Queueing Matching Bandits with Preference Feedback

Jung-hun Kim
Seoul National University
Seoul, South Korea
junghunkim@snu.ac.kr

Min-hwan Oh
Seoul National University
Seoul, South Korea
minoh@snu.ac.kr

Abstract

In this study, we consider multi-class multi-server asymmetric queueing systems consisting of N queues on one side and K servers on the other side, where jobs randomly arrive in queues at each time. The service rate of each job-server assignment is unknown and modeled by a feature-based Multi-nomial Logit (MNL) function. At each time, a scheduler assigns jobs to servers, and each server stochastically serves at most one job based on its preferences over the assigned jobs. The primary goal of the algorithm is to stabilize the queues in the system while learning the service rates of servers. To achieve this goal, we propose algorithms based on UCB and Thompson Sampling, which achieve system stability with an average queue length bound of $\mathcal{O}(\min\{N, K\}/\epsilon)$ for a large time horizon T, where ϵ is a traffic slackness of the system. Furthermore, the algorithms achieve sublinear regret bounds of $\widetilde{\mathcal{O}}(\min\{\sqrt{T}Q_{\max}, T^{3/4}\})$, where $Q_{\max}$ represents the maximum queue length over agents and times. Lastly, we provide experimental results to demonstrate the performance of our algorithms.

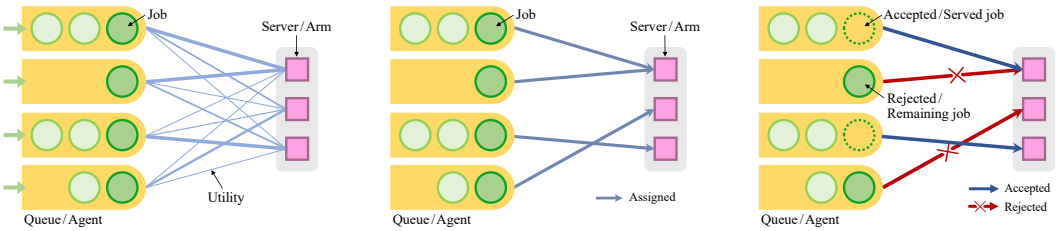
1 Introduction

Multi-class multi-server queueing systems, which have been extensively studied by [20, 19, 37, 8, 4], are motivated by real-world applications such as ride-hailing platforms, where riders are assigned to drivers. Other examples are online job markets, where applicants are recommended for employment, and online labor service markets, where tasks are recommended to freelance workers. In these systems, there are two sides: queues (agents) on one side and servers (arms) on the other. At each time step, jobs stochastically arrive in multiple queues and are then assigned to multiple servers by a matchmaking or scheduler algorithm to stabilize the systems. Importantly, the previous work assumes that the service rates for each server are known for scheduling jobs.

However, real-world observations reveal that the service rate may not be known beforehand. Therefore, learning the service rates associated with the stochastic behavior of servers is essential for real-world applications. Furthermore, the behavior of servers may depend on their (relative) preferences over assigned jobs. For example, in ride-hailing platforms where there are riders and drivers, the preferences of drivers over riders (not necessarily based on personal information but rather a rider's pick-up location, destination, etc.) may not be known in advance to riders or even the systems, necessitating the learning of drivers' preferences through preference feedback over assigned riders.

Learning service rates in an online manner, based on partial feedback from each scheduling, is highly associated with multi-armed bandit problems [33]. These problems are fundamental sequential learning tasks where, for queueing systems, a job is assigned to a server and receives feedback on whether the job is served or not. The utilization of bandit strategies for queueing systems has been recently studied by Choudhury et al. [12], Stahlbuhk et al. [48], Gaitonde and Tardos [17]; Freund et al. [15], Hsu et al. [22], Krishnasamy et al. [30, 31, 32], Sentenac et al. [46], Yang et al. [52], Huang et al. [23], which primary focus is naturally on establishing stability of the systems while learning service rates.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Jobs randomly arrive in queues (agents) and there are unknown different values of utility between queues and servers (arms) closely related to preferences.

(b) Each queue is assigned to a server based on a scheduler.

(c) Each server accepts at most one of the assigned jobs based on its preference. A job of each accepted queue is served while jobs of rejected queues remain.

Figure 1: Illustration of queueing process with 4 queues/agents ($N = 4$) and 3 servers/arms ($K = 3$)

However, there are still gaps between the previous models and the real-world applications. In all previous work, the inherent service rates for each job-server assignment are determined regardless of other jobs assigned to the same server (or it is not allowed to assign multiple jobs to the same server). This differs from real-world scenarios where service rate may depend on (relative) preference over multiple assigned jobs, as seen in online labor markets or ride-hailing platforms. Additionally, in Freund et al. [15], Krishnasamy et al. [31, 32], which is closely related work to our study regarding multi-class multi-server queueing systems with asymmetric service rates, it was allowed to assign a job in an empty queue to a server to obtain feedback (null request), which may not be realistic. Furthermore, all of the previous work considers a simple model without a generalizable structure or utilizing features for service rates.

Another line of related work is matching bandits, in which there are two sides (agents and arms), and the behavior of arms is based on their preferences [36, 45, 35, 9, 54, 29]. However, their focus is on static settings, whereas our work considers dynamic job arrivals in queues. Additionally, it is worth mentioning online matching problems [27, 39, 40, 18, 16, 28], where the sole focus is on optimizing matching rather than learning underlying models from bandit feedback. In contrast, our study concentrates on bandit problems related to learning latent utilities by managing the tradeoff between exploration and exploitation while establishing the stability of the queueing systems. Importantly, neither previous work on matching bandits nor online matching problems addresses the stability of queueing systems, which is our primary focus.

In this work, we propose a novel and practical framework for queueing matching bandits under preference feedback. In the following, we describe our setting accompanied by an illustration in Figure 1. (a) Multiple queues (agents) and servers (arms) are involved, with jobs arriving randomly in queues. Additionally, there are unknown utility values between queues and servers regarding preferences. (b) Nonempty queue are then assigned to servers based on a scheduling policy. (c) Subsequently, each server stochastically accepts at most one of the assigned queues based on its unknown preference and serves one unit of job of the accepted queue. The rejected jobs remain in the queues. The processes of (a), (b), and (c) are repeated over time.

Summary of Our Contributions

- We propose a novel and practical framework for queueing matching bandits, where N agents and K arms are involved, and jobs randomly arrive in agents' queues each round. Subsequently, a scheduler assigns agents to arms, and the service rates for the assigned agents depend on the arms' unknown preferences over them. These service rates are modeled using a feature-based multinomial logit function. To the best of our knowledge, our paper is the first to investigate either a feature-based service function or preference feedback using a multinomial logit function for service rates.

- We propose algorithms based on Upper Confidence Bound (UCB) and Thompson Sampling (TS), which achieve stability with average queue length bounds of $\mathcal{O}(\frac{\min\{N,K\}}{\epsilon})$ for a large time horizon T, under a traffic slackness of ϵ.

- Furthermore, the algorithms achieve sublinear regret bounds of $\widetilde{\mathcal{O}}(\min\{\sqrt{T}Q_{\max}, T^{3/4}\})$, where $Q_{\max}$ represents the maximum queue length over agents and times. It is worth mentioning that regret analysis has not been studied in the closely related queueing bandit literature such as Sentenac et al. [46], Freund et al. [15], Yang et al. [52].
- Finally, we present experimental results demonstrating the stability and regret performance of our algorithms with comparison to previously suggested methods.

2 Related Work

Bandits for Queues. Learning the unknown service rates for queueing systems in an online manner under bandit feedback has recently gained widespread attention [12, 47, 31, 48, 22, 32, 46, 52, 23]. While Stahlbuhk et al. [47], Krishnasamy et al. [30], Choudhury et al. [12] focused on a single queue with multiple servers, aiming to efficiently assign the queue to the optimal server while learning the service rates, subsequent study of Krishnasamy et al. [32] expanded this scope to encompass multiple queues and multiple servers. However, this extension relied on a strong structural assumption that each agent possesses a unique and distinctly optimal server without sharing the same optimal server. Consequently, the algorithms employed in such scenarios do not necessarily consider the dynamic queue lengths when scheduling jobs; instead, they concentrate solely on learning service rates to achieve optimal matching. As the closest work to our study, another line of work of Sentenac et al. [46], Freund et al. [15], Yang et al. [52] has focused more on dynamic queue lengths for multi-queue and multi-server scenarios without assuming the unique and distinct optimal servers for each agent. In this context, all of the previous work aimed to achieve queue length stability.

However, to the best of our knowledge, none of the previous works has focused on a structured model with features for service rates. Furthermore, previous studies either did not allow the assignment of different types of jobs to a server simultaneously [32, 52], or, if they did, as in Sentenac et al. [46], Freund et al. [15], the inherent values of service rates for each job-server assignment remained constant regardless of the entire assignments. Additionally, the server's behavior given multiple assigned jobs was either simple or deterministic, based on the uniform-randomly selected job among the assigned jobs [46], or the job with the highest bid generated from the algorithm [15], respectively. However, these approaches may not reflect real-world scenarios, where the service rates depend on the relative preferences among the assigned jobs. Lastly, in [15, 32], it was allowed to assign a job to a server from an empty queue for obtaining feedback, which may not be realistic.

In our study, we adopt a structured model incorporating features for unknown asymmetric service rates. We enable the assignment of multiple agents to the same server, where the server stochastically selects at most one of the assigned agents based on its undisclosed (relative) preference. Also, the assignment of an agent is available only when the queue of the agent is not empty. This scenario mirrors common situations in real-world applications such as ride-hailing platforms and online labor markets where available multiple riders or tasks can be assigned to a driver or a worker, respectively, and the driver or worker behaves stochastically depending on its unknown preference.

It is noteworthy that we focus not only on the stability of the systems regarding queue lengths but also on regret against an oracle, while the closely related works by [46, 15, 52], which considered multi-class multi-server queueing systems, only addressed stability analysis. Regret analysis has only been studied under some limited scenarios, such as a single-server [23], a single-queue [48], or the strong assumption that each agent has a unique and well-separated optimal server [32].

Matching Bandits. Next, we investigate two-sided matching bandits, a topic initially explored by Liu et al. [36] and subsequently studied by Sankararaman et al. [45], Liu et al. [35], Basu et al. [9], Zhang et al. [54], Kong and Li [29]. The primary goal is to minimize regret by attaining an optimal stable matching by learning agents' side preferences through stochastic reward feedback under static settings and deterministic behavior of arms with known preferences. Our study aligns with the model, wherein agents select arms based on preferences. However, our study sets itself apart from prior work on matching bandits in several key aspects. Firstly, we consider dynamic environments with job arrivals for agents. Secondly, we propose that arm behavior is stochastic, with unknown preferences, necessitating the learning of arms' preferences. Lastly, our main target is to stabilize the queue lengths in the systems rather than find stable matching for stable marriage problems.

MNL Bandits. Lastly, we examine MNL bandits which were initially proposed by Agrawal et al. [6] and followed by Agrawal et al. [7], Chen et al. [11], Oh and Iyengar [43, 44]. In MNL bandits,

the goal is to select assortments of arms to maximize reward, which is based on preferences over the arms in the selected assortment. In our study, we adopt the MNL model for arms' choice preferences in dynamic systems, which, to the best of our knowledge, is the first consideration of bandits for queueing systems.

3 Problem Statement

There are N agents (queues) and K arms (servers). At each time, a job for each agent $n \in [N]$ arrives randomly following a Bernoulli distribution with an unknown arrival rate $\lambda_n \in [0, 1]$ [1]. Then at each time $t \in [T]$ where T is the time horizon, each agent $n \in [N]$ is assigned to an arm $k^\pi_{n,t} \in [K]$ by a policy π. For notational simplicity, we use $k_{n,t}$ for $k^\pi_{n,t}$ when there is no confusion. Let $Q_n(t)$ be the length of the queue for jobs of agent $n \in [N]$ at the beginning of time slot t in the system. We consider that at most L agents can be assigned to each arm $k \in [K]$ at each time and $N \leq KL$ to ensure that all agents can be assigned to arms. Then, we define the set of agents (assortment) assigned to arm k by policy π at time t as $S_{k,t} = \{n \in [N] : k_{n,t} = k, Q_n(t) \neq 0\}$, considering only available agents with nonempty queues.

Now we explain the structure of our model. Each agent n has known d-dimensional feature information of $x_n \in \mathbb{R}^d$, and each arm k has latent (unknown) parameter $\theta_k \in \mathbb{R}^d$ for preference utilities over agents. We adopt the Multi-nomial Logit (MNL) function commonly studied for preference feedback [43, 44, 6, 7, 11]. Then given assortment $S_{k,t}$, arm k serves a job of agent $n \in S_{k,t}$ with a service probability (service rate) determined by the MNL model as

$$\mu(n|S_{k,t}, \theta_k) = \frac{\exp(x_n^\top \theta_k)}{1 + \sum_{m \in S_{k,t}} \exp(x_m^\top \theta_k)}.$$

We note that $\mu(n|S_{k,t}, \theta_k)$ represents the preference of arm k to agent n over the assigned agents $S_{k,t}$. The service rate of each n depends on the assigned agents rather than static by reflecting real-world scenarios, which is different from the conventional queueing problems. Following the MNL function, the arm is allowed not to serve any agents (or allowed to serve null agent n_0) with the probability of $\mu(n_0|S_{k,t}, \theta_k) = \frac{1}{1+\sum_{m \in S_{k,t}} \exp(x_m^\top \theta_k)}$. Under the MNL model for the service rate, at each time, each arm accepts at most one agent and serves that agent's job. The queue lengths for the accepted agents are each reduced by one, while the queue lengths for the refused agents remain the same.

Objective Function. Here we provide the goal of this problem. For describing the stochastic process in the systems, at each time t, let $A_n(t) \in \{0, 1\}$ be a random variable with mean λ_n, which denotes whether a new job arrives in the queue of agent $n \in [N]$ (here, 1 denotes arrival). Also, given that n is assigned to arm $k_{n,t}$, let $D_n(t|S_{k_{n,t},t}) \in \{0, 1\}$ be a random variable with mean $\mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}})$, which represents whether the assigned agent n, given $S_{k_{n,t},t}$, is accepted by arm $k_{n,t}$ (here 1 denotes acceptance). Then the queue length of the agent n evolves as $Q_n(t + 1) = (Q_n(t) + A_n(t) - D_n(t|S_{k_{n,t},t}))^+$ where x^+ denotes $\max\{x, 0\}$ for $x \in \mathbb{R}$.

As in the previous work of queueing bandits for multiple types of agents and servers [46, 15, 52, 23], which are closely related work to ours, our primary focus is on stabilizing the dynamic systems regarding queue lengths. For analyzing the stability of the systems, we define the average queue lengths over horizon time T as

$$\mathcal{Q}(T) = \frac{1}{T} \sum_{t \in [T]} \sum_{n \in [N]} \mathbb{E}\left[Q_n(t)\right].$$

Then the goal of this problem is to design an online matching algorithm to assign the agents to arms for stabilizing the systems by bounding $\mathcal{Q}(T)$. We define the system stability as follows.

Definition 1. *The systems are denoted to be stable when* $\lim_{T \to \infty} \mathcal{Q}(T) < \infty$.

[1] Our study can be easily extended to the multi-unit case where random variable $A_n(t)$ lies in $[0, M]$ for some $M > 0$ with mean $\lambda_n \in [0, M]$.

The same stability definition was considered in Neely [41, 42], Freund et al. [15], Yang et al. [52], Huang et al. [23]. It is noteworthy that according to [42, 41, 15], the system satisfying this stability condition of Definition 1 is denoted to be strongly stable.[2]

For the analyses, we first present the regularity conditions.

Assumption 1. $\|x_n\|_2 \leq 1$ *for all* $n \in [N]$ *and* $\|\theta_k\|_2 \leq 1$ *for all* $k \in [K]$.

Assumption 2. *There exists* $\kappa > 0$ *such that* $\inf_{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1} \mu(n|S, \theta)\mu(n_0|S, \theta) \geq \kappa$ *for any* $n \in S$ *and* $S \subset [N]$.

These regularity conditions are commonly taken into account in the logistic and MNL bandit literature [13, 3, 43, 44]. We note that, in the worst-case, $1/\kappa = O(L^2)$.

We define the set of feasible disjoint assortments given any $\mathcal{N} \subseteq [N]$ as $\mathcal{M}(\mathcal{N}) = \{(S_1, ..., S_K) : S_k \subseteq \mathcal{N}, |S_k| \leq L \ \forall k \in [K], S_k \cap S_l = \varnothing \ \forall k \neq l, \bigcup_{k \in [K]} S_k = \mathcal{N}\}$. Then, we consider the condition of traffic slackness for stability as follows.

Assumption 3. *For some traffic slackness* $0 < \epsilon < 1$*, there exists* $\{S_k\}_{k \in [K]} \in \mathcal{M}([N])$ *such that* $\lambda_n + \epsilon \leq \mu(n|S_k, \theta_k)$ *for all* $n \in S_k$ *and* $k \in [K]$.

We note that similar slackness assumptions have been commonly considered in queueing bandits [15, 52, 23]. As ϵ decreases, achieving stability in our setting becomes more challenging. A discussion for a refined version of ϵ can be found in Appendix A.2.

4 Preliminary Study: An Oracle Algorithm of MaxWeight

For queueing systems, MaxWeight [50] has been proven to have optimal throughput keeping the queues in networks stable under the known service rates [51, 38, 37, 49]. Here, we analyze the oracle policy using MaxWeight under known θ_k's. We define the set of agents having non-empty queues at time t as $\mathcal{N}_t = \{n \in [N]] : Q_n(t) \neq 0\}$. Then, the oracle policy using MaxWeight in our setting is defined as $\{S_{k,t}\}_{k \in [K]} = \text{argmax}_{\{S_k\}_{k \in [K]} \in \mathcal{M}(\mathcal{N}_t)} \sum_{k \in [K]} \sum_{n \in S_k} Q_n(t)\mu(n|S_k, \theta_k)$, where priority in the assortment scheduling is on queues with either large queue lengths or high service rates. We provide an analysis of stability of the MaxWeight oracle for known θ_k's as follows.

Proposition 1. *Given the prior knowledge of* θ_k *for all* $k \in [K]$*, the average queue length of MaxWeight is bounded as* $\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon}\right)$*, which implies that the algorithm achieves stability.*

Proof. The proof is provided in Appendix A.3 □

We note that the term of $\min\{N, K\}$ in the average queue length is a result of the system's variance. Additionally, as ϵ decreases, the stability of the system deteriorates due to the reduced traffic slackness. Further discussion regarding ϵ can be found in Appendix A.2.

Since the oracle algorithm requires prior knowledge of θ_k's for the service rate, we cannot use it directly in our bandit setting. In the following section, we propose algorithms incorporating an efficient learning procedure to achieve stability in our setting.

5 Algorithms and Analyses

5.1 UCB-based Algorithm

We first propose an algorithm based on the UCB strategy, `UCB-QMB` (Algorithm 1). We define the negative log-likelihood as $f_{k,t}(\theta) := -\sum_{n \in S_{k,t} \cup \{n_0\}} y_{n,t} \log \mu(n|S_{k,t}, \theta)$ where $y_{n,t} \in \{0, 1\}$ is observed preference feedback (1 denotes service acceptance, and 0 denotes refusal) and define the gradient of the likelihood as

$$g_{k,t}(\theta) := \nabla_\theta f_{k,t}(\theta) = \sum_{n \in S_{k,t}} (\mu(n|S_{k,t}, \theta) - y_{n,t})x_n. \tag{1}$$

[2] As mentioned in [42, 15], in general, this stability condition is stronger than the mean rate stability of $\lim_{T \to \infty} \frac{\mathbb{E}[\sum_{n \in [N]} Q_n(T)]}{T} = 0$ but weaker than the uniform stability of $\mathbb{E}[\sum_{n \in [N]} Q_n(t)] \leq C$ for all $t > 0$ for some constant $C > 0$.

