde{\lambda}_\rho^{(j_{u_{\nu+1}})}\right] \\ & \quad + \mathbb{E}\left[\left(\mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho \middle| j_{u_{\nu+1}}\right] - \mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right) - 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\sigma_b}(\delta)\right]\right)\mathbf{1}_m^\top \tilde{\lambda}_\rho^{(j_{u_{\nu+1}})}\right].\end{aligned} \tag{50}$$

For the first term, we observe the dual problem of (1),

$$\begin{aligned}\min_{\lambda \in \mathbb{R}_+^m} \quad & \rho \mathbf{1}_m^\top \lambda \\ \text{s.t.} \quad & \lambda^\top \mathbf{b}_k^{\star(j)} \geq u_k^{\star(j)}, \forall j \in [J], \forall k \in [K].\end{aligned} \tag{51}$$

Comparing to the dual problem (49), when $\mathbf{v} = \rho \mathbf{1}_m$ and $j = j_{u_{\nu+1}}$, (51) has more constraints than (49) with same objective function. Denote $\lambda_\star$ be the solution to (51). Then,

$$\rho \mathbf{1}_m^\top \tilde{\lambda}_{\rho \mathbf{1}_m}^{(j_{u_{\nu+1}})} \leq \rho \mathbf{1}_m^\top \lambda_\star = \frac{OPT}{T},$$

where the last equality holds by strong duality for the oracle problem (1). Thus the first term in (50) is bounded as

$$\begin{aligned}& \mathbb{E}\left[\left(\mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right) + 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right]\right)\mathbf{1}_m^\top \tilde{\lambda}_{\rho \mathbf{1}_m}^{(j_{u_{\nu+1}})}\right] \\ & \leq \left(\mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right) + 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right]\right)\frac{OPT}{\rho T}.\end{aligned}$$

For the second term in (50), we observe that $\tilde{\lambda}_{\mathbb{E}\left[\rho_{u_{\nu+1}} \middle| j_{u_{\nu+1}}\right]}^{(j_{u_{\nu+1}})}$ is a feasible solution to (49) when $\mathbf{v} = \rho \mathbf{1}_m$ and $j = j_{u_{\nu+1}}$. Thus

$$\begin{aligned}& \mathbb{E}\left[\left(\mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho \middle| j_{u_{\nu+1}}\right] - 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right] - \mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right)\right)\mathbf{1}_m^\top \tilde{\lambda}_{\rho \mathbf{1}_m}^{(j_{u_{\nu+1}})}\right] \\ & \leq \mathbb{E}\left[\left\{\left(\mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho \middle| j_{u_{\nu+1}}\right] - 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right] - \mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right)\right) \vee 0\right\}\mathbf{1}_m^\top \tilde{\lambda}_{\rho \mathbf{1}_m}^{(j_{u_{\nu+1}})}\right] \\ & \leq \mathbb{E}\left[\left\{\left(\mathbb{E}\left[(u_{\nu+1} - u_\nu)\,\rho \middle| j_{u_{\nu+1}}\right] - 2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu} \in [K]\right) \middle| u_\nu\right]}\gamma_{u_\nu - 1,\lambda}(\sigma_b)\right] - \mathbb{P}\left(\mathcal{G}_{u_\nu}^c\right)\right) \vee 0\right\}\mathbf{1}_m^\top \tilde{\lambda}_{\mathbb{E}\left[\rho_{u_{\nu+1}} \middle| j_{u_{\nu+1}}\right]}^{(j_{u_{\nu+1}})}\right] \\ & \leq \mathbb{E}\left[\left\{\mathbb{E}\left[\rho_{u_{\nu+1}} \middle| j_{u_{\nu+1}}\right] \vee \mathbf{0_m}\right\}^\top \tilde{\lambda}_{\mathbb{E}\left[\rho_{u_{\nu+1}} \middle| j_{u_{\nu+1}}\right]}^{(j_{u_{\nu+1}})}\right],\end{aligned}$$

where the last inequality holds by (47). Because $u_{\nu+1} \in \mathcal{U}$, we have $\rho_{u_{\nu+1}} > 0$ and

$$\begin{aligned}\mathbb{E}\left[\left\{\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right] \vee \mathbf{0_m}\right\}^\top \tilde{\lambda}^{(j_{u_\nu+1})}_{\mathbb{E}\left[\rho_{u_{\nu+1}}|j_{u_{\nu+1}}\right]}\right] \leq & \mathbb{E}\left[\mathbb{E}\left[\rho_{u_{\nu+1}} \vee \mathbf{0_m}\middle| j_{u_{\nu+1}}\right]^\top \tilde{\lambda}^{(j_{u_\nu+1})}_{\mathbb{E}\left[\rho_{u_{\nu+1}}|j_{u_{\nu+1}}\right]}\right] \\ = & \mathbb{E}\left[\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right]^\top \tilde{\lambda}^{(j_{u_\nu+1})}_{\mathbb{E}\left[\rho_{u_{\nu+1}}|j_{u_{\nu+1}}\right]}\right].\end{aligned}$$

Collecting the bounds, we have

$$\begin{aligned}\mathbb{E}\left[(u_{\nu+1}-u_\nu)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right]^\top \tilde{\lambda}^{(j_{u_\nu+1})}_{\mathbb{E}\left[\rho_{u_{\nu+1}}|j_{u_{\nu+1}}\right]}\right] \\ & + \left(2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu}\in[K]\right)\middle| u_\nu\right]}\gamma_{u_\nu-1,\sigma_b}(\delta)\right] + \mathbb{P}\left(\mathcal{G}^c_{u_\nu}\right)\right)\frac{OPT}{\rho T}.\end{aligned}$$

Similar to Step 4, by strong duality,

$$\begin{aligned}&\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right]^\top \tilde{\lambda}^{(j_{u_\nu+1})}_{\mathbb{E}\left[\rho_{u_{\nu+1}}|j_{u_{\nu+1}}\right]} \\ &= \max_{\sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\leq 1, \pi_k^{(j_{u_\nu+1})}\geq 0} \min_{\lambda\in\mathbb{R}^m_+} \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})} u_k^{\star(j_{u_\nu+1})} + \left(\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right] - \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\mathbf{b}_k^{\star(j_{u_\nu+1})}\right)^\top \lambda \\ &\leq \min_{\lambda\in\mathbb{R}^m_+} \max_{\sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\leq 1, \pi_k^{(j_{u_\nu+1})}\geq 0} \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})} u_k^{\star(j_{u_\nu+1})} + \left(\mathbb{E}\left[\rho_{u_{\nu+1}}\middle| j_{u_{\nu+1}}\right] - \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\mathbf{b}_k^{\star(j_{u_\nu+1})}\right)^\top \lambda \\ &\leq \min_{\lambda\in\mathbb{R}^m_+} \mathbb{E}\left[\max_{\sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\leq 1, \pi_k^{(j_{u_\nu+1})}\geq 0} \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})} u_k^{\star(j_{u_\nu+1})} + \left(\rho_{u_{\nu+1}} - \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\mathbf{b}_k^{\star(j_{u_\nu+1})}\right)^\top \lambda \middle| j_{u_{\nu+1}}\right] \\ &\leq \mathbb{E}\left[\max_{\sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\leq 1, \pi_k^{(j_{u_\nu+1})}\geq 0, \rho_{u_{\nu+1}} - \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})}\mathbf{b}_k^{\star(j_{u_\nu+1})}\geq 0} \sum_{k=1}^K \pi_k^{(j_{u_\nu+1})} u_k^{\star(j_{u_\nu+1})} \middle| j_{u_{\nu+1}}\right] \\ &= \sum_{k=1}^K \tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}} u_k^{\star(j_{u_\nu+1})}.\end{aligned}$$

