Learning-based Optimal Admission Control for Erlang–B Queuing Systems

Shubhhi Singh[1] **Shubhanshu Shekhar**[1] **Vijay G. Subramanian**[1]

[1]Electrical and Computer Engineering Division, Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

This paper studies a learning-based optimal admission control policy for an Erlang–B queuing system under partial observation. Only at each arrival, the dispatcher observes the current occupancy and decides whether to accept or reject the job. A completed job yields a fixed reward but incurs a cost proportional to the service duration. The objective is to design an admission control policy that maximizes the long-term average reward. The arrival and service rates are unknown, and unobserved departures complicate parameter estimation. The reward structure induces an asymmetry in the optimal decision across parameter regimes. We establish an instance-dependent asymptotic lower bound, explicitly dependent on the system parameters, demonstrating that the worst-case regret is logarithmic in the number of arrivals in one regime. We complement the bound with a matching upper bound via an explicit sequential-test based policy.

1 INTRODUCTION

Queuing models provide a rigorous mathematical framework for analyzing congestion, delay, and resource allocation arising from stochastic variability in communication networks, computing systems, supply chains, and inventory systems. In this paper, we study learning-based admission control for a specific queuing model—the $M/M/k/k$ queue, popularly known as an Erlang–B queuing system [Kelly, 2011, Harchol-Balter, 2013, Srikant and Ying, 2013]. This model has applications in healthcare, ride-sharing, cloud computing, networks, and inventory management, where there is a finite number of homogeneous processing elements [Gans et al., 2003, Marbach et al., 2011].

An Erlang–B system is characterized by $k \in \mathbb{N}$ homogeneous servers and Poisson arrivals with rate $\lambda > 0$, and independent exponential service time with rate $\mu > 0$. The system has zero buffer capacity, so arrivals finding all k servers busy are dropped. We study the learning-based admission control problem for this system under partial observation and unknown parameters, as introduced by Adler et al. [2022]. At each arrival, the dispatcher observes the system state and decides whether to accept the job into an available server or reject it. Completed jobs yield reward R and incur holding cost c per unit time while in service; rejected jobs incur no penalty. The objective is to design an admission-control policy that maximizes the long-run average reward per unit time. Define the threshold $\mu^* := c/R$. In the genie-aided setting, where μ is known to the dispatcher, the optimal policy π^* has a threshold structure: reject all arrivals when $\mu < \mu^*$, and accept whenever a server is available when $\mu > \mu^*$; at $\mu = \mu^*$, π^* is indifferent between the two actions. With $t \in \{0, 1, \ldots\}$ indexing the arrivals, let $\{N_t\}_{t \in \mathbb{Z}_+}$ denote the system state immediately before arrival t in the controlled $M/M/k/k$ Erlang–B system.

While the bulk of the queuing theory literature assumes that system parameters such as the service rate μ are known, in most real-world scenarios, such parameters are unknown. Further, in many systems, it is also difficult to form an estimate of parameters like the service rate—for example, by taking an empirical average of job service times—because although job arrivals may be observed, job completions are usually hidden. For example, public Wi-Fi systems observe connection requests but often infer departures only through inactivity timeouts; transit systems may record passenger boarding but not exits. Similar issues arise in cloud computing, parking systems, and IoT networks. Moreover, continuous monitoring of server availability may be infeasible due to communication, memory, or computational constraints, so systems often rely on event-driven observations at arrival epochs. Both these aspects are modeled by the following assumptions that we discussed earlier: 1) neither the service rate nor the arrival rate is *known* to the dispatcher; and 2) the dispatcher can only observe the inter-arrival times and system state just before each arrival.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6410–6444.

Our problem can be viewed as one of adaptive control, where the system dynamics depend on an unknown parameter that must be learned while making control decisions. In the classical *certainty-equivalent (CE) paradigm*, the controller estimates the unknown parameter from past observations and applies the optimal policy as if that estimate were correct. Mandl [1974] showed that, under identifiability and irreducibility conditions of the transition kernel for every stationary policy, the maximum likelihood estimate (MLE) converges to the true parameter, and the CE policy is asymptotically optimal. Agrawal et al. [1989] and Graves and Lai [1997] extended this to finite and general state spaces. Our setting differs from this classical framework in three key ways. First, the standard theory assumes transition densities under every fixed control policy are uniformly bounded away from 0 and 1 over the parameter space; which fails here. Second, hidden departures complicate estimation; although departure counts can be recovered from consecutive system state observations, the resulting likelihood has no closed-form maximizer [Adler et al., 2022]. Third, and more critically, the optimal control policy is discontinuous in μ. In the "always reject" regime, a CE policy stops learning because the chain is absorbed at the empty state, making the transition law identical across parameters. Thus, a CE policy can prematurely converge to an incorrect policy and remain there indefinitely.

To address this, Adler et al. [2022] proposed an MLE-based certainty-equivalent policy with forced exploration. This policy converges almost surely to the optimal policy asymptotically: it eventually accepts all arrivals when $\mu > \mu^*$, and rejects them with a probability approaching one when $\mu < \mu^*$. They also prove finite-time regret bounds: logarithmic regret in the number of arrivals (n) when $\mu < \mu^*$, and constant regret when $\mu > \mu^*$. However, they only conjectured the matching $\Omega(\log n)$ lower bound when $\mu < \mu^*$.

Contributions: The remainder of the paper is organized as follows. In Section 2, we provide our first main result (Theorem 2.2): an instance-dependent asymptotic lower bound on the regret (Definition 1.1) of any *uniformly efficient* policy with an explicit dependence on the system parameters. Uniformly efficient policies are informally defined as policies that are asymptotically optimal uniformly across the parameter space; see Definition 2.1 for a formal definition. This result states that when $\mu < \mu^*$, any uniformly efficient policy must incur a regret that is at least $C^*(\lambda, \mu; \mu^*)\log(n)(1 + o(1))$, for a system-dependent constant $C^*(\lambda, \mu; \mu^*)$. This explicit characterization of the constant allows for comparative statics in three critical regimes not captured by the results of Adler et al. [2022]: 1) as μ increases to μ^*, the problem becomes increasingly difficult at the boundary, even though there is indifference at μ^*; 2) as λ decreases to 0, the difficulty in learning is due to the low arrival rate; and 3) as λ increases without bound, the problem still suffers from the lack of parameter knowledge. In Section 3, we complement the lower bound result by providing a hypothesis testing-based scheme for a given fixed horizon; the regret of this policy matches the lower bound up to constants when $\mu < \mu^*$, and is constant when $\mu \geq \mu^*$. We extend this scheme to the anytime setting in Section 4 using the doubling trick; the resulting policy's regret matches the lower bound when $\mu < \mu^*$ but incurs an extra $\log\log n$ factor when $\mu \geq \mu^*$.