Then, we construct the estimator of $\hat{\theta}_{k,t}$ from online updates applying online newton step studied by [21, 53, 24, 14, 44] as $\hat{\theta}_{k,t} = \text{argmin}_{\theta\in\Theta} g_{k,t-1}(\hat{\theta}_{k,t-1})^\top\theta + \frac{1}{2}\|\theta - \hat{\theta}_{k,t-1}\|^2_{V_{k,t}}$, where $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$ and $V_{k,t} = \lambda I_d + \frac{\kappa}{2}\sum_{s=1}^{t-1}\sum_{n\in S_{k,s}} x_n x_n^\top$. Using the estimator, we define the UCB index for agent n in assortment S_k as

$$\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) = \frac{\exp(h_{n,k,t}^{UCB})}{1 + \sum_{m\in S_k}\exp(h_{m,k,t}^{UCB})}, \tag{2}$$

where $h_{n,k,t}^{UCB} := x_n^\top\hat{\theta}_{k,t} + \beta_t\|x_n\|_{V_{k,t}^{-1}}$ with $\beta_t = C_1\sqrt{\lambda + \frac{d}{\kappa}\log(1 + \frac{tLK}{d\lambda})}$ for some constant $C_1 > 0$. We utilize the MaxWeight with the UCB indexes by setting $\lambda = 1$.

Algorithm 1 UCB-Queueing Matching Bandit (`UCB-QMB`)

Input: λ, κ, $C_1 > 0$
for $t = 1, \dots, T$ **do**
  for $k \in [K]$ **do**
    $\hat{\theta}_{k,t} \leftarrow \text{argmin}_{\theta\in\Theta} g_{k,t-1}(\hat{\theta}_{k,t-1})^\top\theta + \frac{1}{2}\|\theta - \hat{\theta}_{k,t-1}\|^2_{V_{k,t}}$ with (1)
  $\{S_{k,t}\}_{k\in[K]} \leftarrow \underset{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)}{\text{argmax}} \sum_{k\in[K]}\sum_{n\in S_k} Q_n(t)\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t})$ with (2)
  Offer $\{S_{k,t}\}_{k\in[K]}$ and observe preference feedback $y_{n,t} \in \{0, 1\}$ for all $n \in S_{k,t}$, $k \in [K]$

5.1.1 Stability Analysis of `UCB-QMB`

Here we provide an analysis for the stability of `UCB-QMB` (Algorithm 1).

Theorem 1. *The average queue length of Algorithm 1 is bounded as* $\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2}{\kappa^4\epsilon^6}\frac{\text{polylog}(T)}{T}\right)$, *which implies that the algorithm achieves stability as*

$$\lim_{T\to\infty}\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon}\right).$$

We note that Algorithm 1 achieves the same average queue length bound of $\mathcal{O}(\frac{\min\{N,K\}}{\epsilon})$ with the oracle of MaxWeight (Proposition 1) when T is large enough.

Proof sketch. Here we provide a proof sketch and the full version is provided in Appendix A.4. We define the set of queues $\mathbf{Q}(t) = [Q_n(t) : n \in [N]]$ and a Lyapunov function as $\mathcal{V}(\mathbf{Q}(t)) = \sum_{n\in[N]} Q_n(t)^2$. For simplicity, we use $D_n(t)$ for $D_n(t|S_{k_{n,t},t})$ and $D_n^*(t)$ for $D_n(t|S_{k^*_{n,t},t})$ when there is no confusion. Then we analyze the Lyapunov drift as follows:

$$\begin{aligned}
&\sum_{t\in[T]}\mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t)) \\
&= \sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t) + A_n(t) - D_n^*(t))^2 - Q_n(t)^2 \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t) + A_n(t) - D_n(t))^{+2} - \sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t) + A_n(t) - D_n^*(t))^2. \qquad (3)
\end{aligned}$$

For the first two terms in Eq.(3), with Assumption 3, we can show that

$$\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(Q_n(t) + A_n(t) - D_n^*(t))^2 - Q_n(t)^2] \le -\sum_{t\in[T]}\sum_{n\in[N]} 2\epsilon\mathbb{E}[Q_n(t)] + 2\min\{N,K\}T, \tag{4}$$

For the last two terms in Eq.(3), we have

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t) + A_n(t) - D_n(t))^{+2} - \sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t) + A_n(t) - D_n^*(t))^2\Big] \\
&\le 2\mathbb{E}\Big[\sum_{t\in[T]}\sum_{n\in[N]}(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\Big] + 5\min\{N,K\}T, \qquad (5)
\end{aligned}$$

where the first term of the last inequality is closely related to the regret analysis of the bandit strategy. In the following, we focus on analyzing the last term in Eq.(5). We define events $E_t^1 = \{\|\hat{\theta}_{k,t} - \theta_k^*\|_{V_{k,t}} \leq \beta_t \text{ for all } k \in [K]\}$ and $E_{n,t}^2 = \{\max_{m \in S_{k,t}} \|x_m\|_{V_{k,t}^{-1}} \leq C_2 \epsilon / 2\beta_t \text{ for } k = k_{n,t}^{\pi}\}$ for some constant $C_2 > 0$. We can show that E_t^1 holds with high probability so here we only consider the case when E_t^1 holds. Then we have

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t\in[T]}\sum_{n\in[N]} \mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\Big] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)(\mathbb{1}(E_{n,t}^2) + \mathbb{1}((E_{n,t}^2)^c))]. \quad (6)
\end{aligned}$$

Then for the first term of Eq.(6), from the UCB strategy and $E_{n,t}^2$, we can show that

$$\sum_{t\in[T]} \mathbb{E}\Big[\sum_{n\in[N]} (\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_{n,t}^2)\Big] \leq C_2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right]. \quad (7)$$

Now we provide a bound for the second term of Eq.(6). For some constant $C_3 > 0$, we can show that

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_{n,t}^2)^c)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} (\epsilon/C_3)\mathbb{E}\left[Q_n(t)\right] + \mathcal{O}\left(\frac{N^2K^2\beta_T^4}{\kappa^2\epsilon^5}\right). \quad (8)
\end{aligned}$$

By putting the results of Eqs. (3), (4), (5), (6), (7), (8) altogether, we can obtain

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t\in[T]} \mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))\Big] \\
&\leq 7\min\{N,K\}T + 2(C_2 + (1/C_3) - 1)\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(N\log(T)) + \mathcal{O}\left(\frac{N^2K^2\beta_T^4}{\kappa^2\epsilon^5}\right).
\end{aligned}$$

Finally, with positive constants $C_2, C_3 > 0$ satisfying $C_2 + (1/C_3) < 1$, from $\mathcal{V}(\mathbf{Q}(1)) = 0$ and $\mathcal{V}(\mathbf{Q}(T+1)) \geq 0$, by using telescoping for the above inequality and rearrangement, we can conclude the proof by $\frac{1}{T}\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)] = \mathcal{O}(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2}{\kappa^4\epsilon^6}\frac{\text{polylog}(T)}{T})$. □

5.1.2 Regret Analysis of `UCB-QMB`

In addition to the stability analysis, we examine the cumulative regret of `UCB-QMB` (Algorithm 1). The regret is defined as the discrepancy between the performance of the oracle policy of MaxWeight π^*, which operates with the knowledge of the true parameters θ_k's, and that of our policy π. Given the queue lengths at each time t, we denote the oracle assignments as

$$\{S_{k,t}^*\}_{k\in[K]} = \underset{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)}{\text{argmax}} \sum_{k\in[K]}\sum_{n\in S_k} Q_n(t)\mu(n|S_k, \theta_k).$$

We show that this oracle policy achieves stability in Proposition 1. For simplicity, we use $k_{n,t}^*$ for $k_{n,t}^{\pi^*}$. Then, the cumulative regret under π is defined as

$$\mathcal{R}^{\pi}(T) = \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k_{n,t}^*,t}^*, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\right]. \quad (9)$$

We define $Q_{\max} = \mathbb{E}[\max_{t\in[T],n\in[N]} Q_n(t)]$. Then the algorithm achieves the following regret bound.

Theorem 2. *The policy π of Algorithm 1 achieves a regret bound of*

$$\mathcal{R}^{\pi}(T) = \widetilde{\mathcal{O}}\Big(\min\Big\{\frac{d}{\kappa}\sqrt{KT}Q_{\max}, \Big(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\Big)^{1/4}T^{3/4}\Big\}\Big).$$

We emphasize that our algorithms achieve a sublinear regret bound, even in the worst-case scenario regarding queue lengths from the minimum in regret. In contrast, [23] achieves a regret bound of $\widetilde{O}(\max\{\sqrt{T}Q_{\max}, T^{3/4}\})$ for a stationary setting, where the worst-case bound is not guaranteed to be sublinear from the maximum in regret.

Proof sketch. Here we provide a proof sketch and the full version is provided in Appendix A.5. We first provide the proof for regret bound of $\mathcal{R}^\pi(T) = \widetilde{\mathcal{O}}(\frac{d}{\kappa}\sqrt{KT}Q_{\max})$. We define event $E_t^1 = \{\|\hat{\theta}_{k,t} - \theta_k^*\|_{V_{k,t}} \leq \beta_t \ \forall k \in [K]\}$ which holds with a high probability. Therefore, here we only consider the case when E_t holds. Then we can show that $\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))] \leq \sum_{n\in[N]} \mathbb{E}[Q_n(t)(\widetilde{\mu}_t^{UCB}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))] \leq \sum_{k\in[K]} \mathbb{E}[2\beta_t \max_{n\in S_{k,t}} \|x_n\|_{V_{k,t}^{-1}} Q_n(t)]$. From the inequality, with $\sum_{t=1}^T \max_{n\in S_{k,t}} \|x_n\|^2_{V_{k,t}^{-1}} \leq (4d/\kappa)\log(1 + (TL/d\lambda))$, we have

$$\begin{aligned}\mathcal{R}^\pi(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))] \\ &\leq 2\mathbb{E}\Big[\max_{t\in[T],n\in[N]} Q_n(t)\beta_T \sqrt{KT \sum_{t\in[T]}\sum_{k\in[K]} \max_{l\in S_{k,t}} \|x_l\|^2_{V_{k,t}^{-1}}}\Big] = \widetilde{\mathcal{O}}\Big(\frac{d}{\kappa}\sqrt{KT}Q_{\max}\Big). \quad (10)\end{aligned}$$

Now we provide the proof for the worst-case regret bound of $\mathcal{R}^\pi(T) = \widetilde{\mathcal{O}}\left(\left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right)$ in the following. We additionally define event $E_{n,t}^2 = \{\max_{m\in S_{k,t}} \|x_m\|_{V_{k,t}^{-1}} \leq \zeta \text{ for } k = k_{n,t}\}$ for some constant $C_2 > 0$. Under E_t^1, we have

$$\begin{aligned}\mathcal{R}^\pi(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)] \\ &\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)(\mathbb{1}(E_{n,t}^2) + \mathbb{1}((E_{n,t}^2)^c))].\end{aligned} \quad (11)$$

Then for the first term of Eq.(11), we have

$$\begin{aligned}&\sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_{n,t}^2)\right] \\ &\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} 2\beta_t \|x_n\|_{V^{-1}_{k_{n,t},t}} Q_n(t)\mathbb{1}(E_{n,t}^2)\right] \leq \sum_{t\in[T]}\sum_{n\in[N]} 2\beta_t\zeta\mathbb{E}\left[Q_n(t)\right], \quad (12)\end{aligned}$$

where the last inequality is obtained from $E_{n,t}^2$. By analyzing the selected number of agent n with $(E_{n,t}^2)^c$, we can show that

$$\begin{aligned}&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_{n,t}^2)^c)] \\ &\leq \sum_{t\in[T]}\sum_{n\in[N]} \zeta\beta_T\mathbb{E}\left[Q_n(t)\right] + \mathcal{O}\left(\frac{NK}{\kappa\zeta^3\beta_T}\right). \quad (13)\end{aligned}$$

By putting the results of Eqs. (11), (12), (13), and Theorem 1, by setting $\zeta = (\epsilon NK/\min\{N,K\}T\kappa\beta_T^2)^{1/4}$, for large enough T, we have

$$\begin{aligned}\mathcal{R}^\pi(T) &= \mathcal{O}\left(\zeta\beta_T \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)] + \frac{NK}{\kappa\zeta^3\beta_T} + N\log(T)\right) \\ &= \widetilde{\mathcal{O}}\left(N\left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right), \quad (14)\end{aligned}$$

which conclude the proof combined with Eq.(10). □

Algorithm 2 Thompson Sampling-Queueing Matching Bandit (`TS-QMB`)

Input: λ, M, κ, $C_1 > 0$
for $t = 1, \ldots, T$ **do**
  for $k \in [K]$ **do**
    $\hat{\theta}_{k,t} \leftarrow \mathrm{argmin}_{\theta \in \Theta} g_{k,t-1}(\hat{\theta}_{k,t-1})^\top \theta + \frac{1}{2}\|\theta - \hat{\theta}_{k,t-1}\|^2_{V_{k,t}}$ with (1)
    Sample $\{\widetilde{\theta}^{(i)}_{k,t}\}_{i \in [M]}$ independently from $\mathcal{N}(\hat{\theta}_{k,t}, \beta_t^2 V_{k,t}^{-1})$
  $\{S_{k,t}\}_{k \in [K]} \leftarrow \underset{\{S_k\}_{k \in [K]} \in \mathcal{M}(\mathcal{N}_t)}{\mathrm{argmax}} \sum_{k \in [K]} \sum_{n \in S_k} Q_n(t) \widetilde{\mu}_t^{TS}(n | S_k, \{\widetilde{\theta}^{(i)}_{k,t}\}_{i \in [M]})$ with (15)
  Offer $\{S_{k,t}\}_{k \in [K]}$ and observe preference feedback $y_{n,t} \in \{0, 1\}$ for all $n \in S_{k,t}$, $k \in [K]$

5.2 Thompson Sampling-based Algorithm

Here, we propose an algorithm based on Thompson Sampling, `TS-QMB` (Algorithm 2). As in the previous algorithm, we construct the estimator as $\hat{\theta}_{k,t} = \mathrm{argmin}_{\theta \in \Theta} g_{k,t-1}(\hat{\theta}_{k,t-1})^\top \theta + \frac{1}{2}\|\theta - \hat{\theta}_{k,t-1}\|^2_{V_{k,t}}$. To facilitate exploration, we sample several $\widetilde{\theta}^{(i)}_{k,t}$ for $i \in [M]$ from a Gaussian distribution of $\mathcal{N}(\hat{\theta}_{k,t}, \beta_t^2 V_{k,t}^{-1})$ and construct the Thompson Sampling (TS) index for assortment S_k as

$$\widetilde{\mu}_t^{TS}(n | S_k, \{\widetilde{\theta}^{(i)}_{k,t}\}_{i \in [M]}) = \frac{\exp(h^{TS}_{n,k,t})}{1 + \sum_{m \in S_k} \exp(h^{TS}_{m,k,t})}, \tag{15}$$

where $h^{TS}_{n,k,t} = \max_{i \in [M]} x_n^\top \widetilde{\theta}^{(i)}_{k,t}$ and $\beta_t = C_1 \sqrt{\lambda + \frac{d}{\kappa} \log(1 + \frac{tLK}{d\lambda})}$ for some constant $C_1 > 0$. Then we utilize the MaxWeight with TS indexes. We set $\lambda = 1$ and $M = \lceil 1 - \frac{\log(KL)}{\log(1 - 1/4\sqrt{e\pi})} \rceil$.

5.2.1 Stability Analysis of `TS-QMB`

Here we provide stability analysis for `TS-QMB` (Algorithm 2).

Theorem 3. *The average queue length of Algorithm 2 is bounded as* $\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^4 N^2 K^2}{\kappa^4 \epsilon^6} \frac{\mathrm{polylog}(T)}{T}\right)$, *which implies that the algorithm achieves stability as*

$$\lim_{T \to \infty} \mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N, K\}}{\epsilon}\right).$$

Proof. The proof is provided in Appendix A.6 □

5.2.2 Regret Analysis of `TS-QMB`

We provide a regret analysis of `TS-QMB` (Algorithm 2) for the regret definition of (9) in the following.

Theorem 4. *The policy π of Algorithm 2 achieves a regret bound of*

$$\mathcal{R}^\pi(T) = \widetilde{\mathcal{O}}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT} Q_{\max}, \left(\frac{d^2 NK \min\{N,K\}^3}{\kappa^2 \epsilon^3}\right)^{1/4} T^{3/4}\right\}\right).$$

Proof. The proof is provided in Appendix A.7 □

We note that the performance of Algorithm 1 and Algorithm 2 in the analysis results shows similar trends. However, the TS-based Algorithm 2 incurs a loss with respect to d compared to the UCB-based Algorithm 1, as commonly seen in previous TS-based algorithms [5, 2, 43].

Here, we briefly discuss the combinatorial optimization of $\mathrm{argmax}_{\{S_k\}_{k \in [K]} \in \mathcal{M}(\mathcal{N}_t)} \sum_{k \in [K]} f_k(S_k)$ for some function $f_k : S \subset [N] \to \mathbb{R}$ in our algorithms. The exact optimization can be expensive due to its NP-hard nature. To address this, we can utilize the technique of α-approximation oracle with $0 \leq \alpha \leq 1$, first introduced in Kakade et al. [25], which is deferred to Appendix A.9.

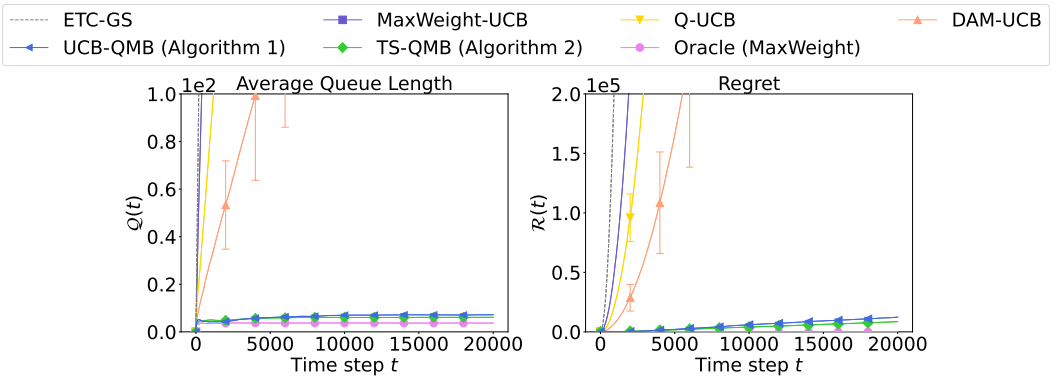


Figure 2: Experimental results for (left) average queue length and (right) regret

6 Experiments

Here, we provide experimental results to demonstrate the performance of our algorithms.[3] For the synthetic experiments, we consider $N = 4$, $K = 2$, $L = 2$, and $d = 2$. Each element in x_n and θ_k is uniformly generated from $[0, 1]$ and then normalized, and λ_n's are determined with $\epsilon = 0.1$. Even though no dedicated benchmark exists for our queueing matching scenario, we compare our algorithms with previously suggested ones for queueing bandits or matching bandits: `Q-UCB` [32], `DAM-UCB` [15], and `MaxWeight-UCB` [52] for multi-queue multi-server bandits with asymmetric service rates and `ETC-GS` [34] for matching bandits. In Figure 2, we can observe that our algorithms (Algorithms 1 and 2) outperform the previously suggested ones except for the oracle (MaxWeight) operated under known (latent) service rates. We demonstrate that our algorithms achieve stability, similar to the oracle in the left figure, which matches the results of our stability analysis (Theorems 1 and 3). Regarding regret shown in the right figure, the previously suggested algorithms exhibit superlinear performance due to the increasing $Q_n(t)$, while our algorithms show relatively small regret (Theorems 2 and 4). Additional experiments can be found in Appendix A.10.

7 Conclusion

In this paper, we introduce a novel framework for queueing matching bandits with preference feedback. To achieve stability in this framework, we propose UCB and TS-based algorithms utilizing the MaxWeight strategy. The algorithms achieve system stability with an average queue length bound of $\mathcal{O}(\min\{N, K\}/\epsilon)$. Furthermore, the algorithms achieve sublinear regret bounds of $\widetilde{\mathcal{O}}(\min\{\sqrt{T}Q_{\max}, T^{3/4}\})$. Lastly, we demonstrate our algorithms using synthetic datasets.

8 Acknowledgements

JK was supported by the Global-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301976). MO was supported by the NRF grant funded by the Korea government(MSIT) (No. 2022R1C1C1006859 and 2022R1A4A1030579) and by AI-Bio Research Grant through Seoul National University.

References

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

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

[3] The source code is available at `https://github.com/junghunkim7786/Queueing-Matching-Bandits-with-Preference-Feedback`

[3] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

[4] Philipp Afeche, Rene Caldentey, and Varun Gupta. On the optimal design of a bipartite matching queueing system. *Operations Research*, 70(1):363–401, 2022.