Thus we have

$$\begin{aligned}\mathbb{E}\left[(u_{\nu+1}-u_\nu)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[\sum_{k=1}^K \tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}} u_k^{\star(j_{u_\nu+1})}\right] \\ & + \left(2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_\nu}\in[K]\right)\middle| u_\nu\right]}\gamma_{u_\nu-1,\sigma_b}(\delta)\right] + \mathbb{P}\left(\mathcal{G}^c_{u_\nu}\right)\right)\frac{OPT}{\rho T}.\end{aligned}$$

Under the event $\mathcal{G}_{u_{\nu+1}}$, the policy $\tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}}$ is a feasible solution to the bandit problem (12),

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^K \tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}} u_k^{\star(j_{u_\nu+1})}\right] \leq & \mathbb{E}\left[\sum_{k=1}^K \tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}} u_k^{\star(j_{u_\nu+1})}\mathbb{I}\left(\mathcal{G}_{u_{\nu+1}}\right)\right] + \mathbb{P}\left(\mathcal{G}^c_{u_{\nu+1}}\right) \\ \leq & \mathbb{E}\left[\sum_{k=1}^K \tilde{\pi}^{(j_{u_\nu+1})}_{k,\rho_{u_{\nu+1}}} \tilde{u}^{(j_{u_\nu+1})}_{k,u_{\nu+1}}\mathbb{I}\left(\mathcal{G}_{u_{\nu+1}}\right)\right] + \mathbb{P}\left(\mathcal{G}^c_{u_{\nu+1}}\right) \\ \leq & \mathbb{E}\left[\sum_{k=1}^K \widehat{\pi}^{(j_{u_\nu+1})}_{k,u_{\nu+1}} \tilde{u}^{(j_{u_\nu+1})}_{k,u_{\nu+1}}\mathbb{I}\left(\mathcal{G}_{u_{\nu+1}}\right)\right] + \mathbb{P}\left(\mathcal{G}^c_{u_{\nu+1}}\right).\end{aligned}$$

Thus, for each $\nu \in [|\mathcal{U}| - 1]$,

$$\begin{aligned}\mathbb{E}\left[\left(u_{\nu+1}-u_{\nu}\right)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[\sum_{k=1}^{K}\widehat{\pi}_{k,u_{\nu+1}}^{(j_{u_{\nu+1}})}\tilde{u}_{k,u_{\nu+1}}^{(j_{u_{\nu+1}})}\mathbb{I}\left(\mathcal{G}_{u_{\nu+1}}\right)\right]+\mathbb{P}\left(\mathcal{G}_{u_{\nu+1}}^{c}\right) \\ & +\left(2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_{\nu}}\in[K]\right)|\,u_{\nu}\right]}\gamma_{u_{\nu}-1,\sigma_b}(\delta)\right]+\mathbb{P}\left(\mathcal{G}_{u_{\nu}}^{c}\right)\right)\frac{OPT}{\rho T}.\end{aligned}$$

Summing up over ν,

$$\begin{aligned}\mathbb{E}\left[\sum_{\nu=1}^{|\mathcal{U}|-1}\left(u_{\nu+1}-u_{\nu}\right)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right]+\left(1+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & +\left(\sum_{\nu=1}^{|\mathcal{U}|-1}2\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_{u_{\nu}}\in[K]\right)|\,u_{\nu}\right]}\gamma_{u_{\nu}-1,\sigma_b}(\delta)\right]\right)\frac{OPT}{\rho T} \\ \leq & \mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right]+\left(1+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & +2\left(\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t\in[K]\right)\right]}\gamma_{t-1,\sigma_b}(\delta)\right]\right)\frac{OPT}{\rho T} \\ \leq & \mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right]+\left(1+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & +2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]}\frac{OPT}{\rho T},\end{aligned}$$

where the last inequality holds by Cauchy-Schwartz inequality, By (45),

$$\begin{aligned}\mathbb{E}\left[\sum_{\nu=1}^{|\mathcal{U}|-1}\left(u_{\nu+1}-u_{\nu}\right)\frac{OPT}{T}\right] \leq & \mathbb{E}\left[\sum_{t\in\mathcal{U}}\sum_{k=1}^{K}\widehat{\pi}_{k,t}^{(j_t)}\tilde{u}_{k,t}^{(j_t)}\mathbb{I}\left(\mathcal{G}_t\right)\right]+\left(1+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & +2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]}\frac{OPT}{\rho T} \\ \leq & \mathbb{E}\left[\sum_{t=1}^{T}R_t^{\widehat{\pi}}\right]+\left(2+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & +2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]}\frac{OPT}{\rho T} \\ & 2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]} \\ \leq & \mathbb{E}\left[\sum_{t=1}^{T}R_t^{\widehat{\pi}}\right]+\left(2+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^{c}\right) \\ & 2\left(1+\frac{OPT}{\rho T}\right)\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b\vee\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]}\end{aligned}$$

Because the last choice of the algorithm happens at round τ, we have $\rho_\tau > 0$ and $u_{|\mathcal{U}|} = \tau$. And by definition, $u_1 = \xi$. Thus

$$\mathbb{E}\left[\sum_{\nu=1}^{|\mathcal{U}|-1}(u_{\nu+1}-u_\nu)\frac{OPT}{T}\right] = \mathbb{E}\left[\left(u_{|\mathcal{U}|}-u_1\right)\frac{OPT}{T}\right] = \frac{OPT}{T}\mathbb{E}\left[\tau-\xi\right].$$

Rearranging the terms

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} R_t^{\widehat{\pi}}\right] \geq{}& \mathbb{E}\left[\sum_{\nu=1}^{|\mathcal{U}|-1}(u_{\nu+1}-u_\nu)\frac{OPT}{T}\right] - \left(2+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right) \\
&- 2\left(1+\frac{OPT}{\rho T}\right)\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b\vee\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]} \\
\geq{}& \frac{OPT}{T}\mathbb{E}\left[\tau-\xi\right] - \left(2+\frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{G}_t^c\right) \\
&- 2\left(1+\frac{OPT}{\rho T}\right)\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b\vee\sigma_r}(\delta)^2\mathbb{I}\left(a_t\in[K]\right)\right]},
\end{aligned}$$

completes the proof. □

B.5. Proof of Lemma 5.3

Proof. Let us fix $\delta \in (0, T^{-2})$ throughout the proof.