Related Work: Following seminal result by Mandl [1974] on the optimality of CE policy, Borkar and Varaiya [1979] identified a fundamental limitation: without identifiability, MLE need not converge to the true parameter, but to an observationally equivalent one that induces the same closed-loop transition law. Consequently, a CE policy may prematurely settle on a suboptimal control, showing that purely exploitative strategies are generally insufficient. This motivated exploration mechanisms in adaptive control, including implicit exploration through optimism or posterior sampling [Kumar and Becker, 1982, Kumar and Lin, 1982, Agrawal, 1995, Auer et al., 2002, Gopalan and Mannor, 2015], and explicit (forced) exploration [Agrawal et al., 1989].

Our lower-bound analysis builds on the information-theoretic framework of Lai and Robbins [1985], who showed that uniformly efficient bandit policies must sample suboptimal arms logarithmically often, with the leading constant determined by a Kullback–Leibler (KL) divergence quantity. This framework was extended to multi-parameter and nonparametric bandits [Burnetas and Katehakis, 1996], Markovian bandits [Anantharam et al., 1987], controlled Markov chains [Agrawal et al., 1989, Graves and Lai, 1997], and parametrized MDPs [Gopalan and Mannor, 2015]. Notably, in structured bandits, optimism- and posterior-sampling-based algorithms can be arbitrarily suboptimal relative to the instance-dependent asymptotic lower bound [Lattimore and Szepesvari, 2017, Combes et al., 2017] motivating our departure from index-based policies in favor of sequential hypothesis testing.

Adaptive control in queues differs from classical bandit and MDP formulations because rewards are delayed and trajectory-dependent, and exploration creates carryover costs [Gaitonde and Tardos, 2023]. Despite these differences, our lower bound follows the information-theoretic principle from Lai and Robbins [1985]: it quantifies the minimum rate at which any uniformly efficient policy must acquire information about the service-time distribution.

Related work on learning in queues includes scheduling with unknown service rates [Krishnasamy et al., 2018, 2020] and admission control with unknown arrival and service rates [Cohen et al., 2024]. All of these works assume service completions are observed or inferable. Our setting differs fundamentally in its information structure: departures are never directly observed and the only available information is the occupancy transitions recorded at arrival epochs. Thus, direct empirical estimation is not possible. While Adler et al. [2022] address this using an MLE-based algorithm; we derive an information-theoretic lower bound and use it to propose a sequential hypothesis testing based scheme.

We refer the reader to Appendix A for a more detailed exposition of related work.

1.1 PROBLEM FORMULATION

We use the following notation throughout. Consider a k-server Erlang–B queue with service rate $\mu > 0$ and arrival rate $\lambda > 0$; observed only at arrival epochs, indexed by $t \in \{0, 1, \dots\}$; see Figure 1 for state evolution and observations.

1. $\mathrm{Exp}(\lambda)$ denotes the exponential distribution with rate λ, $\mathrm{Bin}(n, p)$ represents the binomial distribution with n trials and success probability p, and $\mathbf{1}\{\cdot\}$ is the indicator function.
2. $A_t \in \{0, 1\}$ denotes the action at arrival t; $A_t = 1$ denotes acceptance – feasible only if an idle server is available – and $A_t = 0$ denotes rejection.
3. $T_t \sim \mathrm{Exp}(\lambda)$ denotes the inter-arrival time between arrival $t-1$ and t, for $t \geq 1$.
4. $N_t \in \{0, 1, \dots, k\}$ denotes the system-state (that is, the number of occupied servers) immediately prior to arrival t. Without loss of generality, we initialize $N_0 = 0$.
5. $\mathcal{H}_t$ denotes the history of past actions and observations up to arrival t, that is, $\mathcal{H}_t = \{N_0, A_0, T_1, \dots, N_{t-1}, A_{t-1}, T_t, N_t\}$.
6. $\Delta(\mu) := |R - \frac{c}{\mu}|$ denotes the absolute expected net reward per admitted arrival.
7. For $t < n$, $\mathcal{C}_t(n) \subset \{t+1, t+2, \dots, n\}$ denotes the set of arrivals that observe job admitted at arrival t still in service upon their own arrival. Its cardinality is denoted by $C_t(n) := |\mathcal{C}_t(n)|$.
8. Define the inter-arrival KL divergence

$$D^{\lambda}(\mu_0 \parallel \mu_1) := \mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}[d(e^{-\mu_0 \tau} \parallel e^{-\mu_1 \tau})]$$

where $d(p \parallel q)$ is the KL divergence [Cover and Thomas, 2006, Section 2.3] between two Bernoulli distributions with parameters $p, q \in [0, 1]$, respectively. This will be the key information-theoretic quantity governing our regret bounds.

A policy $\pi = \{u_t\}_{t \geq 0}$ is a sequence of decision rules $u_t : \mathcal{H}_t \to \mathcal{P}(\mathcal{A}_t)$, where $\mathcal{P}(\mathcal{A}_t)$ denotes the probability simplex over the action space $\mathcal{A}_t := \{0, 1\}$. We consider behavioral policies, denoted by Π, under which action A_t is randomized only as a function of the history $\mathcal{H}_t$. This

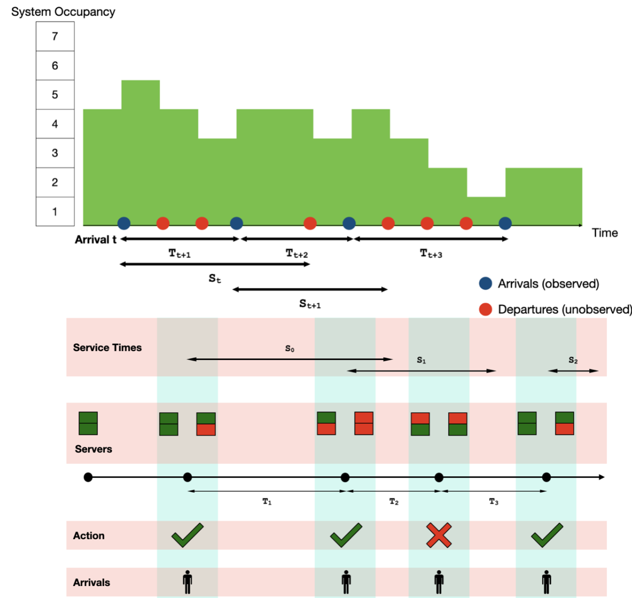


Figure 1: **Top:** System-state evolution observed only at the arrivals. **Bottom:** State of a two-server system before and after an admission decision. Variables S_t and T_t denote the service-time of arrival t and the inter-arrival time between arrivals $t-1$ and t, respectively.

