Algorithm for Contextual Queueing Bandits with Rate-Optimal Queue Length Regret

Seoungbin Bae
Department of Industrial & Systems Engineering
KAIST
sbbae31@kaist.ac.kr

Dabeen Lee
Department of Mathematical Sciences
Seoul National University
dabeenl@snu.ac.kr

Abstract

Contextual queueing bandits provide a framework for learning to schedule heterogeneous jobs under unknown context-dependent service rates. Under stochastic contexts, existing algorithms achieve $\widetilde{\mathcal{O}}(T^{-1/4})$ queue length regret, defined as the expected difference between the learner's and oracle's queue lengths at horizon T. In this paper, we improve this rate to $\widetilde{\mathcal{O}}(T^{-1/2})$. The key observation is that random exploration is needed only up to a carefully chosen cutoff round, rather than throughout the entire horizon. We propose CQB-η-2, a three-phase algorithm: (i) pure random exploration to construct an initial estimator, (ii) η-random exploration combined with a UCB rule to continue learning while maintaining negative drift, and (iii) pure UCB after the exploration cutoff. Our proof decomposes the queue length regret at the cutoff round. Before the cutoff, negative drift suppresses queue length differences caused by suboptimal choices. After the cutoff, the first two phases provide sufficient random exploration samples, ensuring that UCB decisions incur small departure-rate gaps. Combining these two bounds yields queue length regret of order $\widetilde{\mathcal{O}}(T^{-1/2})$. We further prove a minimax lower bound of order $\Omega(T^{-1/2})$. The proof constructs two hard instances that are statistically indistinguishable up to the final service decision, and uses a queue-specific coupling argument to convert the resulting testing error into queue length regret. Together, our upper and lower bounds characterize the minimax dependence on the horizon T up to logarithmic factors.

1 Introduction

Queueing systems play an important role in modern service platforms, including cloud computing (Vilaplana et al., 2014), online service systems (Andrews et al., 2004), call centers (Koole and Mandelbaum, 2002), and multi-LLM services (Mitzenmacher and Shahout, 2025). In these systems, jobs arrive over time, service capacity is limited, and scheduling decisions must be made online. A central challenge is that jobs are heterogeneous and carry contexts, while the corresponding context-dependent service rates are often unknown in advance. This motivates learning-while-scheduling policies that infer unknown service rates from observed departures while keeping the queue stable (Krishnasamy et al., 2016; Kim and Oh, 2024).

Queueing bandits formalize learning-while-scheduling when service rates are unknown (Krishnasamy et al., 2016, 2021). In a discrete-time queueing bandit problem, the learner selects which server or job-server pair to use, observes binary departure feedback, and uses this feedback to improve future scheduling decisions. The objective is not only to identify choices with large departure probabilities, but also to control the queue length, because each suboptimal decision can affect future queue states. However, classical queueing bandits typically do not account for job contexts: each action has a fixed unknown departure probability, independent of the features of the jobs currently in the queue.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Contextual queueing bandits address this limitation by combining queueing bandits with contextual bandit models. In this framework, the departure probability of a job-server pair is modeled as a logistic function of its feature vector and an unknown parameter (Bae et al., 2026a,b), as in logistic and generalized linear bandits (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020).

In this paper, we study contextual queueing bandits under stochastic contexts in a discrete-time system with a single queue and K servers. In each round, the learner observes the contexts of jobs in the queue, selects a job-server pair to process, observes whether the selected job departs, and a new job may arrive. The performance measure is queue length regret, defined as the expected difference between the learner's and oracle's queue lengths at horizon T, where the oracle policy knows the true departure probabilities. Unlike standard bandit regret, which accumulates reward losses over time, queue length regret compares the queue states at horizon T. The closest prior works on contextual queueing bandits (Bae et al., 2026a,b) establish $\widetilde{\mathcal{O}}(T^{-1/4})$ queue length regret under stochastic contexts. This leaves open whether the $T^{-1/4}$ rate is intrinsic to contextual queueing bandits or a consequence of the existing algorithmic design and analysis.

We answer this question by showing that the queue length regret can be improved to $\widetilde{\mathcal{O}}(T^{-1/2})$. Our algorithm, CQB-$\eta$-2, is based on the principle that random exploration should be used to learn the unknown service model, but need not be maintained until the end of the horizon. This differs from previous algorithms, which maintain random exploration throughout the entire horizon. We also prove a minimax lower bound of order $\Omega(T^{-1/2})$, showing that the dependence on T is tight up to logarithmic factors.

Our contributions are summarized as follows:

- We propose CQB-η-2, a three-phase contextual queueing bandit algorithm motivated by the idea of cutting off random exploration after sufficient samples are collected. The algorithm first performs pure random exploration to construct an initial estimator, then uses a mixture of η-random exploration and a UCB rule to continue learning while maintaining negative drift, and finally cuts off random exploration and uses only the UCB rule.
- We prove that CQB-η-2 achieves queue length regret of order $\widetilde{\mathcal{O}}(T^{-1/2})$ under stochastic contexts, improving the previous $\widetilde{\mathcal{O}}(T^{-1/4})$ rate. This shows that random exploration until the end of the horizon is not necessary for achieving the $T^{-1/2}$ order.
- The improvement in queue length regret is due to the clever design of phase transitions. In particular, we stop pure exploration at some cutoff point. First, although the learner may still make suboptimal decisions before the cutoff point due to random exploration, we observe that a negative drift makes the resulting effect on queue length small towards the end of horizon T. After the cutoff point, we would collect $\Theta(T)$ random exploration samples, in which case pure UCB decisions have departure-rate gaps of order $T^{-1/2}$. This separation explains why stopping random exploration can improve the queue length regret rate.
- We establish a minimax lower bound of order $\Omega(T^{-1/2})$. The proof constructs two hard instances that are statistically indistinguishable up to the final service decision and uses a queue-specific coupling argument to convert the resulting testing error into queue length regret. This shows that the dependence on T is tight up to logarithmic factors.
- We provide simulations demonstrating that CQB-η-2 achieves lower empirical queue lengths than baseline algorithms.

2 Preliminaries

This section reviews the contextual queueing bandit model, the definition of queue length regret, and the policy-switching queue/coupling framework, following Bae et al. (2026a,b).

2.1 Model

We consider a discrete-time contextual queueing system with a single queue and K servers, where in each round the agent selects a job-server pair to process and a new job may arrive. At the beginning of round t, the queue state $\mathcal{X}_t$ is the collection of contexts of all remaining jobs. The queue length

is $Q(t) = |\mathcal{X}_t|$. If $\mathcal{X}_t \neq \emptyset$, the agent chooses a job-server pair $(x_t, a_t) \in \mathcal{X}_t \times [K]$. If $\mathcal{X}_t = \emptyset$, the agent selects a fixed dummy pair (x_0, a_0), and the resulting observation is not used for learning. Let $A(t) \in \{0, 1\}$ be the arrival random variable in round t, with $\mathbb{P}(A(t) = 1) = \lambda$, and let $x^{(t)}$ be the newly arriving job context when $A(t) = 1$. Let $D(t) \in \{0, 1\}$ be the departure random variable in round t. We will specify its conditional distribution below. Then the queue state and queue length evolve according to

$$\mathcal{X}_{t+1} = \mathcal{X}_t \setminus \{x_t : D(t) = 1\} \cup \{x^{(t)} : A(t) = 1\}, \quad Q(t+1) = [Q(t) + A(t) - D(t)]^+.$$

Here, $[z]^+ = \max\{0, z\}$. We use $\boldsymbol{A}(t) = (A(t), \widetilde{x}^{(t)})$, where $\widetilde{x}^{(t)} = x^{(t)}$ if $A(t) = 1$ and $\widetilde{x}^{(t)} = \widetilde{x}$ otherwise for a fixed no-arrival symbol $\widetilde{x}$. We also write $\boldsymbol{D}(t) = (D(t), (x_t, a_t))$. Let $E(t) \in \{0, 1\}$ be the exploration random variable used in round $t + 1$. The filtration is $\mathcal{F}_t = \sigma(\mathcal{X}_1, \boldsymbol{A}(1), \boldsymbol{D}(1), E(1), \ldots, \boldsymbol{A}(t-1), \boldsymbol{D}(t-1))$. We assume that arrivals are exogenous with conditional arrival probability $\mathbb{P}(A(t) = 1 \mid \mathcal{F}_t) = \lambda$, and the queue is initially empty, $\mathcal{X}_1 = \emptyset$. For a filtration $\mathcal{G}$ and a random variable Z, we write $\mathbb{E}[\cdot \mid \mathcal{G}, Z]$ for conditioning on $\mathcal{G} \vee \sigma(Z)$. For each job-server pair (x, a), let $\boldsymbol{\phi}(x, a) \in \mathbb{R}^d$ be its feature vector, and let $\boldsymbol{\theta}^* \in \mathbb{R}^d$ be the unknown parameter. Then, the departure probability follows

$$D(t) \mid \mathcal{F}_t, x_t, a_t \sim \mathrm{Bern}(\mu(\langle \boldsymbol{\phi}(x_t, a_t), \boldsymbol{\theta}^* \rangle)), \quad \mu(z) = (1 + \exp(-z))^{-1}.$$

Note that, conditional on $\mathcal{F}_t$ and the selected job-server pair (x_t, a_t), the arrival vector $\boldsymbol{A}(t)$ and the departure indicator $D(t)$ are independent.

2.2 Oracle policy and queue length regret

Let π denote our policy, and let π^* be the oracle policy that knows $\boldsymbol{\theta}^*$. Given a nonempty queue state $\mathcal{Y}$, π^* chooses a job-server pair that maximizes $\mu(\langle \boldsymbol{\phi}(x, a), \boldsymbol{\theta}^* \rangle)$ over $(x, a) \in \mathcal{Y} \times [K]$. The queue length regret at horizon T is $R_T = \mathbb{E}[Q(T) - Q^*(T)]$, where $Q^*(T)$ is the queue length under π^*.

2.3 Policy-switching queues and coupling

Let π be a given policy. For $t \in [0, T-1]$, let π_t be the policy that follows π for rounds $1, \ldots, t$ and then follows π^* for rounds $t+1, \ldots, T-1$. We construct a coupling of the policy-switching queueing processes as follows. All processes use the same arrival random variables $\boldsymbol{A}(i)$ and the same exploration random variables $E(i)$; these exploration random variables affect only rounds in which the process follows π. Let $Q_t(i)$ be the queue length at the beginning of round i in the coupled process governed by π_t, and let $(x_{t,i}, a_{t,i})$ be the job-server pair selected in round i by this process. In each round i, draw a shared random variable $U_i \sim \mathrm{Unif}(0, 1)$. Let $D_t(i) = \mathbf{1}\{U_i \leq \mu(\langle \boldsymbol{\phi}(x_{t,i}, a_{t,i}), \boldsymbol{\theta}^* \rangle)\}$ and write $\boldsymbol{D}_t(i) = (D_t(i), (x_{t,i}, a_{t,i}))$. For notational convenience, we use $\boldsymbol{D}_t(i) = 0$ and $\boldsymbol{D}_t(i) = 1$ as shorthand for the events $D_t(i) = 0$ and $D_t(i) = 1$, respectively, together with the selected job-server pair $(x_{t,i}, a_{t,i})$. This construction gives $\mathbb{E}[Q(T)] = \mathbb{E}[Q_{T-1}(T)]$ and $\mathbb{E}[Q^*(T)] = \mathbb{E}[Q_0(T)]$, while consecutive processes governed by π_t and π_{t-1} have the same queue state at the beginning of round t. Define $\psi_t(T) = Q_t(T) - Q_{t-1}(T)$ for $t \in [T-1]$. Then, we can decompose the queue length regret as

$$R_T = \mathbb{E}[Q_{T-1}(T) - Q_0(T)] = \sum_{t=1}^{T-1} \mathbb{E}[\psi_t(T)].$$

2.4 Assumptions

Throughout the paper, we impose the following standard conditions.