Step 1. Bounding the minimum eigenvalue of $\{F_\nu : \nu \in [n_T]\}$: By Lemma C.3, with probability at least $1 - T\delta$,

$$\begin{aligned}
\frac{1}{2Kd}F_\nu =& \frac{1}{2Kd}\sum_{u=1}^{\nu}\tilde{X}_{k,\tau(u)}\tilde{X}_{k,\tau(u)}^\top + 8\frac{K-1}{K}\log\frac{Jd}{\delta} \\
\succeq& \frac{1}{4Kd}\sum_{u=1}^{\nu}\mathbb{E}\left[\tilde{X}_{k,\tau(u)}\tilde{X}_{k,\tau(u)}^\top\middle|\mathcal{H}_{\tau(u)-1}\right] + 8\frac{K-1}{K}\log\frac{Jd}{\delta} - \log\frac{Jd}{\delta} \\
\succeq& \frac{1}{4Kd}\sum_{u=1}^{\nu}\mathbb{E}\left[\tilde{X}_{k,\tau(u)}\tilde{X}_{k,\tau(u)}^\top\middle|\mathcal{H}_{\tau(u)-1}\right],
\end{aligned}$$

for all $\nu \in [n_T]$. By Assumption 2 and 3,

$$\begin{aligned}
\lambda_{\min}\left(\mathbb{E}\left[\tilde{X}_{k,\tau(u)}\tilde{X}_{k,\tau(u)}^\top\middle|\mathcal{H}_{\tau(u)-1}\right]\right) =& \lambda_{\min}\begin{pmatrix} p_1\mathbb{E}_{X_k\sim\mathbb{F}_1}\left[\sum_{k=1}^K X_kX_k^\top\right] & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_J\mathbb{E}_{\mathbf{x}_k\sim\mathbb{F}_J}\left[\sum_{k=1}^K X_kX_k^\top\right] \end{pmatrix} \\
\geq& p_{\min}\min_{j\in[J]}\lambda_{\min}\left(\mathbb{E}_{X_k\sim\mathbb{F}_j}\left[\sum_{k=1}^K X_kX_k^\top\right]\right) \\
\geq& p_{\min}K\alpha.
\end{aligned}$$

Thus, with probability at least $1-\delta/T$,

$$\begin{aligned}\lambda_{\min}(F_\nu) \geq & \frac{1}{2}\lambda_{\min}\left(\sum_{u=1}^{\nu}\mathbb{E}\left[\tilde{X}_{k,u}\tilde{X}_{k,u}^\top\middle|\mathcal{H}_{u-1}\right]\right)\\ \geq & \frac{1}{2}\sum_{u=1}^{\nu}\lambda_{\min}\left(\mathbb{E}\left[\tilde{X}_{k,u}\tilde{X}_{k,u}^\top\middle|\mathcal{H}_{u-1}\right]\right)\\ \geq & \frac{p_{\min}K\alpha\nu}{2},\end{aligned}$$

for all $\nu \in [n_T]$.

Step 2. Bounding the probability of $\mathcal{M}_t$: Under the event proved in Step 1, the event $\mathcal{M}_t$ is implied by

$$\frac{p_{\min}K\alpha n_t}{2} \geq 4Kd\left\{\sum_{\nu=1}^{n_t}\frac{288\,(K-1)\log\left(\frac{Jd}{\delta}\right)}{\alpha K p_{\min}\nu}+35\log\frac{Jd}{\delta}\right\}, \tag{52}$$

for all $t \in [T]$. The right hand side is bounded as

$$\begin{aligned}4Kd\left\{\sum_{\nu=1}^{n_t}\frac{288\,(K-1)\log\left(\frac{Jd}{\delta}\right)}{\alpha K p_{\min}\nu}+35\log\frac{Jd}{\delta}\right\} \leq & \frac{4\cdot 288Kd\log\left(\frac{Jd}{\delta}\right)\log n_t}{\alpha p_{\min}}+140Kd\log\frac{Jd}{\delta}\\ \leq & \frac{4\cdot 288Kd\log\left(\frac{Jd}{\delta}\right)\log T}{\alpha p_{\min}}+140Kd\log\frac{Jd}{\delta}.\end{aligned}$$

Plugging in (52) and rearranging the terms,

$$n_t \geq d\log\left(\frac{Jd}{\delta}\right)\left\{\frac{8\cdot 288\log T}{\alpha^2 p_{\min}^2}+\frac{280}{\alpha p_{\min}}\right\},$$

implies the event $\mathcal{M}_t$ for all $t \in [T]$ with probability at least $1-T\delta$. In other words,

$$\mathbb{P}\left(\mathcal{M}_t^c\right) \leq \mathbb{P}\left(n_t < dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)\right)+T\delta,$$

for all $t \in [T]$, where $M_{\alpha,p,T} := \left\{\frac{8\cdot 288\log T}{\alpha^2 p_{\min}^2}+\frac{280}{\alpha p_{\min}}\right\}$.

Step 3. Bounding ξ: Let $\tilde{t} = \inf_{t\in[T]}\{\mathcal{M}_t \text{ happens}\}$ be the first round that $\mathcal{M}_t$ happens. After round $\tilde{t}$, the algorithm skips the rounds until $\rho_t > 0$ holds and then pulls an action according to the policy. Thus, for the round $\xi-1$,

$$(\xi-1)\,\rho-\sum_{s=1}^{\xi-2}\mathbf{b}_{a_s,s}^{(j_s)} = (\xi-1)\,\rho-\sum_{s=1}^{\tilde{t}}\mathbf{b}_{a_s,s}^{(j_s)} \leq 0.$$

rearranging the terms, and taking expectation,

$$\mathbb{E}\left[\xi\right] \leq 1+\rho^{-1}\mathbb{E}\left[\sum_{s=1}^{\tilde{t}}\mathbf{b}_{a_s,s}^{(j_s)}\right] \leq 1+\rho^{-1}\mathbb{E}\left[\tilde{t}\right]. \tag{53}$$

Now we need an upper bound for $\tilde{t}$. For $t \in [\tilde{t}-1]$, the event $\mathcal{M}_t$ does not happen and the algorithm admits the arrival for $t \in [\tilde{t}]$. Thus, $n_t = t$ for all $t \in [\tilde{t}]$. For $t = \tilde{t}-1$, the event $\mathcal{M}_{\tilde{t}-1}$ does not happen and

$$\lambda_{\min}\left(F_{n_{\tilde{t}-1}}\right) \leq 4Kd\left\{\sum_{\nu=1}^{n_{\tilde{t}-1}}\frac{144\,(K-1)\log\left(\frac{Jd}{\delta}\right)}{\lambda_{\min}(F_\nu)}+35\log\frac{Jd}{\delta}\right\}.$$

By the fact proved in Step 1, with probability at least $1-\delta/T$,

$$\frac{p_{\min}K\alpha n_{\tilde{t}-1}}{2} \leq 4Kd\left\{\sum_{\nu=1}^{n_{\tilde{t}-1}}\frac{288\left(K-1\right)\log\left(\frac{Jd}{\delta}\right)}{p_{\min}K\alpha\nu}+35\log\frac{Jd}{\delta}\right\}.$$