entails no loss of generality under perfect-recall, since any randomized policy induces the same outcome distribution as some behavioral policy [Kuhn, 2016]. Throughout, $\mathbb{P}^{\pi}_{\mu}$ and $\mathbb{E}^{\pi}_{\mu}$ denote the probability measure and expectation, respectively, induced by the Erlang–B system with service rate μ under policy π. The objective is to maximize the expected net reward, namely,

$$\sup_{\pi \in \Pi} \limsup_{n \to \infty} \frac{\left(R - \frac{c}{\mu}\right) \mathbb{E}^{\pi}_{\mu}\left[\sum_{i=0}^{n-1} A_i\right]}{n}.$$

Adler et al. [2022] showed that the optimal policy π^* has a threshold form: accept arrivals whenever possible if $\mu \geq \mu^*$, and reject all arrivals if $\mu \leq \mu^*$.

Definition 1.1 (Expected regret). *Let $A^{\pi}_t \in \{0, 1\}$ denote the action at arrival t under policy π. Consider a common probability space on which the two systems operated under policies π and π^* are coupled to the same arrival sequence and same service times. The sample-path regret of policy π, up to first n arrivals under service rate μ is defined as:*

$$\mathrm{Reg}(n, \pi; \mu) := \Delta(\mu) \left| \sum_{t=0}^{n-1} (A^{\pi^*}_t - A^{\pi}_t) \right|,$$

where π^* denotes the optimal policy. The expected regret of policy π under μ is then defined as:

$$\mathbb{E}^{\pi}_{\mu}[\mathrm{Reg}(n, \pi; \mu)] := \Delta(\mu) \mathbb{E}^{\pi}_{\mu}\left[\left| \sum_{t=0}^{n-1} (A^{\pi^*}_t - A^{\pi}_t) \right|\right].$$

Assumption 1.2 (Erlang–B queue)**.** *We will work under the following assumptions throughout the paper:*

(A1) The inter-arrival times $\{T_t\}_{t\geq 1}$ are i.i.d. $\mathrm{Exp}(\lambda)$ *random variables.*

(A2) The sequence of inter-arrival times $\{T_t\}_{t\geq 1}$ is independent of the sequence of service times.

2 REGRET LOWER BOUND FOR REJECT-OPTIMAL REGIME

In this section, we establish an instance-dependent asymptotic lower bound on the expected regret in the reject-optimal regime (Theorem 2.2). We restrict our attention to the class of *uniformly efficient policies*, namely policies whose expected regret grows sub-polynomially with horizon uniformly over all problem instances. This is the standard class of policies considered in the adaptive control literature.

Definition 2.1 (Uniformly efficient (UE) policy)**.** *A policy π is said to be uniformly efficient if, for every service rate $\mu > 0$ and for every $a > 0$,*

$$\lim_{n\to\infty} \frac{\mathbb{E}_\mu^\pi[\mathrm{Reg}(n,\pi;\mu)]}{n^a} = 0.$$

The class of UE policies is non-empty. Indeed, Adler et al. [2022] proposed a policy that converges almost surely to the optimal behavior asymptotically. We now present the main result of this section which demonstrates that in the *reject-optimal* regime, any UE policy must incur at least logarithmic expected regret.

Theorem 2.2 (Asymptotic instance-dependent lower bound on expected regret)**.** *Suppose Assumption 1.2 holds. For $\mu, \lambda > 0$, define the complexity parameter*

$$C^*(\mu,\lambda;\mu^*) := \frac{\Delta(\mu)}{D^\lambda(\mu \parallel \mu^*)\left(1+\frac{\lambda}{\mu}\right)}. \tag{1}$$

Then, any uniformly efficient behavioral policy π must satisfy the following for $\mu < \mu^*$:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\mu^\pi[\mathrm{Reg}(n,\pi;\mu)]}{\log n} \;\geq\; C^*(\mu,\lambda;\mu^*).$$

Theorem 2.2 provides several insights into how the asymptotic lower bound scales across three challenging regimes:

1. **High-Arrival Rate Regime ($\lambda \to \infty$):** As arrival rate grows arbitrarily large, we observe that

$$\lim_{\lambda\uparrow\infty} D^\lambda(\mu \parallel \mu^*)\left(1+\frac{\lambda}{\mu}\right) = \frac{\mu^*}{\mu} - 1 - \log\left(\frac{\mu^*}{\mu}\right) = D_{\mathrm{KL}}(\mathrm{Exp}(\mu)\|\mathrm{Exp}(\mu^*)).$$

   Consequently, the complexity parameter in (1) denoted by $C^*_\infty(\mu) := \frac{\Delta(\mu)}{D_{\mathrm{KL}}(\mathrm{Exp}(\mu)\|\mathrm{Exp}(\mu^*))}$. This may appear counterintuitive—one might expect a high arrival rate to facilitate learning by providing more observations. Instead, it precisely captures the core difficulty: regret due to uncertainty about service rate persists even at arbitrarily high arrival rates. This limit corresponds to the system being observed at all times, as in Cohen et al. [2024], so the dispatcher can observe individual service completion times too. Therefore, we now have a one-armed bandit problem—distinguishing an $\mathrm{Exp}(\mu)$ random variable against the null, which is the set of exponential distributions with rate at least μ^*—, and the Lai and Robbins [1985] lower bound applies.

2. **Low-Arrival Rate Regime ($\lambda \to 0$):** Conversely, when the arrival rate vanishes, the complexity parameter denoted by $C^*_0(\mu)$ satisfies

$$\lim_{\lambda\downarrow 0} \lambda C^*_0(\mu) = \frac{\Delta(\mu)}{\frac{-\pi^2}{6}\left(\frac{1}{\mu}-\frac{1}{\mu^*}\right) - \frac{\psi(\mu/\mu^*)+\gamma}{\mu}},$$

   where $\psi(\cdot)$ is the Digamma function and $\gamma \approx 0.5777$ is the Euler-Mascheroni constant. This reveals that sparse arrivals impede learning, making it is a challenging regime for optimal admission control.