Assumption 1. *For all $x \in \mathcal{X}$ and $a \in [K]$, $\|\boldsymbol{\phi}(x, a)\|_2 \leq 1$. Also, for a known $S > 0$, $\boldsymbol{\theta}^* \in \Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq S\}$.*

Assumption 2. *There exist $\kappa, R > 0$ such that $1/\kappa \leq \dot{\mu}(\langle \boldsymbol{\phi}(x, a), \boldsymbol{\theta} \rangle) \leq R$ for all $x \in \mathcal{X}$, $a \in [K]$, and $\boldsymbol{\theta} \in \Theta$.*

Assumption 3. *Newly arriving contexts are i.i.d. from an unknown distribution $\mathcal{D}$. Moreover, for some $\sigma_0^2 > 0$, $\lambda_{\min}(\mathbb{E}_{x \sim \mathcal{D}}[\frac{1}{K} \sum_{a \in [K]} \boldsymbol{\phi}(x, a) \boldsymbol{\phi}(x, a)^\mathsf{T}]) \geq \sigma_0^2$.*

Assumption 4. *There exists a traffic slackness parameter $\epsilon > 0$ such that for every $x \in \mathcal{X}$, there exists $a^*(x) \in [K]$ satisfying $\mu(\langle \boldsymbol{\phi}(x, a^*(x)), \boldsymbol{\theta}^* \rangle) - \lambda \geq \epsilon$.*

Algorithm 1 CQB-η-2

input $d, T, K, S, \kappa, \lambda_0, \lambda, \delta, \epsilon, \sigma_0$
initialize τ_1, τ_2 as in (4), $\mathbf{V}_0 \leftarrow \lambda_0 \mathbf{I}$, $\widehat{\boldsymbol{\theta}}_0 \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  if $t \le \tau_1$ **then** ▷ Phase 1: pure random exploration
    $\eta \leftarrow 1$
  else if $\tau_1 < t \le \tau_2$ **then** ▷ Phase 2: random exploration with probability η and UCB
    $\eta \leftarrow \epsilon/4$
  else ▷ Phase 3: pure UCB
    $\eta \leftarrow 0$
  end if
  Sample $E(t-1) \sim \mathrm{Bern}(\eta)$
  if $A(t-1) = 1$ **and** $E(t-1) = 1$ **then**
    $x_t \leftarrow x^{(t-1)}, \quad a_t \sim \mathrm{Unif}([K])$
  else
    $(x_t, a_t) \leftarrow \arg\max_{x \in \mathcal{X}_t,\, a \in [K]} \mathrm{UCB}_t(x, a)$ $\quad (\mathrm{UCB}_t(x,a)$ as in (1))
  end if
  Choose (x_t, a_t) and observe $D(t)$
  $\mathbf{V}_t \leftarrow \mathbf{V}_{t-1} + \boldsymbol{\phi}(x_t, a_t)\boldsymbol{\phi}(x_t, a_t)^\mathsf{T}$
  Update $\widehat{\boldsymbol{\theta}}_t$ as in (2), β_t as in (3)
end for

Assumptions 1 and 2 are standard in logistic bandits (Filippi et al., 2010). Assumption 3 guarantees sufficient feature diversity, which is also standard in finite-armed logistic bandits (Li et al., 2017). Assumption 4 imposes a traffic slackness condition that guarantees negative drift under the oracle policy, and similar conditions can be found in Krishnasamy et al. (2016); Kim and Oh (2024); Bae et al. (2026a).

3 Improved queue length regret via random exploration cutoff

3.1 Motivation and comparison with previous works

The closest prior works on contextual queueing bandits, due to Bae et al. (2026a,b), obtain $\widetilde{\mathcal{O}}(T^{-1/4})$ queue length regret. There are two reasons to expect that this rate can be improved. First, since queue length regret is evaluated at a terminal horizon rather than accumulated over all rounds, the $\sqrt{T}$ cumulative-regret scale in logistic bandits suggests the possibility of a $T^{-1/2}$ rate for this terminal-horizon performance measure. Second, the algorithms of Bae et al. (2026a,b) both have to maintain random exploration until the end of the horizon. This requirement comes from the proof technique. To elaborate, by the decomposition result in Lemma 6, previous proofs control the departure-rate gap in the first square-root factor by a monotone upper bound, and maintaining random exploration until the end of the horizon preserves this monotone control. From an algorithmic perspective, however, after sufficiently accurate estimation is achieved through early random exploration, it is natural to terminate random exploration and use only the UCB rule. Motivated by these two observations, we propose a three-phase algorithm that cuts off random exploration in the final phase.

3.2 Proposed algorithm

Algorithm 1 presents our CQB-η-2 algorithm. We use the term *random exploration* for the rule that, when a new job arrives ($A(t-1) = 1$) and the exploration flag is set ($E(t-1) = 1$), selects the newly arrived job and chooses a server uniformly at random. The algorithm has three phases. In Phase 1, it performs random exploration with probability one on arrivals. In Phase 2, it combines random exploration with probability η and a UCB rule, where we set $\eta = \epsilon/4$. In Phase 3, it cuts off random exploration and uses only the UCB rule.

The phase transition points τ_1 and τ_2 are defined in the regret analysis in Section 3.3. At a high level, τ_1 ensures that pure random exploration in Phase 1 provides enough information for the rule that

combines random exploration with probability η and UCB in Phase 2 to guarantee a negative drift. The transition point τ_2 ensures that the same Phase 2 rule achieves sufficient estimation accuracy, so that the pure UCB rule is sufficient in Phase 3.

For any job-server pair (x, a), the UCB is calculated as

$$\mathrm{UCB}_t(x,a) = \mu(\langle \boldsymbol{\phi}(x,a), \widehat{\boldsymbol{\theta}}_{t-1}\rangle) + \beta_{t-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{t-1}^{-1}}. \tag{1}$$

We use the regularized logistic maximum likelihood estimator

$$\widehat{\boldsymbol{\theta}}_t \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{t} [-D(i)\log\mu(\langle \boldsymbol{\phi}_i, \boldsymbol{\theta}\rangle) - (1-D(i))\log(1-\mu(\langle \boldsymbol{\phi}_i, \boldsymbol{\theta}\rangle))] + \frac{\lambda_0}{2}\|\boldsymbol{\theta}\|_2^2. \tag{2}$$

and the confidence radius given by

$$\beta_t \leftarrow \kappa\sqrt{0.5d\log(1+t/(\kappa\lambda_0 d)) + 0.5\log(1/\delta)} + S\sqrt{\kappa\lambda_0}. \tag{3}$$

We adopt this simple form of the estimator and confidence radius because they are not the main focus of our contribution.

Finally, we provide the main theorem of Algorithm 1.

Theorem 5. *Suppose that Assumptions 1 to 4 hold. Set $\delta = T^{-4}$. For every T satisfying $\tau_2 - \tau_1 \geq T/2$, the queue length regret of Algorithm 1 satisfies*

$$R_T = \mathcal{O}\left(\left(\frac{d+\log(T)}{\lambda\epsilon\sigma_0^4} + \frac{\kappa^2 d\log(T)}{\lambda\epsilon^3\sigma_0^2} + \frac{\log(T)}{\lambda\epsilon} + \frac{\log(T)}{\epsilon^2}\right)\frac{\kappa\sqrt{d\log(T)}}{\sqrt{\lambda\epsilon}\,\sigma_0\sqrt{T}}\right).$$

3.3 Proof of theorem 5

The improved queue length regret comes from how the policy-switching decomposition is used. The decomposition itself is from Bae et al. (2026a), but previous analyses use it with a monotone departure-rate gap bound over the entire horizon. This requires random exploration until round T. We instead split the regret at the exploration cutoff. For rounds before the cutoff, we control the effect at horizon T of a queue length difference by negative drift. For rounds after the cutoff, we control the departure-rate gap by the estimation accuracy obtained from the $\Theta(T)$ random exploration samples collected before the cutoff. We now recall the decomposition and then make this split explicit. For the queue state $\mathcal{X}_t$ under π, define $(x_t^*, a_t^*) \in \arg\max_{x\in\mathcal{X}_t, a\in[K]} \mu(\langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^*\rangle)$. Let $\boldsymbol{\phi}_t^* = \boldsymbol{\phi}(x_t^*, a_t^*)$ and $\boldsymbol{\phi}_t = \boldsymbol{\phi}(x_t, a_t)$. Also define $\mathcal{F}_t^+ = \mathcal{F}_t \vee \sigma(E(t-1), \boldsymbol{A}(t))$ and

$$\widetilde{\psi}_t(T) = \mathbb{E}[\psi_t(T) \mid \mathcal{F}_t^+, \boldsymbol{D}_t(t) = 0, \boldsymbol{D}_{t-1}(t) = 1].$$

The following decomposition corresponds to Lemma D.2 of Bae et al. (2026a). It separates the effect of the decision in round t into the departure-rate gap in that round and the effect at horizon T of the queue-length difference created in that round.

Lemma 6. *For each $t \in [T-1]$, we have*

$$\mathbb{E}[\psi_t(T)] \leq \sqrt{\mathbb{E}\left[(\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^*\rangle) - \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle))^2\right]} \cdot \sqrt{\mathbb{E}\left[\widetilde{\psi}_t(T)\right]}.$$

In Lemma 6, the first square-root factor is the departure-rate gap, and the second square-root factor is the effect at horizon T of the queue length difference created in round t. The key point is that these two factors do not have to be controlled by the same type of bound in all rounds. Before the cutoff, we use a naive upper bound on the departure-rate gap, but the second factor is small by negative drift. After the cutoff, we use a naive upper bound on the second factor, but the departure-rate gap is small because the first two phases have already collected $\Theta(T)$ random exploration samples. This is the step that leads to the $T^{-1/2}$ rate.

We also introduce a lemma showing that, under our coupling construction, the two consecutive policy-switching processes governed by π_t and π_{t-1} differ in their policy only at round t, and hence their queue states can differ by at most one in all subsequent rounds.