Plugging in $n_{\tilde{t}-1}=\tilde{t}-1$ and rearranging the terms,

$$\begin{aligned}\tilde{t}-1\leq&\frac{4d}{p_{\min}\alpha}\left\{\sum_{\nu=1}^{\tilde{t}-1}\frac{288\left(K-1\right)\log\left(\frac{Jd}{\delta}\right)}{p_{\min}K\alpha\nu}+35\log\frac{Jd}{\delta}\right\}\\ \leq&\frac{4d}{p_{\min}\alpha}\left\{\frac{288\log T}{p_{\min}\alpha}+2\right\}\log\frac{Jd}{\delta}.\end{aligned}$$

Then with probability at least $1-\delta/T$,

$$\begin{aligned}\tilde{t}\leq&1+d\log\left(\frac{Jd}{\delta}\right)\left\{\frac{8\cdot288\log T}{\alpha^2p_{\min}^2}+\frac{140}{\alpha p_{\min}}\right\}\\ :=&1+M_{\alpha,p,T}d\log\left(\frac{Jd}{\delta}\right).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\tilde{t}\right]=&\mathbb{E}\left[\tilde{t}\mathbb{I}\left(\tilde{t}<1+M_{\alpha,p,T}d\log\left(\frac{Jd}{\delta}\right)\right)\right]+\mathbb{E}\left[\tilde{t}\mathbb{I}\left(\tilde{t}\geq1+M_{\alpha,p,T}d\log\left(\frac{Jd}{\delta}\right)\right)\right]\\ \leq&1+dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+T\mathbb{P}\left(\tilde{t}\geq1+M_{\alpha,p,T}d\log\left(\frac{Jd}{\delta}\right)\right)\\ \leq&1+dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+T^2\delta\end{aligned}$$

Plugging in (53),

$$\begin{aligned}\mathbb{E}\left[\xi\right]\leq&1+\rho^{-1}\mathbb{E}\left[\tilde{t}\right]\\ \leq&1+\frac{1+dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+T^2\delta}{\rho}.\end{aligned}$$

Step 4. Proving the lower bound for τ: Let τ be the stopping time of the algorithm. Because the algorithm admits arrival at round τ, we have $\rho_\tau>0$. From the resource constraint in the bandit problem (12),

$$\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\left(\widehat{\mathbf{b}}_{k,\tau}^{(j_\tau)}-\frac{\gamma_{\tau-1,\sigma_b}(\delta)}{\sqrt{p_{j_\tau}}}\mathbf{1}_m\right):=\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\tilde{\mathbf{b}}_{k,\tau}^{(j_\tau)}:\leq\tau\rho-\sum_{s=1}^{\tau-1}\mathbf{b}_{a_s,s}^{(j_s)}$$

Because algorithm stops at round τ, there exists an $r\in[m]$ such that $\sum_{s=1}^{\tau}b_{a_s,s}^{(j_s)}(r)\geq T\rho(r)$. Rearranging the terms,

$$\begin{aligned}\tau\rho\geq&\sum_{s=1}^{\tau-1}b_{a_s,s}^{(j_s)}(r)+\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\tilde{b}_{k,\tau}^{(j_\tau)}(r)\\ \geq&T\rho-b_{a_\tau,\tau}^{(j_\tau)}(r)+\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\tilde{b}_{k,\tau}^{(j_\tau)}(r)\\ =&T\rho-b_{a_\tau,\tau}^{(j_\tau)}(r)+\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}b_{k,\tau}^{\star(j_\tau)}(r)+\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\left(\tilde{b}_{k,\tau}^{(j_\tau)}(r)-b_{k,\tau}^{\star(j_\tau)}(r)\right)\\ \geq&T\rho-b_{a_\tau,\tau}^{(j_\tau)}(r)+\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}b_{k}^{\star(j_\tau)}(r)-\sum_{k=1}^{K}\widehat{\pi}_{k,\tau}^{(j_\tau)}\left\|\tilde{\mathbf{b}}_{k,\tau}^{(j_\tau)}-\mathbf{b}_{k}^{\star(j_\tau)}\right\|_\infty.\end{aligned}$$

Taking expectation on both side,

$$\begin{aligned}
\mathbb{E}\left[\tau\rho\right] \geq & T\rho + \mathbb{E}\left[-b^{(j_\tau)}_{a_\tau,\tau}(r) + \sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} b^{\star(j_\tau)}_{k,\tau}(r)\right] - \mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty\right] \\
= & T\rho - \mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty\right] \\
= & T\rho - \mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty \mathbb{I}\left(\mathcal{E}_\tau \cap \mathcal{M}_{\tau-1}\right)\right] - \mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty \mathbb{I}\left(\mathcal{E}^c_\tau \cup \mathcal{M}^c_{\tau-1}\right)\right] \\
\geq & T\rho - 2\mathbb{E}\left[\gamma_{\tau-1,\sigma_b}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t \in [K]\right)|\,\tau\right]}\right] - \mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty \mathbb{I}\left(\mathcal{E}^c_\tau \cup \mathcal{M}^c_{\tau-1}\right)\right],
\end{aligned} \tag{54}$$

where the last inequality holds by (44). Because $\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau}(r) \leq T\rho$ almost surely,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K} \widehat{\pi}^{(j_\tau)}_{k,\tau} \left\|\tilde{\mathbf{b}}^{(j_\tau)}_{k,\tau} - \mathbf{b}^{\star(j_\tau)}_{k}\right\|_\infty \mathbb{I}\left(\mathcal{E}^c_\tau \cup \mathcal{M}^c_{\tau-1}\right)\right] \leq & T\rho\mathbb{P}\left(\mathcal{E}^c_\tau \cup \mathcal{M}^c_{\tau-1}\right) \\
= & T\rho\mathbb{P}\left(\mathcal{E}^c_\tau\right) \\
\leq & T\rho\left(4(m+1)\delta + 7T^{-1}\right) \\
= & 7\rho + 4(m+1)T\delta,
\end{aligned}$$

where the equality holds because the algorithm takes action according to the policy at round τ and the last inequality holds by Theorem 4.2. from (54),

$$\begin{aligned}
\mathbb{E}\left[\tau\rho\right] \geq & T\rho - 7\rho + 4(m+1)T\rho\delta - 2\mathbb{E}\left[\gamma_{\tau-1,\sigma_b}(\delta)\sqrt{\mathbb{E}\left[\mathbb{I}\left(a_t \in [K]\right)|\,\tau\right]}\right] \\
\geq & T\rho - 7\rho + 4(m+1)T\rho\delta - 2\gamma_{1,\sigma_b}(\delta)
\end{aligned}$$

Rearranging the terms,

$$\mathbb{E}\left[T - \tau\right] \leq \frac{4(m+1)T\delta + 7 + 2\gamma_{\tau-1,\sigma_b}(\delta)}{\rho}.$$