3. **Near-threshold Regime ($\mu \to \mu^*$):** When μ approaches the threshold, expected regret will scale as $C^*_{\mathrm{th}}(\lambda)\log n$ where,

$$\lim_{\mu\uparrow\mu^*} (\mu^*-\mu)C^*_{\mathrm{th}}(\lambda) = \frac{R}{\lambda(\lambda+\mu^*)\sum_{m=1}^\infty(\lambda+m\mu^*)^{-3}}.$$

   Here, R denotes the fixed reward for service completion. Although the system is indifferent at μ^*, distinguishing between the two regimes becomes increasingly difficult as μ approaches μ^*. Consequently, the complexity term diverges; making this a particularly challenging regime for learning-based control.

Proof Strategy: The proof of Theorem 2.2 follows the standard change-of-measure argument for information-theoretic lower bounds; see Lattimore and Szepesvári [2020, Section 16.1]. In the reject-optimal regime ($\mu < \mu^*$), we construct an alternative instance $\nu > \mu^*$, under which accepting is optimal. Applying the Bretagnolle-Huber (BH) inequality relates the regret to the KL divergence between the probability measures induced by the two instances. Lemma 2.3 expresses this divergence is terms of expected number of accepted arrivals. Combining these results and taking the limit inferior as $n \to \infty$, yields the instance-dependent asymptotic lower bound on regret when $\mu < \mu^*$.

Proof of Theorem 2.2. Fix $\mu \in (0,\mu^*)$ as in the statement of Theorem 2.2,and choose any $\nu > \mu^*$. Let $\mathbb{P}_\mu^\pi$ and $\mathbb{P}_\nu^\pi$ denote the probability measures induced by n arrivals under the service rates μ and ν respectively. Since μ and ν are on different sides of the threshold value μ^*, their optimal policies are different. However, any UE policy π must achieve sub-polynomial regret on both problem instances simultaneously. This fact allows us to set up a contradiction that leads to the required lower bound statement as we elaborate below. Fix $\gamma \in (0,1)$ and define the event $\mathcal{E} \coloneqq \{\sum_{j=0}^{n-1} A_j \leq n^\gamma\}$.

Since accepting is suboptimal under μ,

$$\mathrm{Reg}(n,\pi;\mu) \geq \Delta(\mu)\, n^\gamma\, \mathbb{P}_\mu^\pi(\mathcal{E}^c). \tag{2}$$

While under ν, rejecting is suboptimal, so

$$\begin{aligned}\mathrm{Reg}(n,\pi;\nu) &\geq \Delta(\nu)(n-n^\gamma)\mathbb{P}_\nu^\pi(\mathcal{E}) \\ &\geq \Delta(\nu)n^\gamma\mathbb{P}_\nu^\pi(\mathcal{E}),\end{aligned} \tag{3}$$

where the second inequality holds for $n \geq 2e^{\frac{1}{1-\gamma}}$. Let $R_{\mu,\nu} := \mathrm{Reg}(n,\pi;\mu) + \mathrm{Reg}(n,\pi;\nu)$ and $\Delta_{\mu,\nu} := \min\{\Delta(\mu),\Delta(\nu)\}$. Add (2) and (3),

$$\begin{aligned}R_{\mu,\nu} &\geq n^\gamma\Delta_{\mu,\nu}\big(\mathbb{P}_\mu^\pi(\mathcal{E}^c)+\mathbb{P}_\nu^\pi(\mathcal{E})\big) \\ &\overset{(i)}{\geq} n^\gamma\Delta_{\mu,\nu}\left(1-\mathrm{TV}(\mathbb{P}_\mu^\pi,\mathbb{P}_\nu^\pi)\right) \\ &\overset{(ii)}{\geq} n^\gamma\Delta_{\mu,\nu}\frac{1}{2}\exp\big(-D(\mathbb{P}_\mu^\pi \parallel \mathbb{P}_\nu^\pi)\big).\end{aligned}$$

In the above display, (i) follows from the definition of total variation metric, and (ii) uses the BH inequality (recalled in Theorem B.1 in Appendix B).

Taking logarithms, applying the divergence decomposition (Lemma 2.3) $D_{\mathrm{KL}}(\mathbb{P}_\mu^\pi \parallel \mathbb{P}_\nu^\pi) = \mathbb{E}_\mu^\pi[\sum_{t=0}^{n-1} A_t] \times (1+\lambda/\mu)D^\lambda(\mu \parallel \nu)$ and rearranging yields

$$\mathbb{E}_\mu^\pi\left[\sum_{t=0}^{n-1} A_t\right] \geq \frac{\gamma\log n + \log(\Delta_{\mu,\nu}/2) - \log R_{\mu,\nu}}{D^\lambda(\mu \parallel \nu)\,(1+\lambda/\mu)}. \tag{4}$$

Since the policy π is uniformly efficient (Definition 2.1), for every $a > 0$ there exists a constant $K_a < \infty$, such that $R_{\mu,\nu} \leq K_a n^a$ for all sufficiently large n. This implies that $\limsup_{n\to\infty} \log R_{\mu,\nu}/\log n \leq a$. Using this and dividing (4) by $\log n$ we get

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\mu^\pi\left[\sum_{t=0}^{n-1} A_t\right]}{\log n} \geq \frac{\gamma - a}{D^\lambda(\mu \parallel \nu)\left(1+\frac{\lambda}{\mu}\right)}.$$

As $\gamma, a \in (0,1)$ are arbitrary, for every $\delta \in (0,1)$, we can choose γ and a such that the $\gamma - a \geq 1-\delta$. Letting $\delta \downarrow 0$ yields

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\mu^\pi\left[\sum_{t=0}^{n-1} A_t\right]}{\log n} \geq \frac{1}{D^\lambda(\mu \parallel \nu)\left(1+\frac{\lambda}{\mu}\right)}.$$

When $\mu < \mu^*$, the optimal action $A_t^{\pi^*} = 0$ for all $t \geq 1$. Thus,

$$\mathbb{E}_\mu^\pi[\mathrm{Reg}(n,\pi;\mu)] = \Delta(\mu)\,\mathbb{E}_\mu^\pi\left[\sum_{t=0}^{n-1} A_t\right].$$