Lemma 7. *For each $t \in [T-1]$, we have $\psi_t(T) \in \{-1, 0, 1\}$.*

Next, define the phase transition points τ_1, τ_2 as follows: For sufficiently large constants $c_1, c_2 > 0$,

$$\tau_1 = \frac{2}{\lambda}\left[\left(\frac{c_1\sqrt{d} + c_2\sqrt{\log(1/\delta)}}{\sigma_0^2}\right)^2 + \frac{128\beta_T^2}{\epsilon^2\sigma_0^2}\right] + \frac{8\log(1/\delta)}{\lambda}, \quad \tau_2 = T - \frac{4\tau_1}{\epsilon} - \frac{128\log(T)}{\epsilon^2} - 1. \tag{4}$$

For the upper-bound proof, we set $\delta = T^{-4}$ in β_t and τ_1. We now bound the two terms in Lemma 6. We first show that the first term, the departure-rate gap term, is upper bounded at order $T^{-1/2}$ after τ_2, that is, after Phase 2 has collected enough samples of order $\Theta(T)$.

Lemma 8. *If $\tau_2 - \tau_1 \geq T/2$, then for every $t \in [\tau_2 + 1, T-1]$, we have*

$$\sqrt{\mathbb{E}[(\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^*\rangle) - \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle))^2]} \leq \sqrt{T^{-4} + \nu_{\mathrm{ph2}} + \frac{128\beta_T^2}{\lambda\epsilon\sigma_0^2 T}}.$$

where $\nu_{\mathrm{ph2}} = \exp(-\frac{\lambda\epsilon(\tau_2-\tau_1)}{32}) + d\exp(-\frac{\lambda\epsilon(\tau_2-\tau_1)\sigma_0^2}{64})$.

Next, we show that the second term in the decomposition in Lemma 6, the effect of a queue length difference created in round t, is upper bounded by an exponentially increasing sequence clipped at one by Lemma 7.

Lemma 9. *For $t > T - 4\tau_1/\epsilon - 1$, we have $\sqrt{\mathbb{E}[\widetilde{\psi}_t(T)]} \leq 1$. For $t \leq T - 4\tau_1/\epsilon - 1$, we have*

$$\sqrt{\mathbb{E}\left[\widetilde{\psi}_t(T)\right]} \leq \min\left\{1, \sqrt{3T^{-4} + 19\epsilon^{-2}\exp\left(-\frac{\epsilon^2}{32}(T - t - 1 - 4\tau_1/\epsilon)\right)}\right\}.$$

The number of rounds with positive drift up to round t affects the threshold value $4\tau_1/\epsilon$ that separates the two cases in Lemma 9. Notice that this threshold value does not depend on τ_2. This is because our choice of τ_1 and the exploration rule in Phase 2 allow the algorithm to collect sufficiently many samples of order $\Theta(T)$ through random exploration while maintaining negative drift during Phase 2. This is the key reason why, unlike previous works, our algorithm can cut off random exploration in the final phase. Now, we are ready to start the proof of Theorem 5.

Proof of Theorem 5. Recall the queue length regret decomposition of $R_T = \sum_{t=1}^{T-1}\mathbb{E}[\psi_t(T)]$. We analyze the queue length regret before and after τ_2 by writing $R_T = A_1 + A_2$, where $A_1 = \sum_{t=1}^{\tau_2}\mathbb{E}[\psi_t(T)]$ and $A_2 = \sum_{t=\tau_2+1}^{T-1}\mathbb{E}[\psi_t(T)]$. This split follows the two different uses of Lemma 6. In A_1, we control the effect of a queue-length difference by negative drift. In A_2, we control the departure-rate gap by the estimation accuracy obtained before the cutoff. For A_1, we use the decomposition in Lemma 6 and consider the departure-rate gap term and the effect of a queue length difference term for $t \in [1, \tau_2]$. For the departure-rate gap term, we use a naive upper bound of one as $\sqrt{\mathbb{E}[(\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^*\rangle) - \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle))^2]} \leq 1$. For the effect of a queue length difference term,

$$T - t - 1 - 4\epsilon^{-1}\tau_1 \geq T - \tau_2 - 1 - 4\epsilon^{-1}\tau_1 = 128\epsilon^{-2}\log(T)$$

for every $t \leq \tau_2$. Thus Lemma 9 gives $\sqrt{\mathbb{E}[\widetilde{\psi}_t(T)]} \leq \sqrt{3 + 19\epsilon^{-2}}\,T^{-2}$. Therefore,

$$A_1 \leq \sum_{t=1}^{\tau_2}\sqrt{3 + 19\epsilon^{-2}}\,T^{-2} \leq \sqrt{3 + 19\epsilon^{-2}}\,T^{-1}. \tag{5}$$

Next, consider A_2. We again apply Lemma 6 for $t \in [\tau_2 + 1, T-1]$. For the departure-rate gap term, we use Lemma 8 to upper bound it by

$$\sqrt{\mathbb{E}[(\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^*\rangle) - \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle))^2]} \leq \sqrt{T^{-4} + \nu_{\mathrm{ph2}} + \frac{128\beta_T^2}{\lambda\epsilon\sigma_0^2 T}}.$$

For the effect of a queue length difference term, we use the naive bound from Lemma 9 to upper bound it by one as $\sqrt{\mathbb{E}[\widetilde{\psi}_t(T)]} \leq 1$. Therefore,

$$A_2 \leq \sum_{t=\tau_2+1}^{T-1}\left(\sqrt{T^{-4} + \nu_{\mathrm{ph2}} + \frac{128\beta_T^2}{\lambda\epsilon\sigma_0^2 T}} \cdot 1\right) = \left(\frac{4\tau_1}{\epsilon} + \frac{128\log(T)}{\epsilon^2}\right)\sqrt{T^{-4} + \nu_{\mathrm{ph2}} + \frac{128\beta_T^2}{\lambda\epsilon\sigma_0^2 T}}. \tag{6}$$

For ν_{ph2}, by the assumption of Theorem 5, we have $\tau_2 - \tau_1 \geq T/2$. Thus $\nu_{\mathrm{ph2}} \leq \exp(-\frac{\lambda\epsilon T}{64}) + d\exp(-\frac{\lambda\epsilon\sigma_0^2 T}{128})$, which is negligible. Since we set $\delta = T^{-4}$, $\beta_T = \mathcal{O}(\kappa\sqrt{d\log(T)})$, and $\tau_1 = \mathcal{O}(\frac{d+\log(T)}{\lambda\sigma_0^4} + \frac{\kappa^2 d\log(T)}{\lambda\epsilon^2\sigma_0^2} + \frac{\log(T)}{\lambda})$. Substituting these bounds into (5) and (6) completes the proof. □

4 Queue length regret lower bound analysis

We now prove that the $T^{-1/2}$ dependence in Theorem 5 is tight in terms of the horizon T. The proof uses a two-instance testing argument. The two instances have the same arrival process and context distribution, but their optimal server sets are complementary. They are statistically difficult to distinguish from the observed history before round $T-1$, while the service decision in round $T-1$ can still change the terminal queue length. We compare an arbitrary policy with a comparison policy that changes only the server chosen in round $T-1$, and then use a testing argument conditioned on the event that a job arrives in round $T-2$ to convert the testing error into queue length regret.

For an instance ν and a policy π, let $Q^\pi(T)$ be the queue length at horizon T under π, and let $Q^*(T)$ be the queue length at horizon T under the oracle policy. Write $R_T(\pi;\nu) = \mathbb{E}_\nu[Q^\pi(T) - Q^*(T)]$. For fixed $d, K, S, \sigma_0^2, \lambda$, and ϵ, let $\mathcal{C}(d, K, S, \sigma_0^2, \lambda, \epsilon)$ be the class of contextual queueing bandit instances satisfying Assumptions 1 to 4 with dimension d, K servers, parameter radius S, feature diversity parameter σ_0^2, arrival rate λ, and traffic slackness parameter ϵ.

Theorem 10. *Fix $d \geq 2$, $K \geq 2$, $\epsilon \in (0,1)$, and $\lambda \in (0, 1-\epsilon)$. Let $0 < \sigma_0^2 \leq 1/d$, and let $B_{\lambda,\epsilon} = \log((8 - 3(1-\lambda-\epsilon))/(3(1-\lambda-\epsilon)))$. Assume $S \geq B_{\lambda,\epsilon}\sqrt{d}$. Then, for every $T \geq 3$,*

$$\inf_{\pi} \sup_{\nu \in \mathcal{C}(d,K,S,\sigma_0^2,\lambda,\epsilon)} R_T(\pi;\nu) \geq \frac{\lambda(1-\lambda-\epsilon)e^{-1/6}}{16\sqrt{T}}.$$

Remark 11. *The condition $S \geq B_{\lambda,\epsilon}\sqrt{d}$ is used to cover the full range $\lambda + \epsilon < 1$. When $\lambda + \epsilon$ is close to one, the hard instances require service probabilities close to one, which correspond to large logistic logits and therefore require a large parameter radius S.*

4.1 Proof of theorem 10

We first introduce an explicit hard instance construction.

Lemma 12. *Under the conditions of Theorem 10, fix $T \geq 3$ and set*

$$\gamma = 1 - \lambda - \epsilon, \quad \Delta_T = \gamma/(8\sqrt{T}), \quad p_T = 1 - \gamma/2 + \Delta_T, \quad q_T = 1 - \gamma/2 - \Delta_T.$$

Let $z_p = \log(p_T/(1-p_T))$ and $z_q = \log(q_T/(1-q_T))$. Let the context set be

$$\mathcal{X}_{\mathrm{lb}} = \{x(\sigma) = d^{-1/2}(1, \sigma_2, \ldots, \sigma_d)^\mathsf{T} : \sigma_j \in \{-1,+1\},\ j = 2, \ldots, d\},$$

and let the context distribution be uniform on $\mathcal{X}_{\mathrm{lb}}$. Let $\mathbf{M} = \mathrm{diag}(1, -1, 1, 1, \ldots, 1)$, and define $\boldsymbol{\phi}(x, 1) = x$ and $\boldsymbol{\phi}(x, a) = \mathbf{M}x$ for $a = 2, \ldots, K$. Let e_j be the j-th unit vector, and define

$$\boldsymbol{\theta}_+^* = 0.5\sqrt{d}((z_p + z_q)e_1 + (z_p - z_q)e_2), \qquad \boldsymbol{\theta}_-^* = 0.5\sqrt{d}((z_p + z_q)e_1 - (z_p - z_q)e_2).$$

Let ν_T^+ and ν_T^- be the two instances with arrival rate λ, this context distribution and feature map, and unknown parameters $\boldsymbol{\theta}_+^*$ and $\boldsymbol{\theta}_-^*$, respectively. Then $\nu_T^+, \nu_T^- \in \mathcal{C}(d, K, S, \sigma_0^2, \lambda, \epsilon)$.

The next proposition characterizes the optimal server sets and departure-probability properties of the hard instances.

Proposition 13. *For the hard instances in Lemma 12, define, for each $x \in \mathcal{X}_{\mathrm{lb}}$, $\xi(x) = \sqrt{d}\, x_2 \in \{-1, +1\}$ and*

$$\mathcal{S}_+(x) = \begin{cases} \{1\}, & \xi(x) = +1, \\ \{2, \ldots, K\}, & \xi(x) = -1, \end{cases} \qquad \mathcal{S}_-(x) = [K] \setminus \mathcal{S}_+(x).$$

For every $x \in \mathcal{X}_{\mathrm{lb}}$ and $s \in \{+,-\}$, servers in $\mathcal{S}_s(x)$ have departure probability p_T under ν_T^s, and servers outside $\mathcal{S}_s(x)$ have departure probability q_T under ν_T^s. Moreover, $p_T > q_T > \lambda + \epsilon$, $p_T - q_T = (1-\lambda-\epsilon)/(4\sqrt{T})$, and $u(1-u) \geq 3(1-\lambda-\epsilon)/16$ for all $u \in [q_T, p_T]$.