Step 5. Proving a bound for the sum of probabilities Because the algorithm admits the arrival when $\mathcal{M}_{t-1}$ does not happen,

$$\mathcal{M}^c_{t-1} = \mathcal{M}^c_{t-1} \cap \left\{a_t \in [K]\right\}.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^c_{t-1}\right) = & \mathbb{P}\left(\mathcal{M}^c_{t-1} \cap \left\{a_t \in [K]\right\}\right) \\
= & \mathbb{P}\left(\mathcal{M}^c_{t-1} \cap \left\{a_t \in [K]\right\} \cap \left\{n_{t-1} \geq M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right) \\
& + \mathbb{P}\left(\mathcal{M}^c_{t-1} \cap \left\{a_t \in [K]\right\} \cap \left\{n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right) \\
\leq & \mathbb{P}\left(\mathcal{M}^c_{t-1} \cap \left\{n_{t-1} \geq M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right) \\
& + \mathbb{P}\left(\left\{a_t \in [K]\right\} \cap \left\{n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right) \\
\leq & T\delta + \mathbb{P}\left(\left\{a_t \in [K]\right\} \cap \left\{n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right),
\end{aligned}$$

where the last inequality holds by the fact proved in Step 2. Summing over $t \in [T]$,

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\left(\mathcal{M}_{t-1}^{c}\right) \leq & T^2\delta + \sum_{t=1}^{T} \mathbb{P}\left(\{a_t \in [K]\} \cap \left\{n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right\}\right) \\ = & T^2\delta + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t \in [K]\right) \mathbb{I}\left(n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right)\right].\end{aligned}$$

Set $\mu := M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)$ and suppose

$$\sum_{t=1}^{T} \mathbb{I}\left(a_t \in [K]\right) \mathbb{I}\left(n_{t-1} < \mu\right) > \mu. \tag{55}$$

Let $\tau(1) < \tau(2) < \cdots < \tau(|\mathcal{A}|)$ be the ordered admitted round in $\mathcal{A} := \{t \in [T] : a_t \in [K]\}$. By definition, $n_{\tau(\nu)} = \nu$ for $\nu \in [|\mathcal{A}|]$. By (55), the event$\{a_t \in [K]\}$ happens at least $\mu + 1$ times over the horizon $[T]$ and $|\mathcal{A}| > \mu$. For any $\nu \in (\mu, |\mathcal{A}|]$,the number of admitted round is $n_{\tau(\nu)} > \mu$ and

$$\begin{aligned}\sum_{t=\varepsilon}^{T-1} \mathbb{I}\left(n_{t-1} < \mu\right) \mathbb{I}\left(a_t \in [K]\right) = & \sum_{\nu=1}^{|\mathcal{A}|} \mathbb{I}\left(n_{\tau(\nu)-1} < \mu\right) \mathbb{I}\left(a_{\tau(\nu)} \in [K]\right) \\ \leq & \sum_{\nu=1}^{|\mathcal{A}|} \mathbb{I}\left(n_{\tau(\nu)-1} < \mu\right) \mathbb{I}\left(n_{\tau(\nu)} = n_{\tau(\nu)-1} + 1\right) \\ = & \sum_{\nu=1}^{|\mathcal{A}|} \mathbb{I}\left(n_{\tau(\nu)-1} < \mu\right) \mathbb{I}\left(\nu = n_{\tau(\nu)-1} + 1\right) \\ \leq & \sum_{\nu=1}^{|\mathcal{A}|} \mathbb{I}\left(\nu - 1 < \mu\right), \\ = & \sum_{\nu=1}^{|\mathcal{A}|} \mathbb{I}\left(\nu < \mu + 1\right) \\ = & \mu,\end{aligned}$$

which contradicts with (55). Thus

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t \in [K]\right) \mathbb{I}\left(n_{t-1} < M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right)\right)\right] \leq \mu := M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right),$$

which proves,

$$\sum_{t=1}^{T} \mathbb{P}\left(\mathcal{M}_{t-1}^{c}\right) \leq T^2\delta + M_{\alpha,p,T} d \log\left(\frac{Jd}{\delta}\right).$$

□

B.6. Proof of Theorem 5.1

Proof. From Lemma 5.2, rearranging the terms,

$$\begin{aligned}
\mathcal{R}_T^{\widehat{\pi}} :=& OPT - \mathbb{E}\left[\sum_{t=1}^{T} R_t^{\widehat{\pi}}\right] \\
\leq& \frac{OPT}{T}\left\{T - \mathbb{E}\left[\tau - \xi\right]\right\} \\
&+ \left(2 + \frac{OPT}{\rho T}\right)\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{M}_{t-1}^c \cup \mathcal{E}_t^c\right) \\
&+ 2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^2\mathbb{I}\left(a_t \in [K]\right)\right]} \\
&+ 2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b}(\delta)^2\mathbb{I}\left(a_t \in [K]\right)\right]}\frac{OPT}{\rho T}.
\end{aligned}$$

By Lemma 5.3,

$$\begin{aligned}
\mathbb{E}\left[\xi\right] &\leq 1 + \frac{1 + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right) + T^2\delta}{\rho}, \\
\mathbb{E}\left[T - \tau\right] &\leq \frac{4(m+1)T\delta + 7 + 2\gamma_{\tau-1,\sigma_b}(\delta)}{\rho}.
\end{aligned}$$

By definition of $\gamma_{t,\sigma}(\delta)$,

$$\begin{aligned}
\mathbb{E}\left[T - \tau\right] \leq& \frac{4(m+1)T\delta + 7 + 32\sqrt{J\log JKT} + 12\beta_{\sigma_b}(\delta)}{\rho} \\
=& \frac{4(m+1)T\delta + 7 + 32\sqrt{J\log JKT} + C_\sigma(\delta)\sqrt{Jd}}{\rho}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\frac{OPT}{T}\left\{T - \mathbb{E}\left[\tau - \xi\right]\right\} \\
&\leq \frac{OPT}{T\rho}\left(\rho + 4(m+1)T\delta + 8 + 32\sqrt{J\log\left(\frac{JK}{\delta}\right)} + C_{\sigma_b}(\delta)\sqrt{Jd} + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right) + T^2\delta\right) \\
&\leq \frac{OPT}{T\rho}\left(\rho + 8 + \left(5mT + T^2\right)\delta + 32\sqrt{J\log\left(\frac{JK}{\delta}\right)} + C_{\sigma_b}(\delta)\sqrt{Jd} + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)\right).
\end{aligned}$$

[Step 3. Bounding the sum of probability] Because $T \geq 8d\alpha^{-1}p_{\min}^{-1}\log JdT$, by Theorem 4.2 and Lemma 5.3,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{M}_{t-1}^{c}\cup\mathcal{E}_{t}^{c}\right) =& \sum_{t=1}^{T}\left\{\mathbb{P}\left(\mathcal{M}_{t-1}^{c}\right)+\mathbb{P}\left(\mathcal{M}_{t-1}\cap\mathcal{E}_{t}^{c}\right)\right\}\\
\leq& T^{3}\delta + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{M}_{t-1}\cap\mathcal{E}_{t}^{c}\right)\\
\leq& T^{3}\delta + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+8d\alpha^{-1}p_{\min}^{-1}\log JdT\\
&+\sum_{t=8d\alpha^{-1}p_{\min}^{-1}\log JdT}^{T}\mathbb{P}\left(\mathcal{M}_{t-1}\cap\mathcal{E}_{t}^{c}\right)\\
\leq& T^{3}\delta + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)+8d\alpha^{-1}p_{\min}^{-1}\log JdT+4(m+1)T\delta+7.
\end{aligned}$$

