LEARNING TO ROUTE AND SCHEDULE LLMS FROM USER RETRIALS VIA CONTEXTUAL QUEUEING BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explosive demands for LLMs often cause user queries to accumulate in server queues, requiring efficient routing (query-LLM matching) and scheduling (query prioritization) mechanisms. Several online algorithms are being deployed, but they overlook the following two key challenges inherent to conversational LLM services: (1) unsatisfied users may retry queries, increasing the server backlog, and (2) requests for "explicit" feedback, such as ratings, degrade user experiences. In this paper, we develop a joint routing and scheduling algorithm that leverages "implicit" feedback inferred from user retrial behaviors. The key idea is to propose and study the framework of contextual queueing bandits with multinomial logit feedback (CQB-MNL). CQB-MNL models query retrials, as well as context-based learning for user preferences over LLMs. Our algorithm, anytime CQB (**ACQB**), achieves efficient learning while maintaining queue stability by combining Thompson sampling with forced exploration at a decaying rate. We show that ACQB simultaneously achieves a cumulative regret of $\widetilde{\mathcal{O}}(\sqrt{t})$ for routing and a queue length regret of $\widetilde{\mathcal{O}}(t^{-1/4})$ for any large t. For experiments, we refine query embeddings via contrastive learning while adopting a disjoint parameter model to learn LLM-specific parameters. Experiments on synthetic data, offline routing datasets (SPROUT, EMBEDLLM, and ROUTERBENCH), and real user conversation logs (WILDCHAT-1M) confirm that our methods improve routing, scheduling, and queue stability against strong online and offline-trained baselines.

1 INTRODUCTION

Recent advances in large language models (LLMs) have revolutionized various domains, driving explosive demands for LLM-based applications (Achiam et al., 2023; Team et al., 2023; Grattafiori et al., 2024). However, running LLMs is computationally intensive and resource-demanding. Therefore, a flood of user queries, coupled with limited server capacity, inevitably leads to significant query accumulation in system queues. Hence, efficient query routing and scheduling is crucial to ensure high-quality user experiences and system sustainability (Kwon et al., 2023; Agrawal et al., 2024).

To mitigate such system congestion, LLM routing and scheduling frameworks have emerged as necessary components. In a typical system of multiple LLMs with varying types and levels of capabilities, performance, and costs, the system must decide which model to serve a given query (routing) and in what order to process backlogged queries (scheduling) (Mitzenmacher & Shahout, 2025). The first approaches have relied on policies learned offline (Ong et al., 2024; Feng et al., 2024; Fu et al., 2024), but these methods often fail to adapt to the dynamic and non-stationary nature of online environments. This limitation has motivated the adoption of online learning frameworks (Chiang et al., 2025; Jitkrittum et al., 2025), which learn optimal policies through sequential interactions with the environment.

However, existing online learning approaches are often limited in practice as they overlook two key challenges inherent to conversational LLM services. First, they fail to account for the subsequent impact of user dissatisfaction. When users encounter unsatisfactory responses, they often retry queries. This retrial behavior increases server backlog, exacerbating congestion. This introduces a trade-off between exploration and system stability: exploring potentially sub-optimal models may

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

induce user retrials that increase queue length. Second, most algorithms rely on explicit user feedback (e.g., ratings, preference information) to update their policies. In practice, however, such feedback is sparse, as users are often reluctant to provide it. Furthermore, mandating explicit feedback can disrupt conversation flows and degrade user experiences.

In this paper, we address these challenges by presenting a novel formulation for the joint routing and scheduling problem via contextual queueing bandits with multinomial logit feedback (CQB-MNL). Specifically, CQB-MNL considers a discrete-time queueing system with N LLMs, where in each round a query with some context information arrives at the queue. An algorithm for CQB-MNL chooses a pending query with the highest priority (scheduling) and recommends an assortment of K LLMs for response generation (routing) (See Figure 1). Here, the case $K = 1$ corresponds to the standard single-response setting, which is essentially a dynamic matching problem. The case $K = 2$ represents the pairwise comparison setting, commonly employed to collect human preferences for reinforcement learning from human feedback (RLHF), where a user receives two candidate responses and is asked to select one. In CQB-MNL, the choice behavior of a user is modeled with the multinomial logit (MNL) framework (Agrawal et al., 2019; Oh & Iyengar, 2019). Basically, the user accepts one of K candidate responses, indicating satisfaction, or rejects all and retries the query, indicating dissatisfaction, and the choice decision is governed by the MNL model. This formulation enables the system to learn optimal policies based on implicit feedback inferred from user retrials.

However, an online learning algorithm for CQB-MNL requires exploration, which may involve suggesting suboptimal models, and as a result, it inherits the risk of triggering retrials, causing system congestion. Therefore, we need an algorithm that cleverly balances exploration and system stability. To address this requirement, we propose an algorithm, anytime CQB (ACQB), that achieves efficient learning of the underlying MNL model while maintaining queue stability. The main backbone of ACQB is Thompson sampling for MNL, while it enforces uniform exploration steps.

Our contributions are summarized as follows:

- We introduce CQB-MNL, a novel online learning framework designed to address the two critical challenges inherent to online learning for LLM routing and scheduling: system congestion due to user dissatisfaction triggering retrials and impracticality of relying on explicit feedback.
- We propose ACQB for CQB-MNL that allows efficient learning of the unknown MNL model while guaranteeing queue stability. In each time step, Thompson sampling or uniform exploration is performed. Here, the probability of running uniform exploration decreases as time goes on.
- ACQB is an "anytime" algorithm suitable for continuous LLM services. It works without prior knowledge of the time horizon or the traffic slackness parameter. These parameters are usually unavailable a priori. We establish that ACQB simultaneously achieves a cumulative regret of $\widetilde{\mathcal{O}}(\sqrt{t})$ for learning the MNL model and a queue length regret of $\widetilde{\mathcal{O}}(t^{-1/4})$, ensuring convergence to optimal queue length, for any large t.
- For experiments, we employ disjoint parameterization to capture individual LLM-specific characteristics (Section 4.1). Moreover, we refine query embeddings via contrastive learning. Experiments on synthetic data and offline routing datasets (EMBEDLLM, SPROUT, ROUTERBENCH) demonstrate that our methods consistently outperform queueing bandit baselines and offline-trained routing baselines. We also test the scheduling component of our method on real user conversation logs from WILDCHAT-1M by replaying timestamped conversations, where retrials are extracted from repeated or highly similar queries by the same user.

2 PROBLEM FORMULATION

In this section, we formulate the online learning of LLM routing and scheduling with users' retrials as a *contextual queueing bandit* problem. We consider a discrete-time queueing environment with a single queue and N LLMs. At each round t, a user query (together with its context) arrives and joins the queue. The agent then performs a joint operation: it selects a pending query to process from the queue (scheduling) and assigns an appropriate LLM (or an assortment of LLMs) to the selected query (routing). Formally, at round t, the agent observes a queue state $\mathcal{X}_t \subseteq \mathcal{X} \subset \mathbb{R}^d$, given by the set of contexts of the remaining queries, and chooses a query context $x_t \in \mathcal{X}_t$. The corresponding

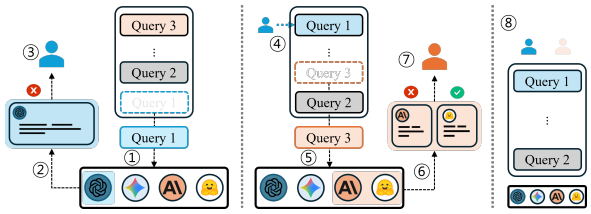


Figure 1: Illustration demonstrating retrial and departure dynamics. ($K = 1$): ① The agent schedules a query (Query 1). ② An assortment of size $K = 1$ is assigned. ③ The user is dissatisfied with the response, which ④ triggers a retrial. ($K = 2$): ⑤ The agent schedules a query (Query 3). ⑥ An assortment of size $K = 2$ is assigned. ⑦ The user selects one of the responses (satisfaction), and consequently, ⑧ the query departs the queue.

feature vector of x_t for each LLM $j \in [N]$ is denoted as $x_{t,j}$ (or x_{-j} for an arbitrary context x). The agent then chooses an assortment of servers $S_t \in \mathcal{C}$, where $\mathcal{C} := \{S \subseteq [N] : |S| = K\}$ and K is the assortment size. For simplicity, we fix K throughout the paper; in practice, one may deploy a policy where the agent primarily uses a single LLM ($K = 1$) and occasionally generates two answers ($K = 2$) when additional feedback is required, and the same analysis applies.

We characterize user satisfaction through follow-up behavior. If the user is satisfied—specifically, accepting the answer in the single-response setting ($K = 1$) or selecting a preferred answer in the pairwise setting ($K = 2$)—the corresponding query departs the queue. Conversely, if the user is dissatisfied, they exhibit retrials, which we model as the query remaining, or re-entering the queue. We describe these dynamics formally as follows: Let $Q(t) = |\mathcal{X}_t|$ denote the queue length at the beginning of round t. Let $A(t) \in \{0, 1\}$ indicate the random arrival of a new job, and $D(t) \in \{0, 1\}$ denote the random departure at time t. When $A(t) = 1$, we denote the context of the newly arriving job as $x^{(t)}$. The queue state then evolves as:

$$\mathcal{X}_{t+1} = \mathcal{X}_t \setminus \{x_t : D(t) = 1\} \cup \{x^{(t)} : A(t) = 1\},$$

and the queue length evolves according to $Q(t + 1) = [Q(t) + A(t) - D(t)]^+$, where $[z]^+ = \max\{0, z\}$. For technical convenience, if the queue is empty ($Q(t) = 0$), the agent selects a dummy job $x_0 \in \mathbb{R}^d$, and the resulting feedback is not used for learning.

We model the user choice probability using the MNL model (Agrawal et al., 2019; Oh & Iyengar, 2019). The agent observes a binary choice vector $y_t = [y_{t0}, y_{t1}, \ldots, y_{tK}] \in \{0, 1\}^{K+1}$, where $\sum_{j \in [K] \cup \{0\}} y_{tj} = 1$ and $y_{tj} = 1$ if the user chooses the j-th LLM in the assortment S_t, and 0 otherwise. We designate the outside option (no choice) as $y_{t0} = 1$, representing user dissatisfaction. The probability that a user chooses item j from assortment S given context x is defined as $p_j(x, S, \theta^*)$, where $\theta^* \in \mathbb{R}^d$ is the unknown parameter, and $p_j(x, S, \theta) := \frac{\exp(x_{-j}^\top \theta)}{1+\sum_{j' \in S} \exp(x_{-j'}^\top \theta)}$ for $j \in \mathcal{S}$, and $p_0(x, S, \theta) := \frac{1}{1+\sum_{j' \in S} \exp(x_{-j'}^\top \theta)}$ for the outside option ($j = 0$). Therefore, a departure $D(t) = 1$ occurs if the user selects any valid option in S_t (i.e., avoids the outside option). Thus, $D(t)$ is a random variable with mean $R(x_t, S_t, \theta^*)$, where the success probability is given by

$$R(x, S, \theta) := \sum_{j \in S} p_j(x, S, \theta) = \frac{\sum_{j \in S} \exp(x_{-j}^\top \theta)}{1 + \sum_{j' \in S} \exp(x_{-j'}^\top \theta)}.$$

Finally, we assume the arrival $A(t)$ is a random variable with mean λ.

We evaluate the performance of the agent using two key measures: (i) *queue length regret* and (ii) *cumulative regret*. Let π denote the agent's policy and π^* denote the optimal policy that has prior knowledge of the true parameter θ^*. Given a set of remaining features $\mathcal{Y} \subseteq \mathcal{X}$, π^* selects the query-assortment pair (x^*, S^*) that maximizes the expected departure rate, i.e., $(x^*, S^*) \in \arg\max_{x \in \mathcal{Y}, S \in \mathcal{C}} R(x, S, \theta^*)$. Now, we define the queue length regret at round t as

$$R_t := \mathbb{E}[Q(t) - Q^*(t)],$$

where $Q^*(t)$ is the queue length under the optimal policy π^*. To minimize R_t, the agent must learn both which query to serve and which LLM assortment to assign. Accordingly, this metric

Algorithm 1 ACQB

initialize design matrix $V_0 = \lambda_0 \mathbf{I}$, assortment size K, sample size M, combination set $\mathcal{C} = \{S \subset [N] : |S| = K\}$, combination counter $c = 0$, exploration parameter $\eta(t)$, confidence radius α_t

1: **for** $t = 1, \ldots$ **do**
2: **if** $A(t-1) = 1$ **and** $E(t-1) = 1$ **then** ▷ on arrivals, $\eta(t)$-exploration
3: Set $x_t \leftarrow x^{(t-1)}$, $S_t \leftarrow \mathcal{C}[c+1]$
4: $c \leftarrow c + 1 \pmod{|\mathcal{C}|}$
5: **else**
6: Sample $\{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^{M} \sim \mathcal{N}(\widehat{\theta}_{t-1}, \alpha_{t-1}^2 V_{t-1}^{-1})$
7: Set $x_t, S_t \leftarrow \arg\max_{x \in \mathcal{X}_t, S \in \mathcal{C}} \widetilde{R}(x, S)$
8: **end if**
9: Assign x_t to S_t, receive $y_t = (y_{t0}, y_{t1}, \ldots, y_{tK})$
10: Update $\widehat{\theta}_t$ as in (1)
11: $V_t \leftarrow V_{t-1} + \sum_{j \in S_t} x_{tj} x_{tj}^{\mathsf{T}}$
12: Sample $E(t) \sim \mathrm{Bern}(\eta(t))$
13: **end for**

simultaneously evaluates the routing and scheduling performance. Moreover, to evaluate the routing performance in terms of departure-rate optimality, we consider cumulative regret, a standard measure in the bandit literature. Let $S_i^* = \arg\max_{S \in \mathcal{C}} R(x_i, S, \theta^*)$ be the optimal assortment for the fixed query x_i chosen by the agent. The standard cumulative regret is defined as

$$\mathrm{Regret}_t := \sum_{i=1}^{t} \mathbb{E}[R(x_i, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)].$$

Lastly, we introduce the assumptions as follows:

Assumption 1. $\|x_{-j}\|_2 \leq 1$ *for all* $x \in \mathcal{X}, j \in [N]$. *Also,* $\|\theta^*\| \leq 1$.

Assumption 2. *There exist* $\kappa > 0$ *such that for all* $x \in \mathcal{X}, j \in [N]$, $\inf_{S \in \mathcal{C}, \theta \in \mathbb{R}^d} p_j(x, S, \theta) p_0(x, S, \theta) \geq 1/\kappa$.

Assumption 3. *The features of newly arriving jobs are assumed to be independently and identically distributed (i.i.d.) from an unknown distribution* $\mathcal{D}$. *Moreover, there exists* $\sigma_0 > 0$ *such that* $\lambda_{\min}(\mathbb{E}_{x \sim \mathcal{D}}[\frac{1}{N} \sum_{j \in [N]} x_{-j} x_{-j}^{\top}]) \geq \sigma_0^2$.

Assumption 4. *There exists some traffic slackness* $\epsilon > 0$ *such that for each* $x \in \mathcal{X}$, *there exists an corresponding assortment* $S^*(x) \in \mathcal{C}$ *with* $R(x, S^*(x), \theta^*) - \lambda \geq \epsilon$.

Assumption 1 states that the norms of the feature vector x_{-j} and the unknown parameter θ^* are bounded. Assumption 2 introduces problem-dependent parameters that control the local behavior of $\dot{\mu}(\cdot)$. Assumption 3 imposes a regularity assumption on the underlying distribution. Assumption 4 specifies a traffic slack condition to guarantee stability, which is standard in the queueing bandit literature (Krishnasamy et al., 2016; Bae et al., 2026).

3 Algorithm For Contextual Queueing Bandits With Multinomial Logit Feedback

In this section, we introduce our proposed algorithm, ACQB, which is illustrated in Algorithm 1. The algorithm consists of two branches: (i) random exploration and (ii) a Thompson sampling-based optimistic rule. Let $E(t) \sim \mathrm{Bern}(\eta(t))$ be the random variable that indicates whether we run random exploration in round $t+1$, where $\eta(t) = \min\{1, c_1(t+1)^{-1/2}\}$ is the exploration parameter for some absolute constant $c_1 > 0$. Then, in each round t, if there is a new job arrival ($A(t-1) = 1$), we perform random exploration with probability $\eta(t-1)$ (i.e., when $E(t-1) = 1$) by selecting the newly arriving job $x^{(t-1)}$ and choosing the LLM assortment in a round-robin manner. If this event (i.e., $A(t-1) = 1$ and $E(t-1) = 1$) does not occur, we choose the query-assortment pair according to the Thompson sampling-based optimistic rule (Oh & Iyengar, 2019): (Line 6) sample $\{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^{M}$

from the Gaussian distribution $N(\widehat{\theta}_{t-1}, \alpha_{t-1}^2 V_{t-1}^{-1})$, where $M = \lceil 1 - \frac{\log(K)}{\log(1-1/(4\sqrt{e\pi}))} \rceil$ and α_{t-1} is a confidence radius defined as

$$\alpha_l = \frac{\kappa}{2}\sqrt{d\log(1+lK/(d\lambda_0)) + 4\log l} + \kappa\sqrt{\lambda_0},$$

where $\lambda_0 > 0$ is a regularization parameter. (Line 7) We then choose the query and LLM assortment pair with the largest optimistic departure rate estimate $\widetilde{R}(x, S)$, where

$$\widetilde{u}_{tj}(x) := \max_i x_{-j}^\intercal \widetilde{\theta}_{t-1}^{(i)}, \quad \widetilde{R}(x, S) := \sum_{k\in S} \frac{\exp(\widetilde{u}_{tk}(x))}{1 + \sum_{j\in S}\exp(\widetilde{u}_{tj}(x))}.$$

After assigning the query x_t to the LLM assortment S_t, we observe the reward vector y_t. Finally, we update the maximum likelihood estimator (MLE) $\widehat{\theta}_t$ by minimizing the regularized cross-entropy loss

$$\mathcal{L}_t(\theta) = \frac{\lambda_0}{2}\|\theta\|_2^2 - \sum_{i=1}^{t}\sum_{j\in S_i\cup\{0\}} y_{ij}\log p_j(x_i, S_i, \theta). \tag{1}$$

Now we introduce our main results, which provide the queue length regret and cumulative regret bound of Algorithm 1. The proof sketches are provided in Section D, and the full proofs are deferred to Section E.

Theorem 5. *For any large t (Condition 20), we have*

$$R_t = \mathcal{O}\left(\frac{d^5 t^{-1/4}\log^5(t)}{\sigma_0^9\epsilon^5} + \frac{d^{11/2}t^{-1}\log^5(t)}{\sigma_0^{12}\epsilon^5}\right).$$

Theorem 6. *For any $t \geq 1$, we have*

$$\textit{Regret}_t = \widetilde{\mathcal{O}}(d^{3/2}\sqrt{t}).$$

Comparison with previous works. Closely related works in queueing bandits literature include Krishnasamy et al. (2016); Kim & Oh (2024); Bae et al. (2026). First, Krishnasamy et al. (2016) pioneered the analysis of queue length regret. However, their multi-armed bandit (MAB)-based framework ignores query contexts and assumes fixed departure rates, making it unsuitable for our setting. Later, Kim & Oh (2024) address contexts but restrict it to a single fixed type per queue, unlike our arbitrary setting, and they lack an analysis of queue length regret. The most comparable study is Algorithm 1 of Bae et al. (2026), which accommodates arbitrary contexts and analyzes queue length regret. Notably, it yields a tighter leading term of $\mathcal{O}(d^{3/2}T^{-1/4}\log^{3/2}(T)\sigma_0^{-5}\epsilon^{-3})$ compared to the $\mathcal{O}(d^5T^{-1/4}\log^5(T)\sigma_0^{-9}\epsilon^{-5})$ term of our Algorithm 1. The heavier dependence on d mainly comes from the Thompson-sampling analysis and the anytime exploration design, rather than from the MNL feedback model itself. If we replace Thompson sampling with a UCB rule in our framework, the queue length regret improves to $\widetilde{\mathcal{O}}(d^{5/2}T^{-1/4})$, and the cumulative regret improves to $\widetilde{\mathcal{O}}(d\sqrt{T})$. In contrast, their approach relies on an exploration schedule dependent on the horizon T and slackness ϵ, which fails to satisfy the anytime property, making it unsuitable for continuous LLM serving (see Remark 15). Finally, their model considers only the logistic function, requiring non-trivial extension to the MNL model.

4 Algorithmic Extensions

We introduce two algorithmic adaptations designed to address the challenges of real-world deployment and to enhance performance.

4.1 Capturing LLM Heterogeneity Via Disjoint Parameterization

Many prior works (Chiang et al., 2025; Shirkavand et al., 2025) use a joint feature vector $x_{t,j}$ with a shared parameter vector θ^*. Although query embeddings x_t can be obtained from pre-trained embedding models (Reimers & Gurevych, 2019; Sentence-Transformers, 2021), constructing $x_{t,j}$ requires additional feature engineering for each LLM. This can make it unclear whether the gain comes from the routing algorithm or from the feature construction.

Therefore, for the experiments, we adopt a disjoint parameter model to bypass the reliance on manual feature engineering. In this setting, the query context x_t is shared, but each server j is governed by a unique unknown parameter θ_j^*. This allows the agent to learn the unique characteristics of heterogeneous LLMs through online learning, thereby removing the confounding effects of feature engineering and enabling a fair assessment of algorithmic performance. Accordingly, define the collection of parameters as $\Theta^* = [\theta_1^*, \theta_2^*, \ldots, \theta_N^*] \in \mathbb{R}^{d \times N}$. When a user with context x is presented with an assortment S, the probability $p_j(x, S, \Theta^*)$ of selecting item j and the corresponding departure rate $R(x, S, \Theta^*)$ are defined as $p_j(x, S, \Theta^*) := \frac{\exp(x^\top \theta_j^*)}{1+\sum_{j' \in S} \exp(x^\top \theta_{j'}^*)}$ for $j \in S$, $p_0(x, S, \Theta^*) := \frac{1}{1+\sum_{j' \in S} \exp(x^\top \theta_{j'}^*)}$ for $j = 0$, and $R(x, S, \Theta^*) := \sum_{j \in S} p_j(x, S, \Theta^*)$. Under this disjoint model, the agent maintains separate statistics for each model j: a design matrix $V_{t,j}$, a maximum likelihood estimator $\hat{\theta}_{t,j}$, and a confidence radius $\alpha_{t,j}$. The detailed algorithm is outlined in Section B.1.

Remark 7. *Notice that this disjoint parameterization is a specific instance of the shared parameter setting used in our theoretical analysis. Let $e_j \in \mathbb{R}^N$ be the standard basis vector with a 1 at the j-th position and 0 elsewhere. Then, for all $j \in [N]$, we can write $x^\mathsf{T}\theta_j^* = (x \otimes e_j)^\mathsf{T} vec(\Theta^*)$ where $\otimes$ denotes the Kronecker product and $vec(\Theta^*)$ represents the vectorization of a matrix. By viewing $(x \otimes e_j) \in \mathbb{R}^{dN}$ as a feature vector and $vec(\Theta^*) \in \mathbb{R}^{dN}$ as the shared parameter, our theoretical analysis remains valid.*

4.2 Utility-Aligned Query Embeddings Via Contrastive Learning

Standard query encoders are typically trained to align representations with linguistic semantics (Reimers & Gurevych, 2019), resulting in an embedding space where proximity reflects categorical or paraphrase-level similarity (Chiang et al., 2025). However, effective multi-LLM routing requires embeddings to capture utility alignment rather than semantic closeness. That is, two queries should be considered similar if they exhibit similar routing utilities (i.e., performance-cost trade-offs) across varying LLMs, even if they are semantically distinct (Chen et al., 2024). To address this, we propose ACQB-CL (Algorithms 2 and 3), which realigns the query representation space. Specifically, given a raw prompt ξ, we freeze the backbone encoder $E(\cdot)$ and train a two-layer MLP projection head $B(\cdot; \theta)$ to obtain a utility-aligned representation $z = B(E(\xi); \theta) \in \mathbb{R}^{d'}$. The projection head is optimized via an InfoNCE loss (Oord et al., 2018) on an offline dataset to cluster queries with similar utilities while separating divergent ones. Comprehensive details regarding offline data construction, pair selection, and the precise loss formulation are provided in Section B.2.

5 Experiments

In this section, we empirically evaluate the performance of our proposed algorithms.

5.1 Experiments On Synthetic Data

Experimental setup. We generate random instances for $K = \{1, 2\}$ by setting the arrival rates to $\lambda = 0.7$ with the parameters $d = 5$, $\epsilon = 0.03$, $N = 5$, and $T = 1{,}000$. The context vectors $x \in \mathbb{R}^d$ and disjoint parameters $\theta_j^* \in \mathbb{R}^d$ for $j \in [N]$ are sampled from $\mathrm{Unif}(-1, 1)$, ensuring that the contexts satisfy the slackness condition, i.e., $\max_{S \in \mathcal{C}} R(x, S, \Theta^*) \geq \lambda + \epsilon$. We perform 10 independent runs and report the average queue length regret and average cumulative regret with ± 1 standard deviation. We benchmark ACQB against five baseline algorithms for $K = 1$: (i) the Optimal policy; (ii) the Random policy, which selects a random query and assortment; (iii–iv) two MAB-based queueing bandit algorithms from (Krishnasamy et al., 2021); and (v) the contextual queueing bandit algorithm (Algorithm 1 from Bae et al. (2026)). To the best of our knowledge, aside from these queueing bandit-based approaches, no other existing baselines are capable of learning from user retrials (implicit feedback). For the $K = 2$ setting, since baselines (iii) and (iv) are incompatible with the MNL model, we compare ACQB only with (i), (ii), and the MNL-adapted version of (v). Detailed descriptions of each baseline policy are provided in Section B.3.

Performance comparison. As shown in Figure 2, ACQB has lower queue length and cumulative regret values than the baselines. To isolate the scheduling rule, we also fix the routing rule of ACQB

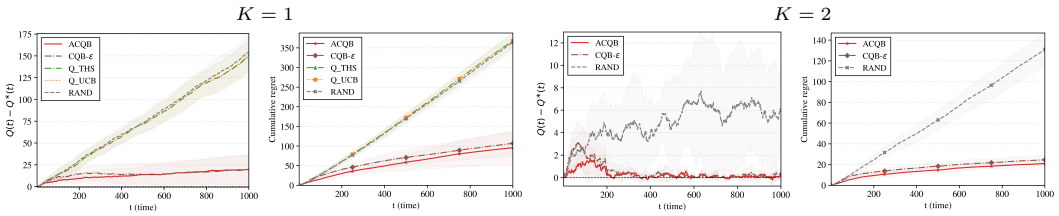


Figure 2: Queue length and cumulative regret on synthetic data with $\lambda = 0.7$, $\epsilon = 0.03$, and $N = 5$.

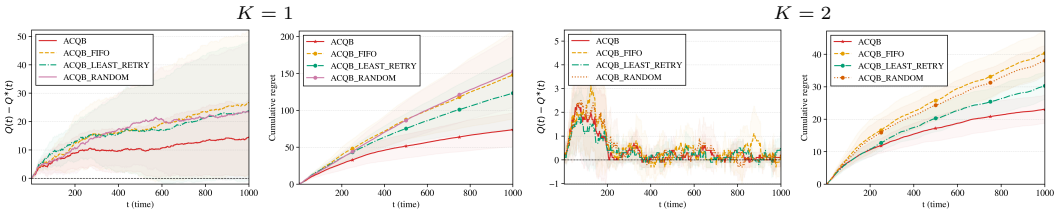


Figure 3: Ablation on ACQB scheduling on synthetic data with $\lambda = 0.7$, $\epsilon = 0.03$, and $N = 5$.

Table 1: Throughput (T.P.) and wall-clock time (W.C., seconds per round, excluding LLM inference) on synthetic data at $T = 500$ and $T = 1000$.