This implies the lower bound

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\mu^\pi[\mathrm{Reg}(n,\pi;\mu)]}{\log n} \geq \frac{\Delta(\mu)}{D^\lambda(\mu \parallel \nu)\left(1+\frac{\lambda}{\mu}\right)}.$$

Since the inequality holds for all $\nu > \mu^*$, it remains valid after taking a supremum of the right-hand side over all $\nu > \mu^*$. By Lemma 2.4 this supremum is attained in the limit $\nu \downarrow \mu^*$. Substituting this value yields the desired lower bound

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\mu^\pi[\mathrm{Reg}(n,\pi;\mu)]}{\log n} \geq \frac{\Delta(\mu)}{D^\lambda(\mu \parallel \mu^*)\left(1+\frac{\lambda}{\mu}\right)}.$$

This completes the proof. □

SUPPORTING TECHNICAL LEMMAS

This section gathers statements of the intermediate results used to prove the lower bound.

Lemma 2.3 (Divergence decomposition)**.** *Suppose Assumption 1.2 holds. For any behavioral policy π and service rate μ, let $\mathbb{P}_\mu^\pi$ denote the probability measure on the history $\mathcal{H}_n$, induced by operating the Erlang–B system under policy π with service rate μ. Then, for any two distinct service rate μ_0 and μ_1,*

$$D_{\mathrm{KL}}(\mathbb{P}_{\mu_0}^\pi \parallel \mathbb{P}_{\mu_1}^\pi) \leq \left(1+\frac{\lambda}{\mu_0}\right)\mathbb{E}_{\mu_0}^\pi\left[\sum_{t=0}^{n-1} A_t\right] D^\lambda(\mu_0 \| \mu_1).$$

Proof. We defer the proof to Appendix C.1 □

Lemma 2.4. *For every $\mu < \mu^*$,*

$$\inf_{\nu \geq \mu^*} D^\lambda(\mu \parallel \nu) = D^\lambda(\mu \parallel \mu^*).$$

Proof. We defer the proof to Appendix C.2. □

3 FIXED HORIZON ACHIEVABILITY

Theorem 2.2 establishes an information-theoretic limit stating that every uniformly efficient admission policy must incur at least logarithmic expected regret in the reject-optimal regime. In this section, we show that this lower bound is tight by constructing an explicit policy that matches the optimal leading-order coefficient when the maximum number of arrivals n is known *apriori*. Throughout, n denotes arrival horizon, that is, the total number of arrivals, rather than elapsed time. Because the optimal policy has a threshold structure, the central learning task is to determine whether the unknown service rate satisfies $\mu < \mu^*$ or $\mu > \mu^*$. Our proposed idea is to always admit whenever capacity until there is sufficient evidence to rule out $\mu > \mu^*$. If $\mu > \mu^*$, always admitting incurs no regret, and hence the key task is to detect the $\mu < \mu^*$ regime as quickly as possible, and then reject all subsequent arrivals. This leads naturally to the sequential hypothesis test

$$H_0 : \mu \geq \mu^*, \quad \text{versus} \quad H_1 : \mu < \mu^*. \tag{5}$$

Formally, given a sequence of state observations $\{N_t : t \geq 1\}$ generated under an always-admit policy, our goal is to design a *level-α power one* [Darling and Robbins, 1967, Robbins and Siegmund, 1974] test for the above hypotheses. That is, for a prescribed parameter $\alpha \in (0, 1]$ we construct a stopping time τ satisfying:

$$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha, \quad \text{and} \quad \mathbb{P}_{H_1}(\tau < \infty) = 1. \tag{6}$$

Here, τ represents the number of admissions required for sequential test to conclude $\mu < \mu^*$.

A natural approach to constructing tests (or stopping times) satisfying (6) is by thresholding a stochastic process $\{W_t : t \geq 1\}$ that quantifies the accumulated evidence against H_0. Due to the parametric nature of our problem, such a process can be constructed using likelihood ratios. For a policy π and service rate μ, let

$$L_t^\pi(\mu) := p_\mu^\pi(\mathcal{H}_t),$$

where p_μ^π is the joint density (likelihood) of the observed history $\mathcal{H}_t$ under (π, μ); its explicit construction is given in Appendix C.1. Since the policy selects actions solely as a function of observed history, its decision rule given the history is independent of the service rate. Consequently, the policy-dependent terms cancel in the likelihood ratio. Using the factorization in (C.4), we obtain

$$\frac{L_t^\pi(\mu)}{L_t^\pi(\mu^*)} = \prod_{i=0}^{t-1} \frac{P_\mu(N_{i+1} \mid N_i, A_i, T_{i+1})}{P_{\mu^*}(N_{i+1} \mid N_i, A_i, T_{i+1})}. \tag{7}$$

Henceforth, we suppress the dependence on the policy and write $L_t(\mu)$ in place of $L_t^\pi(\mu)$.

As the alternative in (5) is composite consisting of all parameters $\mu \in (0, \mu^*)$, we construct a mixture likelihood ratio with uniform prior on $(0, \mu^*)$:

$$W_0 = 1 \text{ and } \forall t > 0, \quad W_t := \frac{1}{\mu^*} \int_0^{\mu^*} \frac{L_t(\nu)}{L_t(\mu^*)} d\nu. \tag{8}$$

This mixture statistic aggregates the evidence for any alternative parameter μ against the extremal null parameter μ^*. Consequently, under the alternative, we expect this process to grow exponentially fast. We define the stopping time τ by thresholding this process as follows:

$$\tau = \inf\left\{t \geq 1 : W_t \geq \frac{1}{\alpha}\right\}. \tag{9}$$

The choice of the threshold $1/\alpha$ is justified by the fact that $\{W_t : t \geq 1\}$ is a nonnegative supermartingale under the null ($\mu > \mu^*$), and hence is subject to a time-uniform analog of Markov's inequality called Ville's inequality (recalled in Theorem B.2 in Appendix B). With the stopping time defined in (9), we now define our proposed admission control policy for the known-horizon setting.