Next, we state a pathwise comparison lemma for the hard instances.

Lemma 14. *Fix $s \in \{+,-\}$ and consider the hard instance ν_T^s in Lemma 12. For any deterministic policy, there exists a coupling under which the oracle queue is pathwise no larger than the queue generated by that policy at every round.*

Finally, let $F = \{A(T-2) = 1\}$. The following proposition is a conditional consequence of the KL chain rule and the Bretagnolle–Huber inequality (Lemma 26).

Proposition 15. *Fix a deterministic policy π and two instances ν^+, ν^- with the same arrival process and context distribution. Let $\mathcal{H}_{T-1} = \sigma(\boldsymbol{A}(1), \boldsymbol{D}(1), \ldots, \boldsymbol{A}(T-2), \boldsymbol{D}(T-2))$. Write $\mathbb{P}_+$ and $\mathbb{P}_-$ for probabilities under ν^+ and ν^-, respectively. Suppose that, up to round $T-2$, the conditional departure distributions under the two instances are either identical or equal to $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$ in some order. If it holds that*

$$(T-2)\max\{D_{\mathrm{KL}}(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)), D_{\mathrm{KL}}(\mathrm{Bern}(q)\|\,\mathrm{Bern}(p))\} \le 1/6$$

, then for every event $G \in \mathcal{H}_{T-1}$, we have

$$\mathbb{P}_+(G^c \mid F) + \mathbb{P}_-(G \mid F) \ge 0.5e^{-1/6}.$$

Now, we are ready to start the proof of Theorem 10.

Proof of Theorem 10. Fix a deterministic policy π. Let ν_T^+, ν_T^-, p_T, and q_T be as in Lemma 12, and let $\mathcal{S}_+(\cdot)$ and $\mathcal{S}_-(\cdot)$ be as in Proposition 13. For $s \in \{+,-\}$, write $\mathbb{P}_s$ and $\mathbb{E}_s$ for probability and expectation under ν_T^s. For each $s \in \{+,-\}$, define a comparison policy π^s that agrees with π in rounds $1, \ldots, T-2$ and, in round $T-1$, selects the same job as π but chooses the smallest server in $\mathcal{S}_s(x)$ for that job context x. Let $Q^s(T)$ and $D^s(T-1)$ be the queue length and departure random variable under π^s. By Lemma 14, under ν_T^s, $Q^*(T) \le Q^s(T)$, and hence

$$R_T(\pi; \nu_T^s) \ge \mathbb{E}_s[Q^\pi(T) - Q^s(T)].$$

Recall that $F = \{A(T-2) = 1\}$. On F, the queue under π is nonempty at the beginning of round $T-1$. Since π and π^s have the same queue state at the beginning of round $T-1$ and the same arrival in round $T-1$,

$$R_T(\pi; \nu_T^s) \ge \mathbb{E}_s[\mathbf{1}\{F\}(D^s(T-1) - D^\pi(T-1))].$$

Let (x_{T-1}, a_{T-1}) be the job-server pair selected by π in round $T-1$, with an arbitrary fixed value if the queue is empty, and set

$$G = \{a_{T-1} \in \mathcal{S}_+(x_{T-1})\}.$$

On $F \cap G^c$, π chooses a suboptimal server under ν_T^+. On $F \cap G$, it chooses a suboptimal server under ν_T^-. Since F depends only on the arrival process, $\mathbb{P}_+(F) = \mathbb{P}_-(F) = \lambda$. By Proposition 13, under ν_T^+, the conditional expectation of $D^+(T-1) - D^\pi(T-1)$ is $p_T - q_T$ on $F \cap G^c$ and zero on $F \cap G$. Under ν_T^-, the same conditional expectation is $p_T - q_T$ on $F \cap G$ and zero on $F \cap G^c$. Thus we have

$$R_T(\pi; \nu_T^+) \ge \lambda(p_T - q_T)\,\mathbb{P}_+(G^c \mid F), \qquad R_T(\pi; \nu_T^-) \ge \lambda(p_T - q_T)\,\mathbb{P}_-(G \mid F). \tag{7}$$

Let $\mathcal{H}_{T-1} = \sigma(\boldsymbol{A}(1), \boldsymbol{D}(1), \ldots, \boldsymbol{A}(T-2), \boldsymbol{D}(T-2))$. Since π is deterministic, $G \in \mathcal{H}_{T-1}$. For each round up to $T-2$, if the queue is empty, the departure random variable is deterministically zero under both instances. Otherwise, by Proposition 13, the conditional departure distributions under ν_T^+ and ν_T^- are either identical or equal to $\mathrm{Bern}(p_T)$ and $\mathrm{Bern}(q_T)$ in some order. By the Bernoulli KL bound (Lemma 25) and Proposition 13,

$$\begin{aligned}&\max\{D_{\mathrm{KL}}(\mathrm{Bern}(p_T)\|\,\mathrm{Bern}(q_T)), D_{\mathrm{KL}}(\mathrm{Bern}(q_T)\|\,\mathrm{Bern}(p_T))\}\\ &\le \frac{(p_T - q_T)^2}{2(3(1-\lambda-\epsilon)/16)} = \frac{1-\lambda-\epsilon}{6T} \le \frac{1}{6T}.\end{aligned}$$

Thus the KL condition of Proposition 15 holds. Applying Proposition 15 gives $\mathbb{P}_+(G^c \mid F) + \mathbb{P}_-(G \mid F) \ge e^{-1/6}/2$. Substituting this bound into (7) gives

$$\begin{aligned}R_T(\pi; \nu_T^+) + R_T(\pi; \nu_T^-) &\ge \lambda(p_T - q_T)(\mathbb{P}_+(G^c \mid F) + \mathbb{P}_-(G \mid F))\\ &\ge \frac{\lambda(p_T - q_T)e^{-1/6}}{2} = \frac{\lambda(1-\lambda-\epsilon)e^{-1/6}}{8\sqrt{T}},\end{aligned}$$

Therefore, $\max\{R_T(\pi; \nu_T^+), R_T(\pi; \nu_T^-)\} \ge \lambda(1-\lambda-\epsilon)e^{-1/6}/(16\sqrt{T})$, which completes the proof. For a randomized policy, conditioning on its internal random seed and averaging gives the same sum lower bound, and hence the same maximum lower bound. □

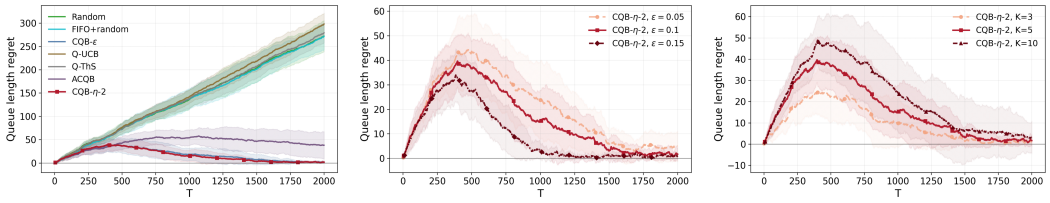


Figure 1: Queue length regret in synthetic experiments. Left: comparison with baselines. Middle: CQB-η-2 with $\epsilon \in \{0.05, 0.1, 0.15\}$. Right: CQB-$\eta$-2 with $K \in \{3, 5, 10\}$.

5 Experiments

We compare the empirical queue length regret of CQB-η-2 with standard baselines. We use $T = 2000$, $d = 10$, $K = 5$, $\lambda = 0.7$, $\epsilon = 0.1$, and $\kappa = 50$. The feature vectors and the unknown parameter are sampled coordinate-wise from $\mathrm{Unif}[-1, 1]$, and instances that do not satisfy the slackness and κ assumptions are rejected. All curves are averaged over 10 independent runs, and the error bands indicate ± 1 standard deviation. For the practical implementation, we use $\delta = T^{-1}$ and scale the confidence radius by 0.1. For CQB-η-2, we set the phase transition points to $\tau_1 = 0.2T$ and $\tau_2 = 0.3T$. For CQB-ε, we set the initial random-exploration length to $\tau = 0.2T$, which plays the same role as τ_1 in CQB-η-2. For ACQB, the random-exploration probability is $\eta(t) = \min\{1, c_1/\sqrt{t+1}\}$, and we set $c_1 = 20$. The random policy selects both the job and the server uniformly at random, and FIFO+random selects the oldest job in the queue and a server uniformly at random.

Figure 1 shows that CQB-η-2 and CQB-ε have the smallest final regret in the baseline comparison. The middle panel shows that larger slackness improves performance, and the right panel shows that CQB-η-2 remains stable across the tested values of K. 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.

6 Conclusion

We studied contextual queueing bandits under stochastic contexts and showed that the queue length regret can be improved from the previous $\widetilde{\mathcal{O}}(T^{-1/4})$ rate to $\widetilde{\mathcal{O}}(T^{-1/2})$. The main algorithmic idea is to stop random exploration after enough samples have been collected and use a pure UCB rule in the final phase. We also proved an $\Omega(T^{-1/2})$ lower bound, showing that the dependence on T is tight up to logarithmic factors.

Our bounds match in their dependence on T up to logarithmic factors, while sharp dependence on ϵ, κ, d, and σ_0^{-1} remains open. Hard instances with service rates close to the arrival rate may yield stronger lower bounds in ϵ. Confidence bounds that adapt to the service variance may improve the dependence on κ, but the burn-in analysis must also be refined. Another direction is to remove prior knowledge of ϵ and T, potentially using a lower confidence bound on service slack and a doubling trick that reuses observations and preserves the current queue. Such extensions would require control of negative drift and estimation accuracy as the exploration schedule changes. An anytime policy with pointwise queue length regret $\widetilde{\mathcal{O}}\left(t^{-1/2}\right)$ would imply $\widetilde{\mathcal{O}}\left(\sqrt{T}\right)$ cumulative queue length regret and vanishing regret in average holding cost under unit costs. Extensions to holding costs that depend on job contexts also remain open.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2026-25518141). The authors declare no competing interests.

References

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

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

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.

Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6): 665–688, 2007.

Matthew Andrews, Krishnan Kumaran, Kavita Ramanan, Alexander Stolyar, Rajiv Vijayakumar, and Phil Whiting. Scheduling in a queuing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences*, 18(2):191–217, 2004.

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*, 2026a.

Seoungbin Bae, Junyoung Son, and Dabeen Lee. Learning to route and schedule llms from user retrials via contextual queueing bandits. *arXiv preprint arXiv:2602.02061*, 2026b.

Agrim Bari, Parikshit Hegde, and Gustavo de Veciana. Optimal scheduling algorithms for llm inference: Theory and practice. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 9(3):1–43, 2025.

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*, pages 181–190, 2021.

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

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

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

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.

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

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

Jiatai Huang, Leana Golubchik, and Longbo Huang. When lyapunov drift based queue scheduling meets adversarial bandit learning. *IEEE/ACM Transactions on Networking*, 32(4):3034–3044, 2024.