Time	Metric	ACQB	ACQB-FIFO	ACQB-RR	ACQB-RAND	CQB-ϵ	Q-ThS	Q-UCB	RAND
$T = 500$	T.P.	**0.672**	0.660	0.660	0.655	0.658	0.551	0.551	0.549
	W.C	0.0130	0.0135	0.0151	0.0125	0.0147	0.0000	0.0000	0.0001
$T = 1000$	T.P.	**0.680**	0.667	0.670	0.670	0.667	0.545	0.544	0.540
	W.C	0.0126	0.0114	0.0129	0.0121	0.0142	0.0001	0.0001	0.0001

and replace only the scheduling policy with FIFO, round-robin over the least-served jobs (RR), or random scheduling. Figure 3 shows that ACQB scheduling has lower queue length regret than these variants. Table 1 reports throughput and wall-clock time. ACQB achieves higher throughput at both $T = 500$ and $T = 1{,}000$, with wall-clock time comparable to the other contextual policies. This overhead is small relative to LLM inference time: a response time can be approximated by $\text{TTFT} + (\text{number of output tokens})/(\text{tokens/sec})$, and a public benchmark reports TTFT 2.84s and throughput 30 tokens/sec for GPT-4 Turbo.[1] The near-zero W.C. entries for Q-ThS, Q-UCB, and RAND are rounded values with very small but nonzero overheads. Details of the scheduling variants and additional experiments with varying N and ϵ are provided in Sections B.3 and C.1.

Exploration variants. ACQB triggers random exploration only when a new query arrives, which preserves i.i.d. arriving contexts in the eigenvalue-growth argument. In Section C.2, we compare this design with heuristic variants that trigger exploration independently of arrivals and select the exploration query by FIFO, LIFO, round-robin over the least-served queries, or random scheduling.

5.2 Experiments On Offline Routing Datasets

We next evaluate ACQB and ACQB-CL through simulations on offline LLM routing datasets. We use three datasets: SPROUT-O3MINI (14 LLMs) (Somerstep et al., 2025), EMBEDLLM (112 LLMs) (Zhuang et al., 2024), and ROUTERBENCH (11 LLMs) (Hu et al., 2024). All datasets include benchmark-specific correctness scores (e.g., exact-match or graded accuracy) and inference costs. For ROUTERBENCH, we use the provided cost metadata, while for EMBEDLLM and SPROUT-O3MINI, we construct the cost map using reports from (Shirkavand et al., 2025; Somerstep et al., 2025). For the query embedding, we adopt the `sentence-transformers/all-MiniLM-L6-v2` (Sentence-Transformers, 2021) as the backbone query encoder ($d = 384$). In the online phase, we run the contextual bandit simulation over an online prompt pool of size 1,000 for a horizon of $T = 5{,}000$.

[1] https://benchlm.ai/llm-speed

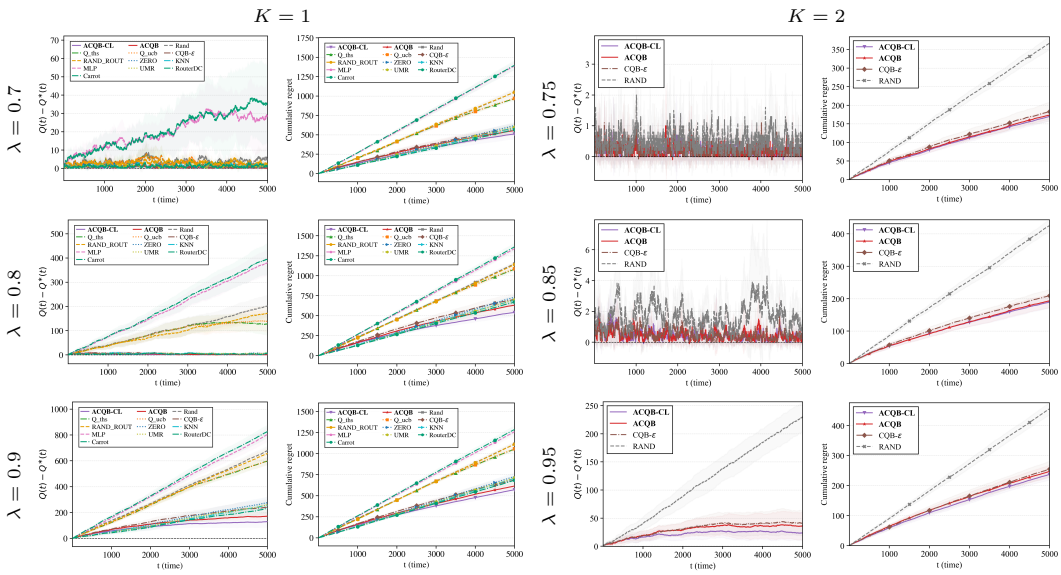


Figure 4: Queue length and cumulative regret on the SPROUT-O3MINI dataset across various arrival rates λ.

Table 2: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 1$ (arrival setting: $\lambda_{\text{Sprout}} = 0.8$, $\lambda_{\text{EmbedLLM}} = 0.4$, $\lambda_{\text{RouterBench}} = 0.6$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ε	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	**1.2**	3.3	101.4	99.9	99.9	6.5	87.8	3.9	2.3	181.6	5.6	192.4	2.2
		σ	1.7	3.2	48.5	46.6	46.6	10.0	23.7	3.4	2.7	38.4	6.5	45.6	2.5
		T.P.	**0.796**	0.795	0.756	0.756	0.756	0.794	0.761	0.795	0.795	0.724	0.794	0.719	0.795
	5000	μ	**0.3**	1.1	202.0	127.8	138.1	2.4	171.8	8.4	6.7	380.3	7.9	397.2	6.5
		σ	0.5	1.7	66.6	66.6	67.2	3.7	48.0	9.4	5.0	76.9	9.9	93.0	5.4
		T.P.	**0.797**	0.797	0.757	0.772	0.770	0.797	0.763	0.796	0.796	0.721	0.796	0.718	0.796
EmbedLLM	2500	μ	**55.0**	59.4	188.3	179.5	179.5	70.5	168.0	255.5	181.4	328.7	182.4	295.7	181.4
		σ	20.3	11.3	27.0	22.9	22.9	28.2	25.5	30.0	36.4	35.8	30.5	34.2	36.4
		T.P.	**0.375**	0.374	0.322	0.326	0.326	0.369	0.330	0.295	0.325	0.266	0.324	0.279	0.325
	5000	μ	**54.6**	60.5	364.4	358.5	358.5	63.4	333.5	506.0	360.6	653.8	355.1	586.0	360.6
		σ	20.8	21.4	38.4	40.0	40.0	43.3	45.4	32.7	50.2	71.0	54.7	69.9	50.2
		T.P.	**0.388**	0.387	0.326	0.327	0.327	0.386	0.332	0.297	0.326	0.268	0.328	0.281	0.326
RouterBench	2500	μ	**6.9**	10.3	160.3	157.3	156.2	29.9	155.7	70.5	68.4	342.8	60.0	287.5	62.9
		σ	6.7	10.3	59.2	62.9	60.6	35.1	26.7	30.6	42.6	40.8	33.3	41.3	43.1
		T.P.	**0.591**	0.590	0.530	0.531	0.532	0.582	0.532	0.566	0.567	0.457	0.570	0.479	0.569
	5000	μ	**3.7**	3.8	330.3	206.3	220.6	22.4	313.3	123.7	118.0	683.1	110.0	586.4	114.8
		σ	3.4	3.3	84.2	85.9	86.2	36.7	56.1	77.1	90.9	84.7	86.1	81.2	93.0
		T.P.	**0.595**	0.595	0.529	0.554	0.551	0.591	0.533	0.571	0.572	0.459	0.573	0.478	0.572

Modeling departure and choice probabilities. For every LLM j and raw prompt ξ, we denote the provided performance score and normalized inference cost as $\text{perf}_j(\xi) \in [0, 1]$ and $\text{cost}_j(\xi) \in [0, 1]$, respectively. The performance score measures the response quality of model j on prompt ξ; for example, a binary score gives an outcome in $\{0, 1\}$, while a graded score represents the expected quality level. To obtain a cost-aware departure probability, we first define the raw utility as $u_j^{\text{raw}}(\xi) = \text{perf}_j(\xi) - \rho \cdot \text{cost}_j(\xi)$. Since $u_j^{\text{raw}}(\xi)$ need not lie in a valid probability range, we apply min-max normalization over models and then rescale it to a clipped interval $[r_{\text{lo}}, r_{\text{hi}}] \subset (0, 1)$, where $r_{\text{lo}} = 0.1$ and $r_{\text{hi}} = 0.99$ in our experiments. We use the resulting $u_j(\xi) \in (0, 1)$ as the departure probability, i.e., a job departs the queue when the served response is accepted, and convert it to the corresponding MNL choice probability as described in Section B.5.

Performance comparison. We benchmark ACQB and ACQB-CL under varying load conditions. We set the arrival rates to $\lambda \in \{0.7, 0.8, 0.9\}$ for $K = 1$ and $\lambda \in \{0.75, 0.85, 0.95\}$ for $K = 2$. We also compare with offline-trained routing baselines combined with FIFO scheduling: CARROT (Somerstep et al., 2025), KNN (Hu et al., 2024), MLP (Hu et al., 2024), ROUTERDC (Chen et al., 2024), UMR (Jitkrittum et al., 2025), and ZERO (Hu et al., 2024). Details of these baselines

are provided in Section B.3. Figure 4 shows the queue length and cumulative regrets on SPROUT-O3MINI. ACQB and ACQB-CL have lower queue length regret than the baselines across arrival rates, and ACQB-CL also has lower cumulative regret than ACQB in most settings. Table 2 reports the queue-length gap and throughput on all three offline routing datasets. ACQB-CL has the lowest mean queue-length gap in all reported cases, and ACQB has the second-lowest gap in most cases. Additional results are provided in Section C.3.

5.3 Experiments On A Real LLM Conversation Dataset

We additionally test the scheduling component on real user conversation logs from WILDCHAT-1M (Zhao et al., 2024). Since publicly available logs from real multi-LLM services are unavailable, we use the conversations served by `gpt-4-1106-preview` and hold routing fixed.

Retrials are not explicitly labeled in WILDCHAT-1M. We infer retrial chains by sorting conversations by inferred user arrival time and linking two consecutive unresolved conversations from the same hashed IP when their first prompts are identical or have cosine similarity at least 0.995 under `paraphrase-multilingual-mpnet-base-v2` embeddings. The user arrival time is estimated by subtracting the response time from the logged response timestamp.

We replay the inferred chains in a discrete-time single-server queue with arrival rate 0.7. A served job departs if it is the last node in its chain, and otherwise its next retrial appears in the next time slot. Since most chains have no retrial, we randomly subsample chains of length one and keep all chains of length at least two. We compare ACQB with FIFO, round-robin scheduling over the least-served queries (RR), random scheduling, and shortest-job-first (SJF) over five runs. Further implementation details are given in Section B.4. As shown in Figure 5 and Table 3, ACQB has lower queue length regret and higher throughput than the other scheduling rules. At $T = 25{,}000$, the mean queue length gap is 2.4 for ACQB and 124.6 for RR.

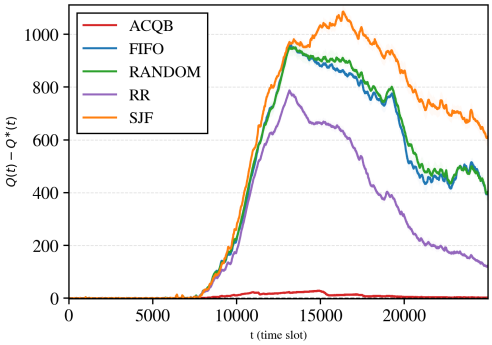


Figure 5: Scheduling evaluation on WILDCHAT-1M.

Table 3: Scheduling ablation results on WILDCHAT-1M.

Time	Stat.	ACQB	FIFO	RR	RANDOM	SJF
12500	μ	**18.6**	800.8	666.4	790.6	845.0
	σ	2.5	19.3	16.4	28.6	16.2
	T.P.	**0.520**	0.457	0.468	0.458	0.454
25000	μ	**2.4**	405.4	124.6	405.0	614.6
	σ	0.5	22.9	9.3	29.1	31.4
	T.P.	**0.584**	0.568	0.579	0.568	0.559

6 Conclusion

In this work, we introduce contextual queueing bandits with MNL feedback (CQB-MNL), an online learning framework for joint LLM routing and scheduling from user retrials. We propose anytime CQB (ACQB), which achieves sublinear cumulative regret and decaying queue length regret without prior knowledge of the time horizon or the traffic slackness parameter. To make the method practical for heterogeneous LLMs, we use disjoint parameterization and refine query embeddings through contrastive learning. Experiments on synthetic data, offline routing datasets, and real user conversation logs show that ACQB improves routing, scheduling, and queue stability against online and offline-trained baselines. Our feedback model focuses on retrials captured through the MNL outside option; extending the model to richer retrial mechanisms involving latency, interface design, reformulation effort, and user patience is an important direction for future work.

References

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp.

3691–3699. PMLR, 2021.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming {Throughput-Latency} tradeoff in {LLM} inference with {Sarathi-Serve}. In *18th USENIX symposium on operating systems design and implementation (OSDI 24)*, pp. 117–134, 2024.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Ruicheng Ao, Gan Luo, David Simchi-Levi, and Xinshang Wang. Optimizing llm inference: Fluid-guided online scheduling with memory constraints. *arXiv preprint arXiv:2504.11320*, 2025.

Seoungbin Bae and Dabeen Lee. Neural logistic bandits. *arXiv preprint arXiv:2505.02069*, 2025.

Seoungbin Bae, Garyeong Kang, and Dabeen Lee. Queue length regret bounds for contextual queueing bandits. *arXiv preprint arXiv:2601.19300*, 2026.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.

Chao-Kai Chiang, Takashi Ishida, and Masashi Sugiyama. Llm routing with dueling feedback. *arXiv preprint arXiv:2510.00841*, 2025.

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*, pp. 181–190, 2021.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834*, 2024.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Yichao Fu, Siqi Zhu, Runlong Su, Aurick Qiao, Ion Stoica, and Hao Zhang. Efficient llm scheduling by learning to rank. *Advances in Neural Information Processing Systems*, 37:59006–59029, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Patrick Jaillet, Jiashuo Jiang, Konstantina Mellou, Marco Molinaro, Chara Podimata, and Zijie Zhou. Online scheduling for llm inference with kv cache constraints. *arXiv preprint arXiv:2502.07115*, 2025.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, et al. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*, 2025.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pp. 5148–5157. PMLR, 2021.

Jung-hun Kim and Min-hwan Oh. Queueing matching bandits with preference feedback. *Advances in Neural Information Processing Systems*, 37:64649–64702, 2024.

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

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Joongkyu Lee and Min-hwan Oh. Improved online confidence bounds for multinomial logistic bandits. *arXiv preprint arXiv:2502.10020*, 2025.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Qingkai Liang and Eytan Modiano. Minimizing queue length regret under adversarial network models. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–32, 2018.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1964–1974, 2024.

Michael Mitzenmacher and Rana Shahout. Queueing, predictions, and large language models: Challenges and open problems. *Stochastic Systems*, 15(3):195–219, 2025.

Alireza Mohammadshahi, Arshad Rafiq Shaikh, and Majid Yazdani. Routoo: Learning to route to large language models effectively. *arXiv preprint arXiv:2401.13979*, 2024.

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

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Sentence-Transformers. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021. Hugging Face model card.

Rana Shahout, Eran Malach, Chunwei Liu, Weifan Jiang, Minlan Yu, and Michael Mitzenmacher. Don't stop me now: Embedding based scheduling for llms. *arXiv preprint arXiv:2410.01035*, 2024.

Reza Shirkavand, Shangqian Gao, Peiran Yu, and Heng Huang. Cost-aware contrastive routing for llms. *arXiv preprint arXiv:2508.12491*, 2025.

Seamus Somerstep, Felipe Maia Polo, Allysson Flavio Melo de Oliveira, Prattyush Mangal, Mírian Silva, Onkar Bhardwaj, Mikhail Yurochkin, and Subha Maity. Carrot: A cost aware rate optimal router. *arXiv preprint arXiv:2502.03261*, 2025.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Mevan Wijewardena, Kamiar Asgari, and Michael J Neely. Bandit-based rate adaptation for a single-server queue. *arXiv preprint arXiv:2512.12016*, 2025.

Fangzhou Wu and Sandeep Silwal. Efficient training-free online routing for high-volume multi-llm serving. *arXiv preprint arXiv:2509.02718*, 2025.

Mingyan Yang, Guanjie Wang, Manqi Luo, Yifei Liu, Chen Chen, Han Zhao, Yu Feng, Quan Chen, and Minyi Guo. Justitia: Fair and efficient scheduling for llm applications. *arXiv preprint arXiv:2510.17015*, 2025.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. *arXiv preprint arXiv:2507.11847*, 2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. Embedllm: Learning compact representations of large language models. *arXiv preprint arXiv:2410.02223*, 2024.

A Related Work

LLM routing and scheduling. LLM routing selects models to balance quality and cost, while LLM scheduling determines the service order of pending requests to minimize latency. Although substantial research has relied on supervised learning with offline datasets, such as Ong et al. (2024); Feng et al. (2024); Lu et al. (2024); Somerstep et al. (2025) for routing and Fu et al. (2024); Shahout et al. (2024); Yang et al. (2025) for scheduling, recent works have emerged that utilize online learning to address the high costs of offline annotation and the non-stationarity of real-world environments (Mohammadshahi et al., 2024; Wu & Silwal, 2025; Chiang et al., 2025; Jitkrittum et al., 2025; Jaillet et al., 2025; Ao et al., 2025). However, these methods neglect the interactive nature of LLMs, either ignoring congestion caused by retrials or disrupting user experience by requiring explicit feedback.

Queueing bandits. The queueing bandit framework, originally designed to address routing problems in systems with unknown service rates, was introduced by Krishnasamy et al. (2016). This seminal work has inspired a significant body of research establishing queue stability—ensuring that expected queue lengths remain bounded—and deriving queue length regret bounds across various settings (Liang & Modiano, 2018; Stahlbuhk et al., 2021; Choudhury et al., 2021; Krishnasamy et al., 2021; Wijewardena et al., 2025). More recently, this framework has been extended to the contextual setting, incorporating job-specific features and adopting logistic models for departure dynamics (Kim & Oh, 2024; Bae et al., 2026). However, the approach of Kim & Oh (2024) is limited by a fixed context assumption. While Bae et al. (2026) accommodate heterogeneous contexts, their reliance on the time horizon and traffic slackness parameter hinders practical deployment.

Logistic Bandits. Starting with the seminal work on generalized linear bandits (Filippi et al., 2010), the (multinomial) logistic bandit problem has been extensively studied due to its effectiveness in modeling discrete user choices and categorical feedback (Li et al., 2017; Agrawal et al., 2019; Oh & Iyengar, 2019; Jun et al., 2021; Abeille et al., 2021; Bae & Lee, 2025; Lee & Oh, 2025; Zhang et al., 2025). In this work, we adopt the MNL model for queue dynamics (Kim & Oh, 2024; Bae et al., 2026) and directly apply these analytical tools to bound the departure rate gap between the optimal policy and our proposed policy.

B Experiment Details

B.1 Algorithm Adaptation For Disjoint Parameterization

In this section, we present the adaptation of Algorithm 1 to the disjoint parameterization setting, as outlined in Algorithm 2. Recall that while the query context x_t is shared, each server j is governed by a distinct unknown parameter vector θ_j^*. Accordingly, we define the collection of parameters as $\Theta^* = [\theta_1^*, \theta_2^*, \dots, \theta_N^*] \in \mathbb{R}^{d\times N}$. When a user with context x is presented with an assortment S, the probability $p_j(x, S, \Theta^*)$ of selecting item j and the corresponding departure rate $R(x, S, \Theta^*)$ are defined as

$$p_j(x, S, \Theta^*) := \frac{\exp(x^\top \theta_j^*)}{1 + \sum_{j'\in S} \exp(x^\top \theta_{j'}^*)} \quad \text{for } j \in S,$$

$$p_0(x, S, \Theta^*) := \frac{1}{1 + \sum_{j'\in S} \exp(x^\top \theta_{j'}^*)} \quad \text{for } j = 0,$$

$$R(x, S, \Theta^*) := \sum_{j\in S} \frac{\exp(x^\top \theta_j^*)}{1 + \sum_{j'\in S} \exp(x^\top \theta_{j'}^*)} = \sum_{j\in S} p_j(x, S, \Theta^*).$$

Under this disjoint model, the agent maintains separate statistics for each server j: a design matrix $V_{t,j}$, a maximum likelihood estimator $\widehat{\theta}_{t,j}$, and a confidence radius $\alpha_{t,j}$. For the design matrices, we initialize $\{V_{t,j}\}_{j=1}^N = \lambda_0 \mathbf{I}$. In each round, we update only the matrices corresponding to the servers included in the assortment S_t, such that $V_{t,j} \leftarrow V_{t-1,j} + x_t x_t^\top$ for all $j \in S_t$ (Line 15). Similarly,

the confidence radius is updated only for $j \in S_t$ (Line 14) as follows:

$$\alpha_{t,j} \leftarrow \frac{\kappa}{2}\sqrt{d\log\left(1+\frac{K\sum_{i=1}^{t}\mathbf{1}\{j\in S_i\}}{d\lambda_0}\right)+4\log\left(\sum_{i=1}^{t}\mathbf{1}\{j\in S_i\}\right)}+\kappa\sqrt{\lambda_0}, \tag{2}$$

where we replace the time horizon t in the original α_t definition with the effective number of times server j has been selected, i.e., $\sum_{i=1}^{t}\mathbf{1}\{j \in S_i\}$.

Regarding the maximum likelihood estimators, we define the regularized negative log-likelihood function as

$$\mathcal{L}_t(\Theta) := \frac{\lambda_0}{2}\sum_{j=1}^{N}\|\theta_j\|_2^2 - \sum_{i=1}^{t}\sum_{j\in S_i\cup\{0\}} y_{ij}\log p_j(x, S_i, \Theta).$$

Then, we update the estimators $\widehat{\theta}_{t,j}$ for the relevant servers $j \in S_t$ (Line 14) by solving the first order stationary point $\nabla_{\theta_j}\mathcal{L}_t(\Theta) = 0$, which is given by

$$0 = \lambda_0\theta_j - \sum_{i=1}^{t}\mathbf{1}\{j\in S_i\}(y_{ij} - p_j(x, S_i, \Theta)). \tag{3}$$

Finally, utilizing these separate statistics, the agent selects a query and an assortment as follows: If $A(t-1) = 1$ and $E(t-1) = 1$, we proceed with the random exploration round (Lines 4–5). Otherwise, for each $j \in [N]$, we sample $\{\widetilde{\theta}_{t-1,j}^{(i)}\}_{i=1}^{M}$ as shown in Line 8. We then compute the optimistic departure rate estimate as

$$\widetilde{u}_{tj}(x) = \max_i x^{\mathsf{T}}\widetilde{\theta}_{t-1,j}^{(i)}, \quad \widetilde{R}(x,S) = \sum_{k\in S}\frac{\exp(\widetilde{u}_{tk}(x))}{1+\sum_{j\in S}\exp(\widetilde{u}_{tj}(x))},$$

and select the pair maximizing this value (Line 10).

B.2 Utility-Aligned Query Embeddings Via Contrastive Learning

This section details the utility-based contrastive learning framework for ACQB-CL, extending the overview in Section 4.2. The goal of this offline phase is to train the projection head $B(\cdot;\theta)$ appended to the frozen encoder $E(\cdot)$. Leveraging an offline dataset containing utility values (i.e., performance and cost) for all servers, we aim to structure the embedding space such that queries with similar routing utilities are clustered together, while those with dissimilar utilities are separated.

We begin the process by constructing a balanced offline dataset from the representation learning split. Specifically, we employ a model-balanced sampling strategy. For each available prompt ξ, we compute the utility vector $\{u_j(\xi)\}_{j=1}^{N}$ across all models and identify the optimal model that yields the highest utility. We then group prompts according to their optimal model index. To ensure diversity, we uniformly sample a fixed number of prompts (e.g., 5 prompts in our experiments for SPROUT data) from each group. This procedure ensures that the offline training set is composed of samples selected evenly across all candidate LLMs.

Offline Training Algorithm. After sampling, we form the offline dataset as $\mathcal{D}_{\text{off}} = \{(\xi_{(i)}, u_{(i)})\}_{i=1}^{n}$, where $\xi_{(i)}$ denotes the selected prompt and $u_{(i)}$ represents its utility vector across all models. We use the frozen encoder outputs $x_{(i)} = E(\xi_{(i)})$ and the utility vectors $u_{(i)}$ to train the two-layer MLP projection head $B(\cdot;\theta)$. Algorithm 3 summarizes the procedure.

At the beginning of each epoch, we compute the projected representation $z_{(i)} = B(x_{(i)};\theta)$ for each offline prompt. To determine similarity targets, we first mean-center each utility vector across models:

$$\bar{u}_{(i)} \leftarrow u_{(i)} - \text{mean}(u_{(i)}).$$

We then compute the pairwise utility similarity $c_{(i,j)}$ for all $i \neq j$ using the cosine similarity of these centered vectors:

$$c_{(i,j)} \leftarrow \cos(\bar{u}_{(i)}, \bar{u}_{(j)}) := \frac{\bar{u}_{(i)}^{\top}\bar{u}_{(j)}}{\|\bar{u}_{(i)}\|_2\,\|\bar{u}_{(j)}\|_2}.$$

Algorithm 2 ACQB/ACQB-CL with Disjoint Parameterization

initialize design matrix $\{V_{0,j}\}_{j=1}^{N} = \lambda_0 \mathbf{I}$, assortment size K, sample size M, combination set $\mathcal{C} = \{S \subset [N] : |S| = K\}$, combination counter $c = 0$, exploration parameter $\eta(t)$, confidence radius $\{\alpha_{t,j}\}_{j=1}^{N}$, projection head $B(\cdot;\theta)$

Define context map $\phi(x) = \begin{cases} x & \text{if ACQB} \\ B(x;\theta) & \text{if ACQB-CL (with } \theta \text{ from Algorithm 3)} \end{cases}$
for $t = 1, \ldots$ **do**
  if $A(t-1) = 1$ **and** $E(t-1) = 1$ **then**
  // on arrivals, $\eta(t)$-exploration
    Set $x_t \leftarrow x^{(t-1)}$, $S_t \leftarrow \mathcal{C}[c+1]$
    $c \leftarrow c + 1 \pmod{|\mathcal{C}|}$
  else
    for $j \in [N]$ **do**
      Sample $\{\widetilde{\theta}_{t-1,j}^{(i)}\}_{i=1}^{M} \sim \mathcal{N}(\widehat{\theta}_{t-1,j}, \alpha_{t-1,j}^2 V_{t-1,j}^{-1})$
    end for
    Set $x_t, S_t \leftarrow \arg\max_{x \in \mathcal{X}_t, S \in \mathcal{C}} \widetilde{R}(\phi(x), S)$
  end if
  Assign x_t to S_t, receive $y_t = (y_{t0}, y_{t1}, \ldots, y_{tK})$
  for $j \in [N]$ **do**
    if $j \in S_t$ **then**
      Update $\widehat{\theta}_{t,j}$ as in (3), $\alpha_{t,j}$ as in (2)
      $V_{t,j} \leftarrow V_{t-1,j} + \phi(x_t)\phi(x_t)^{\mathsf{T}}$
    else
      $\widehat{\theta}_{t,j} \leftarrow \widehat{\theta}_{t-1,j}$, $\alpha_{t,j} \leftarrow \alpha_{t-1,j}$, $V_{t,j} \leftarrow V_{t-1,j}$
    end if
  end for
  Sample $E(t) \sim \mathrm{Bern}(\eta(t))$
end for

Algorithm 3 Utility-Based Contrastive Learning

initialize offline dataset $\mathcal{D}_{\text{off}} = \{(\xi_{(i)}, u_{(i)})\}_{i=1}^{n}$, frozen encoder $E(\cdot)$, projection head $B(\cdot;\theta)$, thresholds $(\iota_{\text{pos}}, \iota_{\text{neg}})$, temperature τ_{temp}, negative cap K_{neg}, learning rate η
for epoch $= 1, \ldots, e$ **do**
  $x_{(i)} \leftarrow E(\xi_{(i)})$, $z_{(i)} \leftarrow B(x_{(i)};\theta)$, $\bar{u}_{(i)} \leftarrow u_{(i)} - \mathrm{mean}(u_{(i)})$ for all i
  $c_{(i,j)} \leftarrow \cos(\bar{u}_{(i)}, \bar{u}_{(j)})$ for all $i \neq j$
  for $i = 1, \ldots, n$ **do**
    Set $P(i) = \{j \neq i : c_{(i,j)} > \iota_{\text{pos}}\}$, $N(i) = \{j \neq i : c_{(i,j)} < \iota_{\text{neg}}\}$
    Pick $j^+ \in \arg\max_{j \in P(i)} c_{(i,j)}$
    Pick $\mathcal{N}_{(i)} \subseteq N(i)$ with $|\mathcal{N}_{(i)}| \leq K_{\text{neg}}$
    Set $\ell_{(i)}(\theta)$ as in (4)
  end for
  $\theta \leftarrow \theta - \eta \nabla_\theta \sum_{(i):\,|P((i))|>0} \ell_{(i)}(\theta)$
end for
return θ

This metric is used for selecting positive and negative pairs. Given a pair of thresholds $(\iota_{\text{pos}}, \iota_{\text{neg}})$, we define the candidate positive set $P(i)$ and negative set $N(i)$ as

$$P(i) = \{j \neq i : c_{(i,j)} > \iota_{\text{pos}}\}, \quad N(i) = \{j \neq i : c_{(i,j)} < \iota_{\text{neg}}\}.$$

Any candidate j with similarity falling between the thresholds (i.e., $\iota_{\text{neg}} \leq c_{(i,j)} \leq \iota_{\text{pos}}$) is ignored. From the positive set $P(i)$, we select the single best positive example j^+ that maximizes utility similarity, i.e. $j^+ \in \arg\max_{j \in P(i)} c_{(i,j)}$. For the negative set, we construct a subset $\mathcal{N}_{(i)} \subseteq N(i)$ to limit computational cost. Specifically, we select up to K_{neg} hard negatives (those with the smallest similarity values $c_{(i,j)}$). If $P(i)$ is empty or no negatives remain, we skip the update for query i.