Definition 3.1 (Fixed horizon policy π^{test})**.** *Consider an Erlang–B system with k identical servers with an unknown service rate μ. Let n denote the maximum number of jobs that may arrive, and let τ denote the stopping time* (9) *instantiated with $\alpha = 1/n$. Define the policy π^{test} that takes the following action at the t^{th} arrival:*

$$A_t^{\pi^{\text{test}}} = \mathbf{1}\{t < \tau,\ N_t < k\} = \begin{cases} 1, & \textit{if } t < \tau \ \&\ N_t < k, \\ 0, & \textit{otherwise}. \end{cases}.$$

That is, this policy accepts all jobs till the sequential test rejects the null ($\mu > \mu^*$), subject to server availability.

In the remainder of this section, we establish that π^{test} (Algorithm 1) is optimal with respect to the derived lower bound. We break down the analysis into two parts, corresponding to the $\mu > \mu^*$ (the "always accept" regime) and the $\mu < \mu^*$ (the "always reject" regime). For the former regime, we show in Theorem 3.2 that the expected regret is $O(1)$, while in the latter, we show that the regret incurred by π^{test} matches the lower bound established in Theorem 2.2 as $n \to \infty$.

3.1 CASE 1: OPTIMAL DECISION IS TO ACCEPT

When the service rate satisfies $\mu \geq \mu^*$, the optimal policy π^* accepts every arrival whenever a server is available. Under this regime, any arrival rejected by a policy while a server was idle accrues directly to the expected regret. Theorem 3.2 demonstrates that in the accept-optimal regime, π^{test} restricts the expected regret to a finite constant depending on the service rate.

Theorem 3.2 (Regret bound in accept-optimal regime)**.** *For $\mu \geq \mu^*$, the policy π^{test} (Definition 3.1) satisfies*

$$\mathbb{E}_\mu^{\pi^{test}}[\text{Reg}(n, \pi^{\text{test}}; \mu)] \leq \Delta(\mu).$$

Proof Strategy: The proof proceeds by analyzing the sample-path coupling between the two policies – π^* always accepts arrivals (subject to capacity), and π^{test} behaves identically to π^*, accepting arrivals up until the sequential test

Algorithm 1 Fixed-Horizon Test Policy π^{test}

Input: number of servers k, horizon n, threshold μ^*
Initialize:
$N_0 \leftarrow 0, A_0 \leftarrow 1, \tau \leftarrow n$
$\zeta_0(\nu) \leftarrow 0,\ \forall \nu \in [0, \mu^*]$.
for each arrival $t \in \{1, \ldots n-1\}$ while $\{t < \tau\}$ **do**
  Observe N_t, T_t.
  For all $\nu \in [0, \mu^*]$ update log-likelihood

$$\zeta_t(\nu) = \zeta_{t-1}(\nu) + l_{t-1}(\nu) - l_{t-1}(\mu^*)$$

where,

$$l_{t-1}(\nu) = -\nu T_t N_t + (N_{t-1} + A_{t-1} - N_t) \log(1 - e^{-\nu T_t}).$$

  Compute the mixture statistic

$$W_t := \frac{1}{\mu^*} \int_0^{\mu^*} \exp(\zeta_t(\nu)) d\nu.$$

  if $W_t \geq n$ **then**
    $\tau \leftarrow t$
    $A_i \leftarrow 0,\ \forall i \in \{t, t+1, \ldots, n-1\}$.
  else
    if $N_t < k$ **then**
      $A_t \leftarrow 1$
    else
      $A_t \leftarrow 0$.
    end if
  end if
end for

rejects the null. Consequently, the sample paths of the two systems match exactly on the event $\{t < \tau\}$, and regret can only accumulate after the test terminates (on the event $\{\tau < n\}$). Establishing that the mixture process $\{W_t\}_{t \geq 0}$ is a supermartingale under the null hypothesis in Lemma 3.3, we apply Ville's inequality (recalled in Theorem B.2) to bound the probability of termination, which yields the constant expected regret upper bound.

Lemma 3.3. *For $\mu \geq \mu^*$, the process $\{W_t\}_{t \geq 0}$ defined in* (8) *is a non-negative supermartingale with respect to the natural filtration $\{\sigma(\mathcal{H}_t, A_t)\}_{t \geq 0}$.*

This result, proved in Appendix D.1, provides the key technical step needed to prove Theorem 3.2 as we show next.

Proof of Theorem 3.2. Using Lemma 3.3 and by Ville's inequality (Theorem B.2) we conclude that for $\alpha = 1/n$:

$$\mathbb{P}_\mu^{\pi^{\text{test}}}(\tau < \infty) = \mathbb{P}_\mu^{\pi^{\text{test}}}(\exists t : W_t \geq n) \leq \frac{\mathbb{E}[W_0]}{n} = \frac{1}{n}. \quad (10)$$

This statement implies that if $\mu \geq \mu^*$, then with probability at least $1 - 1/n$, our proposed policy π^{test} agrees with the optimal policy π^*.

Under the policy π^{test} (Definition 3.1), the regret after n arrivals satisfies

$$\begin{aligned}\frac{\text{Reg}(n, \pi^{\text{test}}; \mu)}{\Delta(\mu)} &= \sum_{t=0}^{n-1} \mathbf{1}\{A_t^{\pi^*} = 1, A_t^{\pi^{\text{test}}} = 0\} \\ &\leq (n - \tau)_+ \mathbf{1}\{\tau < n\} \leq n \mathbf{1}\{\tau < \infty\},\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ represents the indicator function. This follows because the two policies π^{test} and π^* differ only after the random stopping time τ. Taking expectations, we get

$$\begin{aligned}\mathbb{E}_\mu^{\pi^{\text{test}}}[\text{Reg}(n, \pi^{\text{test}}; \mu)] &\leq \Delta(\mu)\, n\, \mathbb{P}_\mu^{\pi^{\text{test}}}(\tau < \infty) \\ &\overset{(i)}{\leq} \Delta(\mu)\, n\, \frac{1}{n} = \Delta(\mu),\end{aligned}$$

where (i) follows from (10). This completes the proof. □

3.2 CASE 2: OPTIMAL DECISION IS TO REJECT

When the service rate satisfies $\mu < \mu^*$, the optimal policy π^* rejects all arrivals. Theorem 3.4 shows that, in this reject-optimal regime, the expected regret incurred by our proposed policy π^{test} grows at most logarithmically with the arrival horizon n, matching the information-theoretic lower bound of Theorem 2.2 up to leading-order coefficient.