Neharika Jali, Guannan Qu, Weina Wang, and Gauri Joshi. Efficient reinforcement learning for routing jobs in heterogeneous queueing systems. In *International Conference on Artificial Intelligence and Statistics*, pages 4177–4185. PMLR, 2024.

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*, pages 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.

Ger Koole and Avishai Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59, 2002.

G Krishnakumar and Abhishek Sinha. Minimizing queue length regret for arbitrarily varying channels. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

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.

Jiung Lee, Hongseok Namkoong, and Yibo Zeng. Design and scheduling of an ai-based queueing system. *arXiv preprint arXiv:2406.06855*, 2024a.

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

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024b.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 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.

Woei Lin and P Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic control*, 29(8):696–703, 1984.

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

Yashaswini Murthy, Isaac Grosof, Siva Theja Maguluri, and R Srikant. Performance of npg in countable state-space average-cost rl. *arXiv preprint arXiv:2405.20467*, 2024.

Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers, 2010.

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.

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

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

Jordi Vilaplana, Francesc Solsona, Ivan Teixidó, Jordi Mateo, Francesc Abella, and Josep Rius. A queuing theory model for cloud computing. *The Journal of Supercomputing*, 69(1):492–507, 2014.

Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. Mixllm: Dynamic routing in mixed large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10912–10922, 2025.

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

Yuqing Yang, Yuedong Xu, and Lei Jiao. A queueing theoretic perspective on low-latency llm inference with variable token length, 2025. URL https://arxiv.org/abs/2407.05347.

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

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.

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

Queueing and contextual queueing bandits. Queueing bandits study learning-while-scheduling problems in which unknown service rates must be learned while queue lengths are controlled (Krishnasamy et al., 2016, 2021). This line of work includes queue length regret, dispatching under unknown service rates, decentralized queueing systems, MaxWeight-type learning algorithms, adversarial or nonstationary queueing models, and related queue-aware learning problems (Stahlbuhk et al., 2021; Choudhury et al., 2021; Sentenac et al., 2021; Freund et al., 2022; Yang et al., 2023; Liang and Modiano, 2018; Huang et al., 2024; Krishnakumar and Sinha, 2025; Wijewardena et al., 2025; Gaitonde and Tardos, 2020; Hsu et al., 2022). Classical queueing-control and scheduling works provide stability, drift, and routing tools that underlie these models (Lin and Kumar, 1984; Neely, 2010; Koole and Mandelbaum, 2002; Andrews et al., 2004; Aksin et al., 2007; Vilaplana et al., 2014). Recent queueing and scheduling studies also consider modern service systems, including LLM inference and multi-LLM routing (Mitzenmacher and Shahout, 2025; Yang et al., 2025; Fu et al., 2024; Lee et al., 2024a; Jali et al., 2024; Bari et al., 2025; Murthy et al., 2024). Contextual queueing bandit variants incorporate job-specific features into departure models (Kim and Oh, 2024; Bae et al., 2026a,b). The closest prior works Bae et al. (2026a,b) obtain $\widetilde{\mathcal{O}}(T^{-1/4})$ queue length regret under stochastic contexts, whereas we prove a $\widetilde{\mathcal{O}}(T^{-1/2})$ upper bound and an $\Omega(T^{-1/2})$ lower bound that match in T up to logarithmic factors.

Logistic bandits. Our logistic departure model is related to generalized linear bandits, logistic bandits, and multinomial-logit bandits (Filippi et al., 2010; Abbasi-Yadkori et al., 2011; Li et al., 2017; Faury et al., 2020; Jun et al., 2021; Abeille et al., 2021; Lee et al., 2024b; Bae and Lee, 2025; Lee and Oh, 2025; Zhang et al., 2025; Agrawal et al., 2019; Oh and Iyengar, 2019). Related choice-model and routing bandits also study categorical feedback under logistic-type models (Chiang et al., 2025; Wang et al., 2025; Ong et al., 2024; Shirkavand et al., 2025; Chen et al., 2024; Zhuang et al., 2024). Unlike these works, contextual queueing bandits measure queue length regret, and the available job-server pairs depend on the queue state; this requires queue-specific arguments rather than a direct application of standard logistic-bandit regret analyses.

B Deferred proofs for section 3.3

B.1 Good events

Throughout this appendix, set $\delta = T^{-4}$ in the confidence radius. For $t \in [T]$, define

$$\mathcal{E}_{\mathrm{pred}}(t) = \{\forall s \le t, (x,a) \in \mathcal{X} \times [K], \\ |\mu(\langle \boldsymbol{\phi}(x,a), \widehat{\boldsymbol{\theta}}_{s-1}\rangle) - \mu(\langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^*\rangle)| \le \beta_{s-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{s-1}^{-1}}\}.$$

Also define

$$\mathcal{E}_{\mathrm{burn}}(t) = \{\forall s \in \{\tau_1+1, \dots, t\}, (x,a) \in \mathcal{X}_s \times [K],\ \beta_{s-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{s-1}^{-1}} \le \epsilon/8\}.$$

Let $\mathcal{E}_{\mathrm{drift}}(t) = \mathcal{E}_{\mathrm{pred}}(t) \cap \mathcal{E}_{\mathrm{burn}}(t)$. We write $\mathcal{E}_{\mathrm{pred}} = \mathcal{E}_{\mathrm{pred}}(T)$, $\mathcal{E}_{\mathrm{burn}} = \mathcal{E}_{\mathrm{burn}}(T)$, and $\mathcal{E}_{\mathrm{drift}} = \mathcal{E}_{\mathrm{drift}}(T)$. Then $\mathcal{E}_{\mathrm{pred}}(t)$, $\mathcal{E}_{\mathrm{burn}}(t)$, and $\mathcal{E}_{\mathrm{drift}}(t)$ are $\mathcal{F}_t$-measurable. If $t_1 \le t_2$, then $\mathcal{E}_{\mathrm{drift}}(t_2) \subseteq \mathcal{E}_{\mathrm{drift}}(t_1)$.

Let

$$N_{\mathrm{ph2}} = \sum_{s=\tau_1+1}^{\tau_2} \mathbf{1}\{A(s-1) = 1,\ E(s-1) = 1\}$$

and

$$\mathbf{S}_{\mathrm{ph2}} = \sum_{s=\tau_1+1}^{\tau_2} \mathbf{1}\{A(s-1) = 1,\ E(s-1) = 1\}\boldsymbol{\phi}_s\boldsymbol{\phi}_s^{\mathsf{T}}.$$

Define

$$\mathcal{E}_{\mathrm{ph2}} = \left\{\lambda_{\min}(\mathbf{S}_{\mathrm{ph2}}) \ge \frac{\lambda\epsilon(\tau_2-\tau_1)\sigma_0^2}{16}\right\}.$$

Finally, let

$$\nu_{\mathrm{ph2}} = \exp\left(-\frac{\lambda\epsilon(\tau_2-\tau_1)}{32}\right) + d\exp\left(-\frac{\lambda\epsilon(\tau_2-\tau_1)\sigma_0^2}{64}\right).$$

Lemma 16. *With the confidence radius in (3) and the phase transition points in (4), for sufficiently large absolute constants $c_1, c_2 > 0$, we have*

$$\mathbb{P}(\mathcal{E}_{\mathrm{pred}}^c) \le T^{-4}, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{drift}}^c) \le 3T^{-4}, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{ph2}}^c) \le \nu_{\mathrm{ph2}}.$$

Proof. The first bound follows from Lemma 18 with confidence level $\delta = T^{-4}$. Lemma 20, with the same confidence level, gives $\mathbb{P}(\mathcal{E}_{\mathrm{burn}}^c) \le 2T^{-4}$. Therefore, by the union bound,

$$\mathbb{P}(\mathcal{E}_{\mathrm{drift}}^c) \le \mathbb{P}(\mathcal{E}_{\mathrm{pred}}^c) + \mathbb{P}(\mathcal{E}_{\mathrm{burn}}^c) \le 3T^{-4}.$$

For $s \in \{\tau_1+1, \dots, \tau_2\}$, set $Z_s = \mathbf{1}\{A(s-1) = 1,\ E(s-1) = 1\}$. Let $\mathcal{G}_s$ be the sigma-field generated before $A(s-1)$ and $E(s-1)$ are drawn. Since the Phase 2 random exploration probability is $\epsilon/4$, we have $\mathbb{E}[Z_s \mid \mathcal{G}_s] = \lambda\epsilon/4$. We apply Lemma 23 with $n = \tau_2 - \tau_1$ and $p = \lambda\epsilon/4$. This gives

$$\mathbb{P}\left(N_{\mathrm{ph2}} < \frac{\lambda\epsilon(\tau_2-\tau_1)}{8}\right) \le \exp\left(-\frac{\lambda\epsilon(\tau_2-\tau_1)}{32}\right). \tag{8}$$

Conditional on the random exploration times, the corresponding feature vectors are i.i.d. with the same distribution as $\boldsymbol{\phi}(X,J)$, where $X \sim \mathcal{D}$ and $J \sim \mathrm{Unif}([K])$. By Assumption 3, $\mathbb{E}[\boldsymbol{\phi}(X,J)\boldsymbol{\phi}(X,J)^{\mathsf{T}}] = \mathbb{E}_{X\sim\mathcal{D}}[K^{-1}\sum_{a\in[K]}\boldsymbol{\phi}(X,a)\boldsymbol{\phi}(X,a)^{\mathsf{T}}] \succeq \sigma_0^2\mathbf{I}$. Also, Assumption 1 gives $\|\boldsymbol{\phi}(X,J)\|_2 \le 1$. On the event $N_{\mathrm{ph2}} \ge \lambda\epsilon(\tau_2-\tau_1)/8$, there are at least $\lceil\lambda\epsilon(\tau_2-\tau_1)/8\rceil$ Phase 2 random exploration samples. We apply Lemma 24 to the first $\lceil\lambda\epsilon(\tau_2-\tau_1)/8\rceil$ such samples in chronological order. Adding positive semidefinite matrices cannot decrease the minimum eigenvalue, so $\lambda_{\min}(\mathbf{S}_{\mathrm{ph2}}) \ge \lambda\epsilon(\tau_2-\tau_1)\sigma_0^2/16$ with conditional probability at least $1 - d\exp(-\lambda\epsilon(\tau_2-\tau_1)\sigma_0^2/64)$. Combining this bound with (8) gives $\mathbb{P}(\mathcal{E}_{\mathrm{ph2}}^c) \le \nu_{\mathrm{ph2}}$. □

Lemma 17. *Fix $t > \tau_1$ and suppose that the queue is nonempty at the beginning of round t. Let $(\bar{x}_t, \bar{a}_t)$ be a job-server pair selected by the UCB rule, that is,*

$$(\bar{x}_t, \bar{a}_t) \in \operatorname*{arg\,max}_{x \in \mathcal{X}_t,\, a \in [K]} \left\{ \mu(\langle \boldsymbol{\phi}(x,a), \widehat{\boldsymbol{\theta}}_{t-1} \rangle) + \beta_{t-1} \|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{t-1}^{-1}} \right\}, \qquad \bar{\boldsymbol{\phi}}_t = \boldsymbol{\phi}(\bar{x}_t, \bar{a}_t).$$

On $\mathcal{E}_{\mathrm{drift}}(t)$,

$$\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) - \mu(\langle \bar{\boldsymbol{\phi}}_t, \boldsymbol{\theta}^* \rangle) \le \frac{\epsilon}{4}.$$

Proof. On $\mathcal{E}_{\mathrm{pred}}(t)$,

$$\begin{aligned}
\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) &\le \mu(\langle \boldsymbol{\phi}_t^*, \widehat{\boldsymbol{\theta}}_{t-1} \rangle) + \beta_{t-1} \|\boldsymbol{\phi}_t^*\|_{\mathbf{V}_{t-1}^{-1}} \\
&\le \mu(\langle \bar{\boldsymbol{\phi}}_t, \widehat{\boldsymbol{\theta}}_{t-1} \rangle) + \beta_{t-1} \|\bar{\boldsymbol{\phi}}_t\|_{\mathbf{V}_{t-1}^{-1}} \\
&\le \mu(\langle \bar{\boldsymbol{\phi}}_t, \boldsymbol{\theta}^* \rangle) + 2\beta_{t-1} \|\bar{\boldsymbol{\phi}}_t\|_{\mathbf{V}_{t-1}^{-1}}.
\end{aligned}$$

The second inequality follows from the definition of the UCB rule. On $\mathcal{E}_{\mathrm{burn}}(t)$, the last term is at most $\epsilon/4$. □

B.2 Proof of lemma 8

Proof of Lemma 8. Let

$$g_t = \mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) - \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle).$$

If the queue is empty in round t, we set $g_t = 0$. Hence it suffices to consider the case where the queue is nonempty.