Using the selected positive j^+ and negative set $\mathcal{N}_{(i)}$, we compute the representation similarity $\mathrm{sim}(z_{(i)}, z_{(j)}) = z_{(i)}^\top z_{(j)}$ and define the InfoNCE loss with temperature τ_{temp} as

$$\ell_{(i)}(\theta) = -\log \frac{\exp(\mathrm{sim}(z_{(i)}, z_{(j^+)})/\tau_{\mathrm{temp}})}{\exp(\mathrm{sim}(z_{(i)}, z_{(j^+)})/\tau_{\mathrm{temp}}) + \sum_{j \in \mathcal{N}_{(i)}} \exp(\mathrm{sim}(z_{(i)}, z_{(j)})/\tau_{\mathrm{temp}})}. \tag{4}$$

Finally, we update the projection head parameters θ via gradient descent over all valid queries:

$$\theta \leftarrow \theta - \eta \nabla_\theta \sum_{i:\,|P(i)|>0} \ell_{(i)}(\theta).$$

After offline pretraining, we freeze the projection head $B(\cdot;\theta)$ and proceed to the online learning stage described in Algorithm 2.

Experimental Settings. We use $\tau_{\mathrm{temp}} = 0.07$ and a negative cap $K_{\mathrm{neg}} = 64$. The projection head is a two-layer MLP with an output dimension of 384, matching the encoder dimension. We set distinct threshold pairs $(\iota_{\mathrm{pos}}, \iota_{\mathrm{neg}})$ for each dataset: $(0.6, 0.3)$ for SPROUT-O3MINI, $(0.4, 0.1)$ for EMBEDLLM, and $(0.5, 0.3)$ for ROUTERBENCH. For contrastive pretraining of B in an offline setting, we use 5 prompts per model for SPROUT-O3MINI and train for 10 epochs; for EMBEDLLM, we use 10 prompts per model (30 prompts per epoch) and train for 50 epochs.

B.3 BASELINE ALGORITHMS

We compare our proposed ACQB and ACQB-CL algorithms against the following baselines:

(i) **Optimal Policy:** The optimal policy has full access to the true set of disjoint parameters $\Theta^* = \{\theta_j^*\}_{j=1}^N$. In every round t, given the set of backlogged queries $\mathcal{X}_t^*$, it selects the query-assortment pair (x^*, S^*) that maximizes the expected departure rate:

$$(x^*, S^*) \in \underset{x \in \mathcal{X}_t^*, S \in \mathcal{C}}{\arg\max}\, R(x, S, \Theta^*).$$

(ii) **Random Policy (RAND):** In every round t, the random policy selects a query x_t uniformly at random from the current backlog and chooses an assortment S_t uniformly at random from the feasible set $\mathcal{C}$.

- **Random Router Policy (RAND_ROUT):** In every round t, the random router policy selects a query x_t that arrived first among the current backlog and chooses an assortment S_t uniformly at random from the feasible set $\mathcal{C}$.

(iii) **Q-UCB** (Krishnasamy et al., 2021, Algorithm 1): This is a MAB-based algorithm designed for the single-item setting ($K = 1$). Since this algorithm treats all queries as identical ignoring their contexts, it employs a first-in-first-out (FIFO) scheduling rule, simply selecting the oldest job in the queue as x_t. The algorithm explores with probability $\min\{1, 3N(\log^2 t)/t\}$. For the assortment selection, it chooses the item that maximizes the upper confidence bound (UCB):

$$S_t = \underset{S \in \mathcal{C}, |S|=1}{\arg\max} \left(\widehat{\mu}_S(t) + \sqrt{\frac{\log^2 t}{2T_S(t-1)}} \right),$$

where $\widehat{\mu}_S(t) = \sum_{i=1}^{t-1} \mathbf{1}\{S_i = S\} y_i / T_S(t-1)$ is the empirical mean reward and $T_S(t-1) = \sum_{i=1}^{t-1} \mathbf{1}\{S_i = S\}$is the number of times assortment S has been played up to time $t-1$.

(iv) **Q-ThS** (Krishnasamy et al., 2021, Algorithm 2): Similar to Q-UCB, this is an MAB-based approach for $K = 1$ using FIFO scheduling. It explores with the same probability schedule as Q-UCB. For every assortment selection S_t, it samples a departure rate estimate from a Beta posterior:

$$\widetilde{r}_t(x_t, S) \sim \mathrm{Beta}\Big(\widehat{\mu}_S(t) T_S(t-1) + 1, (1 - \widehat{\mu}_S(t)) T_S(t-1) + 1\Big).$$

It then selects the assortment S_t that maximizes this sampled value, i.e. $S_t := \arg\max_{K=1, S \in \mathcal{C}} \widetilde{r}_t(x_t, S)$, where we use the same definitions of $\widehat{\mu}_S(t)$ and $T_S(t-1)$ as Q-UCB.

(v) **CQB-ε** (Adapted from Bae et al. (2026)): This baseline is an adaptation of the contextual queueing bandit algorithm from Bae et al. (2026). While the original algorithm was developed as a UCB-based approach under logistic feedback, we modified it to utilize Thompson Sampling with MNL feedback. This adaptation ensures a fair and consistent comparison with our proposed framework under the same environmental assumptions.

Let τ denote the length of the pure-exploration phase. If $A(t-1) = 1$ and the agent is either in the exploration phase ($t \leq \tau$) or triggers a random exploration with probability $T^{-1/2}$, it selects $x_t \leftarrow x^{(t-1)}$ (the most recent job) and chooses S_t in a round-robin manner. Otherwise, it follows the optimistic rule described in ACQB (Line 6–9 of Algorithm 2).

Notice that the theoretical requirement for the exploration length of CQB-ε, $\tau = \mathcal{O}(\frac{d\log(T)}{\sigma_0^4\epsilon^2})$, is often prohibitively large for practical experiments. Therefore, we adopt distinct practical heuristics depending on the experimental setting. For synthetic data experiments, following previous work (Bae et al., 2026), we set a fixed exploration length of $\tau = T/10$. For offline routing datasets, we determine τ by aligning it with the stabilization point of our exploration rate $\eta(t)$. Specifically, we set $\tau = \min\{t : c_1(t+1)^{-1/2} \leq 1\}$, ensuring that the pure-exploration phase ends when ACQB's exploration probability drops below 1.

Scheduling variants. To isolate the effect of scheduling from routing, we evaluate variants that keep the ACQB routing rule fixed while replacing only the query-selection rule. In the synthetic experiments, all variants use the same exploration schedule, arrival process, MNL feedback generation, and Thompson-sampling routing rule as ACQB. The difference appears only in exploitation rounds: FIFO selects the oldest query, round-robin selects uniformly among the least-served queries, and random scheduling selects a query uniformly from the current queue. After the query is selected, each variant chooses the assortment by the same optimistic ACQB routing rule for that query.

Offline-trained routing baselines. For simulations on offline routing datasets, we further compare against offline-trained routing baselines combined with FIFO scheduling. We include CARROT (Somerstep et al., 2025), KNN (Hu et al., 2024), MLP (Hu et al., 2024), ROUTERDC (Chen et al., 2024), UMR (Jitkrittum et al., 2025), and ZERO (Hu et al., 2024). For each run, the baselines are trained on the offline split excluding all evaluation prompts and then evaluated on the same online prompt pool as ACQB. They use the same query embeddings and utility target $\mathrm{perf}_j(\xi) - \rho\mathrm{cost}_j(\xi)$: ZERO chooses the best fixed model from the training split; KNN and MLP predict per-model performance from the query embedding and subtract the model cost; CARROT predicts both performance and cost; UMR clusters query embeddings and routes by cluster-level error and cost; and ROUTERDC trains a neural router with query and model embeddings. Since these routers are not designed for queueing systems with retrial feedback, we combine each router with FIFO scheduling and use the same MNL departure-probability construction as in Section B.5.

B.4 WildChat-1M Experiment Details

For the WildChat-1M experiment, we use conversations served by `gpt-4-1106-preview`. Since the logged timestamp is the assistant response timestamp, we estimate the user arrival time by subtracting $2.84 + \mathrm{tokens}/30$ seconds from the timestamp of the first assistant response. We then sort conversations by the inferred arrival time and construct retrial chains separately for each hashed IP. A conversation is linked to the previous unresolved conversation from the same hashed IP if the first user prompts are identical or their normalized `paraphrase-multilingual-mpnet-base-v2` embeddings have cosine similarity at least 0.995, and if the inferred arrival time is after the previous assistant timestamp. Since most conversations do not have a retrial, we randomly keep a subset of singleton chains so that the singleton-chain fraction is 0.8, while keeping all non-singleton chains. For the queue replay, root conversations are assigned to discrete time bins according to the inferred arrival times with arrival rate 0.7 and service capacity one. When a served node is not terminal in its chain, the next node enters the queue in the next time slot; otherwise, the job departs. ACQB uses 32-dimensional projected prompt embeddings with the chain state variables. For SJF, we set the job length to the number of tokens in the first assistant response.

B.5 Modeling Departure And Choice Probabilities For Offline Routing Dataset Experiments

This section details how we transform raw utility scores into departure and choice probabilities, complementing the experimental setup in Section 5.2. Recall that for a raw prompt ξ and model j, we define the (raw) utility as $u_j^{\text{raw}}(\xi) = \text{perf}_j(\xi) - \rho \cdot \text{cost}_j(\xi)$, where we use $\rho = 5$ in our experiments. Since this value may fall outside the feasible probability range $[0, 1]$ depending on ρ, we apply a two-step transformation. First, we apply min-max normalization over the models to obtain a value in $[0, 1]$:

$$u_j^{\text{norm}}(\xi) = \frac{u_j^{\text{raw}}(\xi) - \min_k u_k^{\text{raw}}(\xi)}{\max_k u_k^{\text{raw}}(\xi) - \min_k u_k^{\text{raw}}(\xi)}.$$

Next, to prevent extreme probabilities (0 or 1) that can cause numerical instability, we linearly rescale this normalized value to a bounded range $[r_{\text{lo}}, r_{\text{hi}}] \subset (0, 1)$. We define the final departure probability $u_j(\xi)$ as:

$$u_j(\xi) := r_{\text{lo}} + (r_{\text{hi}} - r_{\text{lo}}) \cdot u_j^{\text{norm}}(\xi),$$

where we set $r_{\text{lo}} = 0.1$ and $r_{\text{hi}} = 0.99$ in our experiments.

We define the true departure probability for prompt ξ on model j as this transformed value $u_j(\xi)$. To be consistent with our MNL choice framework, let $x = E(\xi)$ denote the query context. We construct the latent reward $x^\top \theta_j^*$ such that its logistic function matches the departure probability:

$$u_j(\xi) = \frac{\exp(x^\top \theta_j^*)}{1 + \exp(x^\top \theta_j^*)} \quad \Longrightarrow \quad \exp(x^\top \theta_j^*) = \frac{u_j(\xi)}{1 - u_j(\xi)}.$$

Substituting this relationship into the MNL choice model for an assortment S, we derive the final choice probabilities entirely in terms of the derived utilities:

$$p(j \mid x, S) = \frac{\exp(x^\top \theta_j^*)}{1 + \sum_{k \in S} \exp(x^\top \theta_k^*)} = \frac{\frac{u_j(\xi)}{1-u_j(\xi)}}{1 + \sum_{k \in S} \frac{u_k(\xi)}{1-u_k(\xi)}}.$$

C Additional Experimental Results

This section provides additional experiments omitted from the main text. We first report synthetic experiments with varying N and ϵ, then show the exploration-scheduling ablation, and finally provide additional simulations on offline routing datasets. All experiments were conducted on a server equipped with an AMD EPYC 9354 32-Core Processor, 251 GiB of RAM, and one NVIDIA RTX A6000 GPU.

C.1 Additional Synthetic Data Results

We use the same setup as Section 5.1 and vary one parameter at a time. Figure 6 changes the number of LLMs N, while Figure 7 changes the slackness parameter ϵ.

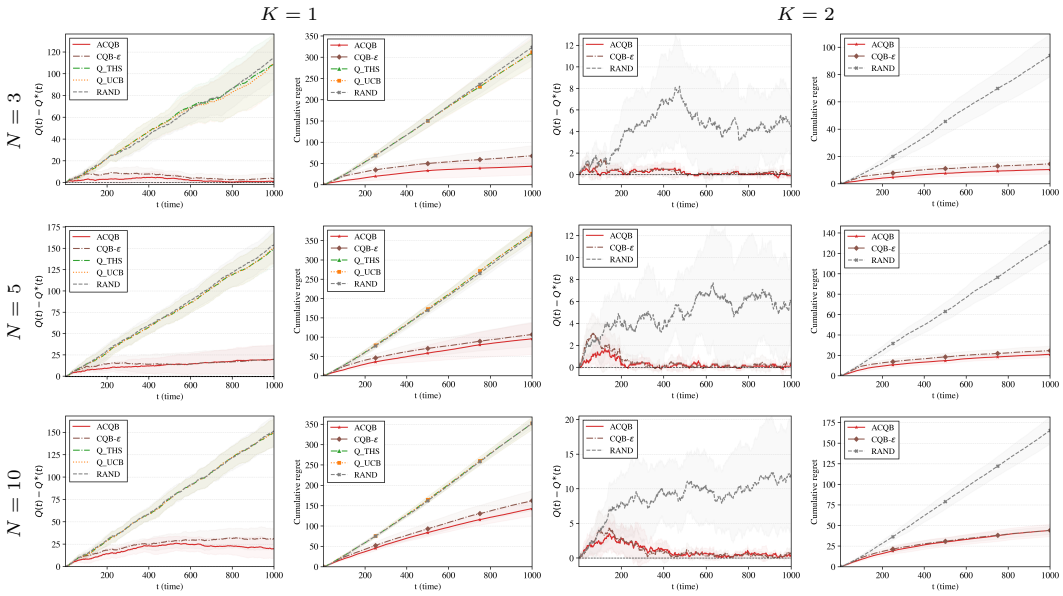


Figure 6: Queue length and cumulative regret on synthetic data with $\lambda = 0.7$, $\epsilon = 0.03$, and varying $N \in \{3, 5, 10\}$.

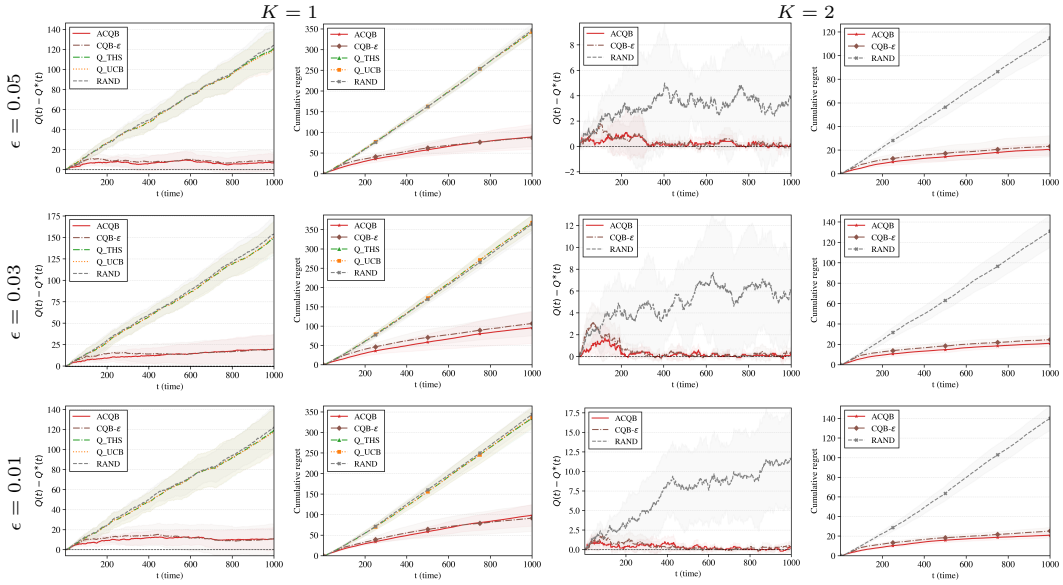


Figure 7: Queue length and cumulative regret on synthetic data with $\lambda = 0.7$, $N = 5$, and varying $\epsilon \in \{0.05, 0.03, 0.01\}$.

C.2 Additional Exploration Scheduling Results

We compare ACQB with heuristic exploration variants that trigger exploration independently of arrivals. Once exploration is triggered, the exploration query is selected by FIFO, round-robin over the least-served jobs, or random scheduling, while the routing rule is kept fixed. The results in Figure 8 show that arrival-based exploration remains stable across both $K = 1$ and $K = 2$.

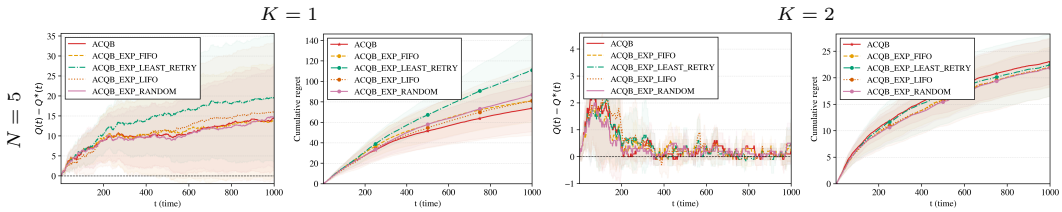


Figure 8: Queue length and cumulative regret for exploration-scheduling variants on synthetic data with $N = 5$, $\epsilon = 0.03$, and $K \in \{1, 2\}$.

C.3 Additional Offline Routing Dataset Results

We provide additional results on EmbedLLM and RouterBench. The figures report queue length and cumulative regret across arrival rates, and the tables report the queue length gap and throughput at intermediate and final time steps.

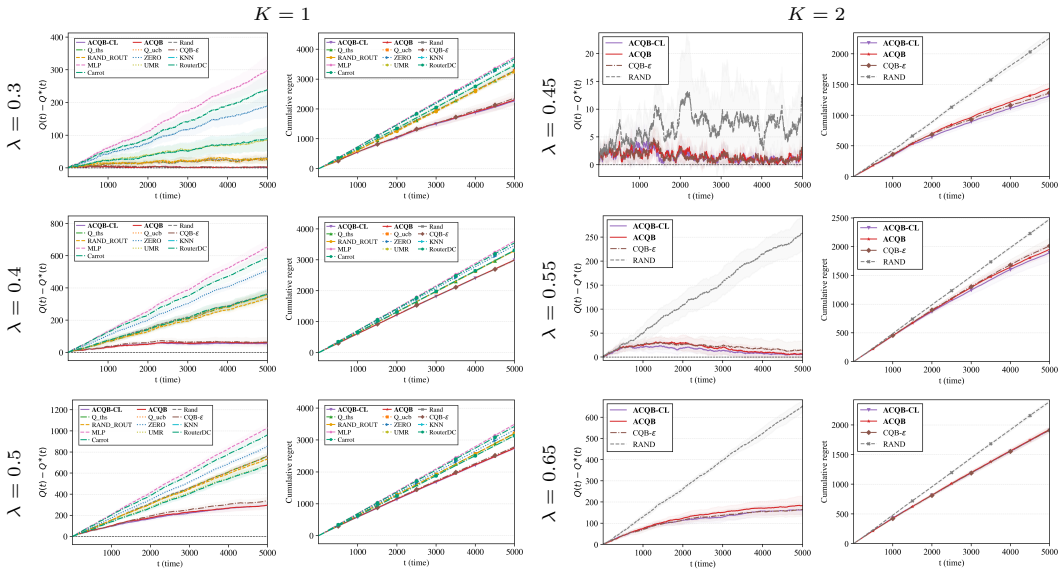


Figure 9: Queue length and cumulative regret on the EmbedLLM dataset.

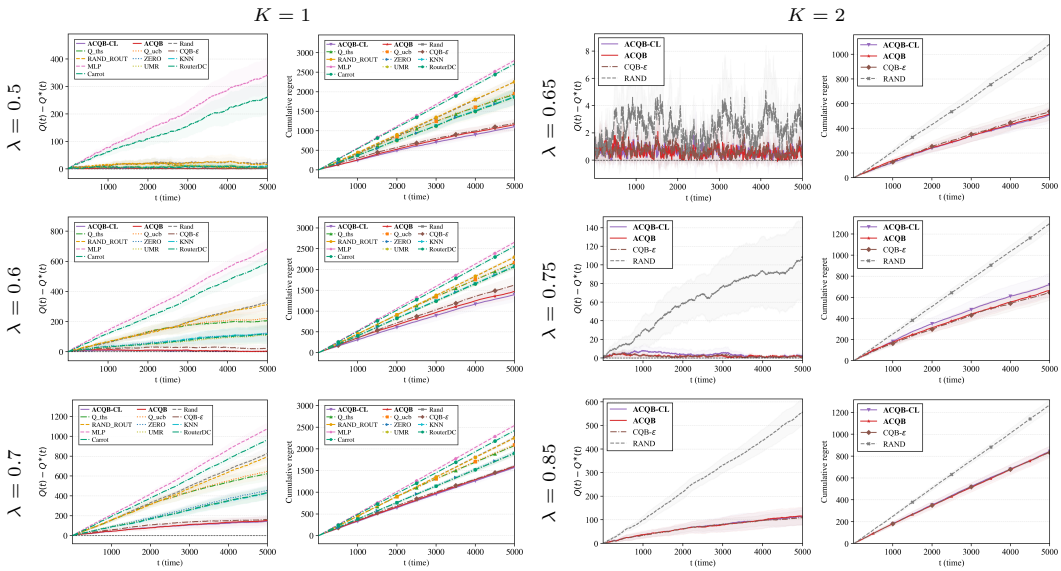


Figure 10: Queue length and cumulative regret on the ROUTERBENCH dataset.

Table 4: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 1$ (low-arrival setting: $\lambda_{\text{Sprout}} = 0.7$, $\lambda_{\text{RouterBench}} = 0.5$, $\lambda_{\text{EmbedLLM}} = 0.3$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ε	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	0.4	**0.2**	2.0	2.7	2.7	0.7	3.2	1.7	1.4	19.0	1.1	19.0	1.1
		σ	1.0	0.4	1.8	3.0	3.0	1.1	4.3	2.2	2.3	16.1	1.7	17.7	1.3
		T.P.	0.696	**0.696**	0.695	0.695	0.695	0.696	0.695	0.695	0.696	0.689	0.696	0.689	0.696
	5000	μ	**0.3**	0.6	5.0	0.6	1.3	0.7	1.6	0.5	0.6	28.3	1.0	34.1	0.9
		σ	0.5	1.3	7.3	0.8	1.3	0.8	1.3	1.6	1.1	25.1	1.9	30.4	1.7
		T.P.	**0.699**	0.698	0.698	0.698	0.698	0.698	0.698	0.699	0.698	0.693	0.698	0.692	0.698
EmbedLLM	2500	μ	3.9	3.4	14.9	16.1	16.1	**2.6**	20.3	98.7	47.3	146.6	50.6	117.6	47.3
		σ	4.4	3.8	14.2	14.4	14.4	3.0	12.1	25.9	31.1	36.3	23.9	29.8	31.1
		T.P.	0.297	0.297	0.292	0.292	0.292	**0.297**	0.290	0.259	0.280	0.240	0.278	0.251	0.280
	5000	μ	2.4	2.5	30.7	25.2	25.2	**1.8**	28.4	189.0	88.9	297.0	84.5	238.3	88.9
		σ	2.2	2.3	18.5	18.6	18.6	2.5	22.4	51.1	54.1	64.3	46.3	65.8	54.1
		T.P.	0.300	0.300	0.295	0.296	0.296	**0.300**	0.295	0.263	0.283	0.241	0.284	0.253	0.283
RouterBench	2500	μ	**0.8**	1.2	23.1	17.3	17.2	2.1	21.7	8.3	8.7	172.5	8.9	130.5	6.4
		σ	1.1	1.4	28.2	24.8	24.9	2.2	23.1	9.3	12.8	50.9	11.1	50.5	9.3
		T.P.	**0.501**	0.501	0.492	0.494	0.494	0.500	0.492	0.498	0.498	0.432	0.498	0.449	0.499
	5000	μ	**0.7**	1.2	20.5	4.0	3.2	1.5	15.2	4.2	9.1	341.5	4.5	260.2	5.9
		σ	1.3	1.4	24.4	3.7	4.5	2.3	26.4	6.0	14.1	93.2	6.9	91.8	8.1
		T.P.	**0.498**	0.498	0.494	0.497	0.498	0.498	0.495	0.497	0.496	0.430	0.497	0.446	0.497

Table 5: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 1$ (high-arrival setting: $\lambda_{\text{Sprout}} = 0.9$, $\lambda_{\text{RouterBench}} = 0.7$, $\lambda_{\text{EmbedLLM}} = 0.5$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ϵ	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	**100.6**	125.4	337.9	326.2	326.2	153.2	330.6	138.1	121.8	400.9	125.9	412.8	115.6
		σ	39.3	20.8	31.1	32.6	32.6	40.0	43.2	41.2	33.5	45.2	16.8	48.4	32.4
		T.P.	**0.856**	0.846	0.761	0.766	0.766	0.835	0.765	0.842	0.848	0.736	0.846	0.732	0.851
	5000	μ	**130.3**	172.6	679.1	598.4	599.1	241.4	656.4	276.5	256.3	803.1	261.1	826.2	231.3
		σ	64.4	40.5	49.6	42.9	46.6	85.1	57.9	63.3	42.5	82.3	31.2	72.6	54.3
		T.P.	**0.873**	0.864	0.763	0.779	0.779	0.851	0.768	0.843	0.848	0.738	0.847	0.734	0.853
EmbedLLM	2500	μ	**201.0**	208.7	393.1	386.5	386.5	232.3	381.0	436.1	340.4	523.1	350.2	492.2	340.4
		σ	30.1	29.1	19.9	21.5	21.5	35.4	37.1	24.1	35.8	36.1	21.9	35.2	35.8
		T.P.	**0.420**	0.417	0.343	0.346	0.346	0.407	0.348	0.326	0.365	0.292	0.361	0.304	0.365
	5000	μ	**294.8**	295.3	762.4	759.0	759.0	336.1	731.9	852.0	676.5	1028.3	680.1	961.5	676.5
		σ	48.7	55.7	38.9	47.1	47.1	54.7	42.7	39.1	49.4	65.2	42.5	76.9	49.4
		T.P.	**0.439**	0.439	0.346	0.346	0.346	0.431	0.352	0.328	0.363	0.293	0.362	0.306	0.363
RouterBench	2500	μ	**92.3**	94.3	407.2	376.4	377.0	123.1	393.0	227.4	209.3	524.1	211.4	469.8	212.8
		σ	47.8	45.2	59.7	69.6	68.5	58.2	46.4	41.7	53.2	44.5	45.9	33.3	53.4
		T.P.	**0.659**	0.658	0.533	0.545	0.545	0.646	0.539	0.605	0.612	0.486	0.611	0.508	0.611
	5000	μ	**143.1**	150.5	826.9	622.1	645.8	161.9	791.1	453.5	427.8	1075.0	429.1	962.0	431.9
		σ	83.5	64.5	89.4	105.3	99.0	51.6	76.6	75.0	92.7	85.5	93.5	77.8	99.4
		T.P.	**0.670**	0.668	0.533	0.574	0.569	0.666	0.540	0.608	0.613	0.483	0.613	0.506	0.612