Theorem 3.4 (Regret bound in reject-optimal regime)**.** *For $\mu < \mu^*$, the policy π^{test} (Definition 3.1) satisfies*

$$\mathbb{E}_\mu^{\pi^{\textit{test}}}[\text{Reg}(n, \pi^{\textit{test}}; \mu)] \leq \frac{\Delta(\mu) \log(n)}{D^\lambda(\mu \parallel \mu^*)(1 + \lambda/\mu)} (1 + o(1)),$$

where $o(1)$ denotes a term that vanishes as $n \to \infty$.

Proof Strategy: The proof proceeds by establishing an upper bound on the expected stopping time of the sequential test under the alternative. The optimal policy π^* rejects all arrivals while π^{test} continues accepting arrivals (subject to room) up until τ, after which it rejects all arrivals and coincides with π^*. Sample-path regret ceases to accumulate after τ, so an upper bound on $\mathbb{E}_\mu^{\pi^{\text{test}}}[\tau]$ immediately yields an upper bound on the expected regret. The analysis proceeds by introducing an auxiliary likelihood process $\{\widetilde{W}_t(\mu)\}_{t \geq 0}$ and associated stopping time $\widetilde{\tau}(\mu)$ that dominates τ almost surely. Consequently, $\mathbb{E}_\mu^{\pi^{\text{test}}}[\tau] \leq \mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$. We establish that $\{\widetilde{W}_t(\mu)\}_{t \geq 0}$ is a submartingale and apply Doob's Optional Stopping Theorem (recalled in Theorem B.3).

Supporting Technical Lemmas
Here, we gather statements of intermediate results used in proving Theorem 3.4.

Definition 3.5 (Erlang–B blocking probability)**.** *The steady-state probability that an arrival finds all k servers occupied is given by the Erlang–B formula:*

$$B(k, \lambda/\mu) = \frac{(\frac{\lambda}{\mu})^k / k!}{\sum_{i=0}^k (\frac{\lambda}{\mu})^i / i!},$$

which follows from the stationary distribution of the associated birth-death process.

Lemma 3.6 (Stopping-time mixing bound)**.** *Let f be a bounded function on the state space satisfying $\|f\|_\infty \le m$, and let T be a stopping time with $\mathbb{E}[T] < +\infty$. Let N_w be a random variable distributed according to the stationary distribution w of the Erlang–B Markov chain. Then*

$$\left|\mathbb{E}\left[\sum_{i=0}^{T-1} f(N_i)\right] - \mathbb{E}[T]\mathbb{E}[f(N_w)]\right| \le \frac{mG}{1-\phi},$$

where $G > 0$ and $\phi \in (0,1)$ are the mixing constants appearing in the Convergence Theorem (recalled in Theorem B.5 in Appendix B).

This ergodic mixing bound, proved in Appendix F.1, will be used to establish Lemma 3.7 and consequently Theorem 3.4.

Lemma 3.7. *Suppose $\mu < \mu^*$. Let τ be the stopping time defined in 9 with respect to the natural filtration $\{\sigma(\mathcal{H}_t, A_t)\}_{t\ge 0}$. Then, the expected stopping time satisfies:*

$$\mathbb{E}_\mu^{\pi^{test}}[\tau] \le \frac{\log(1/\alpha)}{\gamma_*}(1+o(1))$$

where the constant γ_* is given by:

$$\gamma_* = D^\lambda(\mu \parallel \mu^*)\left(1 - B\left(k, \frac{\lambda}{\mu}\right)\right)\left(1+\frac{\lambda}{\mu}\right)$$

This result, proved in Appendix E, is the main technical result of Section 3, and immediately yields Theorem 3.4.

Proof of Theorem 3.4. The optimal policy rejects all jobs when $\mu < \mu^*$. Under the policy π^{test} (Definition 3.1), the regret after n arrivals satisfies

$$\mathrm{Reg}(n, \pi^{\text{test}}; \mu) = \Delta(\mu)\sum_{t=0}^{\tau-1}\mathbf{1}\{A_t^{\pi^{\text{test}}} = 1\}.$$

Under π^{test}, arrivals are accepted prior to τ whenever capacity is available. Thus, for all $t < \tau$, $\mathbf{1}\{A_t^{\pi^{\text{test}}} = 1\} = \mathbf{1}\{N_t < k\}$. Taking expectations and invoking the ergodic mixing bound of Lemma 3.6, together with the stationarity result $\mathbb{E}_w[\mathbf{1}\{N_t < k\}] = 1 - B(k, \lambda/\mu)$, we obtain

$$\begin{aligned}\frac{\mathbb{E}_\mu^{\pi^{\text{test}}}[\mathrm{Reg}(n,\pi^{\text{test}};\mu)]}{\Delta(\mu)} &= \mathbb{E}_\mu^{\pi^{\text{test}}}\left[\sum_{t=0}^{\tau-1}\mathbf{1}\{N_t < k\}\right] \\ &\le \mathbb{E}_\mu^{\pi^{\text{test}}}[\tau]\big(1 - B(k,\lambda/\mu)\big) + \frac{G}{1-\phi}.\end{aligned}$$

Applying Lemma 3.7 yields

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[\mathrm{Reg}(n,\pi^{\text{test}};\mu)] \le \frac{\Delta(\mu)\log(n)}{D^\lambda(\mu \parallel \mu^*)(1+\lambda/\mu)}(1+o(1)).$$

This completes the proof. □

Remark 3.8. *We address the computational tractability of the finite-horizon testing policy π^{test} in Appendix H by replacing the integral-based likelihood test with an efficient covering-grid approximation, π^{alg}, whose validity is established there. Simulation results for this approximation are presented in Appendix H.4.*

4 ANYTIME ACHIEVABILITY

This section considers the unknown-horizon setting, where the number of arrivals is not known *apriori*. Proposition 4.2 shows that the proposed policy π^∞ (Definition 4.1) preserves logarithmic regret when $\mu < \mu^*$, whereas the price of removing horizon dependence is merely an additional $\log\log n$ factor when $\mu > \mu^*$.