[5] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[6] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *arXiv preprint arXiv:1706.03880*, 2017.

[7] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pages 76–78. PMLR, 2017.

[8] Urtzi Ayesta, Peter Jacko, and Vladimir Novak. Scheduling of multi-class multi-server queueing systems with abandonments. *Journal of Scheduling*, 20:129–145, 2017.

[9] Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

[10] Gruia Calinescu, Chandra Chekuri, Martin Pal, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6): 1740–1766, 2011.

[11] Xi Chen, Akshay Krishnamurthy, and Yining Wang. Robust dynamic assortment optimization in the presence of outlier customers. *Operations Research*, 2023.

[12] Tuhinangshu Choudhury, Gauri Joshi, Weina Wang, and Sanjay Shakkottai. Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 181–190, 2021.

[13] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

[14] Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

[15] Daniel Freund, Thodoris Lykouris, and Wentao Weng. Efficient decentralized multi-agent learning in asymmetric queuing systems. In *Conference on Learning Theory*, pages 4080–4084. PMLR, 2022.

[16] Bernhard Fuchs, Winfried Hochstättler, and Walter Kern. Online matching on a line. *Theoretical Computer Science*, 332(1-3):251–264, 2005.

[17] Jason Gaitonde and Éva Tardos. Stability and learning in strategic queuing systems. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 319–347, 2020.

[18] Buddhima Gamlath, Michael Kapralov, Andreas Maggiori, Ola Svensson, and David Wajc. Online matching with general arrivals. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 26–37. IEEE, 2019.

[19] Kevin D Glazebrook and José Niño-Mora. Parallel scheduling of multiclass m/m/m queues: Approximate and heavy-traffic optimization of achievable performance. *Operations Research*, 49(4):609–623, 2001.

[20] J Michael Harrison. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *The Annals of Applied Probability*, 8(3):822–848, 1998.

[21] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

[22] Wei-Kang Hsu, Jiaming Xu, Xiaojun Lin, and Mark R Bell. Integrated online learning and adaptive control in queueing systems with uncertain payoffs. *Operations Research*, 70(2): 1166–1181, 2022.

[23] Jiatai Huang, Leana Golubchik, and Longbo Huang. When lyapunov drift based queue scheduling meets adversarial bandit learning. *IEEE/ACM Transactions on Networking*, 2024.

[24] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

[25] Sham M Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 546–555, 2007.

[26] Michael Kapralov, Ian Post, and Jan Vondrák. Online submodular welfare maximization: Greedy is optimal. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1216–1225. SIAM, 2013.

[27] Richard M Karp, Umesh V Vazirani, and Vijay V Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358, 1990.

[28] Thomas Kesselheim, Klaus Radke, Andreas Tönnis, and Berthold Vöcking. An optimal online algorithm for weighted bipartite matching and extensions to combinatorial auctions. In *European symposium on algorithms*, pages 589–600. Springer, 2013.

[29] Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1512–1522. SIAM, 2023.

[30] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Regret of queueing bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

[31] Subhashini Krishnasamy, Ari Arapostathis, Ramesh Johari, and Sanjay Shakkottai. On learning the cμ rule: Single and multi-server settings. *Available at SSRN 3123545*, 2018.

[32] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations research*, 69(1):315–330, 2021.

[33] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[34] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

[35] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1):9612–9645, 2021.

[36] Nan Liu, Yuhang Ma, and Huseyin Topaloglu. Assortment optimization under the multinomial logit model with sequential offerings. *INFORMS Journal on Computing*, 32(3):835–853, 2020.

[37] Avishai Mandelbaum and Alexander L Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

[38] Nick McKeown, Adisak Mekkittikul, Venkat Anantharam, and Jean Walrand. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications*, 47(8): 1260–1267, 1999.

[39] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

[40] Aranyak Mehta et al. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science*, 8(4):265–368, 2013.

[41] Michael J Neely. Queue stability and probability 1 convergence via lyapunov optimization. *arXiv preprint arXiv:1008.3519*, 2010.

[42] Michael J Neely. Stability and capacity regions or discrete time queueing networks. *arXiv preprint arXiv:1003.3396*, 2010.

[43] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9205–9213, 2021.

[45] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits with uniform valuation. *arXiv preprint arXiv:2006.15166*, 2020.

[46] Flore Sentenac, Etienne Boursier, and Vianney Perchet. Decentralized learning in online queuing systems. *Advances in Neural Information Processing Systems*, 34:18501–18512, 2021.

[47] Thomas Stahlbuhk, Brooke Shrader, and Eytan Modiano. Learning algorithms for scheduling in wireless networks with unknown channel statistics. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 31–40, 2018.

[48] Thomas Stahlbuhk, Brooke Shrader, and Eytan Modiano. Learning algorithms for minimizing queue length regret. *IEEE Transactions on Information Theory*, 67(3):1759–1781, 2021.

[49] Alexander L Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[50] Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *29th IEEE Conference on Decision and Control*, pages 2130–2132. IEEE, 1990.

[51] Leandros Tassiulas and Anthony Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39(2):466–478, 1993.

[52] Zixian Yang, R Srikant, and Lei Ying. Learning while scheduling in multi-server systems with unknown statistics: Maxweight with discounted ucb. In *International Conference on Artificial Intelligence and Statistics*, pages 4275–4312. PMLR, 2023.

[53] Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pages 392–401. PMLR, 2016.

[54] Yirui Zhang, Siwei Wang, and Zhixuan Fang. Matching in multi-arm bandit with collision. *Advances in Neural Information Processing Systems*, 35:9552–9563, 2022.

A Appendix

A.1 Limitations & Discussion

We leave several questions open for future research. Firstly, it remains an open problem to establish lower bounds for queue lengths and regret. However, we believe that constructing a lower bound would be much more challenging compared to other parametric bandit problems. Additionally, improving the dependency on κ for queue length and regret bounds from the structure of MNL would be an interesting avenue for future work.

Regarding computation efficiency, we note that updating estimators in our algorithms is computationally efficient because they are based on online updates with convex optimization for updating estimators. Concerning combinatorial optimization in our algorithms, we can alleviate computational costs using α-approximation oracle algorithms, as discussed in Appendix A.9. Note that almost all combinatorial bandit problems, including our proposed matching bandit framework, involve a combinatorial optimization step which often relies on some type of approximation optimization oracles. Hence, this is not a particular limitation specific to our study. However, developing an improved approximation oracle would also be an interesting direction for future research.

A.2 Discussion on Traffic Slackness Parameter ϵ

As in the closely related works of [15, 52, 23], the traffic slackness remains constant regardless of the number of agents N and the number of servers K as $\epsilon = \epsilon_0$ for some $0 < \epsilon_0 < 1$. However, this may not align with intuition, as a large K might be beneficial in terms of traffic slackness, while a large N could have the opposite effect. To address this, we can consider $\epsilon = \epsilon_0 \frac{\min(N,K)}{N}$. This reflects that when $N \geq K$, due to the lack of servers, increasing K is critical for increasing traffic slackness, while increasing N could decrease it. When $N < K$ implying there are enough servers, however, the value of N doesn't impact the traffic slackness because each agent can be assigned to at most one server at each time, and there are sufficient servers to handle the agents.

If we consider the traffic slackness as $\epsilon = \epsilon_0 \frac{\min\{N,K\}}{N}$, the oracle strategy of MaxWeight achieves $\mathcal{Q}(T) = \mathcal{O}(\frac{N}{\epsilon_0})$ from Proposition 1. This result shows that as N increases, causing the traffic slackness to decrease, the average queue length increases. Meanwhile, the positive influence of K on traffic slackness is neutralized in the average queue length bound due to system variance.

A.3 Proof of Proposition 1

Define the set of queues $\mathbf{Q}(t) = [Q_n(t) : n \in [N]]$ and a Lyapunov function as $\mathcal{V}(\mathbf{Q}(t)) = \sum_{n\in[N]} Q_n(t)^2$. For simplicity, we use $D_n(t)$ for $D_n(t|S_{k_{n,t},t})$ when there is no confusion. We observe that $\sum_{n\in[N]} \mathbb{E}[A_n(t)] \leq \sum_{n\in[N]} \lambda_n \leq \sum_{n\in[N]} \mu(n|S_k, \theta_k) \leq K$ for some S_k from Assumption 3 and $\sum_{n\in[N]} \lambda_n \leq N$. This implies $\sum_{n\in[N]} [A_n(t)] \leq \min\{N, K\}$. We also have $\sum_{n\in[N]} \mathbb{E}[D_n(t)] = \sum_{n\in[N]} \mathbb{E}[\mu(n|S_{k,t}, \theta_k)] \leq \min\{K, N\}$. Then we have

$$
\begin{aligned}
&\mathbb{E}[\mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))] \\
&= \mathbb{E}\left[\sum_{n\in[N]} (Q_n(t) - A_n(t) + D_n(t))^{+2} - Q_n(t)^2\right] \\
&\leq \mathbb{E}\left[\sum_{n\in[N]} (Q_n(t) - A_n(t) + D_n(t))^2 - Q_n(t)^2\right] \\
&\leq \mathbb{E}\left[2\sum_{n\in[N]} Q_n(t)A_n(t) - 2\sum_{n\in[N]} Q_n(t)D_n(t) - 2\sum_{n\in[N]} A_n(t)D_n(t) + \sum_{n\in[N]} A_n(t)^2 + \sum_{n\in[N]} D_n(t)^2\right] \\
&\leq \mathbb{E}\left[2\sum_{n\in[N]} Q_n(t)A_n(t) - 2\sum_{n\in[N]} Q_n(t)D_n(t) + \sum_{n\in[N]} A_n(t) + \sum_{n\in[N]} D_n(t)\right] \\
&\leq \mathbb{E}\left[2\sum_{n\in[N]} Q_n(t)A_n(t) - 2\sum_{n\in[N]} Q_n(t)D_n(t) + 2\min\{N, K\}\right] \qquad (16)
\end{aligned}
$$

From Assumption 3, we define the corresponding assortments as $\{S'_k\}_{k\in[K]} \in \mathcal{M}(\mathcal{N})$, which satisfies $\lambda_n + \epsilon \le \mu(n|S'_k, \theta_k)$ for all $n \in S'_k$ and $k \in [K]$. Then we define the set of non-empty queues in S'_k at time t as $S'_{k,t} = \{n \in S'_k : Q_n(t) \neq 0\}$. From the property of the MNL function, we can observe that $\mu(n|S'_k, \theta_k) \le \mu(n|S'_{k,t}, \theta_k)$ for all $n \in S'_{k,t}$. We also note that $\{S'_{k,t}\}_{k\in[K]} \in \mathcal{M}(\mathcal{N}_t)$. Then we have

$$\begin{aligned}\sum_{n\in[N]} (\lambda_n + \epsilon) Q_n(t) &\le \sum_{k\in[K]} \sum_{n\in S'_k} \mu(n|S'_k, \theta_k) Q_n(t) \\ &\le \sum_{k\in[K]} \sum_{n\in S'_{k,t}} \mu(n|S'_{k,t}, \theta_k) Q_n(t) \\ &\le \sum_{k\in[K]} \sum_{n\in S_{k,t}} \mu(n|S_{k,t}, \theta_k) Q_n(t),\end{aligned}$$

where the second inequality is obtained from $\mu(n|S'_k, \theta_k)Q_n(t) \le \mu(n|S'_{k,t}, \theta_k)Q_n(t)$ when $n \in \mathcal{N}_t$, and otherwise $\mu(n|S'_k, \theta_k)Q_n(t) = \mu(n|S'_{k,t}, \theta_k)Q_n(t) = 0$.

This implies

$$\begin{aligned}\mathbb{E}\left[\sum_{k\in[K]} \sum_{n\in S_{k,t}} Q_n(t) D_n(t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{k\in[K]} \sum_{n\in S_{k,t}} Q_n(t) D_n(t) \middle| \mathbf{Q}(t)\right]\right] \\ &= \mathbb{E}\left[\sum_{k\in[K]} \sum_{n\in S_{k,t}} Q_n(t) \mu(n|S_{k,t}, \theta_k)\right] \\ &\ge \mathbb{E}\left[\sum_{k\in[K]} \sum_{n\in S'_{k,t}} Q_n(t) \mu(n|S'_{k,t}, \theta_k)\right] \\ &\ge \sum_{n\in[N]} (\lambda_n + \epsilon) \mathbb{E}[Q_n(t)]. \qquad (17)\end{aligned}$$

From Eqs.(16) and (17), we have

$$\begin{aligned}&\mathbb{E}[\mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))] \\ &\le \mathbb{E}\left[2 \sum_{n\in[N]} Q_n(t) A_n(t) - 2 \sum_{n\in[N]} Q_n(t) D_n(t)\right] + 2\min\{N, K\} \\ &\le \mathbb{E}\left[\left[2 \sum_{n\in[N]} Q_n(t) A_n(t) - 2 \sum_{n\in[N]} Q_n(t) D_n(t) \middle| \mathbf{Q}(t)\right]\right] + 2\min\{N, K\} \\ &\le \mathbb{E}\left[2 \sum_{n\in[N]} \lambda_n Q_n(t) - 2(\lambda_n + \epsilon) Q_n(t)\right] + 2\min\{N, K\} \\ &\le -\mathbb{E}\left[2\epsilon \sum_{n\in[N]} Q_n(t)\right] + 2\min\{N, K\},\end{aligned}$$

which implies from $\mathcal{V}(\mathbf{Q}(T+1)) \ge 0$ and $\mathcal{V}(\mathbf{Q}(1)) = 0$,

$$\begin{aligned}\sum_{t\in[T]} \mathbb{E}[\mathcal{V}(\mathbf{Q}(t+1) - \mathcal{V}(\mathbf{Q}(t))] &= \mathcal{V}(\mathbf{Q}(T+1)) \\ &\le -\sum_{t\in[T]} \mathbb{E}[2\epsilon \sum_{n\in[N]} Q_n(t)] + 2\min\{N, K\}T.\end{aligned}$$

Finally, we can conclude that $\mathcal{Q}(T) = (1/T)\sum_{t\in[T]} \mathbb{E}[\sum_{n\in[N]} Q_n(t)] \le 2\min\{N, K\}/\epsilon$.

A.4 Proof of Theorem 1

We first define the set of queues $\mathbf{Q}(t) = [Q_n(t) : n \in [N]]$ and a Lyapunov function as $\mathcal{V}(\mathbf{Q}(t)) = \sum_{n\in[N]} Q_n(t)^2$. For simplicity, we use $D_n(t)$ for $D_n(t|S_{k_{n,t},t})$ and $D_n^*(t)$ for $D_n(t|S_{k^*_{n,t},t})$ when there is no confusion. Then we analyze the Lyapunov drift as follows.

$$
\begin{aligned}
&\sum_{t\in[T]} \mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t)) \\
&= \sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n(t))^{+2} - Q_n(t)^2 \\
&= \sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n^*(t))^2 - Q_n(t)^2 \\
&\quad + \sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n(t))^{+2} - \sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n^*(t))^2. \quad (18)
\end{aligned}
$$

We observe that $\sum_{n\in[N]} \mathbb{E}[A_n(t)] \le \sum_{n\in[N]} \lambda_n \le \sum_{n\in[N]} \mu(n|S_k, \theta_k) \le K$ for some S_k from Assumption 3 and $\sum_{n\in[N]} \lambda_n \le N$. This implies $\sum_{n\in[N]}[A_n(t)] \le \min\{N, K\}$. We also have $\sum_{n\in[N]} \mathbb{E}[D_n(t)] = \sum_{n\in[N]} \mathbb{E}[\mu(n|S_{k,t}, \theta_k)] \le \min\{K, N\}$. For the first two terms in Eq.(18), by following the same procedure of Eqs.(16) and (17), we can obtain

$$
\begin{aligned}
&\sum_{t\in[T]} \sum_{n\in[N]} \mathbb{E}[(Q_n(t) + A_n(t) - D_n^*(t))^2 - Q_n(t)^2] \\
&\le \sum_{t\in[T]} \sum_{n\in[N]} 2\mathbb{E}[(\lambda_n - \mu(n|S^*_{k,t}, \theta_k))Q_n(t)] + 2\min\{N, K\}T \\
&\le -\sum_{t\in[T]} \sum_{n\in[N]} 2\epsilon\mathbb{E}[Q_n(t)] + 2\min\{N, K\}T, \quad (19)
\end{aligned}
$$

where the last inequality is obtained using Assumption 3. For the last two terms in Eq.(18), we have

$$
\begin{aligned}
&\sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n(t))^{+2} - \sum_{t\in[T]} \sum_{n\in[N]} (Q_n(t) + A_n(t) - D_n^*(t))^2 \\
&\le \sum_{t\in[T]} \sum_{n\in[N]} (D_n^*(t) - D_n(t))(2Q_n(t) + 2A_n(t) - D_n^*(t) - D_n(t)) \\
&\le 2\sum_{t\in[T]} \sum_{n\in[N]} A_n(t)(D_n^*(t) - D_n(t)) - \sum_{t\in[T]} \sum_{n\in[N]} D_n^*(t)^2 + \sum_{t\in[T]} \sum_{n\in[N]} D_n(t)^2 \\
&\quad + 2\sum_{t\in[T]} \sum_{n\in[N]} (D_n^*(t) - D_n(t))Q_n(t) \\
&\le 4\sum_{t\in[T]} \sum_{n\in[N]} A_n(t) + 2\sum_{t\in[T]} \sum_{n\in[N]} (D_n^*(t) - D_n(t))Q_n(t) + \min\{N, K\}T \\
&\le 2\sum_{t\in[T]} \sum_{n\in[N]} (D_n^*(t) - D_n(t))Q_n(t) + 5\min\{N, K\}T. \quad (20)
\end{aligned}
$$

Now we provide a bound for Eq.(20). We first provide some lemmas for the concentration of estimators.

Using the above lemma, we can show the following lemma of the concentration.

Lemma 1 (Lemma 9 in Oh and Iyengar [44]). *For $t \ge 1$ and some constant $C_1 > 0$, with probability at least $1 - 1/t^2$, for all $k \in [K]$ we have*

$$\|\hat{\theta}_{k,t} - \theta_k\|_{V_{k,t}} \le \beta_t.$$

Proof. For the completeness, we provide the proof in Appendix A.8. □

Define event $E_t^1 = \{\|\hat{\theta}_{k,t} - \theta_k^*\|_{V_{k,t}} \leq \beta_t$ for all $k \in [K]\}$ where $\beta_t = C_1\sqrt{\lambda + \frac{d}{\kappa}\log(1 + tLK/d\lambda)}$, which holds with high probability as $\mathbb{P}(E_t^1) \geq 1 - 1/t^2$ from the above lemma. We also define $E_{n,t}^2 = \{\max_{m \in S_{k,t}} \|x_m\|_{V_{k,t}^{-1}} \leq C_2\epsilon/2\beta_t$ for $k = k_{n,t}\}$ for some constant $C_2 > 0$. Then, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(D_n^*(t) - D_n(t)Q_n(t)] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[\mathbb{E}[(D_n^*(t) - D_n(t)Q_n(t)|Q_n(t)]] \\
&= \mathbb{E}[\sum_{t\in[T]}\sum_{n\in[N]} \mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_{n,t}^2)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E_t^1)^c) + \mathbb{1}((E_{n,t}^2)^c))].
\end{aligned}
\tag{21}
$$

Now we provide a lemma for bounding the first term in Eq.(21).