By definition of $\gamma_{t,\sigma}(\delta)$ and $\beta_{\sigma}(\delta)$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^{2}\mathbb{I}\left(a_t\in[K]\right)\right] =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{16\sqrt{J\log JKT}}{\sqrt{t-1}}+\frac{4\sqrt{2}\beta_{\sigma_r}(\delta)}{\sqrt{n_{t-1}}}\right)^{2}\mathbb{I}\left(a_t\in[K]\right)\right]\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\frac{\left(16\sqrt{J\log JKT}+4\sqrt{2}\beta_{\sigma_r}(\delta)\right)^{2}}{n_{t-1}}\mathbb{I}\left(a_t\in[K]\right)\right]\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\frac{\left(16\sqrt{J\log JKT}+4\sqrt{2}\beta_{\sigma_r}(\delta)\right)^{2}}{n_{t-1}}\mathbb{I}\left(n_t=n_{t-1}+1\right)\right]\\
\leq& \left(16\sqrt{J\log JKT}+4\sqrt{2}\beta_{\sigma_r}(\delta)\right)^{2}\log T,
\end{aligned}$$

where the first inequality holds by $n_t \leq t$ almost surely. Thus by definition of $\beta_{\sigma}(\delta) := 8\sqrt{Jd}+96\sigma\sqrt{Jd\log\frac{4}{\delta}}$,

$$\begin{aligned}
2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_r}(\delta)^{2}\mathbb{I}\left(a_t\in[K]\right)\right]} &\leq \left(32\sqrt{J\log JKT}+4\sqrt{6}\beta_{\sigma_r}(\delta)\right)\sqrt{T\log T}\\
&\leq \left(32\sqrt{J\log JKT}+C_{\sigma_r}(\delta)\sqrt{Jd}\right)\sqrt{T\log T},
\end{aligned}$$

where $C_{\sigma}(\delta) := 8\sqrt{2}\cdot\left(8+96\sigma\sqrt{\log\frac{4}{\delta}}\right)$. Similarly,

$$2\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t-1,\sigma_b}(\delta)^{2}\mathbb{I}\left(a_t\in[K]\right)\right]}\frac{OPT}{\rho T} \leq \left(32\sqrt{J\log JKT}+C_{\sigma_b}(\delta)\sqrt{Jd}\right)\sqrt{T\log T}\frac{OPT}{\rho T}$$

Collecting the bounds,

$$\begin{aligned}\mathcal{R}_T^{\widehat{\pi}} &\le \frac{OPT}{T\rho}\left(\rho + 8 + \left(5mT + T^2\right)\delta + 32\sqrt{J\log JKT} + C_{\sigma_b}(\delta)\sqrt{Jd} + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)\right)\\ &+ \left(2 + \frac{OPT}{\rho T}\right)\left\{T^3\delta + dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right) + 4d\alpha^{-1}p_{\min}^{-1}\log JdT + 4(m+1)T\delta + 7\right\}\\ &+ 2\left(1 + \frac{OPT}{\rho T}\right)\left\{32\sqrt{J\log JKT} + C_{\sigma_b\vee\sigma_r}(\delta)\sqrt{Jd}\right\}\sqrt{T\log T}\\ &\le \left(2 + \frac{OPT}{\rho T}\right)\left\{\left(96\sqrt{J\log JKT} + 3C_{\sigma_r\vee\sigma_r}(\delta)\sqrt{Jd}\right)\sqrt{T\log T} + 2dM_{\alpha,p,T}\log\left(\frac{Jd}{\delta}\right)\right.\\ &\qquad\left. + 4d\alpha^{-1}p_{\min}^{-1}\log JdT + 15 + 10mT^3\delta\right\},\end{aligned}$$

Plugging in $\delta = m^{-1}T^{-3}$ proves (14). □

B.7. Deriving Regret Bound for Unknown Class Arrival Probabilities

In this section, we provide a direct solution to derive the same theoretical results even if the class prior probabilities are unknown. For each $j \in [J]$, let $\widehat{p}_t^{(j)} := \frac{1}{t}\sum_{s=1}^{t} I(j_s = j)$ denote the empirical estimate for class prior p_j. Then, by Lemma C.3, we have

$$\frac{1}{2}p_j - \frac{\log(2JT^2)}{t} \le \widehat{p}_t^{(j)} \le \frac{3}{2}p_j + \frac{\log(2JT^2)}{t},$$

with probability at least $1 - T^{-1}$ for all $j \in [J]$ and $t \in [T]$. When $t \ge 4p_{\min}^{-1}\log(2JT^2)$, the empirical estimate $\widehat{p}_t^{(j)}$ satisfies $(1/4)p_j \le \widehat{p}_t^{(j)} \le (7/4)p_j$. In addition, replacing p_j with $\widehat{p}_t^{(j)}$ only affects the proof of Step 1 in Lemma 5.2, as detailed in Section B.4. Therefore, with minor adjustments to the constants in $\gamma_{t-1,\sigma}(\delta)$, the regret bound remains the same even when the class prior probabilities are unknown.

C. Technical lemmas

Lemma C.1. *(Azuma-Hoeffding's inequality) Azuma (1967) If a super-martingale $(Y_t; t \ge 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \le c_t$ for some constant c_t, for all $t = 1, \ldots, T$, then for any $a \ge 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \ge a\right) \le e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}.$$

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

$$Y_T - Y_0 \le \sqrt{2\log\frac{1}{\delta}\sum_{t=1}^{T} c_t^2}.$$

Lemma C.2. *For a sequence $u_1 \ge u_2 \ge \cdots \ge u_n \ge 0$ and nonnegative real sequences $\{p_i\}_{i\in[n]}$ and $\{q_i\}_{i\in[n]}$ such that $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$, if $p_1 > q_1$ then*

$$\sum_{i=1}^{n} p_i u_i \ge \sum_{i=1}^{n} q_i u_i.$$

Proof. When $n = 1$, $p_1 u_1 \ge q_1 u_1$, for any $u_1 \ge 0$. Suppose for any sequence $u_1 \ge u_2 \ge \cdots \ge u_{n-1} \ge 0$ and nonnegative real sequences $\{p_i\}_{i\in[n-1]}$ and $\{q_i\}_{i\in[n-1]}$ such that $\sum_{i=1}^{n-1} p_i = \sum_{i=1}^{n-1} q_i$,

$$p_1 > q_1 \Longrightarrow \sum_{i=1}^{n-1} p_i u_i \ge \sum_{i=1}^{n-1} q_i u_i.$$