Table 6: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 2$ (low-arrival setting: $\lambda_{\text{Sprout}} = 0.75$, $\lambda_{\text{RouterBench}} = 0.65$, $\lambda_{\text{EmbedLLM}} = 0.45$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ϵ	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	0.1	0.1	0.3	-	-	**0.0**	-	-	-	-	-	-	-
		σ	0.3	0.3	0.8	-	-	0.0	-	-	-	-	-	-	-
		T.P.	0.745	0.745	0.745	-	-	**0.745**	-	-	-	-	-	-	-
	5000	μ	**0.1**	0.3	0.3	-	-	0.1	-	-	-	-	-	-	-
		σ	0.3	0.7	0.7	-	-	0.3	-	-	-	-	-	-	-
		T.P.	**0.747**	0.747	0.747	-	-	0.747	-	-	-	-	-	-	-
EmbedLLM	2500	μ	1.5	1.3	8.6	-	-	**1.2**	-	-	-	-	-	-	-
		σ	2.0	2.5	9.2	-	-	1.8	-	-	-	-	-	-	-
		T.P.	0.450	0.450	0.447	-	-	**0.450**	-	-	-	-	-	-	-
	5000	μ	**1.6**	2.3	12.3	-	-	2.9	-	-	-	-	-	-	-
		σ	1.3	2.5	10.4	-	-	2.8	-	-	-	-	-	-	-
		T.P.	**0.450**	0.449	0.447	-	-	0.449	-	-	-	-	-	-	-
RouterBench	2500	μ	0.6	**0.4**	1.5	-	-	0.4	-	-	-	-	-	-	-
		σ	0.8	1.0	1.5	-	-	0.8	-	-	-	-	-	-	-
		T.P.	0.645	**0.646**	0.645	-	-	0.646	-	-	-	-	-	-	-
	5000	μ	**0.1**	0.5	2.9	-	-	0.9	-	-	-	-	-	-	-
		σ	0.3	0.7	4.3	-	-	1.2	-	-	-	-	-	-	-
		T.P.	**0.647**	0.646	0.646	-	-	0.646	-	-	-	-	-	-	-

Table 7: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 2$ (middle-arrival setting: $\lambda_{\text{Sprout}} = 0.85$, $\lambda_{\text{RouterBench}} = 0.75$, $\lambda_{\text{EmbedLLM}} = 0.55$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ε	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	0.1	**0.0**	1.0	-	-	0.2	-	-	-	-	-	-	-
		σ	0.3	0.0	1.2	-	-	0.4	-	-	-	-	-	-	-
		T.P.	0.850	**0.850**	0.849	-	-	0.849	-	-	-	-	-	-	-
	5000	μ	**0.1**	0.1	1.5	-	-	0.3	-	-	-	-	-	-	-
		σ	0.3	0.3	1.7	-	-	0.7	-	-	-	-	-	-	-
		T.P.	0.850	**0.850**	0.849	-	-	0.850	-	-	-	-	-	-	-
EmbedLLM	2500	μ	**15.9**	23.4	132.7	-	-	25.3	-	-	-	-	-	-	-
		σ	14.0	21.1	41.4	-	-	26.2	-	-	-	-	-	-	-
		T.P.	**0.542**	0.539	0.495	-	-	0.538	-	-	-	-	-	-	-
	5000	μ	**5.7**	6.4	257.6	-	-	14.2	-	-	-	-	-	-	-
		σ	7.2	12.3	52.3	-	-	27.4	-	-	-	-	-	-	-
		T.P.	**0.546**	0.546	0.496	-	-	0.544	-	-	-	-	-	-	-
RouterBench	2500	μ	3.5	**2.4**	68.3	-	-	2.8	-	-	-	-	-	-	-
		σ	6.3	2.2	31.6	-	-	2.6	-	-	-	-	-	-	-
		T.P.	0.744	**0.744**	0.718	-	-	0.744	-	-	-	-	-	-	-
	5000	μ	2.2	**1.2**	108.4	-	-	1.5	-	-	-	-	-	-	-
		σ	3.6	2.6	61.8	-	-	3.4	-	-	-	-	-	-	-
		T.P.	0.746	**0.747**	0.725	-	-	0.746	-	-	-	-	-	-	-

Table 8: Mean (μ), standard deviation (σ) of queue length gap, and throughput (T.P.) at time steps 2500 and 5000 for $K = 2$ (high-arrival setting: $\lambda_{\text{Sprout}} = 0.95$, $\lambda_{\text{RouterBench}} = 0.85$, $\lambda_{\text{EmbedLLM}} = 0.65$).

Dataset	Time	Stat.	ACQB-CL	ACQB	RAND	Q-ThS	Q-UCB	CQB-ε	RAND_ROUT	ZERO	KNN	MLP	UMR	RouterDC	CARROT
Sprout	2500	μ	**24.4**	34.4	114.0	-	-	36.4	-	-	-	-	-	-	-
		σ	18.2	22.6	32.7	-	-	24.4	-	-	-	-	-	-	-
		T.P.	**0.940**	0.936	0.904	-	-	0.935	-	-	-	-	-	-	-
	5000	μ	**23.6**	36.2	230.2	-	-	41.3	-	-	-	-	-	-	-
		σ	18.5	35.7	38.0	-	-	32.0	-	-	-	-	-	-	-
		T.P.	**0.946**	0.943	0.904	-	-	0.942	-	-	-	-	-	-	-
EmbedLLM	2500	μ	**121.3**	139.3	330.6	-	-	128.8	-	-	-	-	-	-	-
		σ	29.2	33.6	15.1	-	-	29.5	-	-	-	-	-	-	-
		T.P.	**0.597**	0.590	0.513	-	-	0.594	-	-	-	-	-	-	-
	5000	μ	**158.0**	183.2	654.1	-	-	164.4	-	-	-	-	-	-	-
		σ	27.8	64.2	40.3	-	-	45.1	-	-	-	-	-	-	-
		T.P.	**0.615**	0.610	0.516	-	-	0.614	-	-	-	-	-	-	-
RouterBench	2500	μ	68.9	**68.4**	275.4	-	-	70.2	-	-	-	-	-	-	-
		σ	32.7	39.4	38.5	-	-	28.5	-	-	-	-	-	-	-
		T.P.	0.821	**0.822**	0.739	-	-	0.821	-	-	-	-	-	-	-
	5000	μ	113.4	115.2	557.1	-	-	**109.0**	-	-	-	-	-	-	-
		σ	45.1	51.5	78.7	-	-	45.9	-	-	-	-	-	-	-
		T.P.	0.826	0.826	0.738	-	-	**0.827**	-	-	-	-	-	-	-

D REGRET ANALYSIS

In this section, we provide a proof sketch for Theorems 5 and 6. The full proofs are deferred to Section E.

D.1 PROOF SKETCH OF THEOREM 5

Queue state misalignment. We start by addressing *queue state misalignment*, a phenomenon unique to contextual queueing bandits with arbitrary contexts. To analyze the queue length difference $Q(t) - Q^*(t)$, we need to upper bound the expected departure rate gap $\mathbb{E}[D^*(i) - D(i)]$ for each round $i < t$. If the queue states (i.e., the sets of pending queries) were identical under both policies at round i, we could control this gap following an optimistic rule from the standard bandit techniques—specifically, by selecting the query-assortment pair with the highest optimistic estimate of the departure rate (as in Line 7 of Algorithm 1). However, a suboptimal decision at round i affects the queue state for the subsequent round. Consequently, the queue state at round $i+1$ under our policy diverges from that under the optimal policy. We call this phenomenon queue state misalignment. This misalignment invalidates standard bandit analysis because following the optimistic rule over the current queue does not guarantee optimism; the optimal policy's queue may contain a superior query that is currently unavailable to the learner.

Coupling process, and policy switching queues. To address such misalignment, we employ a coupling argument and define policy switching queues. We construct a collection of t coupled queueing processes $\{Q_i\}_{i=0}^{t-1}$, where each process Q_i follows our policy π up to round i and then switches to the optimal policy π^* from rounds $i+1$ to $t-1$. Denote the corresponding queue lengths by $\{Q(i,t)\}_{i=0}^{t-1}$. These t processes are coupled through shared randomness. Specifically, all processes experience identical query arrivals. Furthermore, if the same query and the LLM assortment are selected across processes at any given round, the realized feedback (i.e., departure or retrial) is identical. A formal definition is provided in Section E.1.

Queue length regret decomposition. Utilizing the coupling construction, we decompose the queue length regret via the following telescoping sum: Define $\psi(i,t) := Q(i,t) - Q(i-1,t)$. Then

$$R_t = \mathbb{E}[Q(t) - Q^*(t)] = \mathbb{E}[Q(t-1,t) - Q(0,t)] = \sum_{i=1}^{t-1} \mathbb{E}[\psi(i,t)].$$

Consider the term $\psi(i,t)$. Since both processes Q_i and Q_{i-1} follow the same policy (our algorithm) up to round $i-1$, they share the same trajectory up to round $i-1$. Consequently, at the beginning of round i, the two processes Q_i and Q_{i-1} observe the same queue state, which implies that the queue state misalignment does not exist at round i. This alignment enables us to upper bound the expected departure rate gap at round i using the optimistic rule of our algorithm. Building on this, we can decompose $\mathbb{E}[\psi(i,t)]$ as follows: From the result of Lemma 9, for all $i \in [0, t-1]$,

$$\mathbb{E}[\psi(i,t)] \quad \leq \underbrace{\sqrt{\mathbb{E}[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))^2]}}_{:= e_i} \underbrace{\sqrt{\mathbb{E}\left[\widetilde{\psi}(i,t)\right]}}_{:= p_i}$$

where $(x_i^*, S_i^*) = \arg\max_{x \in \mathcal{X}_i, S \in \mathcal{C}} R(x, S, \theta^*)$, is the optimal choice under current queue state. The term $\widetilde{\psi}(i,t)$ is defined as $\mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i) = 1]$, where the conditioning represents the history $\mathcal{F}_i^+$ combined with the specific divergence event where the process Q_{i-1} successfully departure a query ($D(i-1,i) = 1$) while Q_i fails ($D(i,i) = 0$). In this decomposition, e_i quantifies the per-round departure rate gap at round i, representing the probability that the two processes diverge at round $i+1$. Meanwhile, p_i captures the expected long-term impact of this divergence on the queue length at time t. Substituting this back yields the final bound $R_t \leq \sum_{i=1}^{t-1} e_i p_i$.

Bounding strategy. We proceed to bound the regret term using Chebyshev's sum inequality (Lemma 27). Suppose there exist bounding sequences $\{E_i\}$ and $\{P_i\}$ such that $e_i \leq E_i$ and $p_i \leq P_i$

for all i. Provided that $\{E_i\}$ is non-increasing and $\{P_i\}$ is non-decreasing (with $E_i, P_i > 0$), applying the Chebyshev's sum inequality yields

$$R_t \le \sum_{i=0}^{t-1} E_i P_i \le \frac{1}{t-1}\left(\sum_{i=0}^{t-1} E_i\right)\left(\sum_{i=0}^{t-1} P_i\right).$$

Finally, by demonstrating that the product of $\sum_{i=0}^{t-1} E_i$ and $\sum_{i=0}^{t-1} P_i$ is sublinear, we establish the decaying queue length regret.

Construction of the monotonic sequence E_i. We first establish an upper bound on e_i by constructing a monotonically decreasing sequence $\{E_i\}$. (case 1) If random exploration is triggered at round i (with probability $\lambda\eta(i-1)$), the departure rate gap is trivially bounded by 1, i.e., $R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*) \le 1$. Since the exploration rate $\eta(i-1)$ decreases as i increases, the expected regret contribution from this case monotonically decreases in i. (case 2) Otherwise, the agent selects x_i and S_i according to the optimistic rule. In this case, the Thompson Sampling analysis yields (Proposition 10)

$$R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*) \le 49\beta_{i-1}\lambda_{\min}^{-1/2}(V_{i-1}),$$

where $\beta_l = \alpha_l \min\{\sqrt{6d\log(Mt)},\ \sqrt{2\log(2M)} + \sqrt{6\log(Klt)}\}$ and $\lambda_{\min}(V)$ denotes the minimum eigenvalue of V. Furthermore, leveraging the forced $\eta(i)$-exploration and the feature distribution assumption (Assumption 3), we guarantee that $\mathbb{E}[\lambda_{\min}^{-1/2}(V_{i-1})]$ is monotonically decreasing in i. Combining these two cases, we obtain

$$e_i \le \min\left\{1,\ \sqrt{\underbrace{c_0 t^{-2}}_{\text{(bad)}} + \underbrace{2\lambda\eta(i-1)}_{\text{(case 1)}} + \underbrace{(49\beta_{t-1})^2\nu(i)}_{\text{(case 2)}}}\right\}.$$

where the term (bad) accounts for the failure of the high-probability event in the bandit analysis. $\nu(i)$ corresponds to the upper bound on $\mathbb{E}[\lambda_{\min}^{-1/2}(V_{i-1})]$, which is a decreasing function of i (defined in Lemma 11). We define E_i as the right-hand side of this inequality, which forms a monotonically decreasing sequence by construction.

Construction of the monotonic sequence P_i. Next, we upper bound p_i by constructing a monotonically increasing sequence $\{P_i\}$. Recall that $\widetilde{\psi}(i,t)$ is the conditional expectation of the queue length difference $\psi(i,t)$, given that Q_i fails to depart while Q_{i-1} succeeds at round i. This event creates an one-job discrepancy in the queue states at round $i+1$. If Q_{i-1} hits queue length 0 at any round $l \in [i+1, t]$, then the shared randomness in our coupling construction ensures that Q_i also hits queue length 0 at round l. From that round onward, Q_{i-1} and Q_i follow the same trajectory, resulting in $\psi(i,t) = 0$. Therefore, the event $\{\widetilde{\psi}(i,t) = 1\}$ implies that Q_i never hits queue length 0 over rounds $i+1, \dots, t$. Consequently, it suffices to upper bound $\widetilde{\psi}(i,t)$ with the probability that Q_i never hits queue length 0 over rounds $i+1, \dots, t$.

Since Q_i follows the optimal policy from round $i+1$ onward, the traffic slackness assumption (Assumption 4) guarantees that at each step, the agent can select an assortment with an expected departure rate exceeding the arrival rate by at least ϵ. This induces a negative drift of $-\epsilon$ in the queue length. Specifically, by analyzing the probability that the queue length $Q(i, i+1)$ reduces to 0 over the remaining $t-i-1$ rounds, we obtain the following bound: if $Q(i,i+1) \le (t-i-1)\epsilon + 1$, then

$$\widetilde{\psi}(i,t) \le 2\exp\left(-\frac{(Q(i,i+1) - (t-i-1)\epsilon - 1)^2}{8(t-i-1)}\right)$$

where the lemma can be found in Lemma 17.

It remains to bound $Q(i, i+1)$ on the right-hand side, where $Q(i, i+1)$ is the queue length after following our policy up to round i. Unlike the optimal policy, our policy may incur *bad rounds* where the expected departure rate is not sufficiently close to the optimal, causing the queue length to grow. To address this, we carefully choose $\tau(t)$ as an upper bound on the number of such bad rounds (see Proposition 14), so that for all other rounds $l \in [\tau(t)+1, t]$, the departure rate gap is controlled as

$$R(x_l^*, S_l^*, \theta^*) - R(\widetilde{x}_l, \widetilde{S}_l, \theta^*) \le \epsilon/2 - \eta(\tau(t)), \tag{5}$$

thereby guaranteeing the negative drift and decreasing the queue length. Here, $(\widetilde{x}_l, \widetilde{S}_l)$ is a query-assortment pair with the highest optimistic departure-rate estimate. (5) allows us to derive a tail bound on $Q(i, i+1)$ with $\mathcal{B}(i)$ the number of bad rounds up to round i (Lemma 18)

$$\mathbb{P}(Q(i, i+1) \geq a\mathcal{B}(i) + b,\ \mathcal{E}_g) \leq 17\epsilon^{-2} \exp(-b\epsilon/2),$$

where $\mathcal{E}_g$ denotes the event where the bandit analysis holds. a and b are parameters satisfying $a \geq 1 + \epsilon/4$ and $b \geq 0$. Combining the upper bound on $\widetilde{\psi}(i, t)$ with the tail bound on $Q(i, i+1)$ yields the final result: Let $\omega = 4\tau(t)/\epsilon$. Then,

$$p_i \leq \min\left\{1,\ \sqrt{\frac{c_0}{t^2} + \frac{19}{\epsilon^2} \exp\left(-\frac{\epsilon^2}{32}(t - i - 1 - \omega)\right)}\right\},$$

where the lemma can be found in Lemma 19. We denote the right-hand side as P_i. Since the term $-(t - i - 1 - \omega)$ increases as i increases, P_i is monotonically increasing in i.

Completing the proof. We have constructed sequences $\{E_i\}$ and $\{P_i\}$ that upper bound e_i and p_i, respectively, and satisfy the conditions required for Chebyshev's sum inequality. In Section E.7, we demonstrate that the cumulative per-round regret shows $\sum_{i=1}^{t-1} E_i = \widetilde{\mathcal{O}}\left(t^{3/4}\right)$, while the cumulative long-term impact shows $\sum_{i=1}^{t-1} P_i = \widetilde{\mathcal{O}}\left(1\right)$. Consequently, applying the inequality (which introduces a t^{-1} factor) yields a decaying queue length regret of $\widetilde{\mathcal{O}}(t^{-1/4})$.

D.2 Proof Sketch Of Theorem 6

Recall the definition of the cumulative regret as $\text{Regret}_t = \sum_{i=1}^{t} \mathbb{E}[R(x_i, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)]$. Since the optimal pair S_i^* is selected from the current queue state $\mathcal{X}_i$ of the agent (i.e., $S_i^* = \arg\max_{S \in \mathcal{C}} R(x_i, S, \theta^*)$), the departure-rate comparison is performed within the same queue state. Consequently, unlike the queue length regret analysis, this definition allows us to avoid queue state misalignment, allowing us to employ standard bandit analysis techniques.

At each round i, the agent performs random exploration with probability $\eta(i-1)$ (indicated by $E(i-1) = 1$) and otherwise follows the Thompson sampling-based optimistic rule (with probability $1 - \eta(i-1)$). Since the per-round regret is bounded by 1, the expected cumulative regret from random exploration is bounded by $\sum_{i=1}^{t} \eta(i-1) = \mathcal{O}(\sqrt{t})$. For rounds using the optimistic choice, we follow the analysis for MNL-bandits, which yields a regret bound of $\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{t}\right)$. Combining these two components, we obtain an overall cumulative regret bound of $\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{t}\right)$. A detailed proof is provided in Section E.8.

E Formal Regret Analysis

In this section, following Section D, we introduce the formal proof for the queue length regret Theorem 5, and for the cumulative regret Theorem 6. The outline of this section is as follows:

1. Section E.1: Coupling Process and Policy-Switching Queues
2. Section E.2: Queue Length Regret Decomposition
3. Section E.3: Events
4. Section E.4: Monotonic Decrease in Per-Round Regret
5. Section E.5: How to Set Bad Rounds $\tau(t)$
6. Section E.6: Monotonic Increase in the Effect of Queue Length Difference
7. Section E.7: Queue Length Regret Analysis (Proof of Theorem 5)
8. Section E.8: Cumulative Regret Analysis (Proof of Theorem 6)

E.1 Coupling Process And Policy-Switching Queues

We use the following definition of the policy switching queue, where $\widetilde{Q}(t', t'')$ denotes the length of the queue at the beginning of time step t'' under our policy applied from time steps from 1 to t' and the optimal policy applied from $t'+1$ to $t''-1$. We denote such a queueing process and the corresponding policy as $\widetilde{Q}_{t'}$ and $\widetilde{\pi}_{t'}$. By definition $\widetilde{Q}(t'-1, t') = Q(t')$ and $\widetilde{Q}(0, t'') = Q^*(t'')$. Then, we can see that the queue length regret can be decomposed as the telescoping sum of the difference between two queueing processes as $R_t = \sum_{i=1}^{t-1} \mathbb{E}[\widetilde{Q}(i,t) - \widetilde{Q}(i-1,t)]$.

Next, we construct a coupling process for each $\{\widetilde{Q}_i\}_{i=0}^{t-1}$. We denote these as $\{Q_i\}_{i=0}^{t-1}$ and its corresponding queue length and switching policy as $\{Q(i,t)\}_{i=0}^{t-1}$ and $\{\pi_i\}_{i=0}^{t-1}$. First, for the arrival, random exploration, and parameter sampling (which will appear in Thompson sampling), we assume that all processes $\{Q_i\}_{i=0}^{t-1}$ share the same results of randomness, e.g. if in round t a new job $x^{(t)}$ arrives in Q_i, then every other $Q_{i'}$ also has a same event. We conveniently reuse the same notation of $A(t)$, x_t, $x^{(t)}$, $E(t)$, $\{\widetilde{\theta}_t^{(i)}\}_{i=1}^M$ for these coupling processes.

Now, for the departure, we use the definition of $D(i,t')$ for the random departure of Q_i in round $t' \in [t]$. In each round t', we draw a shared random variable $U \sim \mathrm{Unif}(0,1)$. Say that for some i ,t', Q_i choose x'_t and S'_t. Assume that S'_t is sorted in order of the servers, $S'_t[j]$ defines the j-th item of S'_t. Then

$$\text{if } U \le \frac{\exp(x_{t'S_{t'}[1]}^\mathsf{T}\theta^*)}{1+\sum_{j'\in S_{t'}}\exp(x_{t'j'}^\mathsf{T}\theta^*)}, \text{ then } D(i,t')=1 \text{ and } y_{t1}=1$$

$$\text{if } \sum_{k=1}^{j-1}\frac{\exp(x_{t'S[k]}^\mathsf{T}\theta^*)}{1+\sum_{j'\in S_{t'}}\exp(x_{t'j'}^\mathsf{T}\theta^*)} < U \le \sum_{k=1}^{j}\frac{\exp(x_{t'S[k]}^\mathsf{T}\theta^*)}{1+\sum_{j'\in S_{t'}}\exp(x_{t'j'}^\mathsf{T}\theta^*)} \text{ for } j\in[2,K],$$
$$\text{then } D(i,t')=1 \text{ and } y_{tj}=1$$

$$\text{if } U > R(x_{t'}, S_{t'}, \theta^*), \text{ then } D(i,t')=0 \text{ and } y_{t0}=1$$

We can see that these coupling processes preserves the marginals, therefore,

$$R_t = \sum_{i=1}^{t-1}\mathbb{E}[\widetilde{Q}(i,t)-\widetilde{Q}(i-1,t)] = \sum_{i=1}^{t-1}\mathbb{E}[\underbrace{Q(i,t)-Q(i-1,t)}_{:=\psi(i,t)}],$$

and we continue to use this definition afterwards.

Now we introduce the filtration corresponding to these coupling processes of $\{\pi_i\}_{i=0}^{t-1}$: We define the arrival tuple $\boldsymbol{A}(t)$, the departure tuple $\boldsymbol{D}(t) = \{\boldsymbol{D}(i,t)\}_{i=0}^{t-1}$, and the sampled parameter tuple $\boldsymbol{\theta}(t)$ (which comes from the Thompson sampling algorithm) as

$$\boldsymbol{A}(t) := (A(t), \widetilde{x}^{(t)}), \quad \boldsymbol{D}(t) := \{\boldsymbol{D}(i,t)\}_{i=0}^{t-1} = \{(y_t^{(i)}, (x_t^{(i)}, S_t^{(i)}))\}_{i=1}^{t-1}, \quad \boldsymbol{\theta}(t) := (\{\widetilde{\theta}_t^{(i)}\}_{i=1}^M).$$

Here, $\widetilde{x}^{(t)}$ is a masked feature defined as $\widetilde{x}^{(t)} = \widetilde{x}$ if $A(t) = 0$ where $\widetilde{x} \in \mathbb{R}^d$ is a fixed symbol for the sign of no arrival, and $\widetilde{x}^{(t)} = x^{(t)}$ if $A(t) = 1$. $\boldsymbol{D}_i(t)$ denotes to the departure tuple of π_i, and $(y_t^{(i)}, (x_t^{(i)}, S_t^{(i)}))$ denotes its corresponding choice of context, assortment, and reward in round t. We omit the superscript of $\cdot^{(i)}$ if the situation is clear. $\{\widetilde{\theta}_t^{(i)}\}_{i=1}^M$ is M sampled parameters in round t. Then, we define the filtration as

$$\mathcal{F}_t := \sigma(\mathcal{X}_1, \boldsymbol{A}(1), \boldsymbol{D}(1), \boldsymbol{\theta}(1), E(1), \ldots, \boldsymbol{A}(t-1), \boldsymbol{D}(t-1), \boldsymbol{\theta}(t-1)).$$

Notice that $\mathcal{F}_t$ is short of $E(t-1)$. Also, for the notational convenience, if we write $\boldsymbol{D}(i,t') = 0$ (or $\boldsymbol{D}(i,t') = 1$), it means the observations of $D(i,t') = 0$ (or $D(i,t') = 1$) including $y_{t'}^{(i)}$, $x_{t'}^{(i)}$, and $S_{t'}^{(i)}$.

Finally, for the augmented σ-algebra $\mathcal{G}$ with the filtration $\mathcal{F}$ and the random variable X, which is $\mathcal{G} := \mathcal{F} \vee \sigma(X)$, we use the shorthand notation of $\mathcal{F}, X$, e.g. $\mathbb{E}[Z \mid \mathcal{F}, X]$ denotes the conditional expectation of Z with respect to $\mathcal{F} \vee \sigma(X)$ and $\mathbb{E}[Z \mid \mathcal{F}, X = 1]$ denotes the conditional expectation conditioning on the event $\{X = 1\}$ in addition to $\mathcal{F}$.

E.2 Queue Length Regret Decomposition

Now we focus on this quantity $\psi(i,t) := Q(i,t) - Q(i-1,t)$. As the two coupled queues follow the same randomness and the same policy up to round $t-1$, their queue states at time step t are identical. With this alignment, the following lemma characterize $\psi(i,t)$:

Lemma 8. *We have $\psi(i,t) \in \{-1,0,1\}$ for all $i \in [0,t-1]$. Especially, if $D(i,t) = D(i-1,t) = 0$ or 1, we have $\psi(i,t) \in \{-1,0\}$, and if $D(i,t) = 0$, $D(i-1,t) = 1$, we have $\psi(i,t) \in \{0,1\}$ for all $i \in [0,t-1]$.*

Moreover, the expected value of $\psi(i,t)$ can be decomposed even further as follows:

Lemma 9. *Let $x_l^*, S_l^* = \arg\max_{x \in \mathcal{X}_l, S \in \mathcal{C}} R(x,S,\theta^*)$. For all $l \in [t]$, $i \in [0,t-1]$, define*

$$\mathcal{F}_l^+ := \mathcal{F}_l \vee \sigma(E(l-1), \boldsymbol{A}(l), \boldsymbol{\theta}(l)),$$
$$\widetilde{\psi}(i,t) := \mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i) = 1].$$

Then, we have

$$\mathbb{E}[\psi(i,t)] \le \sqrt{\mathbb{E}[(R(x_i^*,S_i^*,\theta^*) - R(x_i,S_i,\theta^*))^2]}\sqrt{\mathbb{E}\left[\widetilde{\psi}(i,t)\right]}.$$

E.3 Events

We use the following definition of confidence radius: for all $l \in [t]$

$$\alpha_l = \frac{\kappa}{2}\sqrt{d\log(1 + lK/(d\lambda_0)) + 4\log l} + \kappa\sqrt{\lambda_0}$$
$$\beta_l = \alpha_l \min\left\{\sqrt{6d\log(Mt)},\ \sqrt{2\log(2M)} + \sqrt{6\log(Klt)}\right\}$$

where $M = \lceil 1 - \frac{\log(K)}{\log(1-1/(4\sqrt{e\pi}))}\rceil$. We use the following definition of the events: for all $l \in [t]$

$$\mathcal{E}_0(l) := \{A(l-1) = 1, E(l-1) = 1\}$$
$$\mathcal{E}_1 := \{\forall l \in [t],\ \|\widehat{\theta}_{l-1} - \theta^*\|_{V_{l-1}} \le \alpha_{l-1}\}$$
$$\mathcal{E}_2(l) := \{\forall x \in \mathcal{X}_l, j \in [N],\ \widetilde{u}_{lj}(x) - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1} \le \beta_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}\}$$
$$\mathcal{E}_3 := \left\{\lambda_{\min}(V_{\tau(t)}) \ge \frac{4(\alpha_{t-1} + 48\beta_{t-1})^2}{(\epsilon - 2\eta(\tau(t))}\right\}$$

By definition, $\mathbb{P}(\mathcal{E}_0(l)) = \mathbb{P}(A(l) = 1)\mathbb{P}(E(l-1) = 1) = \lambda\eta(l-1)$. By Lemmas 33 and 35, we have $\mathbb{P}(\mathcal{E}_1) = 1 - \mathcal{O}(1/t^2)$ and $\mathbb{P}(\mathcal{E}_2(l)) = 1 - \mathcal{O}(1/t^3)$. Denote $\mathcal{E}_2 := \bigcap_{l=1}^t \mathcal{E}_2(l)$. For $\mathcal{E}_2(l)$, taking union bound for $l \in [t]$, we have $\mathbb{P}(\mathcal{E}_2) = 1 - \mathcal{O}(1/t^2)$. By Proposition 13, $\mathbb{P}(\mathcal{E}_3) \ge 1 - \mathcal{O}(1/t^2)$. Finally, for simplicity, we denote

$$\mathcal{E}_g := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3, \quad \mathbb{P}(\mathcal{E}_g) \ge 1 - c_0 t^{-2} \tag{6}$$

for some absolute constant $c_0 > 0$.