Lemma 2. *Under E_t^1, for any $n \in [N]$, we have $\widetilde{\mu}_t^{UCB}(n|S_{k_{n,t}}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}) \leq 2\beta_t\|x_n\|_{V_{k_{n,t},t}^{-1}}$.*

Proof. Let $u_{n,k,t} = x_n^\top\theta_k$. Under E_t^1, for any $n \in [N]$ and $k \in [K]$ we have $x_n^\top\hat{\theta}_{k,t} - \beta_t\|x_n\|_{V_{k,t}^{-1}} \leq x_n^\top\theta_k \leq x_n^\top\hat{\theta}_{k,t} + \beta_t\|x_n\|_{V_{k,t}^{-1}}$, which implies $0 \leq h_{n,k,t}^{UCB} - u_{n,k,t} \leq 2\beta_t\|x_n\|_{V_{k,t}^{-1}}$. Then by the mean value theorem, there exists $\bar{u}_{n,k,t} = (1-c)h_{n,k,t}^{UCB} + cu_{n,k,t}$ for some $c \in (0,1)$ satisfying, for any $n \in S_k$, $S_k \subset [N]$, and $k \in [K]$,

$$
\begin{aligned}
\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) - \mu(n|S_k, \theta_k) &= \frac{\exp(h_{n,k,t}^{UCB})}{1 + \sum_{m\in S_k}\exp(h_{m,k,t}^{UCB})} - \frac{\exp(u_{n,k,t})}{1 + \sum_{m\in S_k}\exp(u_{m,k,t})} \\
&= \nabla_{v_n}\left(\frac{\exp(v_n)}{1 + \sum_{m\in S_k}\exp(v_m)}\right)\Big|_{v_n = \bar{u}_{n,k,t}}(h_{n,k,t}^{UCB} - u_{n,k,t}) \\
&\leq \frac{\exp(\bar{u}_{n,k,t})(h_{n,k,t}^{UCB} - u_{n,k,t})}{1 + \sum_{n\in S_k}\exp(\bar{u}_{n,k,t})} \\
&\leq h_{n,k,t}^{UCB} - u_{n,k,t} \\
&\leq 2\beta_t\|x_n\|_{V_{k,t}^{-1}}.
\end{aligned}
$$

□

Since $x/(1+x)$ is a non-decreasing function for $x > -1$ and $x_n^\top\theta_{k_{n,t}^*} \leq x_n^\top\hat{\theta}_{k_{n,t}^*,t} + \beta_t\|x_n\|_{V_{k_{n,t}^*,t}^{-1}}$ under E_t^1, we have $\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) \leq \widetilde{\mu}_t^{UCB}(n|S_{k_{n,t}^*}, \hat{\theta}_{k_{n,t}^*,t})$. Then for the first term of Eq.(21),

we have

$$
\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_{n,t})\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\widetilde{\mu}^{UCB}_t(n|S_{k^*_{n,t}},\hat{\theta}_{k^*_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_{n,t})\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\widetilde{\mu}^{UCB}_t(n|S_{k_{n,t}},\hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_{n,t})\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} 2\beta_t \|x_n\|_{V^{-1}_{k_{n,t},t}} Q_n(t)\mathbb{1}(E^1_t \cap E^2_{n,t})\right] \\
&\leq C_2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right], \qquad (22)
\end{aligned}
$$

where the second inequality comes from the UCB strategy of the algorithm, the last second inequality is obtained from Lemma 2, and the last inequality is obtained from $E^2_{n,t}$.

Now we provide a bound for the second term of Eq.(21). We first have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^1_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\mathbb{1}((E^1_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^1_t)^c) \\
&= \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)), \qquad (23)
\end{aligned}
$$

where the second inequality is obtained from $Q_n(t) \leq t$.

Here we utilize some techniques introduced in Freund et al. [15]. Let $\mathcal{T}_n$ be the set of time steps $t \in [T]$ such that $Q_n(t) \neq 0$ and let $e_T = \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}((E^2_{n,t})^c)$ and $h = \lceil C_3 e_T/\epsilon \rceil$ for some constant $C_3 > 0$. Then if $t \leq h$, we have $Q_n(t) \leq t \leq h$. Otherwise, we have

$$
Q_n(t) \leq \sum_{s=t-h+1}^{t} (1/h)(Q_n(s) + (t-s)) \leq (1/h)\sum_{s=1}^{t} Q_n(s) + (1/h)h^2 = (1/h)\sum_{s=1}^{t} Q_n(s) + h.
$$

Then, from the above inequality, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_{n,t})^c)]\\
&=\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_{n,t})^c)]\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[Q_n(t)\mathbb{1}((E^2_{n,t})^c)]\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[((\epsilon/C_3e_T)\sum_{s=1}^{T}Q_n(s)+2C_3e_T/\epsilon)\mathbb{1}((E^2_{n,t})^c)]\\
&\le\mathbb{E}\left[(\epsilon/C_3)\sum_{n\in[N]}\sum_{s=1}^{T}Q_n(s)\right]+\mathbb{E}\left[2C_3e_T^2/\epsilon\right]\\
&=\sum_{t\in[T]}\sum_{n\in[N]}(\epsilon/C_3)\mathbb{E}\left[Q_n(t)\right]+2C_3\mathbb{E}[e_T^2]/\epsilon.
\end{aligned}
\tag{24}
$$

Now we provide a bound for $\mathbb{E}[e_T^2]$. Define $N_{n,k}(t)=\sum_{s=1}^{t-1}\mathbb{1}(n\in S_{k,s})$ and $\widetilde{V}_{k_{n,t},t}=\frac{\kappa}{2}\sum_{s=1}^{t-1}\sum_{n\in S_{k,s}}x_nx_n^\top$. Then, we have

$$
\begin{aligned}
e_T&=\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}((E^2_{n,t})^c)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(\|x_n\|_{V^{-1}_{k_{n,t},t}}\ge C_2\epsilon/2\beta_t)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(\|x_n\|_{\widetilde{V}^{-1}_{k_{n,t},t}}\ge C_2\epsilon/2\beta_t)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(1/N_{n,k_{n,t}}(t)\ge(\kappa/2)(C_2\epsilon/2\beta_t)^2)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(N_{n,k_{n,t}}(t)\le(2/\kappa)(2\beta_t/C_2\epsilon)^2)\\
&\le\sum_{n\in[N]}\sum_{k\in[K]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(N_{n,k}(t)\le(2/\kappa)(2\beta_T/C_2\epsilon)^2\text{ and }k_{n,t}=k)\\
&\le NK(2/\kappa)(2\beta_T/C_2\epsilon)^2.
\end{aligned}
$$

From the above we have $\mathbb{E}[e_T^2]\le 64N^2K^2\beta_T^4/\kappa^2(C_2\epsilon)^4$. Then from Eq.(24), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_{n,t})^c)]\\
&\le\sum_{t\in[T]}\sum_{n\in[N]}(\epsilon/C_3)\mathbb{E}\left[Q_n(t)\right]+\mathcal{O}(N^2K^2\beta_T^4/\kappa^2\epsilon^5).
\end{aligned}
\tag{25}
$$

By putting the results of Eqs. (18), (19), (20), (21), (22), (23), (25) altogether, we can obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]} \mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))\right] \\
&\leq 7\min\{N,K\}T - 2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + 2\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(D_n(t|S_{k^*_{n,t},t}) - D_n(t|S_{k_{n,t},t}))Q_n(t)\right] \\
&\leq 7\min\{N,K\}T - 2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + 2C_2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] \\
&\quad + \mathcal{O}(N\log(T)) + (2\epsilon/C_3)\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(N^2K^2\beta_T^4/\kappa^2\epsilon^5) \\
&\leq 7\min\{N,K\}T + 2(C_2 + (1/C_3) - 1)\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(N\log(T)) + \mathcal{O}(N^2K^2\beta_T^4/\kappa^2\epsilon^5).
\end{aligned}
$$

Finally, with positive constants $C_2, C_3 > 0$ satisfying $C_2 + (1/C_3) < 1$, from $\mathcal{V}(\mathbf{Q}(1)) = 0$ and $\mathcal{V}(\mathbf{Q}(T+1)) \geq 0$, by using telescoping for the above inequality, we can conclude the proof by

$$
\frac{1}{T}\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)] = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon} + \frac{1}{T}\frac{d^2N^2K^2}{\kappa^4\epsilon^6}\text{polylog}(T)\right). \tag{26}
$$

A.5 Proof of Theorem 2

We first provide the proof for regret bound of $\mathcal{R}^\pi(T) = \widetilde{\mathcal{O}}\left(\frac{d}{\kappa}\sqrt{KT}Q_{\max}\right)$.

We define event $E_t = \{\|\hat{\theta}_{k,t} - \theta^*_k\|_{V_{k,t}} \leq \beta_t \ \forall k \in [K]\}$ which holds at least probability of $1 - 1/t^2$ from Lemma 1.

Lemma 3. *Under E_t^1, for any $S_k \subset [N]$, we have*

$$
\sum_{n\in S_k} (\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) - \mu(n|S_k,\theta_k))Q_n(t) \leq 2\beta_t \max_{n\in S_k} \|x_n\|_{V_{k,t}^{-1}} Q_n(t).
$$

Proof. Let $u_{n,k,t} = x_n^\top \theta_k$. Under E_t^1, for any $n \in [N]$ and $k \in [K]$ we have $x_n^\top \hat{\theta}_{k,t} - \beta_t\|x_n\|_{V_{k,t}^{-1}} \leq x_n^\top\theta_k \leq x_n^\top\hat{\theta}_{k,t} + \beta_t\|x_n\|_{V_{k,t}^{-1}}$, which implies $0 \leq h_{n,k,t}^{UCB} - u_{n,k,t} \leq 2\beta_t\|x_n\|_{V_{k,t}^{-1}}$. Then by the mean value theorem, there exists $\bar{u}_{n,k,t} = (1-c)h_{n,k,t}^{UCB} + cu_{n,k,t}$ for some $c \in (0,1)$ satisfying, for any $S \subset [N]$,

$$
\begin{aligned}
&\sum_{n\in S_k} (\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) - \mu(n|S_k,\theta_k))Q_n(t) \\
&= \sum_{n\in S_k}\left(\frac{\exp(h_{n,k,t}^{UCB})}{1+\sum_{m\in S_k}\exp(h_{m,k,t}^{UCB})} - \frac{\exp(u_{n,k,t})}{1+\sum_{m\in S_k}\exp(u_{m,k,t})}\right)Q_n(t) \\
&= \sum_{n\in S_k}\nabla_{v_n}\left(\frac{\exp(v_n)}{1+\sum_{m\in S_k}\exp(v_m)}\right)\Big|_{v_n=\bar{u}_{n,k,t}}(h_{n,k,t}^{UCB} - u_{n,k,t})Q_n(t) \\
&= \frac{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))(\sum_{n\in S_k}\exp(\bar{u}_{n,k,t})(h_{n,k,t}^{UCB} - u_{n,k,t})Q_n(t))}{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))^2} \\
&\quad - \frac{(\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))(\sum_{n\in S_k}\exp(\bar{u}_{n,k,t})(h_{n,k,t}^{UCB} - u_{n,k,t})Q_n(t))}{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))^2} \\
&\leq \sum_{n\in S_k}\frac{\exp(\bar{u}_{n,k,t})}{1+\sum_{m\in S_k}\exp(\bar{u}_{m,k,t})}(h_{n,k,t}^{UCB} - u_{n,k,t})Q_n(t) \\
&\leq \max_{n\in S_k}(h_{n,k,t}^{UCB} - u_{n,k,t})Q_n(t) \\
&\leq 2\beta_t \max_{n\in S_k}\|x_n\|_{V_{k,t}^{-1}}Q_n(t).
\end{aligned}
$$

□

Under E_t, from Lemma 3, we have

$$\sum_{n \in S_{k,t}} (\widetilde{\mu}_t^{UCB}(n|S_{k,t}, \hat{\theta}_{k,t}) - \mu(n|S_{k,t}, \theta_k))Q_n(t) \leq 2\beta_t \max_{n \in S_{k,t}} \|x_n\|_{V_{k,t}^{-1}} Q_n(t), \tag{27}$$

and since $x/(1+x)$ is a non-decreasing function for $x > -1$ and $x_n^\top \theta_{k_{n,t}^*} \leq x_n^\top \hat{\theta}'_{k_{n,t}^*,t} + \beta_t \|x_n\|_{V_{k_{n,t}^*,t}^{-1}}$ under E_t^1, we have

$$\mu(n|S_{k_{n,t}^*,t}, \theta_{k_{n,t}^*}) \leq \widetilde{\mu}_t^{UCB}(n|S_{k_{n,t}^*,t}, \hat{\theta}_{k_{n,t}^*,t}). \tag{28}$$

Now we provide an elliptical potential lemma.

Lemma 4. *For any $k \in [K]$, we have*

$$\sum_{t=1}^{T} \max_{n \in S_{k,t}} \|x_n\|^2_{V_{k,t}^{-1}} \leq (4d/\kappa) \log(1 + (TL/d\lambda)).$$

Proof. First, we can show that

$$\begin{aligned}
&\det(V_{k,t}) \\
&= \det\left(V_{k,t-1} + (\kappa/2) \sum_{n \in S_{k,t-1}} x_n x_n^\top\right) \\
&= \det(V_{k,t-1}) \det\left(I_d + (\kappa/2) \sum_{n \in S_{k,t-1}} V_{k,t-1}^{-1/2} x_n (V_{k,t-1}^{-1/2} x_n)^\top\right) \\
&\geq \det(V_{k,t-1}) \left(1 + (\kappa/2) \sum_{n \in S_{k,t-1}} \|x_n\|^2_{V_{k,t-1}^{-1}}\right) \\
&\geq \det(\lambda I_d) \prod_{s=1}^{t-1} \left(1 + (\kappa/2) \sum_{n \in S_{k,s}} \|x_n\|^2_{V_{k,s}^{-1}}\right) = \lambda^d \prod_{s=1}^{t-1} \left(1 + (\kappa/2) \sum_{n \in S_{k,s}} \|x_n\|^2_{V_{k,s}^{-1}}\right)
\end{aligned}$$

From the above, using the fact that $x \leq 2\log(1+x)$ for any $x \in [0,1]$ and $(\kappa/2)\|x_n\|^2_{\widetilde{V}_{k,s}^{-1}} \leq (\kappa/2)\|x_n\|_2^2/\lambda \leq 1$ from $\kappa < 1$, we have

$$\begin{aligned}
\sum_{t \in [T]} \max_{n \in S_{k,t}} (\kappa/2)\|x_n\|^2_{V_{k,t}^{-1}} &\leq \sum_{t \in [T]} \min\left\{\max_{n \in S_{k,t}} (\kappa/2)\|x_n\|^2_{V_{k,t-1}^{-1}}, 1\right\} \\
&\leq 2 \sum_{t \in [T]} \log\left(1 + \sum_{n \in S_{k,t}} (\kappa/2)\|x_n\|^2_{V_{k,t-1}^{-1}}\right) \\
&= 2 \log \prod_{t \in [T]} \left(1 + \sum_{n \in S_{k,t}} (\kappa/2)\|x_n\|^2_{V_{k,t-1}^{-1}}\right) \\
&\leq 2 \log\left(\frac{\det(V_{k,T+1})}{\lambda^d}\right).
\end{aligned} \tag{29}$$

From Lemma 10 in Abbasi-Yadkori et al. [1] with $\kappa < 1$, we can show that

$$\det(V_{k,T+1}) \leq (\lambda + (TL/d))^d.$$

Then from the above inequality and Eq.(29), we can conclude the proof. □

Then from Eqs.(27), (28), and Lemma 4, we can conclude that

$$
\begin{aligned}
\mathcal{R}^{\pi}(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E^1_t)^c)] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^1_t)^c)] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\widetilde{\mu}^{UCB}_t(n|S_{k^*_{n,t},t},\hat{\theta}_{k^*_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\widetilde{\mu}^{UCB}_t(n|S_{k_{n,t},t},\hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&= \sum_{t\in[T]} \mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}} Q_n(t)(\widetilde{\mu}^{UCB}_t(n|S_{k,t},\hat{\theta}_{k,t}) - \mu(n|S_{k,t},\theta_k))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\beta_T \sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t)\right] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\max_{t\in[T],n\in[N]} Q_n(t)\beta_T \sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}}\right] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\max_{t\in[T],n\in[N]} Q_n(t)\beta_T \sqrt{KT\sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S_{k,t}} \|x_n\|^2_{V^{-1}_{k,t}}}\right] + \mathcal{O}(N/T) \\
&= \widetilde{\mathcal{O}}\left(\frac{d}{\kappa}\sqrt{KT}Q_{\max}\right),
\end{aligned}
$$

where the last equality comes from Lemma 4.

Now we provide the worst-case regret bound of $\mathcal{R}^{\pi}(T) = \widetilde{\mathcal{O}}\left(\left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right)$ in the following.

We define event $E^1_t = \{\|\hat{\theta}_{k,t} - \theta^*_k\|_{V_{k,t}} \le \beta_t$ for all $k \in [K]\}$ where $\beta_t = C_1\sqrt{\lambda + \frac{d}{\kappa}\log(1 + tLK/d\lambda)}$, which holds with high probability as $\mathbb{P}(E^1_t) \ge 1 - 1/t^2$ from Lemma 1. We also define $E^2_{n,t} = \{\max_{m\in S_{k,t}} \|x_m\|_{V^{-1}_{k,t}} \le \zeta$ for $k = k_{n,t}\}$ for some constant $C_2 > 0$.

Then we have

$$
\begin{aligned}
\mathcal{R}^{\pi}(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_{n,t})] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E^1_t)^c) + \mathbb{1}((E^2_{n,t})^c))].
\end{aligned} \tag{30}
$$

Since $x/(1+x)$ is a non-decreasing function for $x > -1$ and $x_n^\top \theta_{k^*_{n,t}} \leq x_n^\top \hat{\theta}_{k^*_{n,t},t} + \beta_t \|x_n\|_{V^{-1}_{k^*_{n,t},t}}$ under E_t^1, we have $\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) \leq \widetilde{\mu}_t^{UCB}(n|S_{k^*_{n,t}}, \hat{\theta}_{k^*_{n,t},t})$. Then for the first term of Eq.(30), we have

$$
\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_{n,t}^2)\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\widetilde{\mu}_t^{UCB}(n|S_{k^*_{n,t}}, \hat{\theta}_{k^*_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_{n,t}^2)\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} (\widetilde{\mu}_t^{UCB}(n|S_{k_{n,t}}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_{n,t}^2)\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{n\in[N]} 2\beta_t \|x_n\|_{V^{-1}_{k_{n,t},t}} Q_n(t)\mathbb{1}(E_t^1 \cap E_{n,t}^2)\right] \\
&\leq \sum_{t\in[T]} \sum_{n\in[N]} 2\beta_t \zeta \mathbb{E}\left[Q_n(t)\right], \qquad (31)
\end{aligned}
$$

where the second inequality comes from the UCB strategy of the algorithm, the last second inequality is obtained from Lemma 2, and the last inequality is obtained from $E_{n,t}^2$.