For a sequence $u_1 \geq u_2 \geq \cdots \geq u_n \geq 0$ and nonnegative real sequences $\{p_i\}_{i\in[n]}$ and $\{q_i\}_{i\in[n]}$ such that $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$, and $p_1 > q_1$, there exist $k \in [n]\backslash\{1\}$ such that $p_k < q_k$. In case of $k = n$, define a sequence

$$\begin{aligned}\tilde{q}_i &= q_i, \quad \forall i \in [n-2] \\ \tilde{q}_{n-1} &= q_{n-1} - p_n + q_n \geq 0.\end{aligned}$$

Then $\sum_{i=1}^{n-1} \tilde{q}_i = \sum_{i=1}^{n-1} p_i$ and

$$\begin{aligned}\sum_{i=1}^n p_i u_i &= \sum_{i=1}^{n-1} p_i u_i + p_n u_n \\ &\geq \sum_{i=1}^{n-1} \tilde{q}_i u_i + p_n u_n \\ &= \sum_{i=1}^{n-1} q_i u_i + (-p_n + q_n)\, u_{n-1} + p_n u_n \\ &\geq \sum_{i=1}^{n-1} q_i u_i + (-p_n + q_n)\, u_n + p_n u_n \\ &= \sum_{i=1}^{n} q_i u_i.\end{aligned}$$

In case of $k \neq n$, denote a sequence

$$\begin{aligned}\tilde{q}_i &= q_i, \quad \forall i \in [n-1]\backslash\{k\} \\ \tilde{q}_k &= q_k - p_k + q_n.\end{aligned}$$

Then $\sum_{i=1}^{n-1} \tilde{q}_i = \sum_{j\neq k} p_i$ and

$$\begin{aligned}\sum_{i=1}^n p_i u_i &= \sum_{i\neq k} p_i u_i + p_k u_k \\ &\geq \sum_{i=1}^{n-1} \tilde{q}_i u_i + p_k u_k \\ &\geq \sum_{i=1}^{n-1} q_i u_i - p_k u_k + q_n u_k + p_k u_k \\ &= \sum_{i=1}^{n-1} q_k u_k + q_n u_k \\ &\geq \sum_{i=1}^{n} q_k u_k.\end{aligned}$$

By induction, the proof is complete.

□

Lemma C.3. *Let $\{\boldsymbol{X}_\tau : \tau \in [t]\}$ be a $\mathbb{R}^{d\times d}$-valued stochastic process adapted to the filtration $\{\mathcal{F}_\tau : \tau \in [t]\}$, i.e., $\boldsymbol{X}_\tau$ is $\mathcal{F}_\tau$-measurable for $\tau \in [t]$. Suppose $\boldsymbol{X}_\tau$ is a positive definite symmetric matrices such that$\lambda_{\max}(\boldsymbol{X}_\tau) \leq \frac{1}{2}$.Then with probability at least $1-\delta$,*

$$\sum_{\tau=1}^t \boldsymbol{X}_\tau \succeq \frac{1}{2}\sum_{\tau=1}^t \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] - \log\frac{d}{\delta} I_d.$$

In addition, with probability at least $1-\delta$*,*

$$\sum_{\tau=1}^{t} \boldsymbol{X}_\tau \preceq \frac{3}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] + \log\frac{d}{\delta} I_d.$$

Proof. This proof is an adapted version of Lemma 12.2 in Lattimore & Szepesvári (2020) for matrix stochastic process using the argument of Tropp (2012). For the lower bound, It is sufficient to prove that

$$\lambda_{\max}\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right) \le \log\frac{d}{\delta},$$

with probability at least $1-\delta$. By the spectral mapping theorem,

$$\begin{aligned}\exp\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right) \le& \lambda_{\max}\left(\exp\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right)\\ \le& \mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right).\end{aligned}$$

Taking expectation on both side gives,

$$\begin{aligned}&\mathbb{E}\exp\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right)\\ \le& \mathbb{E}\mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right)\\ =& \mathbb{E}\mathrm{Tr}\left(\mathbb{E}\left[\exp\left(-\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] + \log\exp\left(-\boldsymbol{X}_t\right)\right)\middle| \mathcal{F}_{t-1}\right]\right)\\ \le& \mathbb{E}\mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] + \log\mathbb{E}\left[\exp\left(-\boldsymbol{X}_t\right)\middle| \mathcal{F}_{t-1}\right]\right)\right).\end{aligned}$$

The last inequality holds due to Lieb's theorem Tropp (2015). Because $e^x \le 1+\frac{1}{2}x$for all $x \in [-1/2, 0]$, and the eigenvalue of $-\boldsymbol{X}_t$ lies in $[-1/2, 0]$, we have

$$\mathbb{E}\left[\exp\left(-\boldsymbol{X}_t\right)\middle| \mathcal{F}_{t-1}\right] \preceq I - \frac{1}{2}\mathbb{E}\left[\boldsymbol{X}_t \middle| \mathcal{F}_{t-1}\right] \preceq \exp\left(-\frac{1}{2}\mathbb{E}\left[\boldsymbol{X}_t \middle| \mathcal{F}_{t-1}\right]\right),$$

by the spectral mapping theorem. Thus we have

$$\begin{aligned}&\mathbb{E}\exp\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right)\\ \le& \mathbb{E}\mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] + \log\exp\left(-\frac{1}{2}\mathbb{E}\left[\boldsymbol{X}_t \middle| \mathcal{F}_{t-1}\right]\right)\right)\right)\\ =& \mathbb{E}\mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right] - \frac{1}{2}\mathbb{E}\left[\boldsymbol{X}_t \middle| \mathcal{F}_{t-1}\right]\right)\right)\\ =& \mathbb{E}\mathrm{Tr}\left(\exp\left(-\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau + \frac{1}{2}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\boldsymbol{X}_\tau \middle| \mathcal{F}_{\tau-1}\right]\right)\right)\\ \le& \vdots\\ \le& \mathbb{E}\mathrm{Tr}\left(\exp\left(O\right)\right) = d\end{aligned}$$

Now my Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t}\boldsymbol{X}_\tau+\frac{1}{2}\sum_{\tau=1}^{t}\mathbb{E}\left[\boldsymbol{X}_\tau\middle|\mathcal{F}_{\tau-1}\right]\right)>\log\frac{d}{\delta}\right)\\
&\leq\mathbb{E}\exp\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t}\boldsymbol{X}_\tau+\frac{1}{2}\sum_{\tau=1}^{t}\mathbb{E}\left[\boldsymbol{X}_\tau\middle|\mathcal{F}_{\tau-1}\right]\right)\right)\frac{\delta}{d}\\
&\leq\delta.
\end{aligned}$$

For the upper bound, we prove

$$\lambda_{\max}\left(\sum_{\tau=1}^{t}\boldsymbol{X}_\tau-\frac{3}{2}\sum_{\tau=1}^{t}\mathbb{E}\left[\boldsymbol{X}_\tau\middle|\mathcal{F}_{\tau-1}\right]\right)\leq\log\frac{d}{\delta},$$

in a similar way using the fact that $e^x\leq 1+(3/2)x$ on $x\in[0,1/2]$. □

Lemma C.4. *Suppose a random variable X satisfies $\mathbb{E}[X]=0$, and let Y be an σ-sub-Gaussian random variable. If $|X|\leq|Y|$ almost surely, then X is 6σ-sub-Gaussian.*