We introduce new definitions for 'bad round' phase $\tau(t)$:

$$\widetilde{\tau}(M) := \max\left\{\frac{2M}{\lambda}, \frac{16\log(t)}{\lambda}, \frac{4M^2}{c_1^2\lambda^2}, \frac{256\log^2(t)}{c_1^2\lambda^2}\right\}$$
$$\widetilde{M}(u) := c_2\left(\frac{d + \log(t)}{\sigma_0^4} + \frac{8(\alpha_{t-1} + 48\beta_{t-1})^2}{\sigma_0^2(\epsilon - 2\eta(u))^2}\right) \tag{7}$$
$$\tau(t) := \min\left\{u \in \mathbb{N} :\ u \ge \widetilde{\tau}(\widetilde{M}(u))\right\}$$
$$M(t) := \widetilde{M}(\tau(t))$$

Notice that these new definitions are not required in the algorithm and only appear in the analysis.

E.4 Monotonic Decrease In Per-Round Regret

We show the monotonic decrease in per-round regret in the following sequence:

1. Upper bound the *expected* (squared) per-round regret with bonus term
2. We show that our $\eta(l)$-exploration policy can monotonically decrease the bonus term, thereby decrease the upper bound of the expected squared per-round regret

Define new filtration $\mathcal{F}_t^-$ which shorts of $\boldsymbol{\theta}(t-1)$ as follows:

$$\mathcal{F}_l^- := \sigma(\mathcal{X}_1, \boldsymbol{A}(1), \boldsymbol{D}(1), \boldsymbol{\theta}(1), E(1), \ldots, \boldsymbol{A}(l-1), \boldsymbol{D}(l-1)).$$

The following proposition shows the upper bound of the per-round regret and another upper bound with the minium eigenvalue of the design matrix V_{l-1}:

Proposition 10 (Per-round regret)**.** *For all $l \in [t]$, define $x_l^*, S_l^* = \arg\max_{x\in\mathcal{X}_l, S\in\mathcal{C}} R(x, S, \theta^*)$. On the event $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_0(l)^c$. Then, we have*

$$R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \leq 16\sqrt{e\pi}\beta_{l-1}\mathbb{E}\left[\max_{j\in S(x_l, \widetilde{\theta}_{l-1}^{1:M})} \|x_{lj}\|_{V_{l-1}^{-1}} \,\middle|\, \mathcal{F}_l^-\right] + (\alpha_{l-1}+\beta_{l-1}) \max_{j\in S_l} \|x_{lj}\|_{V_{l-1}^{-1}}$$

Also, we have

$$R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \leq (\alpha_{l-1} + 48\beta_{l-1})\lambda_{\min}^{-1/2}(V_{l-1}).$$

Proposition 10 shows that the upper bound of the per-round regret can be decreased by increasing the minimum eigenvalue of the design matrix V_l. However the upper bound inequality only holds when $\mathcal{E}_0(l)^c$. Based on this, the following lemma shows that the $\eta(l)$-exploration policy can handle this problem and monotonically decrease the upper bound of the *expected* squared per-round regret:

Lemma 11. *We have*

$$\sqrt{\mathbb{E}[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2]} \leq \min\left\{1,\ \sqrt{c_0 t^{-2} + 2\lambda\eta(l-1) + (\alpha_{t-1}+48\beta_{t-1})^2\nu(l)}\right\}$$

where

$$\nu(l) := \left(\lambda_0 + \frac{\sigma_0^2\lambda\min\{(l-1), c_1\sqrt{l-1}\}}{4}\right)^{-1} + \frac{1}{\lambda_0}\left(\exp\left(-\frac{\lambda\min\{(l-1), c_1\sqrt{l-1}\}}{8}\right) + d\exp\left(-\frac{\sigma_0^2\lambda\min\{(l-1), c_1\sqrt{l-1}\}}{16}\right)\right).$$

E.5 How To Set Bad Rounds $\tau(t)$

We set $\tau(t)$ so that the uncertainty term $\|x_{lj}\|_{V_{l-1}^{-1}}$ gets small enough to guarantee 'good round' for round $l \in [\tau(t)+1, t]$. We deferred the definition of the 'good round' to a later section.

First, denote $N(l)$ as the total number of random exploration up to round l. Then we have the following lemma showing the high probability lower bound of the random exploration until a certain round:

Lemma 12. *For some $M > 0$, and for $\widetilde{\tau}(M)$ satisfying*

$$\widetilde{\tau}(M) \geq \max\left\{\frac{2M}{\lambda},\ \frac{16\log(t)}{\lambda},\ \frac{4M^2}{c_1^2\lambda^2},\ \frac{256\log^2(t)}{c_1^2\lambda^2}\right\},$$

we have $\mathbb{P}(N(\widetilde{\tau}(M)) \geq M) \geq 1 - t^{-2}$.

The following proposition shows the high probability lower bound of the minimum eigenvalue of the design matrix in round $\tau(t)$, which implies the minimum eigenvalue afterwards will also be larger than that value:

Proposition 13. *With probability at least $1 - 2t^{-2}$, we have*

$$\lambda_{\min}(V_{\tau(t)}) \geq \frac{4(\alpha_{t-1}+48\beta_{t-1})^2}{(\epsilon - 2\eta(\tau(t)))}.$$

The following proposition shows that after round $\tau(t)$, the per-round regret in round that does not explore is small:

Proposition 14. *Consider $l \in [\tau(t)+1, t]$. On the event $\mathcal{E}_g$ and $\mathcal{E}_0(l)^c$. Then with probability at least $1-2\delta$,*

$$R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \le \frac{\epsilon}{2} - \eta(\tau(t)).$$

E.6 Monotonic Increase In The Effect Of Queue Length Difference

Bad rounds. First, we introduce the following definition: for all $l \in [t]$

$$\widetilde{x}_l, \widetilde{S}_l := \arg\max_{x \in \mathcal{X}_l, S \in \mathcal{C}} \widetilde{R}(x, S)$$

We can see that $\widetilde{x}_l$ and $\widetilde{S}_l$ are both $\mathcal{F}_l$-measurable since $\mathcal{F}_t$ is only short for $E(t-1)$, therefore $\mathcal{X}_t$ and $\{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^M$ is still $\mathcal{F}_l$-measurable, which include sufficient information to compute $\widetilde{x}_l$ and $\widetilde{S}_l$. Also, if $\mathcal{E}_0(l)^c$, we have $\widetilde{x}_l = x_l$ and $\widetilde{S}_l = S_l$.

Next, we introduce the formal definition of bad rounds and good rounds: We denote all rounds up to $\tau(t)$ as bad rounds and $\tau(t)+1$ to t as good rounds. Define the set of bad rounds as $\mathcal{B}$. Then by the direct result of Proposition 14, on the event $\mathcal{E}_g$ and $\mathcal{E}_0(l)^c$, for all $l \notin \mathcal{B}$, we have

$$R(x_l^*, S_l^*, \theta^*) - R(\widetilde{x}_l, \widetilde{S}_l, \theta^*) = R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \le \frac{\epsilon}{2} - \eta(\tau(t)), \tag{8}$$

where we use the notation of $l \notin \mathcal{B}$ to denote that l is in good round.

Remark 15. *Bae et al. (2026) also derive a departure rate gap bound of $\epsilon/2 - \eta$ similar to (8), where $\eta = T^{-1/2}$. Their method, however, necessitates a pure-exploration phase at the beginning, calculating the required exploration amount based on the horizon T and slackness ϵ. In contrast, we employ a time-varying exploration rate $\eta(t)$ to bypass this dependence. We show that setting the threshold $\tau(t)$ as defined in (7) ensures that the accumulated random exploration is sufficient to satisfy the departure rate upper bound in (8), thereby achieving an anytime guarantee without prior knowledge.*

Next, we establish that for good rounds there is a negative drift as described in the following lemma:

Lemma 16. *On the event $\mathcal{E}_g$, for all $l \notin \mathcal{B}$, $i \in [0, t-1]$, we have*

$$\mathbb{E}[A(l) - D(i,l) \mid \mathcal{F}_l] \le -\epsilon/2.$$

Queue length difference under disagreement. In this paragraph, we give the upper bound for the expected queue length difference between two consecutive policy-switching queues, which is $\psi(i,t)$. Consider $i \in [0, t-1]$. Recall the definition of $\mathcal{F}_l^+$, which is $\mathcal{F}_i^+ := \sigma(\mathcal{F}_i \cup \{E(i-1), \boldsymbol{A}(i), \boldsymbol{\theta}(i)\})$. Then, the following lemma shows the upper bound of the conditional expected value of $\psi(i,t)$ is related to the queue length in $i+1$ and the remaining round $t-i+1$: Define a new event

$$\mathcal{E}_5(i) := \{Q(i, i+1) \le (t-i-1)\epsilon + 1\}. \tag{9}$$

Lemma 17. *For all $i \in [0, t-1]$, on the event $\mathcal{E}_5(i)$, we have,*

$$\mathbb{E}\big[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i) = 1\big] \le 2\exp\left(-\frac{(Q(i,i+1) - (t-i-l)\epsilon - 1)^2}{8(t-i-1)}\right).$$

Tail bound for $Q(i, i+1)$. The result of Lemma 17 shows that the queue length difference $\psi(i,t)$ under the disagreement event can be upper bounded by the exponential term of remaining rounds $(t-i-1)$ (which suits our goal to upper bound the queue length difference with the exponential ramp) and the queue length $Q(i, i+1)$. Therefore, in this paragraph, we control the value of $Q(i, i+1)$ by giving the exponential tail bound for it.

We use the definition of $\mathcal{B}(l)$ which denotes the total number of bad rounds up to round l. By our definition of bad rounds, for all l, we have

$$\mathcal{B}(l) := \min\{l, \tau(t)\}$$

Lemma 18. *Set $\gamma \in (0, \epsilon/2]$, $\rho = \exp(-\gamma\epsilon/4)$, $\beta = \exp(\gamma)$, and some $a \geq \frac{1}{\gamma}\log(\frac{\beta}{\rho}) = 1 + \epsilon/4$ and $b \geq 0$. For all $l \in [t]$, we have*

$$\mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1) + b,\ \mathcal{E}_g) \leq \left(\rho^{l-1}\mathbb{E}[\exp(\gamma Q(i,1))] + \frac{1}{1-\rho}\right)\exp(-\gamma b).$$

If we assume the initial queue starts with an empty state of $Q(1) = 0$, and set $\gamma = \epsilon/2$, $\rho = \exp(-\epsilon^2/8)$, then we simplify the result as

$$\mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1) + b,\ \mathcal{E}_g) \leq 17\epsilon^{-2}\exp(-\gamma b).$$

Main lemma. Now we are ready to introduce the main lemma:

Lemma 19. *Let $\omega := 4\tau(t)/\epsilon$. Then, we have*

$$\sqrt{\mathbb{E}\left[\widetilde{\psi}(i,t)\right]} \leq \begin{cases} \min\{1,\ \sqrt{c_0 t^{-2} + 19\epsilon^{-2}\exp(-\frac{\epsilon^2}{32}(t-i-1-\omega))}\} & \text{if } i \leq t-\omega-1 \\ 1 & \text{if } i > t-\omega-1 \end{cases}.$$

E.7 Queue Length Regret Analysis

We first specify the conditions on t required for Theorem 5.

Condition 20. *Theorem 5 holds for any t sufficiently large satisfying:*

$$t \geq \tau(t), \quad t \geq \omega \tag{10}$$

$$\epsilon > 2\eta(\tau(t)) \tag{11}$$

Remark 21. *(10) is required to ensure that the number of bad rounds $\tau(t)$ and the threshold round $\omega = 4\tau(t)/\epsilon$ do not exceed the current time step t. Since $4\tau(t)/\epsilon \geq \tau(t)$, it suffices to consider $t \geq 4\tau(t)/\epsilon$. (11) is required to guarantee the positivity of the denominator in the lower bound for $\lambda_{\min}(V_{\tau(t)})$ derived in Proposition 13:*

$$\lambda_{\min}(V_{\tau(t)}) \geq \frac{4(\alpha_{t-1} + 48\beta_{t-1})^2}{(\epsilon - 2\eta(\tau(t)))}.$$

We now verify the validity of these conditions for large t. Recall the definition of $\tau(t)$:

$$\tau(t) = \max\left\{\frac{2M(t)}{\lambda},\ \frac{16\log(t)}{\lambda},\ \frac{4M(t)^2}{c_1^2\lambda^2},\ \frac{256\log^2(t)}{c_1^2\lambda^2}\right\},$$

where $M(t)$ is given by

$$M(t) = c_2\left(\frac{d + \log(t)}{\sigma_0^4} + \frac{8(\alpha_{t-1} + 48\beta_{t-1})^2}{\sigma_0^2(\epsilon - 2\eta(\tau(t)))}\right).$$

Since $\alpha_{t-1} = \mathcal{O}(\log^{1/2} t)$ and $\beta_{t-1} = \mathcal{O}(\log t)$, it follows that $M(t) = \mathcal{O}(\log^2 t)$ and consequently $\tau(t) = \mathcal{O}(\log^4 t)$. Therefore, (10) holds for sufficiently large t, as linear growth dominates polylogarithmic growth (i.e., $t > \mathcal{O}(\log^4 t)$). Regarding (11), recall that $\eta(t) = \min\{1,\ c_1(t+1)^{-1/2}\}$. Since $\tau(t) = \mathcal{O}(\log^4 t)$ grows with t, $\eta(\tau(t))$ decays to zero. Thus, for sufficiently large t, the condition $\epsilon > 2\eta(\tau(t))$ is satisfied, ensuring (11) holds. Finally, we conclude that for any large t satisfying Condition 20, Theorem 5 holds.

Theorem 22. *We have a decaying queue length regret of*

$$R_t = \mathcal{O}\left(\frac{d^5 t^{-1/4}\log^5(t)}{\sigma_0^9\epsilon^5} + \frac{d^{11/2}t^{-1}\log^5(t)}{\sigma_0^{12}\epsilon^5}\right).$$

Proof. We start with the result of the regret decomposition in Lemma 9 and prepare to apply Chebyshev sum inequality (Lemma 27):

$$R_t = \sum_{i=1}^{t-1}\mathbb{E}[\psi(i,t)] = \sum_{i=1}^{t-1}\sqrt{\mathbb{E}[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))^2]}\sqrt{\mathbb{E}\left[\widetilde{\psi}(i,t)\right]}$$

We show that the upper bound of each sequence, $\{\sqrt{\mathbb{E}[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))^2]}\}_{i=1}^{t-1}$ and $\{\sqrt{\mathbb{E}[\widetilde{\psi}(i,t)]}\}_{i=1}^{t-1}$, show monotonical behavior in opposite direction.

First, denote $M_i := \min\{1, \sqrt{c_0 t^{-2} + 2\lambda\eta(i-1) + (\alpha_{t-1} + 48\beta_{t-1})^2\nu(i)}\}$. By the direct result of Lemma 11, $\{\sqrt{\mathbb{E}[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))^2]}\}_{i=1}^{t-1}$ can be upper bounded with $\{M_i\}_{i=1}^{t-1}$.

Next, denote $\Delta_i := \min\{1, \sqrt{c_0 t^{-2} + 19\epsilon^{-2}\exp(-\frac{\epsilon^2}{32}(t-i-1-\omega))}\}$ if $i \le t-\omega-1$ and $\Delta_i := 1$ if $i > t-\omega-1$. Then, by the direct result of Lemma 19, $\{\sqrt{\mathbb{E}[\widetilde{\psi}(i,t)]}\}_{i=1}^{t-1}$ can be upper bounded with $\{\Delta_i\}_{i=1}^{t-1}$.

Since $\{M_i\}_{i=1}^{t-1}$ are monotonically decreasing in i and $\{\Delta_i\}_{i=1}^{t-1}$ are monotonically increasing in i, and both sequences are always positive, we can apply Chebyshev sum inequality, which gives

$$R_t \le \frac{1}{t-1}\left(\sum_{i=1}^{t-1} M_i\right)\left(\sum_{i=1}^{t-1}\Delta_i\right). \tag{12}$$

Now, for the summation of (M_i),

$$\begin{aligned}\sum_{i=1}^{t-1} M_i &\le \sum_{i=1}^{t-1}\sqrt{c_0 t^{-2} + 2\lambda\eta(i-1) + (\alpha_{t-1}+48\beta_{t-1})^2\nu(i)} \\ &\le \underbrace{\sum_{i=1}^{t-1}\sqrt{c_0}t^{-1}}_{A_1} + \underbrace{\sum_{i=1}^{t-1}\sqrt{2\lambda\eta(i-1)}}_{A_2} + \underbrace{\sum_{i=1}^{t-1}\sqrt{(\alpha_{t-1}+48\beta_{t-1})^2\nu(i)}}_{A_3}\end{aligned}$$

For term A_1, we have $A_1 \le \sqrt{c_0}$. For term A_2,

$$A_2 \le \sqrt{2\lambda c_1}\sum_{i=1}^{t-1} i^{-1/4} \le \sqrt{2\lambda c_1}\left(1 + \int_1^{t-1} x^{-1/4}dx\right) \le \frac{4}{3}\sqrt{2\lambda c_1}t^{3/4}.$$

For term A_3, we ignore the $(\alpha_{t-1}+48\beta_{t-1})$ part and consider the $\sqrt{\nu(i)}$ part:

$$\begin{aligned}\sum_{i=1}^{t-1}\sqrt{\nu(i)} &\le \sum_{i=1}^{t-1}\left(\lambda_0 + \frac{\sigma_0^2\lambda\min\{(i-1), c_1\sqrt{i-1}\}}{4}\right)^{-1/2} \\ &\quad + \sum_{i=1}^{t-1}\frac{1}{\sqrt{\lambda_0}}\exp\left(-\frac{\lambda\min\{(i-1), c_1\sqrt{i-1}\}}{16}\right) + \sum_{i=1}^{t-1}\frac{\sqrt{d}}{\sqrt{\lambda_0}}\exp\left(-\frac{\sigma_0^2\lambda\min\{(i-1), c_1\sqrt{i-1}\}}{32}\right) \\ &\le \underbrace{\sum_{i=1}^{t-1}\left(\lambda_0 + \frac{\sigma_0^2\lambda(i-1)}{4}\right)^{-1/2}}_{A_{4\text{-}1}} + \underbrace{\sum_{i=1}^{t-1}\frac{1}{\sqrt{\lambda_0}}\exp\left(-\frac{\lambda(i-1)}{16}\right)}_{A_{5\text{-}1}} + \underbrace{\sum_{i=1}^{t-1}\frac{\sqrt{d}}{\sqrt{\lambda_0}}\exp\left(-\frac{\sigma_0^2\lambda(i-1)}{32}\right)}_{A_{6\text{-}1}} \\ &\quad + \underbrace{\sum_{i=1}^{t-1}\left(\lambda_0 + \frac{\sigma_0^2\lambda c_1\sqrt{i-1}}{4}\right)^{-1/2}}_{A_{4\text{-}2}} + \underbrace{\sum_{i=1}^{t-1}\frac{1}{\sqrt{\lambda_0}}\exp\left(-\frac{\lambda c_1\sqrt{i-1}}{16}\right)}_{A_{5\text{-}2}} + \underbrace{\sum_{i=1}^{t-1}\frac{\sqrt{d}}{\sqrt{\lambda_0}}\exp\left(-\frac{\sigma_0^2\lambda c_1\sqrt{i-1}}{32}\right)}_{A_{6\text{-}2}}\end{aligned}$$

For term $A_{4\text{-}1}$, term $A_{4\text{-}2}$, term $A_{5\text{-}2}$, and term $A_{6\text{-}2}$, we upper bound the summations via an integral comparison for monotone decreasing functions. For term $A_{5\text{-}1}$ and term $A_{6\text{-}1}$, we upper bound the finite summations by the corresponding infinite geometric series:

$$\begin{aligned}A_{4\text{-}1} &\le \frac{1}{\sqrt{\lambda_0}} + \frac{8}{\sigma_0^2\lambda}\sqrt{\lambda_0 + \frac{\sigma_0^2\lambda}{4}t}, \quad A_{4\text{-}2} \le \frac{1}{\sqrt{\lambda_0}} + \frac{8}{3\sigma_0\sqrt{\lambda c_1}}t^{3/4} \\ A_{5\text{-}1} &\le \frac{1}{\sqrt{\lambda_0}}\frac{1}{1-\exp(-\lambda/16)} \le \frac{1}{\sqrt{\lambda_0}}\left(1 + \frac{16}{\lambda}\right), \quad A_{5\text{-}2} \le \frac{1}{\sqrt{\lambda_0}}\left(1 + \frac{512}{\lambda^2 c_1^2}\right) \\ A_{6\text{-}1} &\le \frac{\sqrt{d}}{\sqrt{\lambda_0}}\frac{1}{1-\exp(-\sigma_0^2\lambda/32)} \le \frac{\sqrt{d}}{\sqrt{\lambda_0}}\left(1 + \frac{32}{\sigma_0^2\lambda}\right), \quad A_{6\text{-}2} \le \frac{\sqrt{d}}{\sqrt{\lambda_0}}\left(1 + \frac{2048}{\sigma_0^4\lambda^2 c_1^2}\right)\end{aligned}$$

For term $A_{5\text{-}1}$ and term $A_{6\text{-}1}$ we used the fact that $\frac{1}{1-\exp(-x)} \leq 1 + \frac{1}{x}$. Summing up results for term A_3, we have

$$\begin{aligned} A_3 &\leq (\alpha_{t-1} + 48\beta_{t-1})\mathcal{O}\left(\frac{t^{3/4}}{\sigma_0} + \frac{d^{1/2}}{\sigma_0^4}\right) \\ &= \mathcal{O}\left(\frac{dt^{3/4}\log(t)}{\sigma_0} + \frac{d^{3/2}\log(t)}{\sigma_0^4}\right) \qquad (\alpha_t = \mathcal{O}(\sqrt{d\log(t)}),\ \beta_t = \mathcal{O}(d\log(t))) \end{aligned}$$

Finally, substituting results of term A_1, term A_2, and term A_3 back to the original inequality,

$$\begin{aligned} \sum_{i=1}^{t-1} M_i &= \mathcal{O}(1) + \mathcal{O}\left(t^{3/4}\right) + \mathcal{O}\left(\frac{dt^{3/4}\log(t)}{\sigma_0} + \frac{d^{3/2}\log(t)}{\sigma_0^4}\right) \\ &= \mathcal{O}\left(\frac{dt^{3/4}\log(t)}{\sigma_0} + \frac{d^{3/2}\log(t)}{\sigma_0^4}\right). \end{aligned} \tag{13}$$

Now, for the summation of (Δ_i), recall the definition of the threshold ω from Lemma 19, where we use the naive upper bound of $\Delta_i \leq 1$ when the remaining rounds are smaller than ω, and if not, used an exponentially increasing function to upper bound Δ_i. Therefore,

$$\sum_{i=1}^{t-1} \Delta_i = \underbrace{\sum_{i=1}^{t-\omega-1} \Delta_i}_{A_6} + \underbrace{\sum_{i=t-\omega}^{t-1} \Delta_i}_{A_7}$$

For term A_6,

$$\begin{aligned} A_6 &\leq \sum_{i=1}^{t-\omega-1} \left[\sqrt{c_0}t^{-1} + 5\epsilon^{-1}\exp\left(-\frac{\epsilon^2}{64}(t-i-1-\omega)\right)\right] \\ &\leq c_0 + \frac{5}{\epsilon} \times \frac{1}{1-\exp(-\epsilon^2/64)} && \text{(geometric series sum)} \\ &\leq c_0 + \frac{640}{\epsilon^3} && (1-\exp(-x) \geq x/2,\ x \in (0,1]) \end{aligned}$$

For term A_7, we have $A_7 \leq \omega$. Since $M(t) = \mathcal{O}(\frac{d+\log(t)}{\sigma_0^4} + \frac{d^2\log^2(t)}{\sigma_0^2\epsilon^2}) = \mathcal{O}(\frac{d^2\log^2(t)}{\sigma_0^4\epsilon^2})$, we have

$$A_7 \leq \omega = \frac{4\tau(t)}{\epsilon} = \mathcal{O}\left(\frac{M(t)^2}{\epsilon}\right) = \mathcal{O}\left(\frac{d^4\log^4(t)}{\sigma_0^8\epsilon^5}\right).$$

Substituting results, we have

$$\sum_{i=1}^{t-1} \Delta_i = \mathcal{O}\left(\frac{1}{\epsilon^3}\right) + \mathcal{O}\left(\frac{d^4\log^4(t)}{\sigma_0^8\epsilon^5}\right) = \mathcal{O}\left(\frac{d^4\log^4(t)}{\sigma_0^8\epsilon^5}\right). \tag{14}$$

Finally, plugging (13) and (14) back in to (12), we have

$$\begin{aligned} R_t &\leq \frac{1}{t-1}\left(\mathcal{O}\left(\frac{dt^{3/4}\log(t)}{\sigma_0} + \frac{d^{3/2}\log(t)}{\sigma_0^4}\right)\right)\left(\mathcal{O}\left(\frac{d^4\log^4(t)}{\sigma_0^8\epsilon^5}\right)\right) \\ &= \mathcal{O}\left(\frac{d^5t^{-1/4}\log^5(t)}{\sigma_0^9\epsilon^5} + \frac{d^{11/2}t^{-1}\log^5(t)}{\sigma_0^{12}\epsilon^5}\right), \end{aligned}$$

finishing the proof. □

E.8 Cumulative Regret Analysis

Remark 23. *We analyze a stronger regret definition. Denote the previous definition of the regret as* $\mathit{Regret}_t^{\mathit{prev}}$. *Let* $(x_i^*, S_i^*) = \arg\max_{x\in\mathcal{X}_i, S\in\mathcal{C}} R(x, S, \theta^*)$ *be the optimal pair chosen from the queue state* $\mathcal{X}_i$. *Define*

$$\mathit{Regret}_t := \sum_{i=1}^{t} \mathbb{E}[R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)].$$

Since $\mathit{Regret}_t^{\mathit{prev}} \leq \mathit{Regret}_t$ *holds by definition, establishing a bound on* Regret_t *suffices.*

Theorem 24. *We have a cumulative regret of*

$$\textit{Regret}_t = \widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{t}\right)$$

Proof. We start with the definition of the cumulative regret:

$$\begin{aligned}
\text{Regret}_t &= \sum_{i=1}^{t} \mathbb{E}[R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^c\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))\right] + \sum_{i=1}^{t} \mathbb{E}[\mathbf{1}\left\{\mathcal{E}_g\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))] \\
&\leq c_0 t^{-1} + \sum_{i=1}^{t} \mathbb{E}[\mathbf{1}\left\{\mathcal{E}_g\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))] \quad ((6),\ R(\cdot,\cdot,\cdot) - R(\cdot,\cdot,\cdot) \leq 1)
\end{aligned}$$

Now, for the second term on the right-hand side,

$$\begin{aligned}
&\sum_{i=1}^{t} \mathbb{E}[\mathbf{1}\left\{\mathcal{E}_g\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))] \\
&\leq \sum_{i=1}^{t} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))\right] && (\mathcal{E}_g^{\leq i} \subseteq \mathcal{E}_g) \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)) \,\middle|\, \mathcal{F}_i, E(i-1)\right]\right] && \text{(tower rule)} \\
&= \underbrace{\sum_{i=1}^{t} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}\mathbf{1}\left\{E(i-1)=1\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)) \,\middle|\, \mathcal{F}_i, E(i-1)\right]\right]}_{A_1} \\
&\quad + \underbrace{\sum_{i=1}^{t} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}\mathbf{1}\left\{E(i-1)=0\right\}(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*)) \,\middle|\, \mathcal{F}_i, E(i-1)\right]\right]}_{A_2}
\end{aligned}$$

For term A_1, we have

$$A_1 \leq \sum_{i=1}^{t} \mathbb{P}(E(i-1)=1) = \sum_{i=1}^{t} \eta(i-1) \leq c_1 \sum_{i=1}^{t} i^{-1/2} \leq c_1\left(1 + \int_1^t x^{-1/2} dx\right) \leq 2c_1\sqrt{t}.$$

For term A_2, on the event $E_g^{\leq i}$ and $E(i-1)=0$, we can directly apply the result of Proposition 10, which gives

$$\begin{aligned}
A_2 &\leq \underbrace{\sum_{i=1}^{t} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}\mathbf{1}\left\{E(i-1)=0\right\}\left(16\sqrt{e\pi}\beta_{i-1}\mathbb{E}\left[\max_{j \in S(x_i, \tilde{\theta}_{i-1}^{1:M})} \|x_{ij}\|_{V_{i-1}^{-1}} \,\middle|\, \mathcal{F}_i^-\right]\right)\right]}_{A_3} \\
&\quad + \underbrace{\sum_{i=1}^{t} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_g^{\leq i}\right\}\mathbf{1}\left\{E(i-1)=0\right\}\left((\alpha_{i-1}+\beta_{i-1})\max_{j \in S_i} \|x_{ij}\|_{V_{i-1}^{-1}}\right)\right]}_{A_4}
\end{aligned}$$

For term A_3, let us ignore the constant part with $16\sqrt{e\pi}\beta_{t-1}$ (as β_{i-1} is monotonically increasing in i) and only consider the summation of the expectation part inside the outer expectation term. Then

we can proceed as

$$\sum_{i=1}^{t}\mathbb{E}\left[\max_{j\in S(x_i,\widetilde{\theta}_{i-1}^{1:M})}\|x_{ij}\|_{V_{i-1}^{-1}}\,\middle|\,\mathcal{F}_i^-\right]$$
$$=\underbrace{\sum_{i=1}^{t}\max_{j\in S_i}\|x_{ij}\|_{V_{i-1}^{-1}}}_{A_5}+\underbrace{\sum_{i=1}^{t}\left(\mathbb{E}\left[\max_{j\in S(x_i,\widetilde{\theta}_{i-1}^{1:M})}\|x_{ij}\|_{V_{i-1}^{-1}}\,\middle|\,\mathcal{F}_i^-\right]-\max_{j\in S_i}\|x_{ij}\|_{V_{i-1}^{-1}}\right)}_{A_6}$$

For term A_5, we have

$$A_5\le\sqrt{t\sum_{i=1}^{t}\max_{j\in S_i}\|x_{ij}\|^2_{V_{i-1}^{-1}}}\le\sqrt{2dt\log\left(1+\frac{tK}{d\lambda_0}\right)},$$

where the first inequality follows from Cauchy-Schwarz inequality and the last inequity follows from Lemma 34.