On $\mathcal{E}_{\mathrm{ph2}}$, the matrix $\mathbf{V}_{\tau_2}$ dominates the Phase 2 random-exploration design matrix, and therefore

$$\lambda_{\min}(\mathbf{V}_{\tau_2}) \ge \frac{\lambda \epsilon (\tau_2 - \tau_1) \sigma_0^2}{16}.$$

For every $t \ge \tau_2 + 1$, we have $\mathbf{V}_{t-1} \succeq \mathbf{V}_{\tau_2}$, and hence, using $\|\boldsymbol{\phi}_t\|_2 \le 1$,

$$\|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}}^2 \le \frac{16}{\lambda \epsilon (\tau_2 - \tau_1) \sigma_0^2}.$$

Since $\tau_2 - \tau_1 \ge T/2$, we have

$$\|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}}^2 \le \frac{32}{\lambda \epsilon \sigma_0^2 T}. \tag{9}$$

Now consider a Phase 3 round $t \in \{\tau_2 + 1, \dots, T-1\}$. By construction, Phase 3 uses the pure UCB rule. On $\mathcal{E}_{\mathrm{pred}}$, for the optimal pair (x_t^*, a_t^*) in the current queue and the pair (x_t, a_t) selected by the algorithm,

$$\begin{aligned}
\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) &\le \mu(\langle \boldsymbol{\phi}_t^*, \widehat{\boldsymbol{\theta}}_{t-1} \rangle) + \beta_{t-1} \|\boldsymbol{\phi}_t^*\|_{\mathbf{V}_{t-1}^{-1}} \\
&\le \mu(\langle \boldsymbol{\phi}_t, \widehat{\boldsymbol{\theta}}_{t-1} \rangle) + \beta_{t-1} \|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}} \\
&\le \mu(\langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle) + 2\beta_{t-1} \|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}}.
\end{aligned}$$

The second inequality follows from the definition of the UCB rule. Therefore, on $\mathcal{E}_{\mathrm{pred}} \cap \mathcal{E}_{\mathrm{ph2}}$,

$$g_t \le 2\beta_{t-1} \|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}} \le 2\beta_T \|\boldsymbol{\phi}_t\|_{\mathbf{V}_{t-1}^{-1}}.$$

Using (9), on $\mathcal{E}_{\mathrm{pred}} \cap \mathcal{E}_{\mathrm{ph2}}$,

$$g_t^2 \le \frac{128 \beta_T^2}{\lambda \epsilon \sigma_0^2 T}.$$

Finally, since $0 \le g_t \le 1$, Lemma 16 gives

$$\begin{aligned}
\mathbb{E}[g_t^2] &\le \mathbb{E}[g_t^2 \mathbf{1}\{\mathcal{E}_{\mathrm{pred}} \cap \mathcal{E}_{\mathrm{ph2}}\}] + \mathbb{P}(\mathcal{E}_{\mathrm{pred}}^c) + \mathbb{P}(\mathcal{E}_{\mathrm{ph2}}^c) \\
&\le \frac{128 \beta_T^2}{\lambda \epsilon \sigma_0^2 T} + T^{-4} + \nu_{\mathrm{ph2}}.
\end{aligned}$$

Taking square roots gives the desired result. □

B.3 Proof of lemma 9

Proof of Lemma 9. We first record the negative drift property after Phase 1 for the coupled process governed by π_{T-1}. This process has the same marginal law as the original queue governed by the learning policy. Fix $t > \tau_1$ and suppose that $Q_{T-1}(t) > 0$. Let η_t be an upper bound on the conditional probability that the random exploration rule is used in round t; in Phase 2, $\eta_t = \epsilon/4$, and in Phase 3, $\eta_t = 0$. Let (x_t^*, a_t^*) be an optimal job-server pair in the queue of this process at the beginning of round t. By Assumption 4,

$$\mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) \geq \lambda + \epsilon.$$

If the algorithm uses random exploration, the departure-rate loss relative to the optimal pair is at most one. If it uses the UCB rule, the loss is at most $\epsilon/4$ on $\mathcal{E}_{\text{drift}}(t)$ by Lemma 17. Therefore, on $\mathcal{E}_{\text{drift}}(t) \cap \{Q_{T-1}(t) > 0\}$,

$$\begin{aligned}\mathbb{E}[D_{T-1}(t) \mid \mathcal{F}_t] &\geq \mu(\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle) - \eta_t - \frac{\epsilon}{4} \\ &\geq \lambda + \epsilon - \frac{\epsilon}{4} - \frac{\epsilon}{4} = \lambda + \frac{\epsilon}{2}.\end{aligned}$$

Hence, whenever $Q_{T-1}(t) > 0$ and $t > \tau_1$, on $\mathcal{E}_{\text{drift}}(t)$,

$$\mathbb{E}[A(t) - D_{T-1}(t) \mid \mathcal{F}_t] \leq -\frac{\epsilon}{2}. \tag{10}$$

We use this drift condition to obtain a tail bound for $Q_{T-1}(t)$. Let $\zeta = \epsilon/2$, $\rho = \exp(-\epsilon^2/8)$, and $b = \exp(\zeta)$. Define the deterministic bad-round count $B(t) = \min\{t, \tau_1\}$. Thus rounds $1, \ldots, \tau_1$ are treated as bad rounds, and all later rounds have negative drift on $\mathcal{E}_{\text{drift}}(t)$. Define

$$V(t) = \left(\frac{b}{\rho}\right)^{-B(t-1)} \exp(\zeta Q_{T-1}(t)).$$

We claim that, for every $t \in [T-1]$,

$$\mathbb{E}[\mathbf{1}\{\mathcal{E}_{\text{drift}}(t+1)\} V(t+1)] \leq \rho \mathbb{E}[\mathbf{1}\{\mathcal{E}_{\text{drift}}(t)\} V(t)] + 1. \tag{11}$$

To prove this, first consider a round $t > \tau_1$. If $Q_{T-1}(t) > 0$, then, on $\mathcal{E}_{\text{drift}}(t)$, (10) and Lemma 22 give

$$\mathbb{E}[\exp(\zeta(A(t) - D_{T-1}(t))) \mid \mathcal{F}_t] \leq \exp\big(-\zeta\epsilon/2 + \zeta^2/2\big) = \exp(-\epsilon^2/8) = \rho.$$

If $Q_{T-1}(t) = 0$, then the actual departure from the queue is zero and $Q_{T-1}(t+1) = A(t)$, so

$$\mathbb{E}[\exp(\zeta Q_{T-1}(t+1)) \mid \mathcal{F}_t] \leq \exp(\zeta) \leq 1 + \rho,$$

where the last inequality holds for $\epsilon \in (0, 1)$. Combining the two cases, for every $t > \tau_1$, on $\mathcal{E}_{\text{drift}}(t)$,

$$\mathbb{E}[\exp(\zeta Q_{T-1}(t+1)) \mid \mathcal{F}_t] \leq 1 + \rho \exp(\zeta Q_{T-1}(t)). \tag{12}$$

For $t \leq \tau_1$, we use the crude bound

$$\mathbb{E}[\exp(\zeta Q_{T-1}(t+1)) \mid \mathcal{F}_t] \leq 1 + b \exp(\zeta Q_{T-1}(t)). \tag{13}$$

Since $\mathcal{E}_{\text{drift}}(t+1) \subseteq \mathcal{E}_{\text{drift}}(t)$ and $\mathcal{E}_{\text{drift}}(t) \in \mathcal{F}_t$, (12) and (13) and the definition of $V(t)$ give (11). Solving the recursion and using $Q_{T-1}(1) = 0$,

$$\mathbb{E}[\mathbf{1}\{\mathcal{E}_{\text{drift}}(t)\} V(t)] \leq 1 + \frac{1}{1-\rho} \leq 17\epsilon^{-2}, \tag{14}$$

where we used $1 - e^{-x} \geq x/2$ for $x \in [0, 1]$.

Let $a = 1 + \epsilon/4$. Since $a = \zeta^{-1} \log(b/\rho)$, $\mathcal{E}_{\text{drift}} \subseteq \mathcal{E}_{\text{drift}}(t)$, and $B(t-1) \leq \tau_1$, Markov's inequality and (14) imply that, for every t and every $y \geq 0$,

$$\mathbb{P}(Q_{T-1}(t) \geq a\tau_1 + y,\ \mathcal{E}_{\text{drift}}) \leq 17\epsilon^{-2} \exp(-\zeta y). \tag{15}$$

We next bound the probability that a one-job discrepancy created in a fixed round i survives until round T. Let $n = T - i - 1$. If $n < 4\tau_1/\epsilon$, then the trivial bound $\widetilde{\psi}_i(T) \leq 1$ gives the first part of the lemma. Thus assume $n \geq 4\tau_1/\epsilon$.

Condition on $\mathcal{F}_i^+$ and on the disagreement event $\boldsymbol{D}_i(i) = 0$, $\boldsymbol{D}_{i-1}(i) = 1$. Let $\bar{Q}_i(i+1)$ be the queue length of the process governed by π_i at the beginning of round $i+1$ under this disagreement event. After round i, the process governed by π_i follows the oracle policy. Let H be the first time after round i at which this process hits queue length zero.