Now we provide a bound for the second term of Eq.(30). From Eq. (23), we have

$$
\begin{aligned}
&\sum_{t\in[T]} \sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_t^1)^c)\right] \\
&= \mathcal{O}\left(\sum_{t\in[T]} \sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)). \qquad (32)
\end{aligned}
$$

Let $\mathcal{T}_n$ be the set of time steps $t \in [T]$ such that $Q_n(t) \neq 0$ and let $e_T = \sum_{n\in[N]} \sum_{t\in\mathcal{T}_n} \mathbb{1}((E_{n,t}^2)^c)$ and $h = \lceil 1/\zeta\beta_T \rceil$. Then if $t \leq h$, we have $Q_n(t) \leq t \leq h$. Otherwise, we have

$$
Q_n(t) \leq \sum_{s=t-h+1}^{t} (1/h)(Q_n(s) + (t-s)) \leq (1/h)\sum_{s=1}^{t} Q_n(s) + (1/h)h^2 = (1/h)\sum_{s=1}^{t} Q_n(s) + h.
$$

Then, by following the steps in Eq.(24), we have

$$
\begin{aligned}
&\sum_{t\in[T]} \sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_{n,t}^2)^c)] \\
&\leq \sum_{t\in[T]} \sum_{n\in[N]} \zeta\beta_T \mathbb{E}\left[Q_n(t)\right] + 2\mathbb{E}[e_T]/\zeta\beta_T. \qquad (33)
\end{aligned}
$$

Now we provide a bound for $\mathbb{E}[e_T]$. Define $N_{n,k}(t) = \sum_{s=1}^{t-1} \mathbb{1}(n \in S_{k,s})$ and $\widetilde{V}_{k,t} = \frac{\kappa}{2}\sum_{s=1}^{t-1}\sum_{n\in S_{k,s}} x_n x_n^\top$. Then, we have

$$
\begin{aligned}
e_T &= \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}((E_{n,t}^2)^c) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(\|x_n\|_{V_{k_{n,t},t}^{-1}} \ge \zeta) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(\|x_n\|_{\widetilde{V}_{k_{n,t},t}^{-1}} \ge \zeta) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(1/N_{n,k_{n,t}}(t) \ge (\kappa/2)\zeta^2) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(N_{n,k_{n,t}}(t) \le 2/\kappa\zeta^2) \\
&\le \sum_{n\in[N]}\sum_{k\in[K]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(N_{n,k}(t) \le 2/\kappa\zeta^2 \text{ and } k_{n,t} = k) \\
&\le 2NK/\kappa\zeta^2. \qquad (34)
\end{aligned}
$$

Then from Eqs.(33), (34), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_{n,t}^2)^c)] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \zeta\beta_T \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(NK/\kappa\zeta^3\beta_T). \qquad (35)
\end{aligned}
$$

By putting the results of Eqs. (30), (31), (32), (35), and Theorem 1, by setting $\zeta = (\epsilon NK/\min\{N,K\}T\kappa\beta_T^2)^{1/4}$, for large enough T, we can obtain

$$
\begin{aligned}
\mathcal{R}^\pi(T) &= \mathcal{O}\left(\zeta\beta_T \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)] + \frac{NK}{\kappa\zeta^3\beta_T} + N\log(T)\right) \\
&= \mathcal{O}\left(\frac{\zeta\beta_T \min\{N,K\}T}{\epsilon} + \frac{NK}{\kappa\zeta^3\beta_T} + N\log(T)\right) \\
&= \mathcal{O}\left(\frac{\beta_T^{1/2}T^{3/4}(NK\min\{N,K\}^3)^{1/4}}{\kappa^{1/4}\epsilon^{3/4}} + N\log(T)\right) \\
&= \widetilde{\mathcal{O}}\left(\left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right) \qquad (36)
\end{aligned}
$$

A.6 Proof of Theorm 3

We first define the set of queues $\mathbf{Q}(t) = [Q_n(t) : n \in [N]]$ and a Lyapunov function as $\mathcal{V}(\mathbf{Q}(t)) = \sum_{n\in[N]} Q_n(t)^2$. For simplicity, we use $D_n(t)$ for $D_n(t|S_{k_{n,t},t})$, $D_n^*(t)$ for $D_n(t|S_{k_{n,t}^*,t})$, and $\widetilde{\mu}_t^{TS}(n|S_{k,t})$ for $\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})$ when there is no confusion. Then we analyze the Lya-

punov drift as follows.

$$
\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}\left[\mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))\right] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(Q_n(t)+A_n(t)-D_n(t))^{+2} - Q_n(t)^2\right] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(Q_n(t)+A_n(t)-D_n^*(t))^{2} - Q_n(t)^2\right] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(Q_n(t)+A_n(t)-D_n(t))^{+2} - (Q_n(t)+A_n(t)-D_n^*(t))^2\right] \\
&\le 7\min\{N,K\}T - \sum_{t\in[T]}\sum_{n\in[N]} 2\epsilon\mathbb{E}[Q_n(t)] + 2\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(D_n^*(t)-D_n(t))Q_n(t)\right], \quad (37)
\end{aligned}
$$

where the last inequality can be obtained by following Eqs.(19) and (20).

Define event $E_t^1 = \{\|\hat{\theta}_{k,t} - \theta_k\|_{V_{k,t}} \le \beta_t$ for all $k \in [K]\}$ where $\beta_t = C_1\sqrt{\lambda + \frac{d}{\kappa}\log(1+tLK/d\lambda)}$, which holds with high probability as $\mathbb{P}(E_t^1) \ge 1 - 1/t^2$ from Lemma 1. We let $\gamma_t = \beta_t\sqrt{d\log(MKt)}$ and filtration $\mathcal{F}_{t-1}$ be the σ-algebra generated by random variables before time t.

Lemma 5 (Lemma 10 in Oh and Iyengar [43])**.** *For any given $\mathcal{F}_{t-1}$, with probability at least $1-\mathcal{O}(1/t^2)$, for all $n \in [N]$ and $k \in [K]$, we have*

$$
|h_{n,k,t}^{TS} - x_n^\top \hat{\theta}_{k,t}| \le \gamma_t \|x_n\|_{V_{k,t}^{-1}}.
$$

Lemma 6. *With probability at least $1-\mathcal{O}(1/t^2)$, for all $n \in [N]$ and $k \in [K]$, we have*

$$
\widetilde{\mu}_t^{TS}(n|S_{k,t}) - \mu(n|S_{k,t},\hat{\theta}_{k,t}) \le \gamma_t \|x_n\|_{V_{k,t}^{-1}}.
$$

Proof. From Lemma 5, with probability at least $1-\mathcal{O}(1/t^2)$, we have $|h_{n,k,t}^{TS} - x_n^\top\hat{\theta}_{k,t}| \le \gamma_t\|x_n\|_{V_{k,t}^{-1}}$. Let $u_{n,k,t} = x_n^\top\hat{\theta}_{k,t}$. Then by the mean value theorem, there exists $\bar{u}_{n,k,t} = (1-c)h_{n,k,t}^{TS} + cu_{n,k,t}$ for some $c \in (0,1)$ satisfying, for any $n \in S_{k,t}$ and $k \in [K]$,

$$
\begin{aligned}
\widetilde{\mu}_t^{TS}(n|S_{k,t}) - \mu(n|S_k,\hat{\theta}_{k,t}) &= \frac{\exp(h_{n,k,t}^{TS})}{1+\sum_{m\in S_{k,t}}\exp(h_{m,k,t}^{TS})} - \frac{\exp(u_{n,k,t})}{1+\sum_{m\in S_{k,t}}\exp(u_{m,k,t})} \\
&= \nabla_{v_n}\left(\frac{\sum_{m\in S_{k,t}}\exp(v_m)}{1+\sum_{m\in S_{k,t}}\exp(v_m)}\right)\Big|_{v_n=\bar{u}_{n,k,t}}(h_{n,k,t}^{TS} - u_{n,k,t}) \\
&\le \frac{\exp(\bar{u}_{n,k,t})|h_{n,k,t}^{TS} - u_{n,k,t}|}{1+\sum_{n\in S_{k,t}}\exp(\bar{u}_{n,k,t})} \\
&\le |h_{n,k,t}^{TS} - u_{n,k,t}| \\
&\le \gamma_t\|x_n\|_{V_{k,t}^{-1}}.
\end{aligned}
$$

□

Then we define $E_t^2 = \{\widetilde{\mu}_t^{TS}(n|S_{k,t}) - \mu(n|S_{k,t},\hat{\theta}_{k,t}) \le \gamma_t\|x_n\|_{V_{k,t}^{-1}}; \forall n \in [N], \forall k \in [K]\}$, which holds with probability at least $1-\mathcal{O}(1/t^2)$ from Lemma 6. We also define $E_{n,t}^3 = \{\|x_n\|_{V_{k,t}^{-1}} \le$

$\epsilon/C_2(\gamma_t+\beta_t); \forall k \in [K]\}$ for some constant $C_2 \geq 17\sqrt{e\pi}$. Then, for bounding Eq.(37), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(D_n^*(t) - D_n(t)Q_n(t)] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[\mathbb{E}[(D_n^*(t) - D_n(t)Q_n(t)|Q_n(t)]] \\
&= \mathbb{E}[\sum_{t\in[T]}\sum_{n\in[N]} \mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E_t^1)^c) + \mathbb{1}((E_t^2)^c) + \mathbb{1}((E_{n,t}^3)^c))].
\end{aligned} \tag{38}
$$

We provide a bound for the first term of Eq.(38). We first have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) \\
&\qquad\qquad + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)].
\end{aligned} \tag{39}
$$

Recall that $\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) = \frac{\exp(h_{n,k,t}^{UCB})}{1+\sum_{m\in S_k}\exp(h_{m,k,t}^{UCB})}$. Then we can show that since $x/(1+x)$ is a non-decreasing function for $x > -1$ and $x_n^\top \hat{\theta}_{k_{n,t},t} \leq x_n^\top \hat{\theta}_{k_{n,t},t} + \beta_t \|x_n\|_{V_{k_{n,t},t}^{-1}}$ under E_t^1, with Lemma 2, we have

$$
\mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}) \leq \widetilde{\mu}_t^{UCB}(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}) \leq 2\beta_t \|x_n\|_{V_{k_{n,t},t}^{-1}}.
$$

From the above inequality, the last two terms in Eq.(39) are bounded as

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[\widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\big[(\widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) \\
&\qquad\qquad + \mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)\big] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) + 2\beta_t \|x_n\|_{V_{k_{n,t},t}^{-1}})Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)] \\
&= \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\sum_{n\in S_{k,t}} (\widetilde{\mu}_t^{TS}(n|S_{k,t}) - \mu(n|S_{k,t}, \hat{\theta}_{k,t}) + 2\beta_t \|x_n\|_{V_{k,t}^{-1}})Q_n(t)\mathbb{1}(E_t^1 \cap E_t^2 \cap E_{n,t}^3)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\sum_{n\in S_{k,t}} (\gamma_t+2\beta_t)\|x_n\|_{V_{k,t}^{-1}} Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\right] \\
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\sum_{n\in S_{k,t}} \frac{(\gamma_t+2\beta_t)\epsilon}{C_2(\gamma_t+\beta_t)} Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\right] \\
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\sum_{n\in S_{k,t}} \frac{(\gamma_t+2\beta_t)\epsilon}{C_2(\gamma_t+\beta_t)} Q_n(t)\mathbb{1}(E_t^1)\right] \\
&= \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\mathbb{E}\left[\sum_{n\in S_{k,t}} \frac{(\gamma_t+2\beta_t)\epsilon}{C_2(\gamma_t+\beta_t)} Q_n(t)|E_t^1,\mathcal{F}_{t-1}\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\right]. \qquad (40)
\end{aligned}
$$

Now we provide a bound for the first two terms in Eq.(39).

We define sets

$$
\widetilde{\Theta}_t = \left\{\{\theta_k^{(i)}\}_{i\in[M],k\in[K]} : \left|\max_{i\in[M]} x_n^\top\theta_k^{(i)} - x_n^\top\hat{\theta}_{k,t}\right| \leq \gamma_t\|x_n\|_{V_{k,t}^{-1}};\ \forall n\in[N],\forall k\in[K]\right\} \text{ and}
$$

$$
\begin{aligned}
\widetilde{\Theta}_t^{opt} = \Bigg\{\{\theta_k^{(i)}\}_{i\in[M],k\in[K]} : &\sum_{k\in[K]}\sum_{n\in S_{k,t}} \widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})Q_n(t) \\
&> \sum_{k\in[K]}\sum_{n\in S_{k,t}^*} \mu(n|S_{k,t}^*,\theta_k)Q_n(t)\Bigg\}\cap\widetilde{\Theta}_t.
\end{aligned}
$$

Then we define event $E_t(\widetilde{\theta}) = \{\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M],k\in[K]} \in \widetilde{\Theta}_t^{opt}\}$. Recall $h_{n,k,t}^{TS} = \max_{i\in[M]} x_n^\top\widetilde{\theta}_{k,t}^{(i)}$. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\sum_{k\in[K]}\left(\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t) - \sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})Q_n(t)\right)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1}\right]\right] \\
&\leq \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t) \\
&\quad - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\ \max_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)}\sum_{k\in[K]}\sum_{n\in S_k}\widetilde{\mu}_t^{TS}(n|S_k,\{\theta_k^{(i)}\}_{i\in[M]})Q_n(t)\Bigg)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1}\Bigg]\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t) \\
&\quad - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\ \max_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)}\sum_{k\in[K]}\sum_{n\in S_k}\widetilde{\mu}_t^{TS}(n|S_k,\{\theta_k^{(i)}\}_{i\in[M]})Q_n(t)\Bigg)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\Bigg]\Bigg] \\
&\leq \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}) \\
&\quad - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})\Bigg)Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\Bigg]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[\sup_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\left(\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})\right.\right.\right.\\
&\qquad\left.\left.\left.-\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})\right)Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&\le \mathbb{E}\left[\mathbb{E}\left[\sup_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\left|h_{n,k,t}^{TS}-\max_{j\in[M]}x_n^\top\theta_k^{(j)}\right|Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\sup_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\left|h_{n,k,t}^{TS}-x_n^\top\hat{\theta}_{k,t}+x_n^\top\hat{\theta}_{k,t}-\max_{j\in[M]}x_n^\top\theta_k^{(j)}\right|\right.\right.\\
&\qquad\left.\left.\times Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&\le 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}\mathbb{E}\left[\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&\le 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}\mathbb{E}\left[\frac{\epsilon}{C_2(\gamma_t+\beta_t)}Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&\le 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}\mathbb{E}\left[\frac{\epsilon}{C_2(\gamma_t+\beta_t)}Q_n(t)\mathbb{1}(E_t^1)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right]\\
&= \frac{2\gamma_t\epsilon}{C_2(\gamma_t+\beta_t)}\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}\mathbb{E}\left[Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\times\mathbb{P}(E_t^1|E_t(\widetilde{\theta}),\mathcal{F}_{t-1})\right]\\
&= \frac{2\gamma_t\epsilon}{C_2(\gamma_t+\beta_t)}\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}\mathbb{E}\left[Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\right],
\end{aligned}
\tag{41}
$$

where the first inequality is obtained by the event E_t^2, the second inequality is obtained from $E_t(\widetilde{\theta})$, the third inequality can be easily obtained by following some of the proof steps in Lemma 2, the third last inequality is obtained from the definition of $\widetilde{\Theta}_t$ and event $E_t(\widetilde{\theta})$, and the last equality comes from independence between E_t^1 and $E_t(\widetilde{\theta})$ given $\mathcal{F}_{t-1}$.

We provide a lemma below for further analysis.

Lemma 7. *For all $t\in[T]$, we have*

$$
\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t})Q_n(t)>\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t)|\mathcal{F}_{t-1},E_t^1\right)\ge 1/4\sqrt{e\pi}.
$$

Proof. Given $\mathcal{F}_{t-1}$, $x_n^\top\widetilde{\theta}_{k,t}^{(i)}$ follows Gaussian distribution with mean $x_n^\top\hat{\theta}_{k,t}$ and standard deviation $\beta_t\|x_n\|_{V_{k,t}^{-1}}$. Then we have

$$
\begin{aligned}
\mathbb{P}\left(\max_{i\in[M]}x_n^\top\widetilde{\theta}_{k,t}^{(i)}>x_n^\top\theta_k|\mathcal{F}_{t-1},E_t^1\right)&=1-\mathbb{P}\left(x_n^\top\theta_{k,t}^{(i)}\le x_n^\top\theta_k;\forall i\in[M]|\mathcal{F}_{t-1},E_t^1\right)\\
&=1-\mathbb{P}\left(Z_i\le\frac{x_n^\top\theta_k-x_n^\top\hat{\theta}_{k,t}}{\beta_t\|x_n\|_{V_{k,t}^{-1}}};\forall i\in[M]|\mathcal{F}_{t-1},E_t^1\right)\\
&\ge 1-\mathbb{P}\left(Z\le 1\right)^M,
\end{aligned}
$$

where Z_i and Z are standard normal random variables. Then we can show that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\mu(n|S^*_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E_t^1\right)\\
&\geq \mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\widetilde{\mu}_t^{TS}(n|S^*_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\mu(n|S^*_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E_t^1\right)\\
&\geq \mathbb{P}\left(\max_{i\in[M]} x_n^\top\widetilde{\theta}_{k,t}^{(i)} > x_n^\top\theta_k; \forall n\in S^*_{k,t}, \forall k\in[K]|\mathcal{F}_{t-1},E_t^1\right)\\
&\geq 1-LK\mathbb{P}(Z\leq 1)^M\\
&\geq 1-LK(1-1/4\sqrt{e\pi})^M\\
&\geq \frac{1}{4\sqrt{e\pi}},
\end{aligned}$$

where the second last inequality is obtained from $\mathbb{P}(Z \leq 1) \leq 1 - 1/4\sqrt{e\pi}$ using the anti-concentration of standard normal distribution, and the last inequality comes from $M = \lceil 1 - \frac{\log KL}{\log(1-1/4\sqrt{e\pi})}\rceil$. □

From Lemmas 5 and 7, for $t \geq t_0$ for some constant $t_0 > 0$, we have

$$\begin{aligned}
&\mathbb{P}(E_t(\widetilde{\theta})|\mathcal{F}_{t-1},E_t^1)\\
&=\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t}) > \sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\mu(n|S^*_{k,t},\theta_k) \text{ and } \{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M],k\in[K]}\in\widetilde{\Theta}_t|\mathcal{F}_{t-1},E_t^1\right)\\
&=\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\mu(n|S^*_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E_t^1\right)\\
&\qquad -\mathbb{P}(\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M],k\in[K]}\notin\widetilde{\Theta}_t|\mathcal{F}_{t-1},E_t^1)\\
&\geq 1/4\sqrt{e\pi}-\mathcal{O}(1/t^2)\\
&\geq 1/8\sqrt{e\pi}.
\end{aligned}$$

For simplicity of the proof, we ignore the time steps before (constant) t_0, which does not affect our final result. Hence, we have

$$\begin{aligned}
&\mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}}Q_n(t)|\mathcal{F}_{t-1},E_t^1]\\
&\geq \mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}}Q_n(t)|\mathcal{F}_{t-1},E_t^1,E_t(\widetilde{\theta})]\mathbb{P}(E_t(\widetilde{\theta})|\mathcal{F}_{t-1},E_t^1)\\
&\geq \mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}}Q_n(t)|\mathcal{F}_{t-1},E_t^1,E_t(\widetilde{\theta})]1/8\sqrt{e\pi}. \qquad (42)
\end{aligned}$$

With (41) and (42), we have

$$\begin{aligned}
&\mathbb{E}[(\sum_{k\in[K]}\sum_{n\in S^*_{k,t}}\mu(n|S^*_{k,t},\theta_k)-\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t}))Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1}]\\
&\leq \frac{2\gamma_t\epsilon}{C_2(\gamma_t+\beta_t)}\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}}Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\\
&\leq \frac{16\sqrt{e\pi}\gamma_t\epsilon}{C_2(\gamma_t+\beta_t)}\mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}}Q_n(t)|\mathcal{F}_{t-1},E_t^1]\mathbb{P}(E_t^1|\mathcal{F}_{t-1}). \qquad (43)
\end{aligned}$$

Then for the first term of Eq.(38), from Eqs.(39), (40), (43), for some $C_3 > 0$ we have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) \\
&\qquad\qquad + \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}[\mathbb{E}[\sum_{n\in S_{k,t}} \frac{(17\sqrt{e\pi}\gamma_t + 2\beta_t)\epsilon}{C_2(\gamma_t+\beta_t)} Q_n(t)|\mathcal{F}_{t-1}, E^1_t]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} C_3\epsilon\mathbb{E}[\mathbb{E}[Q_n(t)|E^1_t,\mathcal{F}_{t-1}]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} C_3\epsilon\mathbb{E}[Q_n(t)\mathbb{1}(E^1_t)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} C_3\epsilon\mathbb{E}[Q_n(t)], \qquad (44)
\end{aligned}$$

where the second last inequality comes from $C_2 \geq 17\sqrt{e\pi}$.