For term A_6, we prepare to apply Azuma-Hoeffding inequality (Lemma 42). By construction, $Y_l:=\sum_{i=1}^{l}(\mathbb{E}[\max_{j\in S(x_i,\widetilde{\theta}_{i-1}^{1:M})}\|x_{ij}\|_{V_{i-1}^{-1}}\mid\mathcal{F}_i^-]-\max_{j\in S_i}\|x_{ij}\|_{V_{i-1}^{-1}})$ is a martingale. Next, we have

$$|Y_l-Y_{l-1}|=\left|\mathbb{E}\left[\max_{j\in S(x_l,\widetilde{\theta}_{l-1}^{1:M})}\|x_{lj}\|_{V_{l-1}^{-1}}\,\middle|\,\mathcal{F}_l^-\right]-\max_{j\in S_l}\|x_{lj}\|_{V_{l-1}^{-1}}\right|\le\frac{2\|x\|_2}{\lambda_{\min}^{-1/2}(V_{l-1})}\le\frac{2}{\sqrt{\lambda_0}}.$$

Therefore, by setting $c_t\leftarrow\frac{2}{\sqrt{\lambda 0}}$ and $a\leftarrow\sqrt{\frac{8t}{\lambda_0}\log(2t)}$, and applying Azuma-Hoeffding inequality, we have

$$A_6\le\sqrt{\frac{8t}{\lambda_0}\log(2t)}$$

with probability at least $1-1/t$. Also, if the Azuma-Hoeffding inequality does not hold, we can use the naive bound of

$$A_6\le\sum_{i=1}^{t}\left|\mathbb{E}\left[\max_{j\in S(x_l,\widetilde{\theta}_{l-1}^{1:M})}\|x_{lj}\|_{V_{l-1}^{-1}}\,\middle|\,\mathcal{F}_l^-\right]-\max_{j\in S_l}\|x_{lj}\|_{V_{l-1}^{-1}}\right|\le\frac{2t}{\sqrt{\lambda_0}}$$

Define the event $\mathcal{E}_6$ as the Azuma-Hoeffding inequality holds. Then, substituting results back to term A_3, we have

$$A_3\le 16\sqrt{e\pi}\beta_{t-1}\left(\sqrt{2dt\log\left(1+\frac{tK}{d\lambda_0}\right)}+\mathbb{P}(\mathcal{E}_5)\sqrt{\frac{8t}{\lambda_0}\log(2t)}+\mathbb{P}(\mathcal{E}_5^c)\frac{2t}{\sqrt{\lambda_0}}\right)$$
$$\le 16\sqrt{e\pi}\beta_{t-1}\left(\sqrt{2dt\log\left(1+\frac{tK}{d\lambda_0}\right)}+\sqrt{\frac{8t}{\lambda_0}\log(2t)}+\frac{2}{\sqrt{\lambda_0}}\right)$$

For term A_4, similarly, we can apply Cauchy-Schwarz inequality and Lemma 34, which gives

$$A_4\le(\alpha_{t-1}+\beta_{t-1})\sum_{i=1}^{t}\max_{j\in S_i}\|x_{ij}\|_{V_{i-1}^{-1}}\le(\alpha_{t-1}+\beta_{t-1})\sqrt{2dt\log\left(1+\frac{tk}{d\lambda_0}\right)}.$$

Finally, substituting term A_1, term A_2, term A_3, and term A_4 back to the original inequality, we have

$$\text{Regret}_t\le c_0t^{-1}+2c_1\sqrt{t}+16\sqrt{e\pi}\beta_{t-1}\left(\sqrt{2dt\log\left(1+\frac{tK}{d\lambda_0}\right)}+\sqrt{\frac{8t}{\lambda_0}\log(2t)}+\frac{2}{\sqrt{\lambda_0}}\right)$$
$$+(\alpha_{t-1}+\beta_{t-1})\sqrt{2dt\log\left(1+\frac{tk}{d\lambda_0}\right)}$$
$$=\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{t}\right),\qquad(\alpha_{t-1}=\mathcal{O}(\sqrt{d\log(t)}),\ \beta_{t-1}=\mathcal{O}(d\log(t)))$$

finishing the proof. □

F Deferred Proofs For Section E.2

F.1 Proof Of Lemma 8

To characterize $\psi(t,T) = Q(i,t) - Q(i-1,t)$, we understand the dynamics of the coupled queues of Q_i and Q_{i-1}. Notice that in $l \in [i+1,t]$, both Q_i and Q_{i-1} follows the same optimal policy. Therefore, to track down the queue length difference, we don't need to consider about the context of the job inside the queue x itself; instead, we only need to consider the optimal departure rate of the corresponding job $\max_{S\in\mathcal{C}} R(x,S,\theta^*)$. For example, if there are two different job but with the same departure rate, i.e. for $x_1 \neq x_2$ and $\max_{S\in\mathcal{C}} R(x_1,S,\theta^*) = \max_{S\in\mathcal{C}} R(x_2,S,\theta^*)$, we don't have to distinguish whether the optimal policy choose x_1 or x_2 since the optimal policy does not involve the learning procedure and the queue length will evolve same by the coupling process. Accordingly, we introduce the new definitions: Define $\mathcal{X}(i,t')$ for the queue state of Q_i in round t'. Then we define the departure rate set of the queue state as

$$\mathcal{D}(i,t') := \left\{ d(x),\ \forall x \in \mathcal{X}(i,t') :\ d(x) = \max_{S\in\mathcal{C}} R(x,S,\theta^*) \right\}.$$

For $l \in [i+1,t]$, let us consider the following five states:

$$S_{l,0} = \{\mathcal{D}(i,l) = \mathcal{D}(i-1,l)\}$$
$$S_{l,1} = \{\mathcal{D}(i,l) \setminus \mathcal{D}(i-1,l) = \{d_l^+\},\ \mathcal{D}(i,l) \supset \mathcal{D}(i-1,l)\}$$
$$S_{l,2} = \{\mathcal{D}(i-1,l) \setminus \mathcal{D}(i,l) = \{d_l^-\},\ \mathcal{D}(i-1,l) \supset \mathcal{D}(i,l)\}$$
$$S_{l,3} = \{\mathcal{D}(i,l) \setminus \mathcal{D}(i-1,l) = \{d_l^+\},\ \mathcal{D}(i-1,l) \setminus \mathcal{D}(i,l) = \{d_l^-\},\ d_l^+ > d_l^-\}$$
$$S_{l,4} = \{\mathcal{D}(i,l) \setminus \mathcal{D}(i-1,l) = \{d_l^+\},\ \mathcal{D}(i-1,l) \setminus \mathcal{D}(i,l) = \{d_l^-\},\ d_l^+ < d_l^-\}.$$

Now we proceed with the proof as follows:

1. We show that under optimal policy, $S_{l,s'}$ only transits to $S_{l+1,s''}$ where $s', s'' \in \{0,1,2,3,4\}$.
2. We show that in round $i+1$, the state is inside $S_{l,s'}$, thereby at the end of round t, it will still resides inside those five states, which implies $\psi(i,t) \in \{-1,0,1\}$.

First, we consider each case $S_{l,s'}$ for $s' \in \{0,1,2,3,4\}$

(case 1) If round l is in state $S_{l,0}$, then $\mathcal{D}(i,l) = \mathcal{D}(i-1,l)$, so the optimal policy would choose the same job. As a result, round $l+1$ would be in state $S_{l+1,0}$.

(case 2) If round l is in state $S_{l,1}$, it falls into the following two cases, based on whether the optimal policy chooses d_l^+ for $\mathcal{D}(i,l)$.

(case 2-1) If the optimal policy selects d_i^+ for $\mathcal{D}(i,l)$, this means that $D(i,i) \geq D(i-1,i)$. Therefore, there are three possibilities. When $D(i,i) = D(i,i) = 0$, as the queues keep the same sets of for round $l+1$, we have $S_{l+1,1}$ for round $l+1$. If $D(i,i) = D(i,i-1) = 1$, as the optimal policy would choose another job in $\mathcal{D}(i,i-1)$, we still have state $S_{l+1,1}$ for round $l+1$. When $D(i,i) = 1$ and $D(i,i-1) = 0$, round $i+1$ would be in state $S_{l+1,0}$.

(case 2-2) If the optimal policy does not choose d_l^+ from $\mathcal{D}(i,l)$, then the same for $\mathcal{D}(i,l)$ and $\mathcal{D}(i-1,l)$ will be chosen, which means that round $l+1$ would be in state $S_{l+1,1}$.

(case 3) If round l is in state $S_{l,2}$, by the symmetry between $S_{l,1}$ and $S_{l,2}$, we can argue that we have $S_{l+1,0}$ or $S_{l+1,2}$ in round $l+1$ with a similar argument as in case 2.

(case 4) If round l is in state $S_{l,4}$, it falls into the following two cases, based on whether the optimal policy chooses d_t^+ for $\mathcal{D}(i,l)$.

(case 4-1) If the optimal policy selects d_l^+ for $Q(i,l)$, this means that $D(i,l) \geq D(i-1,l)$. Therefore, there are three possibilities. When $D(i,l) = D(i-1,l) = 0$, as the queues keep the same sets of jobs for round $l+1$, we have $S_{l+1,4}$ for round $l+1$. If $D(i,l) = D(i-1,l) = 1$, the optimal policy would choose another job in $\mathcal{D}(i-1,l)$. If the optimal policy chooses d_l^- from $\mathcal{D}(i-1,l)$, we have state $S_{l+1,0}$ in round $l+1$. If not, round $l+1$ would be in state $S_{1,4}$. When $D(i,l) = 1$ and $D(i-1,l) = 0$, round $l+1$ would be in state $S_{l+1,2}$.

(case 4-2) If the optimal policy does not choose d_l^+ for $\mathcal{D}(i,l)$, then it would not choose d_l^- from $\mathcal{D}(i-1,l)$ either. Hence, the optimal policy chooses the same job for $\mathcal{D}(i,l)$ and $\mathcal{D}(i-1,l)$, so round $l+1$ would be in state $S_{l+1,3}$.

In summary, for (case 4), we have $S_{l+1,0}$ or $S_{l+1,2}$ or $S_{l+1,3}$ in round $l+1$.

(case 5) If round i is in state $S_{l,4}$, by the symmetry between $S_{l,3}$ and $S_{l,4}$, we may argue that we have $S_{l+1,0}$ or $S_{l+1,1}$ or $S_{l+1,4}$ in round $l+1$ with a similar argument as in (case 4).

The results above show that for $l \in [i+1,t]$, $S_{l,s'}$ only transits to $S_{l+1,s''}$ where $s', s'' \in \{0,1,2,3,4\}$.

Now, we consider the state in round $i+1$. Since $Q(i,l)$ and $Q(i-1,l)$ are coupled and follow the same policy, $\mathcal{D}(i,i) = \mathcal{D}(i-1,i)$. In round i, Q_i follows our policy and Q_{i-1} follow the optimal policy, which implies $D(i,i) \le D(i-1,i)$ and we split this in two cases:

(case 1') If $D(i,i) = 0$ and $D(i-1,i) = 1$, then round $i+1$ would be in state $S_{i+1,1}$. Due to our case analysis above, we have $S_{i+2,0}$ or $S_{i+2,1}$ for round $i+2$. If the state of round $i+2$ is $S_{i+2,0}$, then we have $S_{i',0}$ for each round $i' \ge i+2$, in which case $\psi(i,t) = 0$. If we have $S_{i+2,1}$ for round $i+2$, we repeat the same argument as for state $i+1$. If we observe $S_{t,1}$ for round t, then we have $\psi(i,t) = 1$. Otherwise, round t would be in state $S_{t,0}$, in which case $\psi(i,t) = 0$.

(case 2') If $D(i,i) = D(i-1,i)$, then round $i+1$ would be in state $S_{i+1,0}$ or $S_{i+1,3}$. By our case analysis above, we have $S_{i+2,0}$ or $S_{i+2,2}$ or $S_{i+2,3}$ for round $i+2$. If the state of round $i+2$ is $S_{i+2,0}$, then we have $S_{i',0}$ for each round $i' \ge i+2$, in which case $\psi(i,t) = 0$. If we have $S_{i+2,2}$ for round $i+2$, we have $S_{i+3,0}$ or $S_{i+3,2}$ for round $i+3$. If $S_{i+3,0}$ is the state of round $i+3$, then as before, we deduce $\psi(i,t) = 0$. If the state is $S_{i+3,2}$, we repeat the same argument as for state $i+2$. If we observe $S_{t,2}$ for round t, then we have $\psi(i,t) = -1$. Otherwise, round t would be in state $S_{t,0}$, in which case $\psi(i,t) = 0$. If we observe $S_{i+2,3}$ for round $i+2$, then we again repeat the argument as for round $i+1$ to argue that $\psi(i,t) \in \{0,-1\}$.

This finishes the proof.

F.2 Proof Of Lemma 9

Recall the definition of filtration given by

$$\mathcal{F}_l^+ := \sigma(\mathcal{F}_l \cup \{E(l-1), \boldsymbol{A}(l), \boldsymbol{\theta}(l)\})$$

Notice that x_l, S_l are $\mathcal{F}_l^+$-measurable.

By the queue length regret decomposition $\mathbb{R}_t = \sum_{i=1}^{t-1} \mathbb{E}[\psi(i,t)]$, we deduce that

$$\begin{aligned}
R_t &= \sum_{i=1}^{t-1} \mathbb{E}[\mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+]] \\
&= \sum_{i=1}^{t-1} \mathbb{E}[\mathbb{P}(D(i,i) = D(i-1,i) \mid \mathcal{F}_i^+)]\mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+, D(i,i) = D(i-1,i)] \\
&\quad + \sum_{i=1}^{t-1} \mathbb{E}[\mathbb{P}(D(i,i) = 0, D(i-1,i) = 1 \mid \mathcal{F}_i^+)]\mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+, D(i,i) = 0, D(i-1,i) = 1],
\end{aligned}$$

where the first equality holds due to the tower rule and the second equality holds since $D(i,i) \le D(i-1,i)$ by our coupling process. For the first part of the right-hand side, it follows from Lemma 8 that

$$\mathbb{E}[\psi(i,t) \mid \mathcal{F}_t^+, D(i,i) = D(i-1,i)] \le 0.$$

Next, for the second part, notice that the departure disagreement event with $D(i,i) = 0$, $D(i-1,i) = 1$ occurs when

$$R(x_i^{(i)}, S_i^{(i)}, \theta^*) \le U \le R(x_i^{(i-1)}, S_i^{(i-1)}, \theta^*), \quad U \sim \mathrm{Unif}(0,1).$$

Moreover, as the queue state of both Q_i and Q_{i-1} in round i is identical to $\mathcal{X}_i$, we know that

$$x_i^{(i)} = x_t, \quad S_i^{(i)} = S_i, \quad x_i^{(i-1)} = x_i^*, \quad S_i^{(i-1)} = S_t^*$$

where $x_i^*, S_i^* = \arg\max_{x \in \mathcal{X}_i, S \in \mathcal{C}} R(x, S, \theta^*)$. Therefore, we have

$$\mathbb{P}(D(i,i) = 0, D(i-1,i) = 1 \mid \mathcal{F}_t^+) = R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*).$$

Plugging in these results to the above decomposition of R_T, we obtain

$$\begin{aligned} R_T &\le \sum_{i=1}^{t-1} \mathbb{E}\big[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))\mathbb{E}[\psi(i,t) \mid \mathcal{F}_t^+, D(i,i) = 0, D(i-1,i) = 1]\big] \\ &\le \sum_{i=1}^{t-1} \sqrt{\mathbb{E}[(R(x_i^*, S_i^*, \theta^*) - R(x_i, S_i, \theta^*))^2]} \sqrt{\mathbb{E}\big[\mathbb{E}\big[\psi(i,t) \mid \mathcal{F}_t^+, D(i,i) = 0, D(i-1,i) = 1\big]^2\big]}, \end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality. For the second square root term,

$$\begin{aligned} &\mathbb{E}\big[\mathbb{E}\big[\psi(i,t) \mid \mathcal{F}_t^+, D(i,i) = 0, D(i-1,i) = 1\big]^2\big] \\ &\quad \le \mathbb{E}\big[\mathbb{E}\big[\psi(i,t)^2 \mid \mathcal{F}_t^+, D(i,i) = 0, D(i-1,i) = 1\big]\big] && (\mathbb{E}[X \mid \mathcal{F}]^2 \le \mathbb{E}[X^2 \mid \mathcal{F}]) \\ &\quad = \mathbb{E}\big[\mathbb{E}\big[\psi(i,t) \mid \mathcal{F}_t^+, D(i,i) = 0, D(i-1,i) = 1\big]\big] && \text{(Lemma 8)} \\ &\quad = \mathbb{E}\Big[\widetilde{\psi}(i,t)\Big], \end{aligned}$$

where the last inequality follows from the definition. This finishes the proof.

G Deferred Proofs For Section E.4

G.1 Proof Of Proposition 10

The per-round regret can be decomposed as

$$\begin{aligned} &R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \\ &\quad = \underbrace{R(x_l^*, S_l^*, \theta^*) - \widetilde{R}(x_l, S_l)}_{A_1} + \underbrace{\widetilde{R}(x_l, S_l) - R(x_l, S_l, \widehat{\theta}_{l-1})}_{A_2} + \underbrace{R(x_l, S_l, \widehat{\theta}_{l-1}) - R(x_l, S_l, \theta^*)}_{A_3} \end{aligned}$$

Since $\mathcal{E}_1, \mathcal{E}_2$ holds and $\mathcal{E}_0(l)^c$, we can apply Lemma 37 for term A_1:

$$A_1 \le 16\sqrt{e\pi}\beta_{l-1}\mathbb{E}\left[\max_{j \in S(x_l, \widetilde{\theta}_{l-1}^{1:M})} \|x_{lj}\|_{V_{l-1}^{-1}} \,\middle|\, \mathcal{F}_l^-\right].$$

For term A_2, applying Lemmas 32 and 35, we have

$$A_2 \le \max_{j \in S_l} |\widetilde{u}_{lj}(x_l) - x_{lj}^\mathsf{T}\widehat{\theta}_{l-1}| \le \max_{j \in S_l} \beta_{l-1}\|x_{lj}\|_{V_{l-1}^{-1}}.$$

For term A_3, applying Lemmas 32 and 33, we have

$$A_3 \le \max_{j \in S_l} |x_{lj}^\mathsf{T}\widehat{\theta}_{l-1} - x_{lj}^\mathsf{T}\theta^*| \le \max_{j \in S_l} \alpha_{l-1}\|x_{lj}\|_{V_{l-1}^{-1}}.$$

Substituting results, and taking the union bound, we have the desired result.

Also, notice that $16\sqrt{e\pi} \le 47$, and for all $x \in \mathcal{X}$,

$$\|x_{-j}\|_{V_{l-1}^{-1}} \le \frac{\|x\|_2}{\lambda_{\min}^{1/2}(V_{l-1})} \le \lambda_{\min}^{-1/2}(V_{l-1}).$$

Therefore,

$$R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \le (\alpha_{l-1} + 48\beta_{l-1})\lambda_{\min}^{-1/2}(V_{l-1}),$$

finishing the proof.

G.2 Proof Of Lemma 11

Before starting the proof, we introduce a technical result first. By our $\eta(t)$-exploration policy and Assumption 3, we deduce the following lemma:

Lemma 25. *For all $l \in [t]$, we have*

$$\mathbb{P}\left(\lambda_{\min}(V_l) \geq \lambda_0 + \frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{4}\right) \geq 1 - \exp\left(-\frac{\lambda \min\{l, c_1\sqrt{l}\}}{8}\right) - d\exp\left(-\frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{16}\right).$$

Now we are ready to start the proof. Consider $l \in [t]$. Recall the notation of $\mathcal{E}_g := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. Denote $\mathcal{F}_l' := \mathcal{F}_l \vee \sigma(E(t-1))$ for simplicity. Notice that $\mathbf{1}\{\mathcal{E}_0(l)\}$ is $\mathcal{F}_l'$-measurable. We can ignore for $\mathcal{E}_3$. We start by decomposing the expected squared per-round regret as

$$\begin{aligned}
&\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big] && \text{(tower rule)} \\
&= \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g^c\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big] \\
&\quad + \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big] \\
&\leq \mathbb{P}(\mathcal{E}_g^c) + \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big] && (R(\cdot,\cdot,\cdot) - R(\cdot,\cdot,\cdot) \leq 1) \\
&= c_0 t^{-2} + \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big]. && ((6))
\end{aligned}$$

For the last term, we can proceed as

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big] \\
&= \underbrace{\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbf{1}\{\mathcal{E}_0(l)\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big]}_{A_1} \\
&\quad + \underbrace{\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g\}\mathbf{1}\{\mathcal{E}_0(l)^c\}\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2 \mid \mathcal{F}_l'\big]\big]}_{A_2}
\end{aligned}$$

For term A_1, we have

$$A_1 \leq \mathbb{E}[\mathbf{1}\{\mathcal{E}_0(l)\}] = \mathbb{P}(\mathcal{E}_0(l)) = \lambda\eta(t-1).$$

For term A_2, under $\mathcal{E}_0^c(l)$ and $\mathcal{E}_g$, we can directly apply Proposition 10, which yields

$$\begin{aligned}
A_2 &\leq \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c\}\mathbb{E}\big[(\alpha_{l-1} + 48\beta_{l-1})^2 \lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l'\big]\big] \\
&\leq (\alpha_{t-1} + 48\beta_{t-1})^2 \mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c\}\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l'\big]\big]
\end{aligned}$$

where the last inequality follows from the monotonicity of α_l, β_l.

Consider the expectation term. Denote the event $\mathcal{E}_4(l)$ such that Lemma 25 ($\lambda_{\min}(V_{l-1}) \geq \lambda_0 + \frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{4}$) holds. Then, we have

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c\}\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l\big]\big] \\
&= \underbrace{\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c, \mathcal{E}_4(l)\}\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l\big]\big]}_{A_3} + \underbrace{\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c, \mathcal{E}_4(l)^c\}\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l\big]\big]}_{A_4}
\end{aligned}$$

For term A_3, by the definition of $\mathcal{E}_4(l)$, we have

$$A_3 \leq \left(\lambda_0 + \frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{4}\right)^{-1}.$$

For term A_4,

$$\begin{aligned}
A_4 &\leq \mathbb{P}(\mathcal{E}_4(l)^c)\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l\big] \\
&\leq \left(\exp\left(-\frac{\lambda \min\{(l-1), c_1\sqrt{l-1}\}}{8}\right) + d\exp\left(-\frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{16}\right)\right) \times \frac{1}{\lambda_0}
\end{aligned}$$

where the last inequality follows from Lemma 25 and by the naive bound of $V_{t-1} \succeq \lambda_0 \mathbf{I}$. Substituting the results back yields

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}\{\mathcal{E}_g, \mathcal{E}_0(l)^c\}\mathbb{E}\big[\lambda_{\min}^{-1}(V_{l-1}) \mid \mathcal{F}_l\big]\big] \\
&\quad \le \left(\lambda_0 + \frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{4}\right)^{-1} \\
&\qquad + \frac{1}{\lambda_0}\left(\exp\left(-\frac{\lambda \min\{(l-1), c_1\sqrt{l-1}\}}{8}\right) + d\exp\left(-\frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{16}\right)\right).
\end{aligned}$$

Plugging this expectation term into term A_2, and substituting the result of term A_1 into the original inequality, we have the desired result as

$$\begin{aligned}
&\mathbb{E}\big[(R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*))^2\big] \\
&\quad \le c_0 t^{-2} + 2\lambda\eta(l-1) + (\alpha_{t-1} + 48\beta_{t-1})^2 \nu(l)
\end{aligned}$$

where we use the definition of

$$\begin{aligned}
\nu(l) := &\left(\lambda_0 + \frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{4}\right)^{-1} \\
&+ \frac{1}{\lambda_0}\left(\exp\left(-\frac{\lambda \min\{(l-1), c_1\sqrt{l-1}\}}{8}\right) + d\exp\left(-\frac{\sigma_0^2 \lambda \min\{(l-1), c_1\sqrt{l-1}\}}{16}\right)\right).
\end{aligned}$$

Finally, taking $\sqrt{\cdot}$ and applying $\min\{\cdot, 1\}$ to both sides finishes the proof.