By Lemma 21, under the disagreement event, $\psi_i(T) \in \{0,1\}$, and $\psi_i(T) = 1$ implies $H > T$. Therefore,

$$\widetilde{\psi}_i(T) \le \mathbb{P}\big(H > T \mid \mathcal{F}_i^+, \boldsymbol{D}_i(i) = 0, \boldsymbol{D}_{i-1}(i) = 1\big).$$

We now bound this survival probability. Suppose that $\bar{Q}_i(i+1) = q$. For rounds $t = i+1, \dots, T-1$, as long as the queue is nonempty, the oracle policy has departure probability at least $\lambda + \epsilon$. Thus, for such rounds,

$$\mathbb{E}[A(t) - D_i(t) \mid \mathcal{G}_t] \le -\epsilon,$$

where $\mathcal{G}_t$ is the history of the process governed by π_i up to the beginning of round t. By Lemma 22 with $\zeta = \epsilon/2$,

$$\mathbb{E}[\exp(\zeta(A(t) - D_i(t))) \mid \mathcal{G}_t] \le \exp\big(-\zeta\epsilon + \zeta^2/2\big) \le \exp(-\epsilon^2/4).$$

We apply this bound only before the hitting time. In the next display, all probabilities and expectations are conditional on $\bar{Q}_i(i+1) = q$, $\mathcal{F}_i^+$, and the disagreement event $\boldsymbol{D}_i(i) = 0, \boldsymbol{D}_{i-1}(i) = 1$. For $m \in \{i+1, \dots, T\}$, let $\mathcal{H}_m = \{H > m\}$. For $m \in \{i+1, \dots, T-1\}$, on $\mathcal{H}_m$, the queue is nonempty at the beginning of round m, and hence the following one-step stopped inequality holds:

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\{\mathcal{H}_{m+1}\} \exp\left(\zeta \sum_{s=i+1}^{m} (A(s) - D_i(s))\right) \mid \mathcal{G}_m\right] \\
&\quad \le \mathbf{1}\{\mathcal{H}_m\} \exp\left(\zeta \sum_{s=i+1}^{m-1} (A(s) - D_i(s))\right) \mathbb{E}[\exp(\zeta(A(m) - D_i(m))) \mid \mathcal{G}_m] \\
&\quad \le \exp(-\epsilon^2/4) \mathbf{1}\{\mathcal{H}_m\} \exp\left(\zeta \sum_{s=i+1}^{m-1} (A(s) - D_i(s))\right).
\end{aligned}$$

Iterating the last display gives

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{H}_T\} \exp\left(\zeta \sum_{s=i+1}^{T-1} (A(s) - D_i(s))\right)\right] \le \exp\left(-\frac{\epsilon^2}{4} n\right).$$

On $\mathcal{H}_T$, we have

$$q + \sum_{s=i+1}^{T-1} (A(s) - D_i(s)) \ge 1.$$

Therefore,

$$\mathbb{P}\big(H > T \mid \bar{Q}_i(i+1) = q, \mathcal{F}_i^+, \boldsymbol{D}_i(i) = 0, \boldsymbol{D}_{i-1}(i) = 1\big) \le \min\left\{1, \exp\left(\zeta(q-1) - \frac{\epsilon^2}{4} n\right)\right\}. \tag{16}$$

This stopped estimate handles the empty-queue case: no negative drift is asserted after the queue becomes empty.

It remains to average (16) over $\mathcal{F}_i^+$. Set

$$\omega = \frac{4\tau_1}{\epsilon}, \qquad y_i = \frac{\epsilon}{4}(n - \omega).$$

Since $n \ge \omega$, we have $y_i \ge 0$. Also, under the disagreement event, $\bar{Q}_i(i+1) \le Q_{T-1}(i) + 1$. Hence (15) gives

$$\mathbb{P}\big(\bar{Q}_i(i+1) \ge a\tau_1 + 1 + y_i,\ \mathcal{E}_{\text{drift}}\big) \le 17\epsilon^{-2} \exp(-\zeta y_i). \tag{17}$$

On the complementary event $\bar{Q}_i(i+1) < a\tau_1 + 1 + y_i$, (16) yields

$$\mathbb{P}\big(H > T \mid \mathcal{F}_i^+, \boldsymbol{D}_i(i) = 0, \boldsymbol{D}_{i-1}(i) = 1\big) \le \exp\left(\zeta(a\tau_1 + y_i) - \frac{\epsilon^2}{4}n\right)$$
$$\le \exp\left(-\frac{\epsilon^2}{32}(n-\omega)\right).$$

The last inequality follows from $\zeta = \epsilon/2$, $a = 1 + \epsilon/4$, $y_i = \epsilon(n-\omega)/4$, $\omega = 4\tau_1/\epsilon$, and $\epsilon \in (0,1)$.

Combining the bound on the event $\bar{Q}_i(i+1) < a\tau_1 + 1 + y_i$ with (17) and Lemma 16,

$$\mathbb{E}\Big[\widetilde{\psi}_i(T)\Big] \le \mathbb{P}(\mathcal{E}_{\text{drift}}^c) + \exp\left(-\frac{\epsilon^2}{32}(n-\omega)\right) + 17\epsilon^{-2}\exp(-\zeta y_i)$$
$$\le 3T^{-4} + 19\epsilon^{-2}\exp\left(-\frac{\epsilon^2}{32}\left(T - i - 1 - \frac{4\tau_1}{\epsilon}\right)\right),$$

where we used $\zeta y_i = \epsilon^2(n-\omega)/8$ and $\epsilon^{-2} \ge 1$. Taking square roots, using the trivial bound $\widetilde{\psi}_i(T) \le 1$, and relabeling i as t complete the proof. □

C Deferred proofs for section 4

C.1 Proof of lemma 12

Proof of Lemma 12. Let $\gamma = 1-\lambda-\epsilon$, $\Delta_T = \gamma/(8\sqrt{T})$, $p_T = 1-\gamma/2+\Delta_T$, and $q_T = 1-\gamma/2-\Delta_T$. Since $T \ge 3$, we have $p_T \le 1 - 3\gamma/8 < 1$, $q_T \ge 1 - 5\gamma/8 \ge 3/8$, and $q_T - (\lambda+\epsilon) \ge 3\gamma/8 > 0$. Thus z_p and z_q are well-defined, and $p_T > \lambda + \epsilon$.

It remains to verify that $\nu_T^+, \nu_T^- \in \mathcal{C}(d, K, S, \sigma_0^2, \lambda, \epsilon)$. Since $\|x\|_2 = 1$ and $\mathbf{M}$ is orthogonal, $\|\boldsymbol{\phi}(x,a)\|_2 = 1$. Also, if X is uniform on $\mathcal{X}_{\text{lb}}$, then $\mathbb{E}[XX^\mathsf{T}] = d^{-1}\mathbf{I}$ and

$$\mathbb{E}\left[\frac{1}{K}\sum_{a\in[K]}\boldsymbol{\phi}(X,a)\boldsymbol{\phi}(X,a)^\mathsf{T}\right] = \frac{1}{d}\mathbf{I}.$$

Thus Assumption 3 holds for every $\sigma_0^2 \le 1/d$. For every $x \in \mathcal{X}_{\text{lb}}$ and each sign $s \in \{+,-\}$, the logits $\langle\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_s^*\rangle$ over $a \in [K]$ take values in $\{z_p, z_q\}$ and include z_p. Therefore $\max_{a\in[K]}\mu(\langle\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_s^*\rangle) = p_T > \lambda + \epsilon$, and the traffic slackness condition holds. For the parameter radius, let $L_\gamma = \log((8-3\gamma)/(3\gamma))$. The bounds $q_T \ge 3/8$ and $p_T \le 1 - 3\gamma/8$ imply $|z_p|, |z_q| \le L_\gamma$. Therefore

$$\|\boldsymbol{\theta}_\pm^*\|_2 = \sqrt{\frac{d}{2}(z_p^2 + z_q^2)} \le L_\gamma\sqrt{d} = B_{\lambda,\epsilon}\sqrt{d} \le S.$$

Finally, Assumption 2 holds for finite constants because all logits over Θ lie in $[-S, S]$ and $\dot{\mu}$ is positive and continuous on this interval. Hence $\nu_T^+, \nu_T^- \in \mathcal{C}(d, K, S, \sigma_0^2, \lambda, \epsilon)$. □

C.2 Proof of proposition 13

Proof of Proposition 13. Let $\gamma = 1 - \lambda - \epsilon$. For $x \in \mathcal{X}_{\text{lb}}$, let $\xi(x) = \sqrt{d}\,x_2 \in \{-1, +1\}$. If $\xi(x) = +1$, then

$$\langle\boldsymbol{\phi}(x,1), \boldsymbol{\theta}_+^*\rangle = z_p, \qquad \langle\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_+^*\rangle = z_q \quad \text{for all } a \in \{2,\dots,K\}.$$

If $\xi(x) = -1$, then these two logits are reversed. Thus, under ν_T^+, servers in $\mathcal{S}_+(x)$ have departure probability p_T, and servers outside $\mathcal{S}_+(x)$ have departure probability q_T. The same calculation with $\boldsymbol{\theta}_-^*$ gives the corresponding statement under ν_T^-.

It remains to verify the numerical bounds. Since $T \ge 3$, we have $p_T \le 1 - 3\gamma/8 < 1$, $q_T \ge 1 - 5\gamma/8 \ge 3/8$, and $q_T - (\lambda+\epsilon) \ge 3\gamma/8 > 0$. Thus $p_T > q_T > \lambda + \epsilon$ and $p_T - q_T = \gamma/(4\sqrt{T})$. Also, $p_T \ge 1/2$, $1 - p_T \ge 3\gamma/8$, $q_T \ge 3/8$, and $1 - q_T \ge \gamma/2$. Thus

$$p_T(1-p_T) \ge \frac{3\gamma}{16}, \qquad q_T(1-q_T) \ge \frac{3\gamma}{16}.$$

For every $u \in [q_T, p_T]$, the concavity of $u(1-u)$ gives

$$u(1-u) \geq \min\{p_T(1-p_T), q_T(1-q_T)\} \geq \frac{3\gamma}{16}.$$

Substituting $\gamma = 1 - \lambda - \epsilon$ completes the proof. □

C.3 Proof of lemma 14

Proof of Lemma 14. Fix $s \in \{+,-\}$ and a deterministic policy ρ. Let $Q^*(t)$ be the oracle queue under ν_T^s, and let $Q^\rho(t)$ be the queue generated by ρ under the same instance. Couple the two processes by using the same arrival random variables and the same uniform random variables to generate departures. We prove the claim by induction. At $t = 1$, both queues are empty. Assume $Q^*(t) \leq Q^\rho(t)$. If $Q^*(t) = 0$, then $Q^*(t+1) = A(t)$, while $D^\rho(t) \leq Q^\rho(t)$ implies

$$Q^\rho(t+1) = Q^\rho(t) + A(t) - D^\rho(t) \geq A(t).$$

If $Q^*(t) > 0$, then $Q^\rho(t) > 0$ as well. By Proposition 13, the oracle can choose a job-server pair with departure probability p_T, and every feasible job-server pair has departure probability at most p_T. Let $r_t^\rho \leq p_T$ be the departure probability of the job-server pair selected by ρ. Using the same uniform random variable U_t for the departures, $D^*(t) = \mathbf{1}\{U_t \leq p_T\}$ and $D^\rho(t) = \mathbf{1}\{U_t \leq r_t^\rho\}$, so $D^*(t) \geq D^\rho(t)$. Thus

$$Q^*(t+1) = Q^*(t) + A(t) - D^*(t) \leq Q^\rho(t) + A(t) - D^\rho(t) = Q^\rho(t+1).$$