For the second term of Eq.(38), we first have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^1_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\mathbb{1}((E^1_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^1_t)^c) \\
&= \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)), \qquad (45)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\mathbb{1}((E^2_t)^c)\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^2_t)^c) \\
&= \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)). \qquad (46)
\end{aligned}$$

Let $\mathcal{T}_n$ be the set of time steps $t \in [T]$ such that $Q_n(t) \neq 0$ and let $e_T = \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}((E^3_{n,t})^c)$ and $h = \lceil C_4 e_T/\epsilon \rceil$ for some constant $C_4 > 0$. Then if $t \leq h$, we have $Q_n(t) \leq t \leq h$. Otherwise, we have

$$Q_n(t) \leq \sum_{s=t-h+1}^{t} (1/h)(Q_n(s) + (t-s)) \leq (1/h)\sum_{s=1}^{t} Q_n(s) + (1/h)h^2 = (1/h)\sum_{s=1}^{t} Q_n(s) + h.$$

From the above, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^3_{n,t})^c)]\\
&=\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^3_{n,t})^c)]\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[Q_n(t)\mathbb{1}((E^3_{n,t})^c)]\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{E}[((\epsilon/C_4e_T)\sum_{s=1}^{T}Q_n(s)+2C_4e_T/\epsilon)\mathbb{1}((E^3_{n,t})^c)]\\
&\le\mathbb{E}\left[(\epsilon/C_4)\sum_{n\in[N]}\sum_{s=1}^{T}Q_n(s)\right]+\mathbb{E}\left[2C_4e_T^2/\epsilon\right]\\
&\le\sum_{t\in[T]}\sum_{n\in[N]}(\epsilon/C_4)\mathbb{E}\left[Q_n(t)\right]+2C_4\mathbb{E}[e_T^2]/\epsilon. \qquad (47)
\end{aligned}
$$

Now we provide a bound for $\mathbb{E}[e_T^2]$. Define $N_{n,k}(t)=\sum_{s=1}^{t-1}\mathbb{1}(n\in S_{k,s})$ and $\widetilde{V}_{k,t}=(\kappa/2)\sum_{s=1}^{t-1}\sum_{n\in S_{k,s}}x_nx_n^\top$. Then, we have

$$
\begin{aligned}
e_T&=\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}((E^2_{n,t})^c)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(\|x_n\|_{V^{-1}_{k_{n,t},t}}\ge\epsilon/C_2(\gamma_t+\beta_t))\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(\|x_n\|_{\widetilde{V}^{-1}_{k_{n,t},t}}\ge\epsilon/C_2(\gamma_t+\beta_t))\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(1/N_{n,k_{n,t}}(t)\ge(\kappa/2)(\epsilon/C_2(\gamma_t+\beta_t))^2)\\
&\le\sum_{n\in[N]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(N_{n,k_{n,t}}(t)\le(2/\kappa)(C_2(\gamma_t+\beta_t)/\epsilon)^2)\\
&\le\sum_{n\in[N]}\sum_{k\in[K]}\sum_{t\in\mathcal{T}_n}\mathbb{1}(N_{n,k}(t)\le(2/\kappa)(C_2(\gamma_t+\beta_t)/\epsilon)^2\text{ and }k_{n,t}=k)\\
&\le NK(2/\kappa)(C_2(\gamma_T+\beta_T)/\epsilon)^2.
\end{aligned}
$$

From the above we have $\mathbb{E}[e_T^2]\le N^2K^2(4/\kappa^2)(C_2(\gamma_T+\beta_T)/\epsilon)^4$. Then from Eq.(47) we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_{n,t})^c)]\\
&\le\sum_{t\in[T]}\sum_{n\in[N]}(\epsilon/C_4)\mathbb{E}\left[Q_n(t)\right]+\mathcal{O}(N^2K^2(\beta_T+\gamma_T)^4/\kappa^2\epsilon^5). \qquad (48)
\end{aligned}
$$

By putting the results of Eqs. (37), (38), (39), (40), (44), (45), (46), (48) altogether, we can obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]} \mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))\right] \\
&\leq 7\min\{N,K\}T - 2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + 2\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(D_n^*(t) - D_n(t))Q_n(t)\right] \\
&\leq 7\min\{N,K\}T - 2\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + 2C_3\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] \\
&\quad + \mathcal{O}(N\log(T)) + (2\epsilon/C_4)\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(N^2K^2(\beta_T+\gamma_T)^4/\epsilon^5) \\
&\leq 7\min\{N,K\}T + 2(C_3 + (1/C_4) - 1)\epsilon \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[Q_n(t)\right] + \mathcal{O}(N^2K^2(\beta_T+\gamma_T)^4/\epsilon^5).
\end{aligned}
$$

Finally, with positive constants $C_3, C_4 > 0$ satisfying $C_3 + (1/C_4) < 1$, from $\mathcal{V}(\mathbf{Q}(1)) = 0$ and $\mathcal{V}(\mathbf{Q}(T+1)) \geq 0$, by using telescoping for the above inequality, we can conclude the proof by

$$
\frac{1}{T}\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)] = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon} + \frac{1}{T}\frac{d^4N^2K^2}{\kappa^4\epsilon^6}\text{polylog}(T)\right). \tag{49}
$$

A.7 Proof of Theorem 4

We first provide the proof for the regret bound of $\widetilde{\mathcal{O}}\left(\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}\right)$. We define event $E_t^1 = \{\|\hat{\theta}_{k,t} - \theta_k\|_{V_{k,t}} \leq \beta_t; \forall k \in [K]\}$ which holds at least probability of $1 - 1/t^2$ from Lemma 1. We let $\gamma_t = \beta_t\sqrt{d\log(Mt)}$. Then we also define $E_t^2 = \{|h_{n,k,t}^{TS} - x_n^\top\hat{\theta}_{k,t}| \leq \gamma_t\|x_n\|_{V_{k,t}^{-1}}; \forall n \in [N], \forall k \in [K]\}$, which holds at least probability of $1 - 1/t^2$ from Lemma 5.

Then we have

$$
\begin{aligned}
\mathcal{R}^\pi(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E_t^1\cap E_t^2)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E_t^1)^c)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E_t^2)^c)].
\end{aligned} \tag{50}
$$

Lemma 8. *Under E_t^2, for all $k \in [K]$, we have*

$$
\left|\sum_{n\in S_{k,t}} (\widetilde{\mu}_t^{TS}(n|S_{k,t}) - \mu(n|S_{k,t},\hat{\theta}_{k,t}))Q_n(t)\right| \leq \gamma_t \max_{n\in S_{k,t}} \|x_n\|_{V_{k,t}^{-1}} Q_n(t). \tag{51}
$$

Proof. Let $u_{n,k,t} = x_n^\top\hat{\theta}_{k,t}$. Under E_t^2, for any $k \in [K]$ we have $|h_{n,k,t}^{TS} - u_{n,k,t}| \leq \gamma_t\|x_n\|_{V_{k,t}^{-1}}$. Then by the mean value theorem, there exists $\bar{u}_{n,k,t} = (1-c)h_{n,k,t}^{TS} + cu_{n,k,t}$ for some $c \in (0,1)$

satisfying,

$$
\begin{aligned}
&\left|\sum_{n\in S_{k,t}}(\widetilde{\mu}_t^{TS}(n|S_k)-\mu(n|S_k,\theta_k))Q_n(t)\right| \\
&=\left|\sum_{n\in S_{k,t}}\left(\frac{\exp(h_{n,k,t}^{TS})}{1+\sum_{m\in S_k}\exp(h_{m,k,t}^{TS})}-\frac{\exp(u_{n,k,t})}{1+\sum_{m\in S_k}\exp(u_{m,k,t})}\right)Q_n(t)\right| \\
&=\left|\sum_{n\in S_{k,t}}\nabla_{v_n}\left(\frac{\sum_{m\in S_{k,t}}\exp(v_m)}{1+\sum_{m\in S_k}\exp(v_m)}\right)\Big|_{v_n=\bar{u}_{n,k,t}}(h_{n,k,t}^{TS}-u_{n,k,t})Q_n(t)\right| \\
&=\left|\frac{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))(\sum_{n\in S_k}\exp(\bar{u}_{n,k,t})(h_{n,k,t}^{TS}-u_{n,k,t})Q_n(t))}{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))^2}\right. \\
&\quad\left.-\frac{(\sum_{n\in S_{k,t}}\exp(\bar{u}_{n,k,t}))(\sum_{n\in S_{k,t}}\exp(\bar{u}_{n,k,t})(h_{n,k,t}^{TS}-u_{n,k,t})Q_n(t))}{(1+\sum_{n\in S_k}\exp(\bar{u}_{n,k,t}))^2}\right| \\
&\le\sum_{n\in S_{k,t}}\frac{\exp(\bar{u}_{n,k,t})}{1+\sum_{m\in S_{k,t}}\exp(\bar{u}_{m,k,t})}|h_{n,k,t}^{TS}-u_{n,k,t}|Q_n(t) \\
&\le\max_{n\in S_{k,t}}|h_{n,k,t}^{TS}-u_{n,k,t}|Q_n(t)\le\gamma_t\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t).
\end{aligned}
$$

□

Lemma 9. *Under E_t^1, for all $k\in[K]$ we have*

$$
\sum_{n\in S_{k,t}}(\mu(n|S_{k,t},\hat{\theta}_{k,t})-\mu(n|S_{k,t},\theta_k))Q_n(t)\le 2\beta_t\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t).
$$

Proof. Since $x/(1+x)$ is a non-decreasing function for $x>-1$ and $x_n^\top\hat{\theta}_{k_{n,t},t}\le x_n^\top\hat{\theta}_{k_{n,t},t}+\beta_t\|x_n\|_{V_{k_{n,t},t}^{-1}}$ under E_t^1, with the definition of $\widetilde{\mu}_t^{UCB}(n|S_{k,t})$, we have

$$
\sum_{n\in S_{k,t}}(\mu(n|S_{k,t},\hat{\theta}_{k,t})-\mu(n|S_{k,t},\theta_k))Q_n(t)\le\sum_{n\in S_{k,t}}(\widetilde{\mu}_t^{UCB}(n|S_{k,t})-\mu(n|S_{k,t},\theta_k))Q_n(t).
$$

Then let $u_{n,k,t}=x_n^\top\theta_k$. Under E_t^1, for any $n\in[N]$ and $k\in[K]$ we have $x_n^\top\hat{\theta}_{k,t}-\beta_t\|x_n\|_{V_{k,t}^{-1}}\le x_n^\top\theta_k\le x_n^\top\hat{\theta}_{k,t}+\beta_t\|x_n\|_{V_{k,t}^{-1}}$, which implies $0\le h_{n,k,t}^{UCB}-u_{n,k,t}\le 2\beta_t\|x_n\|_{V_{k,t}^{-1}}$. Then by the

mean value theorem, there exists $\bar{u}_{n,k,t} = (1-c)h^{UCB}_{n,k,t} + cu_{n,k,t}$ for some $c \in (0,1)$ satisfying,

$$
\begin{aligned}
&\sum_{n \in S_{k,t}} (\widetilde{\mu}^{UCB}_t(n|S_{k,t}, \hat{\theta}_{k,t}) - \mu(n|S_{k,t}, \theta_k))Q_n(t) \\
&= \sum_{n \in S_k} \left(\frac{\exp(h^{UCB}_{n,k,t})}{1 + \sum_{m \in S_{k,t}} \exp(h^{UCB}_{m,k,t})} - \frac{\exp(u_{n,k,t})}{1 + \sum_{m \in S_{k,t}} \exp(u_{m,k,t})} \right) Q_n(t) \\
&= \sum_{n \in S_{k,t}} \nabla_{v_n} \left(\frac{\exp(v_n)}{1 + \sum_{m \in S_{k,t}} \exp(v_m)} \right) \Big|_{v_n = \bar{u}_{n,k,t}} (h^{UCB}_{n,k,t} - u_{n,k,t})Q_n(t) \\
&= \frac{(1 + \sum_{n \in S_k} \exp(\bar{u}_{n,k,t}))(\sum_{n \in S_k} \exp(\bar{u}_{n,k,t})(h^{UCB}_{n,k,t} - u_{n,k,t})Q_n(t))}{(1 + \sum_{n \in S_k} \exp(\bar{u}_{n,k,t}))^2} \\
&\quad - \frac{(\sum_{n \in S_{k,t}} \exp(\bar{u}_{n,k,t}))(\sum_{n \in S_{k,t}} \exp(\bar{u}_{n,k,t})(h^{UCB}_{n,k,t} - u_{n,k,t})Q_n(t))}{(1 + \sum_{n \in S_{k,t}} \exp(\bar{u}_{n,k,t}))^2} \\
&\le \sum_{n \in S_{k,t}} \frac{\exp(\bar{u}_{n,k,t})}{1 + \sum_{m \in S_{k,t}} \exp(\bar{u}_{m,k,t})} (h^{UCB}_{n,k,t} - u_{n,k,t})Q_n(t) \\
&\le \max_{n \in S_k} (h^{UCB}_{n,k,t} - u_{n,k,t})Q_n(t) \\
&\le 2\beta_t \max_{n \in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t).
\end{aligned}
$$

□

Then we first focus on the first term in the above. From Lemmas 8 and 9, we have

$$
\begin{aligned}
&\sum_{t \in [T]} \sum_{n \in [N]} \mathbb{E}[\widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t)] \\
&\le \sum_{t \in [T]} \sum_{n \in [N]} \mathbb{E}[(\widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) \\
&\qquad\qquad + \mu(n|S_{k_{n,t},t}, \hat{\theta}_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t)] \\
&\le \sum_{t \in [T]} \mathbb{E}[\sum_{k \in [K]} \sum_{n \in S_{k,t}} (\widetilde{\mu}^{TS}_t(n|S_{k,t}) - \mu(n|S_{k,t}, \hat{\theta}_{k,t}) \\
&\qquad\qquad + \mu(n|S_{k,t}, \hat{\theta}_{k,t}) - \mu(n|S_{k,t}, \theta_k))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t)] \\
&\le \sum_{t \in [T]} \sum_{k \in [K]} \mathbb{E}[(\gamma_t + 2\beta_t) \max_{n \in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t)\mathbb{1}(E^1_t)] \\
&= \sum_{t \in [T]} \sum_{k \in [K]} \mathbb{E}[\mathbb{E}[(\gamma_t + 2\beta_t) \max_{n \in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t)|E^1_t, \mathcal{F}_{t-1}]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})]. \qquad (52)
\end{aligned}
$$

We define sets

$$
\begin{aligned}
\widetilde{\Theta}_t = \Bigg\{ \{\theta^{(i)}_k\}_{i \in [M], k \in [K]} : \Bigg| \sum_{n \in S_{k,t}} (\widetilde{\mu}^{TS}_t(n|S_{k,t}, \{\theta^{(i)}_k\}_{i \in [M]}) - \mu(n|S_{k,t}, \hat{\theta}_{k,t}))Q_n(t) \Bigg| \\
\le \gamma_t \max_{n \in S_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t);\ \forall k \in [K] \Bigg\}
\end{aligned}
$$

and

$$
\begin{aligned}
\widetilde{\Theta}^{opt}_t = \Bigg\{ \{\theta^{(i)}_k\}_{i \in [M], k \in [K]} : \sum_{k \in [K]} \sum_{n \in S_{k,t}} \widetilde{\mu}^{TS}_t(n|S_{k,t}, \{\theta^{(i)}_k\}_{i \in [M]})Q_n(t) \\
> \sum_{k \in [K]} \sum_{n \in S^*_{k,t}} \mu(n|S^*_{k,t}, \theta_k)Q_n(t) \Bigg\} \cap \widetilde{\Theta}_t.
\end{aligned}
$$

We define event $E_t(\widetilde{\theta}) = \{\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M],k\in[K]} \in \widetilde{\Theta}_t^{opt}\}$. Then by following the proof steps in Eq.(41), we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t) - \sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})Q_n(t)\right)\mathbb{1}(E_t^1\cap E_t^2)|\mathcal{F}_{t-1}\right]\right] \\
&\le \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\mu(n|S_{k,t}^*,\theta_k)Q_n(t) \\
&\qquad - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})Q_n(t)\Bigg)\mathbb{1}(E_t^1\cap E_t^2)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\Bigg]\Bigg] \\
&\le \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sup_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\Bigg(\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}) \\
&\qquad\qquad - \widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})\Bigg)Q_n(t)\mathbb{1}(E_t^1\cap E_t^2)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\Bigg]\Bigg] \\
&\le \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sup_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}}\Bigg(\widetilde{\mu}_t^{TS}(n|S_{k,t},\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}) - \mu(n|S_{k,t},\hat{\theta}_{k,t}) \\
&\qquad\qquad + \mu(n|S_{k,t},\hat{\theta}_{k,t}) - \widetilde{\mu}_t^{TS}(n|S_{k,t},\{\theta_k^{(i)}\}_{i\in[M]})\Bigg)Q_n(t)\mathbb{1}(E_t^1\cap E_t^2)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\Bigg]\Bigg] \\
&\le 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\mathbb{1}(E_t^1)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right] \\
&= 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\times\mathbb{P}(E_t^1|E_t(\widetilde{\theta}),\mathcal{F}_{t-1})\right] \\
&= 2\gamma_t\mathbb{E}\left[\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\right] \\
&\le 32\sqrt{e\pi}\gamma_t\mathbb{E}\left[\mathbb{E}\left[\sum_{k\in[K]}\max_{n\in S_{k,t}}\|x_n\|_{V_{k,t}^{-1}}Q_n(t)|\mathcal{F}_{t-1},E_t^1\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\right],
\end{aligned}
\tag{53}
$$

where the last inequality is obtained from (42). Then from Eqs.(52) and (53), for some constant $C_2 > 0$, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*}) - \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) + \widetilde{\mu}_t^{TS}(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E_t^1\cap E_t^2)] \\
&\le \sum_{t\in[T]}\sum_{k\in[K]}\mathbb{E}[\mathbb{E}[\max_{n\in S_{k,t}}(33\sqrt{e\pi}\gamma_t + 2\beta_t)\|x_n\|_{V_{k,t}^{-1}}Q_n(t)|E_t^1,\mathcal{F}_{t-1}]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})] \\
&\le C_2\sum_{t\in[T]}\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S_{k,t}}(\gamma_t+\beta_t)\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\mathbb{1}(E_t^1)\right] \\
&\le C_2\sum_{t\in[T]}\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S_{k,t}}(\gamma_t+\beta_t)\|x_n\|_{V_{k,t}^{-1}}Q_n(t)\right].
\end{aligned}
\tag{54}
$$

For the second and third terms of Eq.(50), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^1_t)^c)\right]\\
&\le \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[Q_n(t)\mathbb{1}((E^1_t)^c)\right]\\
&\le \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^1_t)^c)\\
&= \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)),
\end{aligned}
\tag{55}
$$

and

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_t)^c)\right]\\
&\le \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[Q_n(t)\mathbb{1}((E^2_t)^c)\right]\\
&\le \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^2_t)^c)\\
&= \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} t(1/t^2)\right) = \mathcal{O}(N\log(T)).
\end{aligned}
\tag{56}
$$

Finally from Eqs.(50), (54), (55), and (56), we can conclude that

$$
\begin{aligned}
\mathcal{R}^\pi(T) &= \mathcal{O}\left(\sum_{t\in[T]}\sum_{k\in[K]}\mathbb{E}[\max_{n\in S_{k,t}}(\gamma_t+\beta_t)\|x_n\|_{V^{-1}_{k,t}}Q_n(t)] + N\log(T)\right)\\
&= \mathcal{O}\left(\gamma_T\mathbb{E}[\max_{t\in[T],n\in[N]}Q_n(t)\sum_{t\in[T]}\sum_{k\in[K]}\max_{n\in S_{k,t}}\|x_n\|_{V^{-1}_{k,t}}] + N\log(T)\right)\\
&= \mathcal{O}\left(\gamma_T\mathbb{E}[\max_{t\in[T],n\in[N]}Q_n(t)\sqrt{KT\sum_{t\in[T]}\sum_{k\in[K]}\max_{n\in S_{k,t}}\|x_n\|^2_{V^{-1}_{k,t}}}] + N\log(T)\right)\\
&= \widetilde{\mathcal{O}}\left(\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}\right),
\end{aligned}
$$

where the last equality is obtained from Lemma 4.

Here, we provide the proof for the worst-case regret bound of $\widetilde{\mathcal{O}}\left(\left(\frac{d^2NK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right)$.

We define event $E^1_t = \{\|\hat{\theta}_{k,t}-\theta_k\|_{V_{k,t}} \le \beta_t$ for all $k\in[K]\}$ where $\beta_t = C_1\sqrt{\lambda+\frac{d}{\kappa}\log(1+tLK/d\lambda)}$, which holds with high probability as $\mathbb{P}(E^1_t)\ge 1-1/t^2$ from Lemma 1. We also define $E^2_t = \{\widetilde{\mu}^{TS}_t(n|S_{k,t})-\mu(n|S_{k,t},\hat{\theta}_{k,t})\le\gamma_t\|x_n\|_{V^{-1}_{k,t}};\forall n\in[N],\forall k\in[K]\}$, which holds with high probability as $\mathbb{P}(E^2_t)\ge 1-\mathcal{O}(1/t^2)$, and define $E^3_{n,t} = \{\|x_n\|_{V^{-1}_{k,t}}\le\zeta;\forall k\in[K]\}$.

Then we have

$$
\begin{aligned}
&\mathcal{R}^{\pi}(T) \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t\cap E^2_t\cap E^3_{n,t})] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E^1_t)^c) + \mathbb{1}((E^2_t)^c) + \mathbb{1}((E^3_{n,t})^c))].
\end{aligned} \tag{57}
$$

We provide a bound for the first term of Eq.(57). We first have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t\cap E^2_t\cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) \\
&\qquad\qquad + \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t\cap E^2_t\cap E^3_{n,t})].
\end{aligned} \tag{58}
$$

By following the steps in Eq.(40), the last two terms in Eq.(58) are bounded as

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[\widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t\cap E^2_t\cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\mathbb{E}\left[\sum_{n\in S_{k,t}} (\gamma_t + 2\beta_t)\zeta Q_n(t)|E^1_t,\mathcal{F}_{t-1}\right]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})\right].
\end{aligned} \tag{59}
$$

Now we provide a bound for the first two terms in Eq.(58).