G.3 Proof Of Lemma 25

Recall the definition of $N(l)$, which is the total number of random explorations up to round l. We also use the definition of $\widetilde{V}_l$, which is the design matrix, only consists of the feature vector in random exploration round as

$$\widetilde{V}_l := \sum_{i=1}^{l} \sum_{j \in S_l} \mathbf{1}\{A(i-1) = 1, E(i-1)\} x_{ij} x_{ij}^\mathsf{T}.$$

For simplicity, denote

$$N_1 = \mathbb{E}[N(l)]/2, \quad N_2 = \sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}/4.$$

Now, we want the lower bound probability of the following event:

$$\mathbb{P}(N(l) \ge N_1, \lambda_{\min}(\widetilde{V}_l) \ge N_2) = \underbrace{\mathbb{P}(N(l) \ge N_1)}_{A_1} \times \underbrace{\mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \ge N_2 \mid N(l) \ge N_1)}_{A_2}$$

For term A_1, we use the Chernoff bound (Lemma 28). By setting $\delta \leftarrow 1/2, \mu \leftarrow \mathbb{E}[N(l)]$, we have

$$\mathbb{P}(N(l) \ge N_1) = 1 - P(N(l) \ge N_1) \ge 1 - \exp(-\mathbb{E}[N(l)]/8).$$

For the lower bound of $\mathbb{E}[N(l)]$,

$$\mathbb{E}[N(l)] = \sum_{i=1}^{l} \lambda\eta(i-1) \ge \lambda \sum_{i=1}^{l} \min\{1, c_1 l^{-1/2}\} = \lambda \min\{l, c_1\sqrt{l}\} \tag{15}$$

Plugging (15) back to the above inequality, we have

$$\mathbb{P}(N(l) \ge N_1) \ge 1 - \exp\left(-\frac{\lambda \min\{l, c_1\sqrt{l}\}}{8}\right).$$

Next, for term A_2, we have

$$\begin{aligned}
A_2 &\ge \mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \ge N_2 \mid N(l) = N_1) \\
&\ge \mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \ge N_2 \mid N(l) = \lambda \min\{l, c_1\sqrt{l}\}/2). \qquad ((15))
\end{aligned}$$

Now, we use the Matrix Chernoff bound (Lemma 29) on the right-hand side of the inequality. By setting $\delta \leftarrow 1/2$, $n \leftarrow \lambda \min\{l, c_1\sqrt{l}\}/2$, $\sigma_0^2 \leftarrow \sigma_0^2$ we have

$$\mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \geq N_2 \mid N(l) = \lambda \min\{l, c_1\sqrt{l}\}/2) = 1 - \mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \leq N_2 \mid N(l) = \lambda \min\{l, c_1\sqrt{l}\}/2)$$
$$\geq 1 - d\exp\left(-\frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{16}\right).$$

Substituting term A_1 and term A_2 back, we have

$$\mathbb{P}(N(l) \geq N_1, \lambda_{\min}(\widetilde{V}_l) \geq N_2) \geq \left(1 - \exp\left(-\frac{\lambda \min\{l, c_1\sqrt{l}\}}{8}\right)\right)\left(1 - d\exp\left(-\frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{16}\right)\right)$$
$$\geq 1 - \exp\left(-\frac{\lambda \min\{l, c_1\sqrt{l}\}}{8}\right) - d\exp\left(-\frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{16}\right).$$

Finally, we have $\mathbb{P}(\lambda_{\min}(\widetilde{V}_l) \geq N_2) \geq \mathbb{P}(N(l) \geq N_1, \lambda_{\min}(\widetilde{V}_l) \geq N_2)$. Also, $V_l \succeq \lambda_0 \mathbf{I} + \widetilde{V}_l$, therefore

$$\mathbb{P}(\lambda_{\min}(V_l) \geq \lambda_0 + N_2) \geq 1 - \exp\left(-\frac{\lambda \min\{l, c_1\sqrt{l}\}}{8}\right) - d\exp\left(-\frac{\sigma_0^2 \lambda \min\{l, c_1\sqrt{l}\}}{16}\right),$$

finishing the proof.

H Deferred Proofs For Section E.5

H.1 Proof Of Lemma 12

Denote $N(l)$ as the total number of random exploration up to round l. Notice that for every round l, the probability for the random exploration is $\lambda\eta(l-1)$ as it only happens when $A(l-1) = 1$ and $E(l-1) = 1$. Then, we have

$$N(l) = \sum_{i=1}^{l-1} A(i)E(i), \quad \mathbb{E}[N(l)] = \sum_{i=1}^{l-1} \lambda\eta(i)$$

For the lower bound of $\mathbb{E}[N(l)]$,

$$\mathbb{E}[N(\widetilde{\tau}(M))] = \sum_{i=1}^{\widetilde{\tau}(M)} \lambda\eta(i-1) \geq \lambda \sum_{i=1}^{\widetilde{\tau}(M)} \min\{1, c_1 l^{-1/2}\} = \lambda \min\{\widetilde{\tau}(M), c_1\sqrt{\widetilde{\tau}(M)}\} \tag{16}$$

Now, we show the lower bound of the number of random exploration round until $\widetilde{\tau}(M)$. Since $A(l-1)E(l-1) \in \{0, 1\}$, we can apply Multiplicative Chernoff bound (Lemma 28) with $\delta \leftarrow 1/2$, which gives

$$\mathbb{P}\left(N(\widetilde{\tau}(M)) \leq \frac{\mathbb{E}[N(\widetilde{\tau}(M))]}{2}\right) \leq \exp\left(-\frac{\mathbb{E}[N(\widetilde{\tau}(M))]}{8}\right).$$

In order to show

$$\mathbb{P}(N(\widetilde{\tau}(M)) \leq M) \leq t^{-2},$$

we are sufficient to show

$$M \leq \frac{\mathbb{E}[N(\widetilde{\tau}(M))]}{2}, \quad \exp\left(-\frac{\mathbb{E}[N(\widetilde{\tau}(M))]}{8}\right) \leq t^{-2}.$$

Therefore, we need

$$\mathbb{E}[N(\widetilde{\tau}(M))] \geq \max\{2M,\ 16\log(t)\}$$

By the lower bound of $\mathbb{E}[N(\widetilde{\tau}(M))]$ in (16), if

$$\widetilde{\tau}(M) \geq \max\left\{\frac{2M}{\lambda}, \frac{16\log(t)}{\lambda}, \frac{4M^2}{c_1^2\lambda^2}, \frac{256\log^2(t)}{c_1^2\lambda^2}\right\}$$

holds, then it is sufficient to show the lemma holds.

H.2 Proof Of Proposition 13

We prepare to apply Lemma 12. Set, $M(t)$ as

$$M(t) = c_2\left(\frac{d+\log(t)}{\sigma_0^4} + \frac{8(\alpha_{t-1}+48\beta_{t-1})^2}{\sigma_0^2(\epsilon - 2\eta(\tau(t))}\right)$$

and set $\widetilde{\tau}(M(t))$ satisfying the condition in Lemma 12 as

$$\tau(t) = \widetilde{\tau}(M(t)) = \max\left\{\frac{2M(t)}{\lambda}, \frac{16\log(t)}{\lambda}, \frac{4M(t)^2}{c_1^2\lambda^2}, \frac{256\log^2(t)}{c_1^2\lambda^2}\right\}$$

Remark 26. *Notice that $M(t)$ and $\tau(t)$ have circular definition as $M(t)$ requires $\tau(t)$ and $\tau(t)$ requires $M(t)$. Although this is not problematic as both $M(t)$ and $\tau(t)$ are only required in analysis, we avoid this circular definition and remove ambiguity by using the definitions in (7). To check the existence of $\tau(t)$ and $M(t)$, first we can see that $\widetilde{M}(u)$ is monotonically decreasing in u and $\widetilde{\tau}(M)$ is monotonically increasing in M. Therefore $\widetilde{\tau}(\widetilde{M}(u))$ is monotonically decreasing in u, which means there exists a minimum value of $u \in \mathbb{N}$ such that $u \geq \widetilde{\tau}(\widetilde{M}(u))$ which we define $\tau(t)$. Then $M(t)$ is also defined as $\widetilde{M}(\tau(t))$.*

Now, by the direct result of Lemma 12, we have

$$\mathbb{P}(N(\tau(t)) \geq M(t)) \geq 1 - t^{-2}, \tag{17}$$

which means we have $M(t)$ random exploration until round $\tau(t)$ with high probability.

Now, we prepare to apply Proposition 39, which shows high probability lower bound of the design matrix V_l that consists of the i.i.d. sampled features. Recall that our random exploration is done in a round-robin manner. Therefore, we can see that $M(t)$ features inside the design matrix $V_{\tau(t)}$ are i.i.d. samples from the unknown distribution $\mathcal{D}$. Also, for the feature vector, selected under the random exploration round, say $x_{ij'}$, we have

$$\lambda_{\min}\left(\mathbb{E}\left[x_{ij'}x_{ij'}^\mathsf{T}\right]\right) = \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{N}\sum_{j\in[N]} x_{ij}x_{ij}^\mathsf{T}\right]\right) \geq \sigma_0^2. \qquad \text{(Assumption 3)}$$

Finally, assuming (17) holds, and setting

$$B(t) = \frac{4(\alpha_{t-1}+48\beta_{t-1})^2}{(\epsilon - 2\eta(\tau(t)))^2},$$

and applying Proposition 39 yields

$$\lambda_{\min}(V_{\tau(t)}) \geq B(t). \tag{18}$$

with probability at least $1 - t^{-2}$.

Summing up the results, taking union bound on (17) and (18), we have $\lambda_{\min}(V_{\tau(t)}) \geq B(t)$ with probability at least $1 - 2t^{-2}$, finishing the proof.

H.3 Proof Of Proposition 14

For some $l \in [\tau(t)+1, t]$, on the event $\mathcal{E}_g$, we have

$$\begin{aligned} R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) &\leq (\alpha_{l-1}+48\beta_{l-1})\lambda_{\min}^{-1/2}(V_{l-1}) \\ &\leq (\alpha_{l-1}+48\beta_{l-1})\lambda_{\min}^{-1/2}(V_{\tau(t)}) \\ &\leq (\alpha_{l-1}+48\beta_{l-1})\frac{\epsilon - 2\eta(\tau(t))}{2(\alpha_{t-1}+48\beta_{t-1})}, \end{aligned}$$

where the last inequality follows from $\mathcal{E}_3$. Since α_l and β_l are monotonically increasing in l, we have

$$R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \leq (\alpha_{t-1}+48\beta_{t-1})\frac{\epsilon - 2\eta(\tau(t))}{2(\alpha_{t-1}+48\beta_{t-1})} = \frac{\epsilon}{2} - \eta(\tau(t)),$$

finishing the proof.

I Deferred Proofs For Section E.6

I.1 Proof Of Lemma 16

If $i \geq l$, it means the departure is followed by the optimal policy. Then by Assumption 4, we have

$$\mathbb{E}[A(l) - D(i,l) \mid \mathcal{F}_l] \leq -\epsilon \leq -\epsilon/2.$$

If $i < l$, it means the departure is followed by our policy. As the assumption, we consider $\mathcal{E}_g$ occurs. Then, we proceed as

$$\mathbb{E}[A(l) - D(i,l) \mid \mathcal{F}_l] = \underbrace{E[A(l) - D(l-1,l) \mid \mathcal{F}_l]}_{A_1} + \underbrace{\mathbb{E}[D(l-1,l) - D(i,l) \mid \mathcal{F}_l]}_{A_2}$$

First, notice that $D(l-1,l)$ and $D(i,l)$ both follows our policy until round $l-1$, and $D(l-1,l)$ follows the optimal policy in round l, while $D(i,l)$ follows our policy in round l. Therefore, for term A_1, similarly, by Assumption 4, we have $A_1 \leq -\epsilon$.

For term A_2, we have

$$\begin{aligned} A_2 = &\underbrace{\mathbb{P}(E(l-1) = 1 \mid \mathcal{F}_t)\mathbb{E}[D(l-1,l) - D(i,l) \mid \mathcal{F}_l, E(l-1) = 1]}_{A_3} \\ &+ \underbrace{\mathbb{P}(E(l-1) = 0 \mid \mathcal{F}_t)\mathbb{E}[D(l-1,l) - D(i,l) \mid \mathcal{F}_l, E(l-1) = 0]}_{A_4}. \end{aligned}$$

For term A_3, we have $A_3 \leq \mathbb{P}(E(l-1) = 1) = \eta(l-1)$.

For term A_4, we have

$$\begin{aligned} A_4 &\leq \mathbb{E}[D(l-1,l) - D(i,l) \mid \mathcal{F}_l, E(l-1) = 0] \\ &= \mathbb{E}[R(x_l^*, S_l^*, \theta^*) - R(x_l, S_l, \theta^*) \mid \mathcal{F}_l, E(l-1) = 0], \end{aligned}$$

where the equality holds since $D(l-1,l)$ and $D(i,l)$ have the same queue state in round l of $\mathcal{X}_l$. Now, since l is in good rounds, and on the event $\mathcal{E}_g$ and $E(l-1) = 0$ (which implies $\mathcal{E}_0(l)^c$), by (8), we have $A_4 \leq \epsilon/2 - \eta(\tau(t))$. Substituting term A_3 and term A_4 back yields

$$A_2 \leq \eta(l-1) + \frac{\epsilon}{2} - \eta(\tau(t)) \leq \frac{\epsilon}{2},$$

where the last inequality is from the fact that $\eta(l)$ is monotonically decreasing in l, and $l \notin \mathcal{B}$ means that $l \in [\tau(t)+1, t]$, which means $l - 1 \geq \tau(t)$.

Finally, substituting results of term A_1 and term A_2 back finishes the proof.

I.2 Proof Of Lemma 17

We consider the event that $Q(i, i+1) \leq (t-i-1)\epsilon + 1$ occurs ($\mathcal{E}_5(i)$). For simplicity, denote the event of $\boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i) = 1$ as $\mathcal{E}$.

By Lemma 8, under $\mathcal{E}$ (when $\boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i) = 1$), the value of $\psi(i,t)$ is in $\{0,1\}$. We can simply ignore the case when $\psi(i,t) = 0$. Now we consider the case when $\psi(i,t) = 1$, which means the difference of queue length occurred by a disagreement event in round i is preserved until round t. Notice that if the queue with an extra job Q_i hits 0 queue length before t, the queue length difference between Q_i and Q_{i-1} will always become 0 thereafter by our coupling process. This implies that the probability of Q_i never hitting 0 for all round $l \in [i+1, t]$ is larger than the probability that the queue length difference is preserved until round t. Then, we have

$$\mathbb{P}\big(\psi(i,t) = 1 \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big) \leq \mathbb{P}\big(Q(i,j) > 0,\ \forall j \in [i+1,t] \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big).$$

Next, notice that $Q(i, i+1)$ is a realized value under $\mathcal{F}_i^+$ and $\mathcal{E}$. By the queue dynamics, the probability of Q_i never hitting length 0 for all round $l \in [i+1, t]$ can be upper bounded by the probability that the cumulative net service $\sum_{l=i+1}^{t-1}(D(i,l) - A(l))$ cannot exceed $Q(i, i+1) - 1$, which is

$$\mathbb{P}\big(Q(i,j) > 0,\ \forall j \in [i+1,t] \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big) \leq \mathbb{P}\left(Q(i,i+1) + \sum_{l=i+1}^{t-1}(A(l) - D(i,l)) \geq 1 \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right).$$

Combining two inequalities, we have

$$\begin{aligned}&\mathbb{E}\big[\psi(i,t) \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big] \\ &\quad= \mathbb{P}\big(\psi(i,t) = 1 \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big) \\ &\quad\le \mathbb{P}\left(Q(i,i+1) + \sum_{l=i+1}^{t-1} (A(l) - D(i,l)) \ge 1 \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right).\end{aligned}$$

For the last inequality, we prepare to apply Azuma-Hoeffding inequality (Lemma 42). Define a martingale difference sequence for $l \in [i+1, t-1]$ as $X_l := \mathbb{E}[D(i,l) - A(l) \mid \mathcal{F}_l] - (D(i,l) - A(l))$ and $Y_k := \sum_{l=i+1}^{k} X_l$. We have $\mathbb{E}[X_l \mid \mathcal{F}_l] = 0$, and $|Y_k - Y_{k-1}| = |X_k| \le 2$, therefore, by applying Azuma-Hoeffding inequality, setting $a \leftarrow (t-i-1)\epsilon + 1 - Q(i,i+1)$ (where the condition of $a > 0$ holds since we consider on the event $\mathcal{E}_5(i)$), we have

$$\begin{aligned}&\mathbb{P}\left(\sum_{l=i+1}^{t-1} (\mathbb{E}[D(i,l) - A(l) \mid \mathcal{F}_l] - (D(i,l) - A(l))) \ge (t-i-1)\epsilon + 1 - Q(i,i+1) \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right) \\ &\quad\le 2\exp\left(-\frac{((t-i-1)\epsilon + 1 - Q(i,i+1))^2}{8(t-i-1)}\right)\end{aligned}$$

For $l \in [i+1, t-1]$, Q_i always follows the optimal policy, thereby $\mathbb{E}[D(i,l) - A(l) \mid \mathcal{F}_l] \ge \epsilon$ by Assumption 4. This implies that

$$\begin{aligned}&\mathbb{P}\left(\sum_{l=i+1}^{t-1} (\mathbb{E}[D(i,l) - A(l) \mid \mathcal{F}_l] - (D(i,l) - A(l))) \ge (t-i-1)\epsilon + 1 - Q(i,i+1) \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right) \\ &\quad\ge \mathbb{P}\left(\sum_{l=i+1}^{t-1} (-(D(i,l) - A(l))) \ge 1 - Q(i,i+1) \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right) \\ &\quad= \mathbb{P}\left(Q(i,i+1) + \sum_{l=i+1}^{t-1} (A(l) - D(i,l)) \ge 1 \;\middle|\; \mathcal{F}_i^+, \mathcal{E}\right).\end{aligned}$$

Combining results and plugging back to the original inequality, we have

$$\mathbb{E}\big[\psi(i,t) \;\big|\; \mathcal{F}_i^+, \mathcal{E}\big] \le 2\exp\left(-\frac{(Q(i,i+1) - (t-i-1)\epsilon - 1)^2}{8(t-i-1)}\right)$$

as desired.

I.3 Proof Of Lemma 18

We define a prefix good event $\mathcal{E}_g^{\le l}$ where we cut $\mathcal{E}_g$ until round l to make it $\mathcal{F}_l$-measurable. Formally,

$$\begin{aligned}&\mathcal{E}_1^{\le l} := \{\forall i \le l,\ \|\widehat{\theta}_{i-1} - \theta^*\|_{V_{i-1}} \le \alpha_{i-1}\}, \quad \mathcal{E}_2^{\le l} := \bigcap_{i=1}^{l} \mathcal{E}_2(i) \\ &\mathcal{E}_g^{\le l} := \mathcal{E}_1^{\le l} \cap \mathcal{E}_2^{\le l}.\end{aligned}$$

Here, we skipped $\mathcal{E}_3$ since we don't need it in proof. We can see that $\mathcal{E}_g \subseteq \mathcal{E}_g^{\le l}$ for all l.

We start with the one-step bound for the moment generating function (mgf) of $Q(i,l)$. For some $\gamma \in (0, \epsilon/2]$,

$$\begin{aligned}\mathbb{E}[\exp(\gamma Q(i,l+1)) \mid \mathcal{F}_l] &= \mathbb{E}[\exp(\gamma[Q(i,l) + A(l) - D(i,l)]^+) \mid \mathcal{F}_l] \\ &\le 1 + \exp(\gamma Q(i,l))\mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l],\end{aligned}$$

where the inequality follows by considering both cases, where $Q(i,l) + A(l) - D(l) \le 0$ gives $\exp(\gamma[Q(i,l) + A(l) - D(i,l)]^+) = 1$ and $Q(i,l) + A(l) - D(l) > 0$ gives $\exp(\gamma[Q(i,l) + A(l) - D(i,l)]^+) = \exp(\gamma(Q(i,l) + A(l) - D(i,l)))$. We split into two cases for $\mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l]$:

(case 1): If $l \notin \mathcal{B}$, on the event $\mathcal{E}_g^{\le l}$, by Lemma 16, we have

$$\mathbb{E}[A(l) - D(i,l) \mid \mathcal{F}_l] \le -\epsilon/2.$$

Then,

$$\begin{aligned}\mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l] &= \mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l] \\ &\le \exp\big(\gamma \mathbb{E}[A(l) - D(i,l) \mid \mathcal{F}_l] + \gamma^2/2\big) \\ &\le \exp(-\gamma\epsilon/2 + \gamma^2/2)\end{aligned}$$

where for the last inequality, we apply Hoeffding lemma (Lemma 41). Moreover, by our choice of $\gamma \in (0, \epsilon/2]$, we have

$$\mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l] \le \exp(-\gamma\epsilon/4) =: \rho.$$

(case 2): If $l \in \mathcal{B}$, we use a naive bound of $A(l) - D(i,l) \le 1$ as

$$\mathbb{E}[\exp(\gamma(A(l) - D(i,l))) \mid \mathcal{F}_l] \le \exp(\gamma) =: \beta.$$

Combining results for both cases, we have that, on the event $\mathcal{E}_g^{\le l}$,

$$\mathbb{E}[\exp(\gamma Q(i,l+1)) \mid \mathcal{F}_l] \le 1 + \exp(\gamma Q(i,l))(\rho \mathbf{1}\{l \notin \mathcal{B}\} + \beta \mathbf{1}\{l \in \mathcal{B}\}). \tag{19}$$

To control the indicator term $(\rho \mathbf{1}\{l \notin \mathcal{B}\} + \beta \mathbf{1}\{l \in \mathcal{B}\})$, we define a new weighted process $V(i,l)$ as

$$V(i,l) := \left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l-1)} \exp(\gamma Q(i,l)).$$

Now, we can proceed as

$$\begin{aligned}&\mathbb{E}\big[V(i,l+1)\mathbf{1}\{\mathcal{E}_g^{\le l+1}\} \mid \mathcal{F}_l\big] \\ &\quad\le \mathbb{E}\big[V(i,l+1)\mathbf{1}\{\mathcal{E}_g^{\le l}\} \mid \mathcal{F}_l\big] \\ &\quad= \mathbf{1}\{\mathcal{E}_g^{\le l}\}\mathbb{E}\left[\left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l)} \exp(\gamma Q(i,l+1)) \,\middle|\, \mathcal{F}_l\right] && (\mathbf{1}\{\mathcal{E}_g^{\le l}\} \in \mathcal{F}_l) \\ &\quad= \mathbf{1}\{\mathcal{E}_g^{\le l}\}\left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l-1)}\left(\frac{\beta}{\rho}\right)^{-\mathbf{1}\{l\in\mathcal{B}\}}\mathbb{E}[\exp(\gamma Q(i,l+1)) \mid \mathcal{F}_l] \\ &\quad\le \mathbf{1}\{\mathcal{E}_g^{\le l}\}\left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l-1)}\left(\frac{\beta}{\rho}\right)^{-\mathbf{1}\{l\in\mathcal{B}\}}(1 + \exp(\gamma Q(i,l))(\rho \mathbf{1}\{l \notin \mathcal{B}\} + \beta \mathbf{1}\{l \in \mathcal{B}\})). && ((19))\end{aligned}$$

Since

$$\left(\frac{\beta}{\rho}\right)^{-\mathbf{1}\{l\in\mathcal{B}\}}(\rho \mathbf{1}\{l \notin \mathcal{B}\} + \beta \mathbf{1}\{l \in \mathcal{B}\}) = \rho,$$

we have

$$\begin{aligned}\mathbb{E}\big[V(i,l+1)\mathbf{1}\{\mathcal{E}_g^{\le l+1}\} \mid \mathcal{F}_l\big] &\le \mathbf{1}\{\mathcal{E}_g^{\le l}\}\rho\left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l-1)} \exp(\gamma Q(i,l)) + \mathbf{1}\{\mathcal{E}_g^{\le l}\}\left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l)} \\ &\le \rho V(i,l)\mathbf{1}\{\mathcal{E}_g^{\le l}\} + 1,\end{aligned}$$

as the last inequality holds since $\beta/\rho \ge 1$. Taking the expectation on both sides and applying the tower rule on the left-hand side gives

$$\mathbb{E}[V(i,l+1)\mathbf{1}\{\mathcal{E}_g^{\le l+1}\}] \le \rho \mathbb{E}[V(i,l)\mathbf{1}\{\mathcal{E}_g^{\le l}\}] + 1.$$

Solving a linear recursion (with the fact that $\mathbf{1}\{\mathcal{E}_g^{\le 1}\} = 1$), we have

$$\mathbb{E}[V(i,l)\mathbf{1}\{\mathcal{E}_g^{\le l}\}] \le \rho^{l-1}\mathbb{E}[V(i,1)] + \sum_{k=0}^{l-2}\rho^k \le \rho^{l-1}\mathbb{E}[\exp(\gamma Q(i,1))] + \sum_{k=0}^{\infty}\rho^k \le \rho^{l-1}\mathbb{E}[\exp(\gamma Q(i,1))] + \frac{1}{1-\rho}. \tag{20}$$

Since we obtained the tail probability of $V(i,l)$, we can start the proof as

$$\begin{aligned}
&\mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1)+b,\ \mathcal{E}_g) \\
&\leq \mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1)+b,\ \mathcal{E}_g^{\leq l}) \\
&= \mathbb{P}(\exp(\gamma Q(i,l)) \geq \exp(\gamma(a\mathcal{B}(l-1)+b)),\ \mathcal{E}_g^{\leq l}) \\
&= \mathbb{P}\left(V(i,l) \geq \left(\frac{\beta}{\rho}\right)^{-\mathcal{B}(l-1)} \exp(\gamma(a\mathcal{B}(l-1)+b)),\ \mathcal{E}_g^{\leq l}\right)
\end{aligned}$$

Note that by our choice of $a \geq \frac{1}{\gamma}\log(\frac{\beta}{\rho})$, we have

$$\exp(\gamma a\mathcal{B}(l-1)) \geq \exp(\log(\beta/\rho)\mathcal{B}(l-1)) = (\beta/\rho)^{\mathcal{B}(l-1)}.$$

Therefore,

$$\mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1)+b,\ \mathcal{E}_g) \leq \mathbb{P}(V(i,l) \geq \exp(\gamma b),\ \mathcal{E}_g^{\leq l}).$$

Under the condition of $b \geq 0$, applying Markov inequality tot he right hand side yields

$$\mathbb{P}(V(i,l) \geq \exp(\gamma b),\ \mathcal{E}_g^{\leq l}) \leq \frac{\mathbb{E}[V(i,l)\mathbf{1}\{\mathcal{E}_g^{\leq l}\}]}{\exp(\gamma b)} \leq \left(\rho^{l-1}\mathbb{E}[\exp(\gamma Q(i,1))] + \frac{1}{1-\rho}\right)\exp(-\gamma b). \tag{(20)}$$

Substituting this result back gives the desired result.

Lastly, if we assume the initial queue starts with an empty state of $Q(1) = 0$, and set $\gamma = \epsilon/2$, $\rho = \exp(-\epsilon^2/8)$, then we simplify as

$$\rho^{l-1}\mathbb{E}[\exp(\gamma Q(i,1))] + \frac{1}{1-\rho} \leq 1 + \frac{1}{1-\exp(-\epsilon^2/8)} \leq 1 + 16/\epsilon^2,$$

where the inequality follows from the fact that $1 - \exp(-x) \geq x/2$ for $x \in [0,1]$. Substituting the result back yields

$$\mathbb{P}(Q(i,l) \geq a\mathcal{B}(l-1)+b,\ \mathcal{E}_g) \leq (1+16/\epsilon^2)\exp(-\gamma b) \leq 17\epsilon^{-2}\exp(-\gamma b),$$

finishing the proof.