Proof. Because $|X|\leq|Y|$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{X^2}{6\sigma^2}\right)\right]\leq&\mathbb{E}\left[\exp\left(\frac{Y^2}{6\sigma^2}\right)\right]\\
=&1+\mathbb{E}\left[\int_0^\infty\mathbb{I}\left(|Y|\geq x\right)\frac{x}{3\sigma^2}e^{\frac{x^2}{6\sigma^2}}dx\right]\\
\leq&1+\int_0^\infty\mathbb{P}\left(|Y|\geq x\right)\frac{x}{3\sigma^2}e^{\frac{x^2}{6\sigma^2}}dx.
\end{aligned}$$

Because

$$\begin{aligned}
\mathbb{P}\left(|Y|\geq x\right)=&\mathbb{P}\left(Y\geq x\right)+\mathbb{P}\left(-Y\leq x\right)\\
\leq&2e^{-\frac{x^2}{2\sigma^2}},
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{X^2}{6\sigma^2}\right)\right]\leq&1+\int_0^\infty\frac{2x}{3\sigma^2}e^{-\frac{x^2}{3\sigma^2}}dx\\
\leq&2.
\end{aligned}$$

Now for any $\lambda\in\mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda X\right)\right]=&\mathbb{E}\left[\sum_{n=0}^{\infty}\frac{(\lambda X)^n}{n!}\right]\\
=&1+\mathbb{E}\left[\sum_{n=2}^{\infty}\frac{(\lambda X)^n}{n!}\right]\\
\leq&1+\mathbb{E}\left[\frac{\lambda^2X^2}{2}\sum_{n=2}^{\infty}\frac{|\lambda X|^{n-2}}{(n-2)!}\right]\\
\leq&1+\frac{\lambda^2}{2}\mathbb{E}\left[X^2\exp\left(|\lambda X|\right)\right].
\end{aligned}$$

Because $6\sigma^2\lambda^2 + \frac{X^2}{12\sigma^2} \geq |\lambda X|$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq & 1 + \frac{\lambda^2}{2}\exp\left(6\sigma^2\lambda^2\right)\mathbb{E}\left[X^2\exp\left(\frac{X^2}{12\sigma^2}\right)\right] \\
= & 1 + 6\sigma^2\lambda^2\exp\left(6\sigma^2\lambda^2\right)\mathbb{E}\left[\frac{X^2}{12\sigma^2}\exp\left(\frac{X^2}{12\sigma^2}\right)\right] \\
\leq & 1 + 6\sigma^2\lambda^2\exp\left(6\sigma^2\lambda^2\right)\mathbb{E}\left[\exp\left(\frac{X^2}{6\sigma^2}\right)\right] \\
\leq & 1 + 12\sigma^2\lambda^2\exp\left(6\sigma^2\lambda^2\right) \\
\leq & \left(1 + 12\sigma^2\lambda^2\right)\exp\left(6\sigma^2\lambda^2\right) \\
\leq & \exp\left(\frac{36}{2}\sigma^2\lambda^2\right).
\end{aligned}$$

Thus X is 6σ-sub-Gaussian. □

Lemma C.5. *(Lee et al., 2016, Lemma 2.3) Let $\{N_t\}$ be a martingale on a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Then there exists a $\mathbb{R}^2$-valued martingale $\{P_t\}$ such that for any time $t \geq 0$, $\|P_t\|_2 = \|N_t\|_{\mathcal{H}}$ and $\|P_{t+1} - P_t\|_2 = \|N_{t+1} - N_t\|_{\mathcal{H}}$.*

Lemma C.6. *(A dimension-free bound for vector-valued martingales.) Let $\{\mathcal{F}_s\}_{s=0}^t$ be a filtration and $\{\eta_s\}_{s=1}^t$ be a real-valued stochastic process such that η_s is $\mathcal{F}_\tau$-measurable. Let $\{X_s\}_{s=1}^t$ be an $\mathbb{R}^d$-valued stochastic process where X_s is $\mathcal{F}_0$-measurable. Assume that $\{\eta_s\}_{s=1}^t$ are σ-sub-Gaussian as in Assumption 1. Then with probability at least $1-\delta$,*

$$\left\|\sum_{s=1}^t \eta_s X_s\right\|_2 \leq 12\sigma\sqrt{\sum_{s=1}^t \|X_s\|_2^2}\sqrt{\log\frac{4t^2}{\delta}}. \tag{56}$$

Proof. Fix a $t \geq 1$. For each $s = 1, \ldots, t$, we have $\mathbb{E}\left[\eta_s \middle| \mathcal{F}_{s-1}\right] = 0$ and X_s is $\mathcal{F}_0$-measurable. Thus the stochastic process,

$$\left\{\sum_{s=1}^u \eta_s X_s\right\}_{u=1}^t \tag{57}$$

is a $(\mathbb{R}^d, \|\cdot\|_2)$-martingale. Since $(\mathbb{R}^d, \|\cdot\|_2)$ is a Hilbert space, by Lemma C.5, there exists an $\mathbb{R}^2$-martingale $\{M_u\}_{u=1}^t$ such that

$$\left\|\sum_{s=1}^u \eta_s X_s\right\|_2 = \|M_u\|_2, \ \|\eta_u X_u\|_2 = \|M_u - M_{u-1}\|_2, \tag{58}$$

and $M_0 = 0$. Set $M_u = (M_1(u), M_2(u))^\top$. Then for each $i = 1, 2$, and $u \geq 2$,

$$\begin{aligned}
|M_i(u) - M_i(u-1)| \leq & \|M_u - M_{u-1}\|_2 \\
= & \|\eta_u X_u\|_2 \\
= & |\eta_u|\,\|X_u\|_2,
\end{aligned}$$

almost surely. By Lemma C.4, $M_i(u) - M_i(u-1)$ is 6σ-sub-Gaussian. By Lemma C.1, for $x > 0$,

$$\begin{aligned}
\mathbb{P}\left(|M_i(t)| > x\right) = & \mathbb{P}\left(\left|\sum_{u=1}^t M_i(u) - M_i(u-1)\right| > x\right) \\
\leq & 2\exp\left(-\frac{x^2}{72t\sigma^2\sum_{s=1}^t \|X_s\|_2^2}\right),
\end{aligned}$$

for each $i = 1, 2$. Thus, with probability $1 - \delta/2$,

$$M_i(t)^2 \leq 72\left(\sum_{s=1}^t \|X_s\|_2^2\right)\sigma^2\log\frac{4}{\delta}.$$

In summary, with probability at least $1 - \delta/2$,

$$\left\| \sum_{\tau=1}^{t} \eta_s X_s \right\|_2 = \sqrt{M_1(t)^2 + M_2(t)^2} \leq 6\sigma \sqrt{\sum_{s=1}^{t} \|X_s\|_2^2} \sqrt{2 \log \frac{4t^2}{\delta}}.$$

□