We define sets

$$
\widetilde{\Theta}_t = \left\{\{\theta^{(i)}_k\}_{i\in[M],k\in[K]} : \left|\max_{i\in[M]} x_n^\top\theta^{(i)}_k - x_n^\top\hat{\theta}_{k,t}\right| \leq \gamma_t\|x_n\|_{V^{-1}_{k,t}};\ \forall n\in[N],\forall k\in[K]\right\} \text{ and}
$$

$$
\widetilde{\Theta}^{opt}_t = \left\{\{\theta^{(i)}_k\}_{i\in[M],k\in[K]} : \sum_{k\in[K]}\sum_{n\in S_{k,t}} \widetilde{\mu}^{TS}_t(n|S_{k,t},\{\theta^{(i)}_k\}_{i\in[M]})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^*_{k,t}} \mu(n|S^*_{k,t},\theta_k)Q_n(t)\right\}\cap\widetilde{\Theta}_t.
$$

By following the steps in Eq.(41) with Eq. (42), we have

$$
\begin{aligned}
&\mathbb{E}[\mathbb{E}[(\sum_{k\in[K]}\sum_{n\in S^*_{k,t}} \mu(n|S^*_{k,t},\theta_k) - \sum_{k\in[K]}\sum_{n\in S_{k,t}} \widetilde{\mu}^{TS}_t(n|S_{k,t},\{\widetilde{\theta}^{(i)}_{k,t}\}_{i\in[M]}))Q_n(t)\mathbb{1}(E^1_t\cap E^2_t\cap E^3_{n,t})|\mathcal{F}_{t-1}]] \\
&= 2\gamma_t\zeta\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}} \mathbb{E}\left[Q_n(t)\middle|\mathcal{F}_{t-1},\{\widetilde{\theta}^{(i)}_{k,t}\}_{i\in[M],k\in[K]}\in\widetilde{\Theta}^{opt}_t, E^1_t\right]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})\right] \\
&\leq 16\sqrt{e\pi}\gamma_t\zeta\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}} Q_n(t)|\mathcal{F}_{t-1},E^1_t\right]\mathbb{P}(E^1_t|\mathcal{F}_{t-1}).
\end{aligned} \tag{60}
$$

Then for the first term of Eq.(57), from Eqs.(58), (59), for some $C_3 > 0$ we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) \\
&\qquad\qquad + \widetilde{\mu}^{TS}_t(n|S_{k_{n,t},t}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\leq \sum_{t\in[T]}\sum_{k\in[K]} \mathbb{E}\left[\mathbb{E}\left[\sum_{n\in S_{k,t}} (17\sqrt{e\pi}\gamma_t + 2\beta_t)\zeta Q_n(t)|\mathcal{F}_{t-1}, E^1_t\right]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})\right] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} (17\sqrt{e\pi}\gamma_t + 2\beta_t)\zeta\mathbb{E}[\mathbb{E}[Q_n(t)|E^1_t, \mathcal{F}_{t-1}]\mathbb{P}(E^1_t|\mathcal{F}_{t-1})] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} (17\sqrt{e\pi}\gamma_t + 2\beta_t)\zeta\mathbb{E}[Q_n(t)\mathbb{1}(E^1_t)] \\
&\leq (17\sqrt{e\pi}\gamma_T + 2\beta_T)\zeta \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)]. \qquad (61)
\end{aligned}
$$

For the second term of Eq.(57), by following the steps in Eqs.(45), (46) we have

$$
\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^1_t)^c)\right] = \mathcal{O}(N\log(T)), \quad (62)
$$

and

$$
\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}\left[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_t)^c)\right] = \mathcal{O}(N\log(T)). \quad (63)
$$

Let $\mathcal{T}_n$ be the set of time steps $t \in [T]$ such that $Q_n(t) \neq 0$ and let $e_T = \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}((E^3_{n,t})^c)$ and $h = \lceil 1/(17\sqrt{e\pi}\gamma_T + 2\beta_T)\zeta\rceil$. Then if $t \leq h$, we have $Q_n(t) \leq t \leq h$. Otherwise, we have

$$
Q_n(t) \leq \sum_{s=t-h+1}^{t} (1/h)(Q_n(s) + (t-s)) \leq (1/h)\sum_{s=1}^{t} Q_n(s) + (1/h)h^2 = (1/h)\sum_{s=1}^{t} Q_n(s) + h.
$$

Then, by following the steps in Eq.(47), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E^2_{n,t})^c)] \\
&\leq \sum_{t\in[T]}\sum_{n\in[N]} (17\sqrt{e\pi}\gamma_T + 2\beta_T)\zeta\mathbb{E}\left[Q_n(t)\right] + 2\mathbb{E}[e_T]/(17\sqrt{e\pi}\gamma_T + 2\beta_T)\zeta. \qquad (64)
\end{aligned}
$$

Now we provide a bound for $\mathbb{E}[e_T]$. Define $N_{n,k}(t) = \sum_{s=1}^{t-1} \mathbb{1}(n \in S_{k,s})$ and $\widetilde{V}_{k,t} = (\kappa/2)\sum_{s=1}^{t-1}\sum_{n\in S_{k,s}} x_n x_n^\top$. Then, we have

$$
\begin{aligned}
e_T &= \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}((E_{n,t}^2)^c) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(\|x_n\|_{V_{k_{n,t},t}^{-1}} \ge \zeta) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(\|x_n\|_{\widetilde{V}_{k_{n,t},t}^{-1}} \ge \zeta) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(1/N_{n,k_{n,t}}(t) \ge (\kappa/2)\zeta^2) \\
&\le \sum_{n\in[N]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(N_{n,k_{n,t}}(t) \le 2/\kappa\zeta^2) \\
&\le \sum_{n\in[N]}\sum_{k\in[K]}\sum_{t\in\mathcal{T}_n} \mathbb{1}(N_{n,k}(t) \le 2/\kappa\zeta^2 \text{ and } k_{n,t} = k) \\
&\le 2NK/\kappa\zeta^2. \qquad (65)
\end{aligned}
$$

Then from Eqs.(64), (65), we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[(\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}((E_{n,t}^2)^c)] \\
&\le \mathcal{O}\left(\sum_{t\in[T]}\sum_{n\in[N]} \zeta(\beta_T+\gamma_T)\mathbb{E}\left[Q_n(t)\right] + \frac{NK}{\kappa\zeta^3(\beta_T+\gamma_T)}\right). \qquad (66)
\end{aligned}
$$

By putting the results of Eqs. (57), (58), (59), (61), (62), (63), (66), and Theorem 3, by setting $\zeta = (\epsilon NK/\min\{N,K\}\kappa T(\beta_T+\gamma_T)^2)^{1/4}$, for large enough T, we can obtain

$$
\begin{aligned}
\mathcal{R}^\pi(T) &= \mathcal{O}\left(\zeta(\beta_T+\gamma_T)\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)] + \frac{NK}{\kappa\zeta^3(\beta_T+\gamma_T)} + N\log(T)\right) \\
&= \mathcal{O}\left(\frac{\zeta(\beta_T+\gamma_T)\min\{N,K\}T}{\epsilon} + \frac{NK}{\kappa\zeta^3(\beta_T+\gamma_T)} + N\log(T)\right) \\
&= \mathcal{O}\left(\frac{(\beta_T+\gamma_T)^{1/2}T^{3/4}(NK\min\{N,K\}^3)^{1/4}}{\kappa^{1/4}\epsilon^{3/4}} + N\log(T)\right) \\
&= \widetilde{\mathcal{O}}\left(\left(\frac{d^2NK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right).
\end{aligned}
$$

A.8 Proof of Lemma 1

We first define

$$\bar{f}_{k,t}(\theta) = \mathbb{E}_y[f_{k,t}(\theta)|\mathcal{F}_{t-1}] \text{ and } \bar{g}_{k,t}(\theta) = \mathbb{E}_y[g_{k,t}(\theta)|\mathcal{F}_{t-1}],$$

where $\mathcal{F}_{t-1}$ is the filtration contains outcomes for time s such that $s \le t-1$ and $y = \{y_{n,t} : n \in S_{k,t}\}$. From Lemma 10 in Oh and Iyengar [44], by taking expectation over y gives

$$\bar{f}_{k,t}(\hat{\theta}_{k,t}) \le \bar{f}_{k,t}(\theta_k) + \bar{g}_{k,t}(\hat{\theta}_{k,t})^\top(\hat{\theta}_{k,t}-\theta_k) - \frac{\kappa}{2}(\theta_k-\hat{\theta}_{k,t})^\top W_{k,t}(\theta_k-\hat{\theta}_{k,t}),$$

where $W_{k,t} = \sum_{n \in S_{k,t}} x_n x_n^\top$. Then with $\bar{f}_{k,t}(\theta_k) \leq \bar{f}_{k,t}(\hat{\theta}_{k,t})$ from Lemma 12 in Oh and Iyengar [44], we have

$$
\begin{aligned}
0 &\leq \bar{f}_{k,t}(\hat{\theta}_{k,t}) - \bar{f}_{k,t}(\theta_k) \\
&\leq \bar{g}_{k,t}(\hat{\theta}_{k,t})^\top (\hat{\theta}_{k,t} - \theta_k) - \frac{\kappa}{2} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,t}} \\
&= g_{k,t}(\hat{\theta}_{k,t})^\top (\hat{\theta}_{k,t} - \theta_k) - \frac{\kappa}{2} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,t}} + (\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k) \\
&\leq \frac{1}{2} \|g_{k,t}(\hat{\theta}_{k,t})\|^2_{V^{-1}_{k,t+1}} + \frac{1}{2} \|\hat{\theta}_{k,t} - \theta_k\|^2_{V_{k,t+1}} - \frac{1}{2} \|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} \\
&\quad - \frac{\kappa}{2} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,t}} + (\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k) \\
&\leq 2 \max_{n \in S_{k,t}} \|x_n\|^2_{V^{-1}_{k,t+1}} + \frac{1}{2} \|\hat{\theta}_{k,t} - \theta_k\|^2_{V_{k,t+1}} - \frac{1}{2} \|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} \\
&\quad - \frac{\kappa}{2} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,t}} + (\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k) \\
&\leq 2 \max_{n \in S_{k,t}} \|x_n\|^2_{V^{-1}_{k,t+1}} + \frac{1}{2} \|\hat{\theta}_{k,t} - \theta_k\|^2_{V_{k,t}} + \frac{\kappa}{4} \|\hat{\theta}_{k,t} - \theta_k\|^2_{W_{k,t}} - \frac{1}{2} \|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} \\
&\quad - \frac{\kappa}{2} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,t}} + (\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k) \\
&\leq 2 \max_{n \in S_{k,t}} \|x_n\|^2_{V^{-1}_{k,t+1}} + \frac{1}{2} \|\hat{\theta}_{k,t} - \theta_k\|^2_{V_{k,t}} - \frac{\kappa}{4} \|\hat{\theta}_{k,t} - \theta_k\|^2_{W_{k,t}} - \frac{1}{2} \|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} \\
&\quad + (\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k), \qquad (67)
\end{aligned}
$$

where the third inequality comes from Lemma 11 in Oh and Iyengar [44] and the fourth inequality comes from Lemma 13 in Oh and Iyengar [44]. We note that, although our estimator lies in Θ, we can still utilize Lemma 11 from Oh and Iyengar [44] by following the same proof steps.

Hence, from the above, we have

$$
\|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} \leq 4 \max_{n \in S_{k,t}} \|x_n\|^2_{V^{-1}_{k,t+1}} + \|\hat{\theta}_{k,t} - \theta_k\|^2_{V_{k,t}} - \frac{\kappa}{2} \|\hat{\theta}_{k,t} - \theta_k\|^2_{W_{k,t}} + 2(\bar{g}_{k,t}(\hat{\theta}_{k,t}) - g_{k,t}(\hat{\theta}_{k,t}))^\top (\hat{\theta}_{k,t} - \theta_k).
$$

Then using telescoping by summing the above over t, with at least probability $1 - \delta$, we have

$$
\begin{aligned}
\|\hat{\theta}_{k,t+1} - \theta_k\|^2_{V_{k,t+1}} &\leq \lambda + 4 \sum_{s=1}^{t} \max_{n \in S_{k,s}} \|x_n\|^2_{V^{-1}_{k,s+1}} - \frac{\kappa}{2} \sum_{s=1}^{t} \|\hat{\theta}_{k,s} - \theta_k\|^2_{W_{k,s}} \\
&\quad + 2 \sum_{s=1}^{t} (\bar{g}_{k,s}(\hat{\theta}_{k,s}) - g_{k,s}(\hat{\theta}_{k,s}))^\top (\hat{\theta}_{k,s} - \theta_k) \\
&\leq \lambda + 4 \sum_{s=1}^{t} \max_{n \in S_{k,s}} \|x_n\|^2_{V^{-1}_{k,s+1}} - \frac{\kappa}{2} \sum_{s=1}^{t} \|\hat{\theta}_{k,s} - \theta_k\|^2_{W_{k,s}} \\
&\quad + \frac{\kappa}{2} \sum_{s=1}^{t} \|\theta_k - \hat{\theta}_{k,t}\|^2_{W_{k,s}} + \frac{2C_1}{\kappa} \log((\log tL) t^2 K/\delta) \\
&\leq \lambda + 4 \sum_{s=1}^{t} \max_{n \in S_{k,s}} \|x_n\|^2_{V^{-1}_{k,s+1}} + \frac{2C_1}{\kappa} \log((\log tL) t^2 K/\delta) \\
&\leq \lambda + 16(d/\kappa) \log(1 + (tL/d\lambda)) + \frac{2C_1}{\kappa} \log((\log tL) t^2 K/\delta), \qquad (68)
\end{aligned}
$$

where the second inequality is obtained from Lemma 14 in Oh and Iyengar [44] and the last one is obtained from Lemma 4

Then with $\delta = 1/t^2$, we can conclude that

$$
\|\hat{\theta}_{k,t} - \theta_k\|_{V_{k,t}} \leq C_1 \sqrt{\lambda + \frac{d}{\kappa} \log(1 + tLK/d\lambda)}.
$$

A.9 α-approximation Oracle

In this section, we provide a detailed explanation for α-approxiamtion oracle to reduce the computation. Instead of obtaining the exact solution, the α-approximation oracle, denoted by $\mathbb{O}^\alpha$, outputs $\{S_k^\alpha\}_{k\in[K]}$ satisfying $\sum_{k\in[K]} f_k(S_k^\alpha) \geq \max_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)} \sum_{k\in[K]} \alpha f_k(S_k)$. Such an oracle can be constructed using a straightforward greedy policy as outlined in prior work [26, 10]. Then for assortments $\{S_{k,t}^\alpha\}_{k\in[K]}$ for $t\in[T]$ from Algorithms 1 and 2 using $\mathbb{O}^\alpha$, we can obtain the same queue length bounds and regret bounds for α-regret in Theorems 1, 2, 3, and 4 under an α-slackness assumption, respectively.

We first consider the following traffic slackness assumption with $0<\alpha<1$ instead of Assumption 3.

Assumption 4. *For some traffic slackness $0<\epsilon<1$, for each $t\in[T]$, there exists $\{S_{k,t}\}_{k\in[K]}\in\mathcal{M}(\mathcal{N})$ which satisfies $\lambda_n+\epsilon\leq\alpha\mu(n|S_{k,t},\theta_k)$ for all $n\in S_{k,t}$ and $k\in[K]$.*

A.9.1 α-approximation Oracle for Algorithm 1

Here we introduce an algorithm (Algorithm 3) by modifying Algorithm 1 using an α-approximation oracle. We explain the distinct parts of the algorithm as follows. We define an oracle $\mathbb{O}^\alpha$, which outputs $\{S_{k,t}^\alpha\}_{k\in[K]}$ satisfying

$$\max_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)} \sum_{k\in[K]}\sum_{n\in S_k} \alpha Q_n(t)\widetilde{\mu}_t^{UCB}(n|S_k,\hat{\theta}_{k,t}) \leq \sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha} Q_n(t)\widetilde{\mu}_t^{UCB}(n|S_{k,t}^\alpha,\hat{\theta}_{k,t}). \tag{69}$$

Algorithm 3 α-approximated UCB-Queueing Matching Bandit

Input: λ, κ, $C_1>0$
for $t=1,\dots,T$ **do**
  for $k\in[K]$ **do**
    $\hat{\theta}_{k,t}\leftarrow \mathrm{argmin}_{\theta\in\Theta}\, g_{k,t}(\hat{\theta}_{k,t-1})^\top(\theta-\hat{\theta}_{k,t-1})+\frac{1}{2}\|\theta-\hat{\theta}_{k,t-1}\|_{V_{k,t}}^2$
  $\{S_{k,t}^\alpha\}_{k\in[K]}\leftarrow\mathbb{O}^\alpha$ from (69)
  Offer $\{S_{k,t}^\alpha\}_{k\in[K]}$ and observe preference feedback $y_{n,t}\in\{0,1\}$ for all $n\in S_{k,t}$, $k\in[K]$

For the stability analysis, we provide the following theorem.

Theorem 5. *The time average expected queue length of Algorithm 3 is bounded as*

$$\mathcal{Q}(T)=\mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon}+\frac{1}{T}\frac{d^2N^2K^2}{\kappa^4\epsilon^6}\mathrm{polylog}(T)\right),$$

which implies that the algorithm achieves stability as

$$\lim_{T\to\infty}\mathcal{Q}(T)=\mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon}\right).$$

Proof. Here we provide only the proof parts which are different from Theorem 1. We analyze the Lyapunov drift as follows.

$$\begin{aligned}
&\sum_{t\in[T]}\mathcal{V}(\mathbf{Q}(t+1))-\mathcal{V}(\mathbf{Q}(t))\\
&=\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-D_n(t))^{+2}-Q_n(t)^2\\
&=\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-\alpha D_n^*(t))^2-Q_n(t)^2\\
&\quad+\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-D_n(t))^{+2}-\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-\alpha D_n^*(t))^2.
\end{aligned}\tag{70}$$

For the first two terms in Eq.(70), by following the same procedure of Eqs.(16) and (17), under Assumption 4, we can obtain

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(Q_n(t)+A_n(t)-\alpha D_n^*(t))^2-Q_n(t)^2]\\
&\leq \sum_{t\in[T]}\sum_{n\in[N]}2\mathbb{E}[(\lambda_n-\alpha\mu(n|S_{k,t}^*,\theta_k))Q_n(t)]+2NT\\
&\leq -\sum_{t\in[T]}\sum_{n\in[N]}2\epsilon\mathbb{E}[Q_n(t)]+2\min\{N,K\}T,
\end{aligned}\tag{71}$$

For the last two terms in Eq.(18), by following the steps for Eq.(20), we have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-D_n(t))^{+2}-\sum_{t\in[T]}\sum_{n\in[N]}(Q_n(t)+A_n(t)-\alpha D_n^*(t))^2\\
&\leq 4\sum_{t\in[T]}\sum_{n\in[N]}A_n(t)+2\sum_{t\in[T]}\sum_{n\in[N]}(\alpha D_n^*(t)-D_n(t))Q_n(t)+\min\{N,K\}T\\
&\leq 2\sum_{t\in[T]}\sum_{n\in[N]}(\alpha D_n^*(t)-D_n(t))Q_n(t)+5\min\{N,K\}T.
\end{aligned}\tag{72}$$

We also have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha D_n^*(t)-D_n(t)Q_n(t)]\\
&\leq \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*})-\mu(n|S_{k_{n,t},t}^{\alpha},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1\cap E_{n,t}^2)]\\
&\quad+\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*})-\mu(n|S_{k_{n,t},t}^{\alpha},\theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E_t^1)^c)+\mathbb{1}((E_{n,t}^2)^c))].
\end{aligned}\tag{73}$$

Then for the first term of Eq.(73), we have

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}\left[\sum_{n\in[N]}(\alpha\mu(n|S_{k_{n,t}^*,t},\theta_{k_{n,t}^*})-\mu(n|S_{k_{n,t},t}^{\alpha},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1\cap E_{n,t}^2)\right]\\
&\leq \sum_{t\in[T]}\mathbb{E}\left[\sum_{n\in[N]}(\alpha\widetilde{\mu}_t^{UCB}(n|S_{k_{n,t}^*},\hat{\theta}_{k_{n,t}^*,t})-\mu(n|S_{k_{n,t},t}^{\alpha},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1\cap E_{n,t}^2)\right]\\
&\leq \sum_{t\in[T]}\mathbb{E}\left[\sum_{n\in[N]}(\widetilde{\mu}_t^{UCB}(n|S_{k_{n,t},t}^{\alpha},\hat{\theta}_{k_{n,t},t})-\mu(n|S_{k_{n,t},t}^{\alpha},\theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E_t^1\cap E_{n,t}^2)\right]\\
&\leq \sum_{t\in[T]}\mathbb{E}\left[\sum_{n\in[N]}2\beta_t\|x_n\|_{V_{k_{n,t},t}^{-1}}Q_n(t)\mathbb{1}(E_t^1\cap E_{n,t}^2)\right]\\
&\leq C_2\epsilon\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[Q_n(t)\right].
\end{aligned}\tag{74}$$

The rest of the proofs can be easily obtained from the proof steps in Theorem 1. □

Now, we investigate the regret of Algorithm 3. The α-regret regarding policy π is defined as

$$\mathcal{R}^{\alpha,\pi}(T)=\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(\alpha\mu(n|S_{k_{n,t}^*,t}^*,\theta_{k_{n,t}^*})-\mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\right].$$

The algorithm achieves the following regret bound.