I.4 Proof Of Lemma 19

We start as follows:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\psi}(i,t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_g^c\}\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]\right] \\
&\leq c_0 t^{-2} + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_g\}\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]\right]
\end{aligned}$$

((6) and lemma 8)

Now for the second term on the right-hand side, we consider the term $\mathbb{E}[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i) = 0, \boldsymbol{D}(i-1,i)=1]$ under the case when the event $\mathcal{E}_g$ occurs: Recall the result of Lemma 17, which is, on the event $\mathcal{E}_5(i)$,

$$\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right] \leq 2\exp\left(-\frac{(Q(i,i+1)-1-(t-i-1)\epsilon)^2}{8(t-i-1)}\right)$$

Our goal there is to obtain an exponential decay of the right-hand side in terms of the number of remaining rounds $t-l-1$ and avoid the dependence of $Q(i,i+1)$. Therefore, we split into two cases where $Q(i,i+1) \leq \frac{\epsilon(t-l-1)}{2}$ and $Q(i,i+1) > \frac{\epsilon(t-l-1)}{2}$ which gives

$$\begin{aligned}
&\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right] \\
&= \underbrace{\mathbf{1}\left\{Q(i,i+1) \leq \frac{\epsilon(t-l-1)}{2}\right\}\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]}_{A_1} \\
&\quad + \underbrace{\mathbf{1}\left\{Q(i,i+1) > \frac{\epsilon(t-l-1)}{2}\right\}\mathbb{E}\left[\psi(i,t) \mid \mathcal{F}_i^+, \boldsymbol{D}(i,i)=0, \boldsymbol{D}(i-1,i)=1\right]}_{A_2}
\end{aligned}$$

For term A_1, we can see that $\mathbf{1}\{Q(i,i+1) \le \frac{\epsilon(t-l-1)}{2}\}$ implies $\mathcal{E}_5(i)$ (i.e., $\mathbf{1}\{Q(i,i+1) \le (t-i-1)\epsilon+1\}$), therefore we can directly apply the result of Lemma 17, which gives

$$\begin{aligned} A_1 &\le \mathbf{1}\left\{Q(i,i+1) \le \frac{\epsilon(t-i-1)}{2}\right\} 2\exp\left(-\frac{(Q(i,i+1)-1-(t-i-1)\epsilon)^2}{8(t-i-1)}\right) \\ &\le 2\exp\left(-\frac{(\frac{1}{2}(t-i-1)\epsilon+1)^2}{8(t-i-1)}\right) \\ &\le 2\exp\left(-\frac{\epsilon^2(t-i-1)}{32}\right). \end{aligned}$$

For term A_2, by Lemma 8, we have $A_2 \le \mathbf{1}\{Q(i,i+1) > \frac{\epsilon(t-l-1)}{2}\}$. Substituting results to the original inequality gives

$$\begin{aligned} \mathbb{E}\left[\widetilde{\psi}(i,t)\right] &\le c_0 t^{-2} + 2\exp\left(-\frac{\epsilon^2(t-i-1)}{32}\right) + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_g\}\mathbf{1}\left\{Q(i,i+1) > \frac{\epsilon(t-l-1)}{2}\right\}\right] \\ &= c_0 t^{-2} + 2\exp\left(-\frac{\epsilon^2(t-i-1)}{32}\right) + \mathbb{P}\left(Q(i,i+1) > \frac{\epsilon(t-l-1)}{2},\ \mathcal{E}_g\right) \end{aligned}$$

Now, for the last term on the right-hand side, we are going to apply the tail bound of Lemma 18, by setting $l \leftarrow i+1$, which gives

$$\mathbb{P}(Q(i,i+1) \ge a\mathcal{B}(i)+b) \le 17\epsilon^{-2}\exp(-\gamma b))$$

under the condition of $a \ge \frac{1}{\gamma}\log(\frac{\beta}{\rho})$ and $b \ge 0$. Since $\frac{1}{\gamma}\log(\frac{\beta}{\rho}) = 1+\epsilon/4$, we can set $a=2$. Now, we need to assure $b \ge 0$ holds. In order to control this, we set a threshold value ω of the remaining round as

$$\omega := \frac{4\tau(t)}{\epsilon} \ge \frac{2a\mathcal{B}(t)}{\epsilon}.$$

We split into 2 cases:

(case 1): If $(t-i-1) < \omega$, we do not apply Lemma 18, since it means that there are not many rounds remaining to reduce the queue length difference by emptying the queue with an extra job. This leads to a naive bound of $\mathbb{E}[\widetilde{\psi}(i,t)] \le 1$.

(case 2): If $(t-i-1) \ge \omega$, we have

$$\begin{aligned} \mathbb{P}\left(Q(i,i+1) > \frac{\epsilon(t-i-1)}{2},\ \mathcal{E}(g)\right) &= \mathbb{P}\left(Q(i,i+1) > a\mathcal{B}(i) + \left(\frac{\epsilon(t-i-1)}{2} - a\mathcal{B}(i)\right),\ \mathcal{E}(g)\right) \\ &\le \mathbb{P}\left(Q(i,i+1) > a\mathcal{B}(i) + \left(\frac{\epsilon(t-i-1)}{2} - a\mathcal{B}(t)\right),\ \mathcal{E}(g)\right) \\ &\le \mathbb{P}\left(Q(i,i+1) > a\mathcal{B}(i) + \left(\frac{\epsilon(t-i-1-\omega)}{2}\right),\ \mathcal{E}(g)\right) \\ &\qquad (\omega \ge 2a\mathcal{B}(t)/\epsilon) \end{aligned}$$

Since $(\frac{\epsilon(t-i-1-\omega)}{2}) \ge 0$ by the assumption, setting $b \leftarrow (\frac{\epsilon(t-i-1-\omega)}{2})$ and applying Lemma 18 gives

$$\mathbb{P}\left(Q(i,i+1) > \frac{\epsilon(t-i-1)}{2},\ \mathcal{E}(g)\right) \le 17\epsilon^{-2}\exp\left(-\gamma\left(\frac{\epsilon(t-i-1-\omega)}{2}\right)\right)$$

Substituting both cases gives

$$\begin{aligned} \mathbb{E}\left[\widetilde{\psi}(i,t)\right] &\le \min\left\{1,\ c_0 t^{-2} + 2\exp\left(-\frac{\epsilon^2(t-i-1)}{32}\right) + 17\epsilon^{-2}\exp\left(-\gamma\left(\frac{\epsilon(t-i-1-\omega)}{2}\right)\right)\right\} \\ &\le \min\left\{1,\ c_0 t^{-2} + 19\epsilon^{-2}\exp\left(-\frac{\epsilon^2}{32}(t-i-1-\omega)\right)\right\}. \qquad (\gamma = \epsilon/2) \end{aligned}$$

Finally, taking $\sqrt{\cdot}$ on both sides finishes the proof.

J AUXILIARY LEMMAS

Lemma 27 (Chebyshev sum inequality). *If $(a_i)_{i=1}^t$ is nondecreasing and $(b_i)_{i=1}^t$ is nonincreasing, and $a_i, b_i \geq 0$, we have*

$$\sum_{i=1}^{t} a_i b_i \leq \frac{1}{t}\left(\sum_{i=1}^{t} a_i\right)\left(\sum_{i=1}^{t} b_i\right).$$

Lemma 28 (Multiplicative Chernoff bound). *Suppose $X_1, \ldots, X_n \in \{0,1\}$ are independent random variables. Let X denote their sum and $\mu = \mathbb{E}[X]$. Then for any $0 \leq \delta \leq 1$,*

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\bigl(-\delta^2\mu/2\bigr).$$

Lemma 29 (Matrix Chernoff bound). *Let $X \in \mathbb{R}^d$ be a random vector with $\|X\|_2 \leq 1$ and $\mathbb{E}[XX^\mathsf{T}] \succeq \sigma_0^2\mathbf{I}$ for some $\sigma_0 > 0$. Suppose $X_1, \ldots, X_n$ be i.i.d. sampled vectors and define $V_n = \sum_{i=1}^n X_i X_i^\mathsf{T}$. Then for any $0 \leq \delta < 1$,*

$$\mathbb{P}\bigl(\lambda_{\min}(V_n) \leq (1-\delta)n\sigma_0^2\bigr) \leq d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{n\sigma_0^2} \leq d\exp\left(-\frac{\delta^2 n\sigma_0^2}{2}\right)$$

Lemma 30 (Lemma 1 of Oh & Iyengar (2019)). *Suppose $S^*(x)$ is the optimal assortment under context x and true parameter θ^*, i.e., $S^*(x) = \arg\max_{S\in\mathcal{C}} R(x,S,\theta^*)$. Also suppose that $x_{-j}^\mathsf{T}\theta^* \leq x_{-j}^\mathsf{T}\theta'$ for all $j \in S^*(x)$. Then $R(x, S_l^*, \theta^*) \leq R(x, S_l^*, \theta')$.*

Lemma 31 (Modified version of Lemma 2 of Oh & Iyengar (2019)). *Suppose $\|\widehat{\theta}_{t-1} - \theta^*\|_{V_{t-1}} \leq \alpha_{t-1}$. Then we have*

$$\mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta_t^*) \mid \mathcal{F}_t^-) \geq (4\sqrt{e\pi})^{-1}$$

Proof. Recall the definition of filtration F_t^- which shorts of $\boldsymbol{\theta}(t-1)$ as follows:

$$\mathcal{F}_t^- := \sigma(\mathcal{X}_1, \boldsymbol{A}(1), \boldsymbol{D}(1), \boldsymbol{\theta}(1), E(1), \ldots, \boldsymbol{A}(t-1), \boldsymbol{D}(t-1)).$$

Note that $\mathcal{X}_t$ is still $\mathcal{F}_t^-$-measurable, and x_t and $\{\widetilde{\theta}_{t-1}\}_{i=1}^M$ are not $\mathcal{F}_t^-$-measurable. We use the definition of the optimal assortment $S^*(x)$ and the optimistic assortment $S(x)$ given a context $x \in \mathcal{X}_t$ as

$$S^*(x) = \arg\max_{S\in\mathcal{C}} R(x,S,\theta^*), \quad S(x) = \arg\max_{S\in\mathcal{C}} \widetilde{R}(x,S).$$

Note that $S^*(x_t^*) = S_t^*$ and $S(x_t) = S_t$. Given $\mathcal{F}_t^-$, each of Gaussian random variable $x_{-j}^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)}$ has mean $x_{-j}^\mathsf{T}\widehat{\theta}_{t-1}$ and standard deviation $\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}$. Hence, in round t, for all $x \in \mathcal{X}_t$, $j \in S^*(x)$, we have

$$\begin{aligned}
&\mathbb{P}\Bigl(\max_i x_{-j}^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)} > x_{-j}^\mathsf{T}\theta^* \mid \mathcal{F}_t^-\Bigr) \\
&\quad = 1 - \mathbb{P}\Bigl(x_{-j}^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)} \leq x_{-j}^\mathsf{T}\theta^*,\ \forall i \in \{1,\ldots,M\} \mid \mathcal{F}_t^-\Bigr) \\
&\quad = 1 - \mathbb{P}\left(\frac{x_{-j}^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)} - x_{-j}^\mathsf{T}\widehat{\theta}_{t-1}}{\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}} \leq \frac{x_{-j}^\mathsf{T}\theta^* - x_{-j}^\mathsf{T}\widehat{\theta}_{t-1}}{\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}},\ \forall i \in \{1,\ldots,M\} \mid \mathcal{F}_t^-\right) \\
&\quad = 1 - \mathbb{P}\left(Z_j \leq \frac{x_{-j}^\mathsf{T}\theta^* - x_{-j}^\mathsf{T}\widehat{\theta}_{t-1}}{\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}},\ \forall i \in \{1,\ldots,M\} \mid \mathcal{F}_t^-\right)
\end{aligned}$$

where Z_j is a standard normal random variable. Then, we can bound the right-hand side term within the probability as

$$\frac{x_{-j}^\mathsf{T}\theta^* - x_{-j}^\mathsf{T}\widehat{\theta}_{t-1}}{\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}} \leq \frac{\|x_{-j}\|_{V_{t-1}^{-1}}\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}}}{\alpha_{t-1}\|x_{-j}\|_{V_{t-1}^{-1}}} \leq 1,$$

where the inequality follows from the Cauchy-Schwarz inequality and the assumption. Then, it follows that

$$\mathbb{P}\Big(\max_i x_{-j}^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)} > x_{-j}^\mathsf{T}\theta^* \mid \mathcal{F}_t^-\Big) \geq 1 - (\mathbb{P}(Z \leq 1))^M. \tag{21}$$

Now, we are ready to lower-bound the probability of having an expected revenue optimistic under the sampled parameter as follows:

$$\begin{aligned}
&\mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \mid \mathcal{F}_t^-) \\
&\quad = \mathbb{P}(\widetilde{R}(x_t, S(x_t))) > R(x_t^*, S^*(x_t^*), \theta^*) \mid \mathcal{F}_t^-) \\
&\quad \geq \mathbb{P}(\widetilde{R}(x_t^*, S(x_t^*))) > R(x_t^*, S^*(x_t^*), \theta^*) \mid \mathcal{F}_t^-) && \text{(optimistic rule)}
\end{aligned}$$

By the definition, we have $\widetilde{R}(x, S(x)) \geq \widetilde{R}(x, S^*(x))$, therefore

$$\begin{aligned}
&\mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \mid \mathcal{F}_t^-) \\
&\quad \geq \mathbb{P}(\widetilde{R}(x_t^*, S^*(x_t^*)) > R(x_t^*, S^*(x_t^*), \theta^*) \mid \mathcal{F}_t^-) \\
&\quad \geq \mathbb{P}(\max_i (x_{tj}^*)^\mathsf{T}\widetilde{\theta}_{t-1}^{(i)} > (x_{tj}^*)^\mathsf{T}\theta^*,\ \forall j \in S^*(x_t^*), \mid \mathcal{F}_t^-) && \text{(Lemma 30)} \\
&\quad \geq 1 - K(\mathbb{P}(Z \leq 1))^M. && \text{((21), union bound)}
\end{aligned}$$

Using the anti-concentration inequality in Lemma 38, we have $\mathbb{P}(Z \leq 1) \leq 1 - (4\sqrt{e\pi})^{-1}$. Hence, we can have the desired result as

$$\begin{aligned}
\mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \mid \mathcal{F}_t^-) &\geq 1 - K(1 - (4\sqrt{e\pi})^{-1})^M \\
&\geq 1 - (1 - (4\sqrt{e\pi})^{-1}) \\
&= (4\sqrt{e\pi})^{-1}
\end{aligned}$$

where the second inequality follows from our choice of $M = \lceil 1 - \frac{\log(K)}{\log(1-1/(4\sqrt{e\pi}))} \rceil$ which implies $(1 - (4\sqrt{e\pi})^{-1})^M \leq \frac{1}{K}(1 - (4\sqrt{e\pi})^{-1})$. □

Lemma 32 (Lemma 3 of Oh & Iyengar (2019)). *For any two utility parameters $u_t = [u_{t1}, \ldots, u_{tN}]$ and $u_t' = [u_{t1}', \ldots, u_{tN}']$, we have*

$$\frac{\sum_{j\in S}\exp(u_{tj})}{1+\sum_{j\in S}\exp(u_{tj})} - \frac{\sum_{j\in S}\exp(u_{tj}')}{1+\sum_{j\in S}\exp(u_{tj}')} \leq \max_{i\in S}|u_{tj} - u_{tj}'|.$$

In particular, if $u_{tj} \geq u_{tj}'$ for all j, then

$$\frac{\sum_{j\in S}\exp(u_{tj})}{1+\sum_{j\in S}\exp(u_{tj})} - \frac{\sum_{j\in S}\exp(u_{tj}')}{1+\sum_{j\in S}\exp(u_{tj}')} \leq \max_{i\in S}(u_{tj} - u_{tj}').$$

Lemma 33 (Lemma 4 of Oh & Iyengar (2019)). *Define $\alpha_l = \frac{\kappa}{2}\sqrt{d\log(1+\frac{lK}{d\lambda_0}) + 4\log l} + \kappa\sqrt{\lambda_0}$. Then*

$$\|\widehat{\theta}_l - \theta^*\|_{V_l} \leq \alpha_l$$

holds for al all $l \in [t]$ with probability $1 - \mathcal{O}(1/t^2)$.

Lemma 34 (Lemma 6 of Oh & Iyengar (2019)). *Define $V_t := \lambda_0\mathbf{I} + \sum_{i=1}^t \sum_{j\in S_i} x_{ij}x_{ij}^\mathsf{T}$. Then, we have*

$$\sum_{i=1}^t \max_{j\in S_i}\|x_{ij}\|_{V_{t-1}^{-1}} \leq \sqrt{2dt\log\left(1+\frac{tK}{d\lambda_0}\right)}$$

Lemma 35 (Modified version of Lemma 10 of Oh & Iyengar (2019)). *Set $l \in [t]$. Let $\beta_l = \alpha_l \min(\sqrt{6d\log(Mt)}, \sqrt{2\log(2M)} + \sqrt{6\log(Klt)})$. Then for all $x \in \mathcal{X}_l$, $j \in [N]$,*

$$\widetilde{u}_{lj}(x) - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1} \leq \beta_{t-1}\|x_{-j}\|_{V_{l-1}^{-1}}$$

with probability $1 - \mathcal{O}(1/t^3)$.

Proof. We follow the same proof procedure introduced in Section E.1 of Oh & Iyengar (2019). Given $\mathcal{F}_l^-$, for some $x \in \mathcal{X}_l$, each of Gaussian random variable $x_{-j}^\mathsf{T}\widetilde{\theta}_{l-1}^{(i)}$ has mean $x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}$ and standard deviation $\alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}^{-1}$.

$$\begin{aligned}\left|\widetilde{u}_{tj}(x) - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}\right| &= \alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}} \frac{|\max_i x_{-j}^\mathsf{T}\widetilde{\theta}_{l-1}^{(i)} - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}|}{\alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}} \\ &\le \alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}} \max_i \left|\frac{x_{-j}^\mathsf{T}\widetilde{\theta}_{l-1}^{(i)} - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}}{\alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}}\right| \\ &= \alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}} \max_i |Z_j|\end{aligned}$$

where each Z_i is a standard normal random variable. With Lemma 36, we have

$$\max_i |Z_i| \le \sqrt{2\log(2M)} + \sqrt{6\log(t)}$$

with probability at least $1 - t^{-3}$. Then for all $x \in \mathcal{X}_l$ and for all $j \in [N]$,

$$|\widetilde{u}_{lj} - x_{lj}^\mathsf{T}\widehat{\theta}_{l-1}| \le \left(\sqrt{2\log(2M)} + \sqrt{6\log(Klt)}\right)\alpha_{l-1}\|x_{tj}\|_{V_{t-1}^{-1}} \tag{22}$$

with probability at least $1 - t^{-3}$, where we take union bound using the fact that $|\mathcal{X}_l| \le l$. Now, let $m = \arg\max_i x_{lj}^\mathsf{T}\widetilde{\theta}_{l-1}^{(i)}$. Then we can write

$$\begin{aligned}\left|\widetilde{u}_{lj}(x) - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}\right| &= \left|\max_i x_{-j}^\mathsf{T}\widetilde{\theta}_{l-1}^{(j)} - x_{-j}^\mathsf{T}\widehat{\theta}_{l-1}\right| \\ &= \left|x_{-j}^\mathsf{T}(\widetilde{\theta}_{l-1}^{(m)} - \widehat{\theta}_{l-1})\right| \\ &\le \alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}\left\|\alpha_{l-1}^{-1}V_{l-1}^{1/2}(\widetilde{\theta}_{l-1}^{(m)} - \widehat{\theta}_{l-1})\right\| \\ &\le \alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}} \max_i\left\|\alpha_{l-1}^{-1}V_{l-1}^{1/2}(\widetilde{\theta}_{l-1}^{(i)} - \widehat{\theta}_{l-1})\right\| \\ &= \alpha_{l-1}\|x_{-j}\|_{V_{t-1}}^{-1} \max_i \|\zeta_i\|\end{aligned}$$

where each element in $\zeta_i \in \mathbb{R}^d$ is a univariate standard normal variable $\mathcal{N}(0,1)$. Therefore, each $\|\zeta_i\| \le \sqrt{6d\log(t)}$ with probability at least $1 - t^{-3}$. Using the union bound for all $i \in [M]$, with probability at least $1 - t^{-3}$, we have

$$|\widetilde{u}_{lj} - x_{lj}^\mathsf{T}\widehat{\theta}_{l-1}| \le \sqrt{6d\log(Mt)}\alpha_{l-1}\|x_{-j}\|_{V_{l-1}^{-1}}. \tag{23}$$

Finally, taking the minimum for both cases (22) and (23) finishes the proof. □

Lemma 36 (Lemma 13 of Oh & Iyengar (2019))**.** *Let $Z_i \sim \mathcal{N}(0,1)$, $i \in [n]$ be a standard Gaussian random variable. Then we have*

$$\mathbb{P}\left(\max_i |Z_i| \le \sqrt{2\log(2n)} + \sqrt{2\log(1/\delta)}\right) \ge 1 - \delta$$

Lemma 37 (Lemma 12 of Oh & Iyengar (2019))**.** *Assume $E(t-1) = 0$. With probability $1 - \mathcal{O}(1/t^2)$, we have*

$$R(x_t^*, S_t^*, \theta^*) - \widetilde{R}(x_t, S_t) \le 16\sqrt{e\pi}\beta_{t-1}\mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M_t})} \|x_{tj}\|_{V_{t-1}^{-1}} \,\middle|\, \mathcal{F}_t^-\right]$$

Proof. The proof follows the similar steps as the proof of Lemma 12 in Oh & Iyengar (2019) (see Section E.3 of Oh & Iyengar (2019)). For completeness, we provide the full proof here.

First, notice that $E(t-1) = 0$, which means x_t and S_t are selected by Thompson sampling. Next, define $\widetilde{\Theta}_t$ as the set of parameter samples for which the expected revenue concentrates appropriately

to the expected revenue based on the ML parameter. Also, define the set of optimistic parameter samples $\widetilde{\Theta}_t^{\text{opt}}$ which coinciding with $\widetilde{\Theta}_T$ as follows:

$$\widetilde{\Theta}_t := \left\{ \{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^M : \ \widetilde{R}(x_t, S_t) - R(x_t, S_t, \widehat{\theta}_{t-1}) \le \beta_{t-1} \max_{j \in S_t} \|x_{tj}\|_{V_{t-1}^{-1}} \right\}$$
$$\widetilde{\Theta}_t^{\text{opt}} := \left\{ \{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^M : \ \widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \right\} \cap \widetilde{\Theta}_t$$

Define the event $\mathcal{E}_t$ where

$$\mathcal{E}_t = \{x_{-j}^\mathsf{T} \widehat{\theta}_{t-1} - x_{-j}^\mathsf{T} \theta^* \le \alpha_{t-1} \|x_{-j}\|_{V_{t-1}^{-1}}, \ \forall x \in \mathcal{X}_t, j\} \cap \{\widetilde{u}_{tj}(x) - x_{-j}^\mathsf{T} \widehat{\theta}_{t-1} \le \beta_{t-1} \|x_{-j}\|_{V_{t-1}^{-1}}, \ \forall x \in \mathcal{X}_t, j\}.$$

For any $\widetilde{\theta}_{t-1}^{1:M} := \{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^M \in \widetilde{\Theta}_t^{\text{opt}}$, we have

$$\Big(R(x_t^*, S_t^*, \theta^*) - \widetilde{R}(x_t, S_t)\Big) \mathbf{1}\{\mathcal{E}_t\} \le \left(R(x_t^*, S_t^*, \theta^*) - \inf_{\theta_{t-1}^{1:M} \in \widetilde{\Theta}_t} \widetilde{R}(x_t, S_t, \theta_{t-1}^{1:M})\right) \mathbf{1}\{\mathcal{E}_t\}$$

where $\widetilde{R}(x_t, S_t, \theta_{t-1}^{1:M})$ is the optimistic expected revenue under the sampled parameters $\theta_{t-1}^{1:M}$. Then, we can bound $R(x_t^*, S_t^*, \theta_t^*) - \widetilde{R}(x_t, S_t)$ by the expectation over any random choice $\widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}$:

$$\begin{aligned}
R(x_t^*, S_t^*, \theta^*) - \widetilde{R}(x_t, S_t) &\le \mathbb{E}\left[\left(\widetilde{R}(x_t, S_t) - \inf_{\theta_{t-1}^{1:M} \in \widetilde{\Theta}_t} \widetilde{R}(x_t, S_t, \theta_{t-1}^{1:M}) \right) \mathbf{1}\{\mathcal{E}_t\} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \right] \\
&= \mathbb{E}\left[\sup_{\theta_{t-1}^{1:M} \in \widetilde{\Theta}_t} \Big(\widetilde{R}(x_t, S_t) - \widetilde{R}(x_t, S_t, \theta_{t-1}^{1:M}) \Big) \mathbf{1}\{\mathcal{E}_t\} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \right] \\
&\le \mathbb{E}\left[\sup_{\theta_{t-1}^{1:M} \in \widetilde{\Theta}_t} \max_{j \in S_t} \Big| \widetilde{u}_{tj}(x_t) - x_{tj}^\mathsf{T} \theta_{t-1}^{(i)} \Big| \mathbf{1}\{\mathcal{E}_t\} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \right] \\
&\le 2\beta_{t-1} \mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t \right] \mathbb{P}(\mathcal{E}_t)
\end{aligned}$$

where the last inequality follows from the definition of $\widetilde{\Theta}_t$ and $S(x_t, \widetilde{\theta}_{t-1}^{1:M})$ stands for the optimal assortment under the sampled parameters $\widetilde{\theta}_{t-1}^{1:M} = \{\widetilde{\theta}_{t-1}^{(i)}\}_{i=1}^M$.

Now, from Lemma 31, we have $\mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) | \mathcal{F}_t^-, \mathcal{E}_t) \ge (4\sqrt{e\pi})^{-1}$, Therefore, we have

$$\begin{aligned}
\mathbb{P}(\widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \mid \mathcal{F}_t^-, \mathcal{E}_t) &= \mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \text{ and } \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t, \mathcal{E}_t) \\
&\ge \mathbb{P}(\widetilde{R}(x_t, S_t) > R(x_t^*, S_t^*, \theta^*) \mid \mathcal{F}_t^-, \mathcal{E}_t) - \mathbb{P}(\widetilde{\theta}_{t-1}^{1:M} \notin \widetilde{\Theta}_t, \mathcal{E}_t) \\
&\ge (4\sqrt{e\pi})^{-1} - \mathcal{O}\big(t^{-1}\big) \\
&\ge (4\sqrt{e\pi})^{-1}/2.
\end{aligned}$$

Note that we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \ \middle| \ \mathcal{F}_t^-, \mathcal{E}_t \right] \\
&\ge \mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t \right] \mathbb{P}(\widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \mid \mathcal{F}_t^-, \mathcal{E}_t) \\
&\ge \mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \ \middle| \ \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t \right] \cdot (4\sqrt{e\pi})^{-1}/2.
\end{aligned}$$

Substituting results, we have

$$\begin{aligned} R(x_t^*, S_t^*, \theta^*) - \widetilde{R}(x_t, S_t) &\le 2\beta_{t-1}\mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \;\middle|\; \mathcal{F}_t^-, \widetilde{\theta}_{t-1}^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right]\mathbb{P}(\mathcal{E}_t) \\ &\le 4\beta_{t-1}(4\sqrt{e\pi})\mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \;\middle|\; \mathcal{F}_t^-, \mathcal{E}_t\right]\mathbb{P}(\mathcal{E}_t) \\ &\le 16\sqrt{e\pi}\beta_{t-1}\mathbb{E}\left[\max_{j \in S(x_t, \widetilde{\theta}_{t-1}^{1:M})} \|x_{tj}\|_{V_{t-1}^{-1}} \;\middle|\; \mathcal{F}_t^-\right], \end{aligned}$$

as desired. □

Lemma 38 (Abramowitz & Stegun (1948)). *For a Gaussian random variable Z with mean μ and variance σ^2, for any $z \ge 1$,*

$$\frac{1}{2\sqrt{\pi}z}e^{-z^2/2} \le \mathbb{P}(|Z - \mu| > z\sigma) \le \frac{1}{\sqrt{\pi}z}e^{-z^2/2}.$$

Proposition 39 (Proposition 1 of Li et al. (2017)). *Define $V_t = \sum_{i=1}^t x_i x_i^\mathsf{T}$, where x_i is drawn i.i.d. from some unknown distribution ν with support in the unit ball, $\mathbb{B}^d$. Furthermore, let $\Sigma := \mathbb{E}[x_i x_i^\mathsf{T}]$ be the second moment matrix, and B and $\delta > 0$ be two positive constants. Then, there exists absolute constants $c', c'' > 0$ such that $\lambda_{\min}(V_t) \ge B$ with probability at least $1 - \delta$, as long as*

$$t \ge \left(\frac{c'\sqrt{d} + c''\sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)}\right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}.$$

Lemma 40 (Hoeffding inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \le X_i \le b_i$ almost surely. Consider $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$, we have*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Lemma 41 (Hoeffding lemma). *Let X be a real-valued random variable with $a \le X \le b$ almost surely. Then, for all $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}[\exp(\lambda X)] \le \exp\left(\lambda\mathbb{E}[X] + \frac{\lambda^2(b-a)^2}{8}\right).$$

Lemma 42 (Azuma-Hoeffding inequality). *If a supermartingale $(Y_i)_{i \ge 0}$ corresponding to filtration $\mathcal{G}_i$ satisfies $|Y_i - Y_{i-1}| \le c_i$ for all $t \in [t]$, then for any $a \ge 0$, we have*

$$\mathbb{P}(Y_t - Y_0 \ge a) \le 2\exp\left(-\frac{a^2}{2\sum_{i=1}^t c_i^2}\right).$$

Use Of Generative-AI Tools

GPT-5.2 assisted in manuscript polishing, and simulation code development. The authors take responsibility for the final arguments and presentation.