□

C.4 Proof of proposition 15

Proof of Proposition 15. Let P_+^F and P_-^F be the conditional laws on $\mathcal{H}_{T-1}$ under ν^+ and ν^- given F. Since the initial queue is empty, the observations generating $\mathcal{H}_{T-1}$ determine the queue state before each round up to $T-1$. For a deterministic policy, the selected job-server pair is therefore determined by the past history. The two instances have the same arrival process and context distribution, and conditioning on $F = \{A(T-2) = 1\}$ fixes only the shared arrival random variable $A(T-2)$. Thus, in each round, only the conditional distribution of the departure random variable can differ between the two instances. If the queue is empty, the departure random variable is deterministically zero under both instances. Otherwise, by assumption, the two conditional distributions of the departure random variable are either identical or are $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$ in some order. Therefore, the KL chain rule bounds $D_{\mathrm{KL}}(P_+^F \| P_-^F)$ by the sum of the one-round conditional KL divergences. The arrival and context terms contribute zero, and each departure term is at most the larger KL divergence between $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$. Thus,

$$D_{\mathrm{KL}}\left(P_+^F \| P_-^F\right) \leq (T-2)\max\{D_{\mathrm{KL}}(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)), D_{\mathrm{KL}}(\mathrm{Bern}(q)\|\,\mathrm{Bern}(p))\}.$$

By assumption, the right-hand side is at most $1/6$. By the Bretagnolle–Huber inequality (Lemma 26) applied to P_+^F, P_-^F, and G,

$$\mathbb{P}_+(G^c \mid F) + \mathbb{P}_-(G \mid F) = P_+^F(G^c) + P_-^F(G) \geq \frac{1}{2}e^{-1/6}.$$

□

D Auxiliary lemmas

This section collects auxiliary results used in the appendix. The following prediction-error bound is Lemma 5.1 of Bae et al. (2026a), stated in the notation of this paper.

Lemma 18. *With the confidence radius defined by (3), it holds with probability at least* $1-\delta$ *that*

$$\left|\mu(\langle \boldsymbol{\phi}(x,a), \widehat{\boldsymbol{\theta}}_{s-1}\rangle) - \mu(\langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^*\rangle)\right| \leq \beta_{s-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{s-1}^{-1}}$$

for all $s \in [T]$ *and all* $(x,a) \in \mathcal{X} \times [K]$.

The following design-matrix lower bound is Proposition 1 of Li et al. (2017).

Lemma 19. *Let $Y_1, \dots, Y_n$ be i.i.d. random vectors drawn from a distribution supported on the unit ball in $\mathbb{R}^d$. Let $\Sigma = \mathbb{E}[Y_1 Y_1^\mathsf{T}]$. For any $B > 0$ and $\delta > 0$, there exist absolute constants $C_1, C_2 > 0$ such that*

$$\lambda_{\min}\left(\sum_{s=1}^{n} Y_s Y_s^\mathsf{T}\right) \geq B$$

with probability at least $1 - \delta$, provided that

$$n \geq \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)}\right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}.$$

The following burn-in uncertainty bound is adapted from Lemma 5.2 of Bae et al. (2026a).

Lemma 20. *Suppose that τ_1 is chosen according to (4) with sufficiently large absolute constants $c_1, c_2 > 0$. Then, with probability at least $1 - 2\delta$,*

$$\beta_{t-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{t-1}^{-1}} \leq \epsilon/8$$

for all $t \in \{\tau_1 + 1, \dots, T\}$ and all $(x,a) \in \mathcal{X}_t \times [K]$.

Proof of Lemma 20. Let $B_1 = 64\beta_T^2/\epsilon^2$, and let $N_{\mathrm{ph1}} = \sum_{s=1}^{\tau_1} \mathbf{1}\{A(s-1) = 1\}$ be the number of Phase 1 random exploration samples. During Phase 1, every arrival is used for random exploration, and the server is selected uniformly at random. Conditional on the random exploration times, the corresponding feature vectors are i.i.d. with the same distribution as $\boldsymbol{\phi}(X, J)$, where $X \sim \mathcal{D}$ and $J \sim \mathrm{Unif}([K])$. By Assumption 3, $\mathbb{E}[\boldsymbol{\phi}(X,J)\boldsymbol{\phi}(X,J)^\mathsf{T}] \succeq \sigma_0^2 \mathbf{I}$. Set

$$n_{\mathrm{ph1}} = \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta)}}{\sigma_0^2}\right)^2 + \frac{2B_1}{\sigma_0^2}.$$

By taking the absolute constants c_1, c_2 in the definition of τ_1 sufficiently large,

$$\frac{\lambda\tau_1}{2} \geq n_{\mathrm{ph1}}, \qquad \exp\left(-\frac{\lambda\tau_1}{8}\right) \leq \delta.$$

For each $s \in [\tau_1]$, let $\mathcal{G}_s$ be the sigma-field generated by the history before $A(s-1)$ is drawn. Then $\mathbf{1}\{A(s-1) = 1\}$ is $\{0,1\}$-valued and has conditional mean λ given $\mathcal{G}_s$. We apply Lemma 23 with $n = \tau_1$ and $p = \lambda$ to get

$$\mathbb{P}(N_{\mathrm{ph1}} < n_{\mathrm{ph1}}) \leq \mathbb{P}\left(N_{\mathrm{ph1}} < \frac{\lambda\tau_1}{2}\right) \leq \delta.$$

On the event $N_{\mathrm{ph1}} \geq n_{\mathrm{ph1}}$, we apply Lemma 19 to the first $\lceil n_{\mathrm{ph1}} \rceil$ Phase 1 random exploration samples in chronological order, with $n = \lceil n_{\mathrm{ph1}} \rceil$, $B = B_1$, and $\lambda_{\min}(\Sigma) = \sigma_0^2$. Adding positive semidefinite matrices cannot decrease the minimum eigenvalue, so $\lambda_{\min}(\mathbf{S}_{\mathrm{ph1}}) \geq B_1$ with conditional probability at least $1 - \delta$, where

$$\mathbf{S}_{\mathrm{ph1}} = \sum_{s=1}^{\tau_1} \mathbf{1}\{A(s-1) = 1\}\boldsymbol{\phi}_s\boldsymbol{\phi}_s^\mathsf{T}$$

is the design matrix formed by the Phase 1 random exploration samples. By the union bound, this design lower bound holds with probability at least $1 - 2\delta$.

On this event, for every $t \in \{\tau_1 + 1, \dots, T\}$,

$$\mathbf{V}_{t-1} \succeq \mathbf{V}_{\tau_1} \succeq \mathbf{S}_{\mathrm{ph1}}.$$

Using $\|\boldsymbol{\phi}(x,a)\|_2 \leq 1$ and $\beta_{t-1} \leq \beta_T$,

$$\beta_{t-1}\|\boldsymbol{\phi}(x,a)\|_{\mathbf{V}_{t-1}^{-1}} \leq \frac{\beta_T}{\sqrt{B_1}} = \frac{\epsilon}{8}$$

for all $(x,a) \in \mathcal{X}_t \times [K]$. □

The following coupling property follows from Lemmas B.1 and C.5 of Bae et al. (2026a) and Lemma D.10 of Bae et al. (2026b), stated in the notation of this paper.

Lemma 21. *Consider the coupled consecutive policy-switching queues governed by π_t and π_{t-1}. If $\boldsymbol{D}_t(t) = \boldsymbol{D}_{t-1}(t)$, then $\psi_t(T) \in \{-1, 0\}$. If $\boldsymbol{D}_t(t) = 0$ and $\boldsymbol{D}_{t-1}(t) = 1$, then $\psi_t(T) \in \{0, 1\}$. Moreover, under the event $\boldsymbol{D}_t(t) = 0$ and $\boldsymbol{D}_{t-1}(t) = 1$, if the process governed by π_t hits queue length zero in a round after t and no later than T, then $\psi_t(T) = 0$. Equivalently, $\psi_t(T) = 1$ implies that this process does not hit queue length zero before round T.*

We also use the following standard inequalities.

Lemma 22 (Conditional Hoeffding lemma)**.** *Let X be a random variable satisfying $X \in [-1, 1]$ almost surely. If $\mathbb{E}[X \mid \mathcal{G}] \le -a$ for some $a \ge 0$, then for every $\zeta \ge 0$,*

$$\mathbb{E}[\exp(\zeta X) \mid \mathcal{G}] \le \exp\left(-\zeta a + \frac{\zeta^2}{2}\right).$$

Lemma 23 (Multiplicative Chernoff bound)**.** *Let $Z_1, \dots, Z_n$ be $\{0, 1\}$-valued random variables adapted to a filtration $\{\mathcal{G}_i\}_{i=0}^n$. If $\mathbb{E}[Z_i \mid \mathcal{G}_{i-1}] \ge p$ for every i, then*

$$\mathbb{P}\left(\sum_{i=1}^n Z_i < \frac{pn}{2}\right) \le \exp\left(-\frac{pn}{8}\right).$$

Lemma 24 (Matrix Chernoff bound)**.** *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. random vectors satisfying $\|Y_i\|_2 \le 1$ almost surely and $\mathbb{E}[Y_i Y_i^\mathsf{T}] \succeq \sigma_0^2 \mathbf{I}$. Then*

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{i=1}^n Y_i Y_i^\mathsf{T}\right) < \frac{n\sigma_0^2}{2}\right) \le d \exp\left(-\frac{n\sigma_0^2}{8}\right).$$

Lemma 25. *Let $u, v \in (0, 1)$. If $w(1-w) \ge m$ for every $w \in [\min\{u, v\}, \max\{u, v\}]$, then*

$$D_{\mathrm{KL}}(\mathrm{Bern}(u) \| \mathrm{Bern}(v)) \le \frac{(u-v)^2}{2m}.$$

Proof of Lemma 25. Fix $v \in (0, 1)$ and define $f(r) = D_{\mathrm{KL}}(\mathrm{Bern}(r) \| \mathrm{Bern}(v))$ for $r \in (0, 1)$. Then $f(v) = 0$, $f'(v) = 0$, and $f''(r) = 1/(r(1-r))$. By Taylor's theorem, for some point $\bar{r}$ between u and v,

$$D_{\mathrm{KL}}(\mathrm{Bern}(u) \| \mathrm{Bern}(v)) = f(u) = \frac{(u-v)^2}{2\bar{r}(1-\bar{r})}.$$

The point $\bar{r}$ lies between u and v. Thus $\bar{r}(1-\bar{r}) \ge m$ by assumption, and the desired bound follows. □

Lemma 26 (Bretagnolle–Huber inequality)**.** *For any probability measures P, Q and event G,*

$$P(G^c) + Q(G) \ge \frac{1}{2} \exp\{-D_{\mathrm{KL}}(P \| Q)\}.$$

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Our paper provides the full set of assumptions and complete proofs in the main text and appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results in the experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We provide open access to the code.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The paper specifies all the training and test details in the experiments section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Section 5 reports averages over 10 independent runs, with error bands indicating ± 1 standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper provides information on the computer resources needed to reproduce the experiments in the experiments section.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

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

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

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

    Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.