Theorem 6. *The policy π of Algorithm 3 achieves a regret bound of*

$$\mathcal{R}^{\alpha,\pi}(T) = \widetilde{\mathcal{O}}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\max}, \left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right\}\right).$$

Proof. In this proof, we provide only the parts that are different from the proof of Theorem 2.

We first provide the proof for regret bound of $\mathcal{R}^{\alpha,\pi}(T) = \widetilde{\mathcal{O}}\left(\frac{d}{\kappa}\sqrt{KT}Q_{\max}\right)$.

We can show that

$$\begin{aligned}
\mathcal{R}^{\alpha,\pi}(T) &= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E^1_t)^c)] \\
&= \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \sum_{t\in[T]}\sum_{n\in[N]} t\mathbb{P}((E^1_t)^c)] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)(\alpha\widetilde{\mu}^{UCB}_t(n|S_{k^*_{n,t},t},\hat{\theta}_{k^*_{n,t},t}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&\le \sum_{t\in[T]}\sum_{n\in[N]} \mathbb{E}[Q_n(t)((\widetilde{\mu}^{UCB}_t(n|S^\alpha_{k_{n,t},t},\hat{\theta}_{k_{n,t},t}) - \mu(n|S^\alpha_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&\le \sum_{t\in[T]} \mathbb{E}[\sum_{k\in[K]}\sum_{n\in S_{k,t}} Q_n(t)(\widetilde{\mu}^{UCB}_t(n|S^\alpha_{k,t}) - \mu_t(n|S^\alpha_{k,t}))\mathbb{1}(E^1_t)] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\beta_T \sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S^\alpha_{k,t}} \|x_n\|_{V^{-1}_{k,t}} Q_n(t)\right] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\max_{t\in[T],n\in[N]} Q_n(t)\beta_T \sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S^\alpha_{k,t}} \|x_n\|_{V^{-1}_{k,t}}\right] + \mathcal{O}(N/T) \\
&\le 2\mathbb{E}\left[\max_{t\in[T],n\in[N]} Q_n(t)\beta_T \sqrt{KT\sum_{t\in[T]}\sum_{k\in[K]} \max_{n\in S^\alpha_{k,t}} \|x_n\|^2_{V^{-1}_{k,t}}}\right] + \mathcal{O}(N/T) \\
&= \widetilde{\mathcal{O}}\left(\frac{d}{\kappa}\sqrt{KT}Q_{\max}\right).
\end{aligned}$$

By following the similar steps above and the proofs for Theorem 2, we can easily obtain the worst-case regret bound of $\mathcal{R}^{\alpha,\pi}(T) = \widetilde{\mathcal{O}}\left(\left(\frac{dNK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right)$, which conclude the proof. □

A.9.2 α-approximation Oracle for Algorithm 2

We can obtain similar results for Algorithm 2 as in the case of Algorithm 3. Here we introduce an algorithm (Algorithm 4) by modifying Algorithm 2 using an α-approximation oracle. We explain the distinct parts of the algorithm as follows. We define an oracle $\mathbb{O}^\alpha$, which outputs $\{S^\alpha_{k,t}\}_{k\in[K]}$ satisfying

$$\max_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)} \sum_{k\in[K]}\sum_{n\in S_k} \alpha Q_n(t)\widetilde{\mu}^{TS}_t(n|S_k,\{\widetilde{\theta}^{(i)}_{k,t}\}_{i\in[M]}) \le \sum_{k\in[K]}\sum_{n\in S^\alpha_{k,t}} Q_n(t)\widetilde{\mu}^{TS}_t(n|S^\alpha_{k,t},\{\widetilde{\theta}^{(i)}_{k,t}\}_{i\in[M]}). \tag{75}$$

Algorithm 4 α-approximated Thompson Sampling-Queueing Matching Bandit

Input: λ, M, κ, $C_1 > 0$
for $t = 1, \ldots, T$ **do**
  for $k \in [K]$ **do**
    $\hat{\theta}_{k,t} \leftarrow \mathrm{argmin}_{\theta \in \Theta}\, g_{k,t}(\hat{\theta}_{k,t-1})^\top(\theta - \hat{\theta}_{k,t-1}) + \frac{1}{2}\|\theta - \hat{\theta}_{k,t-1}\|^2_{V_{k,t}}$
    Sample $\{\widetilde{\theta}^{(i)}_{k,t}\}_{i\in[M]}$ independently from $\mathcal{N}(\hat{\theta}_{k,t}, \beta_t^2 V_{k,t}^{-1})$
  $\{S^\alpha_{k,t}\}_{k\in[K]} \leftarrow \mathbb{O}^\alpha$ from (75)
  Offer $\{S^\alpha_{k,t}\}_{k\in[K]}$ and observe preference feedback $y_{n,t} \in \{0,1\}$ for all $n \in S_{k,t}$, $k \in [K]$

For the stability analysis, we provide the following theorem.

Theorem 7. *The time average expected queue length of Algorithm 4 is bounded as*

$$\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon} + \frac{1}{T}\frac{d^4N^2K^2}{\kappa^4\epsilon^6}\mathrm{polylog}(T)\right),$$

which implies that the algorithm achieves stability as

$$\lim_{T\to\infty}\mathcal{Q}(T) = \mathcal{O}\left(\frac{\min\{N,K\}}{\epsilon}\right).$$

Proof. Here we provide only the proof parts which are different from Theorem 3. We analyze the Lyapunov drift as follows with Assumption 4.

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}\left[\mathcal{V}(\mathbf{Q}(t+1)) - \mathcal{V}(\mathbf{Q}(t))\right] \\
&= \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(Q_n(t) + A_n(t) - D_n(t))^{+2} - Q_n(t)^2\right] \\
&= \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(Q_n(t) + A_n(t) - \alpha D^*_n(t))^2 - Q_n(t)^2\right] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(Q_n(t) + A_n(t) - D_n(t))^{+2} - (Q_n(t) + A_n(t) - \alpha D^*_n(t))^2\right] \\
&\le 7\min\{N,K\}T - \sum_{t\in[T]}\sum_{n\in[N]} 2\epsilon\mathbb{E}[Q_n(t)] + 2\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(\alpha D^*_n(t) - D_n(t))Q_n(t)\right], \quad (76)
\end{aligned}$$

Then, for bounding Eq.(76), we have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha D^*_n(t) - D_n(t)Q_n(t)] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)(\mathbb{1}((E^1_t)^c) + \mathbb{1}((E^2_t)^c) + \mathbb{1}((E^3_{n,t})^c))]. \quad (77)
\end{aligned}$$

We provide a bound for the first term of Eq.(77). We first have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \mu(n|S^\alpha_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})] \\
&\le \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[(\alpha\mu(n|S_{k^*_{n,t},t}, \theta_{k^*_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S^\alpha_{k_{n,t},t}) \\
&\qquad + \widetilde{\mu}^{TS}_t(n|S^\alpha_{k_{n,t},t}) - \mu(n|S^\alpha_{k_{n,t},t}, \theta_{k_{n,t}}))Q_n(t)\mathbb{1}(E^1_t \cap E^2_t \cap E^3_{n,t})]. \quad (78)
\end{aligned}$$

Now we provide a bound for the first two terms in Eq.(78).

We define sets

$$\widetilde{\Theta}_t = \left\{ \{\theta_k^{(i)}\}_{i\in[M],k\in[K]} : \left| \max_{i\in[M]} x_n^\top \theta_k^{(i)} - x_n^\top \hat{\theta}_{k,t} \right| \le \gamma_t \|x_n\|_{V_{k,t}^{-1}};\ \forall n \in [N], \forall k \in [K] \right\} \text{ and}$$

$$\widetilde{\Theta}_t^{opt} = \left\{ \{\theta_k^{(i)}\}_{i\in[M],k\in[K]} : \sum_{k\in[K]} \sum_{n\in S_{k,t}^\alpha} \widetilde{\mu}_t^{TS}(n|S_{k,t}^\alpha, \{\theta_k^{(i)}\}_{i\in[M]}) Q_n(t) > \sum_{k\in[K]} \sum_{n\in S_{k,t}^*} \alpha\mu(n|S_{k,t}^*, \theta_k) Q_n(t) \right\} \cap \widetilde{\Theta}_t.$$

Then we define event $E_t(\widetilde{\theta}) = \{\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M],k\in[K]} \in \widetilde{\Theta}_t^{opt}\}$. Recall $h_{n,k,t}^{TS} = \max_{i\in[M]} x_n^\top \widetilde{\theta}_{k,t}^{(i)}$.
Then we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\sum_{k\in[K]}\left(\sum_{n\in S_{k,t}^*}\alpha\mu(n|S_{k,t}^*,\theta_k)Q_n(t) - \sum_{n\in S_{k,t}^\alpha}\widetilde{\mu}_t^{TS}(n|S_{k,t}^\alpha,\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})Q_n(t)\right)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\alpha\mu(n|S_{k,t}^*,\theta_k)Q_n(t)\right.\right.\right. \\
&\qquad \left.\left.\left. - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha}\widetilde{\mu}_t^{TS}(n|S_{k,t}^\alpha,\{\theta_k^{(i)}\}_{i\in[M]})Q_n(t)\right)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha}\widetilde{\mu}_t^{TS}(n|S_{k,t}^\alpha,\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})\right.\right.\right. \\
&\qquad \left.\left.\left. - \inf_{\{\theta_l^{(i)}\}_{i\in[M],l\in[K]}\in\widetilde{\Theta}_t}\sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha}\widetilde{\mu}_t^{TS}(n|S_{k,t}^\alpha,\{\theta_k^{(i)}\}_{i\in[M]})\right)Q_n(t)\mathbb{1}(E_t^1\cap E_t^2\cap E_{n,t}^3)|\mathcal{F}_{t-1},E_t(\widetilde{\theta})\right]\right] \\
&= \frac{2\gamma_t\epsilon}{C_2(\gamma_t+\beta_t)}\mathbb{E}\left[\sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha}\mathbb{E}\left[Q_n(t)\big|\mathcal{F}_{t-1},E_t(\widetilde{\theta}),E_t^1\right]\mathbb{P}(E_t^1|\mathcal{F}_{t-1})\right],
\end{aligned} \tag{79}$$

We provide a lemma below for further analysis.

Lemma 10. *For all $t \in [T]$, we have*

$$\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S_{k,t}^\alpha}\widetilde{\mu}_t^{TS}(n|S_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S_{k,t}^*}\alpha\mu(n|S_{k,t}^*,\theta_k)Q_n(t)|\mathcal{F}_{t-1},E_t^1\right) \ge 1/4\sqrt{e\pi}$$

Proof. Given $\mathcal{F}_{t-1}$, $x_n^\top\widetilde{\theta}_{k,t}^{(i)}$ follows Gaussian distribution with mean $x_n^\top\hat{\theta}_{k,t}$ and standard deviation $\beta_t\|x_n\|_{V_{k,t}^{-1}}$. Then we have

$$\begin{aligned}
\mathbb{P}\left(\max_{i\in[M]} x_n^\top\widetilde{\theta}_{k,t}^{(i)} > x_n^\top\theta_k|\mathcal{F}_{t-1},E_t^1\right) &= 1-\mathbb{P}\left(x_n^\top\theta_{k,t}^{(i)} \le x_n^\top\theta_k;\forall i\in[M]|\mathcal{F}_{t-1},E_t^1\right) \\
&= 1-\mathbb{P}\left(Z_i \le \frac{x_n^\top\theta_k - x_n^\top\hat{\theta}_{k,t}}{\beta_t\|x_n\|_{V_{k,t}^{-1}}};\forall i\in[M]|\mathcal{F}_{t-1},E_t^1\right) \\
&\ge 1-\mathbb{P}\left(Z\le 1\right)^M,
\end{aligned}$$

where Z_i and Z are standard normal random variables. Then we can show that

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S^{\alpha}_{k,t}}\widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\alpha\mu(n|S^{*}_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E^1_t\right)\\
&\geq \mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S^{\alpha}_{k,t}}\alpha\widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\alpha\mu(n|S^{*}_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E^1_t\right)\\
&\geq \mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\alpha\widetilde{\mu}^{TS}_t(n|S^{*}_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\alpha\mu(n|S^{*}_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E^1_t\right)\\
&= \mathbb{P}\left(\sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\widetilde{\mu}^{TS}_t(n|S^{*}_{k,t})Q_n(t) > \sum_{k\in[K]}\sum_{n\in S^{*}_{k,t}}\mu(n|S^{*}_{k,t},\theta_k)Q_n(t)|\mathcal{F}_{t-1},E^1_t\right)\\
&\geq \mathbb{P}\left(\max_{i\in[M]} x_n^{\top}\widetilde{\theta}^{(i)}_{k,t} > x_n^{\top}\theta_k; \forall n\in S^{*}_{k,t}, \forall k\in[K]|\mathcal{F}_{t-1},E^1_t\right)\\
&\geq 1 - LK\mathbb{P}(Z\leq 1)^M\\
&\geq 1 - LK(1-1/4\sqrt{e\pi})^M\\
&\geq \frac{1}{4\sqrt{e\pi}},
\end{aligned}
$$

where the second last inequality is obtained from $\mathbb{P}(Z \leq 1) \leq 1 - 1/4\sqrt{e\pi}$ using the anti-concentration of standard normal distribution, and the last inequality comes from $M = \lceil 1 - \frac{\log KL}{\log(1-1/4\sqrt{e\pi})}\rceil$. □

The rest of the proof can be easily obtained by following the proof steps in Theorem 3. □

Now, we investigate the regret of Algorithm 4. The α-regret regarding policy π is defined as

$$
\mathcal{R}^{\alpha,\pi}(T) = \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}\left[(\alpha\mu(n|S^{*}_{k^{*}_{n,t},t},\theta_{k^{*}_{n,t}}) - \mu(n|S_{k_{n,t},t},\theta_{k_{n,t}}))Q_n(t)\right].
$$

The algorithm achieves the following regret bound.

Theorem 8. *The policy π of Algorithm 4 achieves a regret bound of*

$$
\mathcal{R}^{\alpha,\pi}(T) = \widetilde{\mathcal{O}}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}, \left(\frac{d^2NK\min\{N,K\}^3}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right).
$$

Proof. Here we provide only the proof parts which are different from Theorem 4.

We first provide the proof for the regret bound of $\widetilde{\mathcal{O}}\left(\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}\right)$. We have

$$
\begin{aligned}
\mathcal{R}^{\pi}(T) &= \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^{*}_{n,t},t},\theta_{k^{*}_{n,t}}) - \mu(n|S^{\alpha}_{k_{n,t},t},\theta_{k_{n,t}}))]\\
&\leq \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^{*}_{n,t},t},\theta_{k^{*}_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) + \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) - \mu(n|S^{\alpha}_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E^1_t\cap E^2_t)]\\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^{*}_{n,t},t},\theta_{k^{*}_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) + \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) - \mu(n|S^{\alpha}_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E^1_t)^c)]\\
&\quad + \sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^{*}_{n,t},t},\theta_{k^{*}_{n,t}}) - \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) + \widetilde{\mu}^{TS}_t(n|S^{\alpha}_{k_{n,t},t}) - \mu(n|S^{\alpha}_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}((E^2_t)^c)].
\end{aligned}
$$

By following the similar proof steps in Theorem 7, we can obtain

$$\sum_{t\in[T]}\sum_{n\in[N]}\mathbb{E}[Q_n(t)(\alpha\mu(n|S_{k^*_{n,t},t},\theta_{k^*_{n,t}})-\widetilde{\mu}_t^{TS}(n|S^{\alpha}_{k_{n,t},t})$$
$$+\widetilde{\mu}_t^{TS}(n|S^{\alpha}_{k_{n,t},t})-\mu(n|S^{\alpha}_{k_{n,t},t},\theta_{k_{n,t}}))\mathbb{1}(E_t^1\cap E_t^2)]$$
$$\leq C_2\sum_{t\in[T]}\sum_{k\in[K]}\mathbb{E}\left[\max_{n\in S^{\alpha}_{k,t}}(\gamma_t+\beta_t)\|x_n\|_{V^{-1}_{k,t}}Q_n(t)\right].$$

The rest of the proof can be easily obtained from the proof steps in Theorem 4 by incorporating techniques in the proof of Theorem 7. □

A.10 Additional Experiments

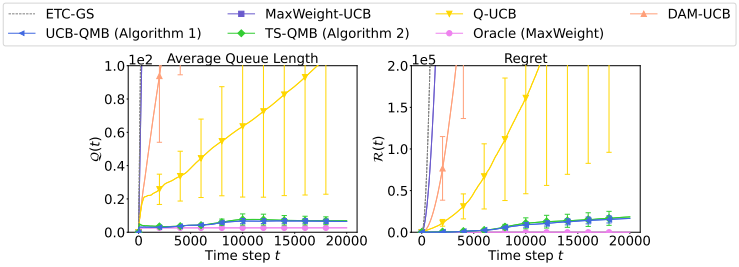


Figure 3: Experimental results with $N = 4, K = 3, L = 2, d = 2$ for (left) average queue length and (right) regret

A.11 Notation Table

Table 1: Notation Table.

T	Time horizon
N	Number of agents (queues)
K	Number of arms (servers)
d	Dimension of model parameters and features
x_n	Feature vector of agent n
θ_k	Model parameter vector of arm k
$Q_n(t)$	Length of agent n at time t
ϵ	Traffic slackness parameter
$\mu(n\|S_k,\theta_k)$	Service rate of arm k with θ_k for a job in agent n given assortment S_k
n_0	Null agent
λ_n	Job arrival rate for agent n
$k^{\pi}_{n,t}(=k_{n,t})$	A server that agent n is assigned at time t according to π (simply $k_{n,t}$)
$A_n(t)$	The number of arrival jobs in agent n at time t; Random variable with mean λ_n
$D_n(t\|S_k)$	The number of departure job in agent n by arm k given S_k; Random variable with mean $\mu(n\|S_k,\theta_k)$
$\mathcal{Q}(T)$	Average queue lengths over horizon time T
$\mathcal{R}^{\pi}(T)$	Cumulative regret under π over T
κ	Regularity parameter for MNL
$\mathcal{N}_t$	Set of non-empty agents at time t
$\mathcal{M}()$	Set of feasible disjoint assortments given a set of agents
$S_{k,t}$	Set of agents assigned to arm k by policy π
$S^*_{k,t}$	Set of agents assigned to arm k by the oracle π^*
$k^{\pi^*}_{n,t}(=k^*_{n,t})$	A server that agent n is assigned by π^* at time t
$y_{n,t}$	Preference feedback, $\{0,1\}$, for assigned agent n at time t

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Through the abstract and introduction, we explain our setting with providing motivation examples and summarize our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the limitations section (Section A.1), we provide several avenues for interesting future work. We also discuss the computational efficiency of our algorithms.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide assumptions in Section 3 and proof sketches of some of the main theorems in the main text, with complete proofs for theorems in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Section 6, we provide all the information necessary for conducting the synthetic experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: There is a link to our code in Section 6.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We use synthetic datasets and provide details on how to generate them in Section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In our experimental results in Section 6, we include error bars of standard deviation along with expectation values.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The conducted experiments do not require significant computing power.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Given that this study primarily focuses on theoretical analysis, we do not foresee any negative social consequences.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.