Definition 4.1 (Horizon-free policy π^∞)**.** *Consider an Erlang–B system with k identical servers and unknown service rate $\mu > 0$. The policy π^∞ operates over successive epochs indexed by $m \ge 0$. Each epoch consists of a test phase followed by a clear phase. Let t denote the arrival index within an epoch, with $t = 0$ at the beginning of every epoch. In epoch m, the test phase runs fixed-horizon policy π^{test} (Definition 3.1) with horizon $n_m = e^{(1+\varepsilon)^m}$, where $\varepsilon > 0$ and associated stopping time $\tau_m = \tau(\alpha_m, n_m)$, where $\alpha_m = 1/n_m$. After the test phase ends, the policy enters clear phase, during which every arrival is accepted until the system becomes empty. When the queue first reaches 0, the clear phase ends and epoch $m+1$ begins. The action taken at the t^{th} arrival of epoch m is*

$$A_t^{\pi^\infty} = \begin{cases}\mathbf{1}\{t < \tau_m,\ N_t < k\}, & t < n_m,\\ \mathbf{1}\{N_t < k\}, & t \ge n_m.\end{cases}$$

Proposition 4.2. *For unknown service rate μ, the policy π^∞ introduced in Definition 4.1 satisfies:*

$$\frac{\mathbb{E}_\mu^{\pi^\infty}[\mathrm{Reg}(n,\pi^\infty;\mu)]}{\Delta(\mu)} \le \begin{cases}\log\log(n), & \mu > \mu^*\\ \dfrac{\log(n)\ (1+o(1))}{D^\lambda(\mu \parallel \mu^*)(1+\lambda/\mu)}, & \mu < \mu^*.\end{cases}$$

The policy π^∞ (Algorithm 2) is obtained from π^{test} using the classical doubling trick [Auer et al., 1995, Besson and Kaufmann, 2018]: each epoch reruns π^{test} over a horizon that doubles (grows exponentially) with the epoch index m. The subsequent clear phase accepts every arrival until the system empties. Since empty state is regenerative, each epoch begins from the same initial condition and can therefore be analyzed independently. This regeneration incurs only a small overhead: Proposition 4.2 establishes that π^∞ attains the optimal logarithmic regret rate when $\mu < \mu^*$; while the repeated restarts introduce an additional $\log\log n$ factor only in the regime $\mu > \mu^*$. Finally, accepting all arrivals during the clear phase (rather than, for example, rejecting all arrivals) is crucial for keeping the regret bounded in the regime $\mu > \mu^*$.

Algorithm 2 Horizon-Free Test Policy π^{∞}

Input: number of servers k, threshold μ^*
Initialize: $m \leftarrow 0$, $N_0 \leftarrow 0$, $\varepsilon > 0$.
while arrival are incoming **do**
Epoch m
Test Phase
  Set horizon $n_m \leftarrow e^{(1+\varepsilon)^m}$, parameter $\alpha_m \leftarrow \dfrac{1}{n_m}$
  Discard history, $t \leftarrow 0$
  Define stopping time $\tau_m = \tau(\alpha_m, n_m)$
  for $t = 0, 1, \ldots, n_m - 1$ **do**
    if $t < \min(\tau_m, n_m)$ **and** $N_t < k$ **then**
      $A_t \leftarrow 1$.
    else
      $A_t \leftarrow 0$.
    end if
  end for
Clear Phase
  while $N_t > 0$ **do**
    if $N_t < k$ **then**
      $A_t \leftarrow 1$.
    else
      $A_t \leftarrow 0$.
    end if
  end while
  Update: $m \leftarrow m + 1$
end while

During the clear phase, the system evolves as a standard birth-death Markov chain corresponding to an $M/M/k/k$ queue. Lemma 4.3 stated below, and proved in Appendix G.1, will provide a key technical step in proving Proposition 4.2 .

Lemma 4.3 (Expected hitting time to state 0)**.** *Let T_0^h denote the maximum expected hitting time to the empty state of an Erlang–B queue operating under the policy that admits every arrival, where the maximum is taken over all initial states. Then, $T_0^h < \infty$.*

Proof of Proposition 4.2 Let T_m^c denote the duration of clear phase of epoch m. Since every clear phase starts from some state in $\{0, \ldots, k\}$ and accepts every arrival until the system empties, Lemma 4.3 implies

$$\mathbb{E}_{\mu}^{\pi^{\infty}}[T_m^c] \leq T_0^h < \infty, \quad \forall m. \tag{11}$$

Let R_m denote the expected regret during epoch m. Since the test horizon in epoch m is $n_m = e^{(1+\varepsilon)^m}$, the number of epochs completed by arrival n is at most $\frac{\log\log n}{\log(1+\varepsilon)}$. Therefore

$$\mathbb{E}_{\mu}^{\pi^{\infty}}[\mathrm{Reg}(n, \pi^{\infty}; \mu)] \leq \sum_{m=0}^{\log\log n/\log(1+\varepsilon)} R_m. \tag{12}$$

First consider the case $\mu > \mu^*$. In this regime, the optimal action is to accept whenever there is an available server. Therefore, regret is incurred only during the test phase after the stopping time τ_m. By Theorem 3.2 $\forall m$, $R_m \leq \Delta(\mu)$. Substituting into (12) yields

$$\mathbb{E}_{\mu}^{\pi^{\infty}}[\mathrm{Reg}(n, \pi^{\infty}; \mu)] \leq \Delta(\mu)\frac{\log\log(n)}{\log(1+\varepsilon)}.$$

Now consider the case $\mu < \mu^*$. In this regime, regret is incurred during the test phase, and also during the clear phase since all arrivals are accepted while clearing the system. From Theorem 3.4 and (11) we get

$$R_m \leq \Delta(\mu)\frac{(1+\varepsilon)^m}{D^{\lambda}(\mu\|\mu^*)(1+\lambda/\mu)} + \Delta(\mu)T_0^h + o(\log n_m).$$

Substituting in (12) yields

$$\mathbb{E}_{\mu}^{\pi^{\infty}}[\mathrm{Reg}(n, \pi^{\infty}; \mu)] \leq \left(\frac{1+\varepsilon}{\varepsilon}\right)\frac{\Delta(\mu)\log(n)\ (1+o(1))}{D^{\lambda}(\mu \parallel \mu^*)(1+\lambda/\mu)}.$$

The constant multiplying the leading-order regret term can be made arbitrarily close to 1 by taking ε sufficiently large. This completes the proof. $\square$

5 CONCLUSION AND FUTURE WORK

We studied learning-based optimal admission control for an Erlang–B queuing system with unknown service-rate and unobserved departures. The reward structure of our problem setting results in two operational regimes: always accept (if room) if the service rate is high enough, or always reject otherwise. In the latter regime, we prove an instance-dependent asymptotic lower bound with explicit system-dependent constants, resolving a conjecture of Adler et al. [2022] and showing that logarithmic regret is unavoidable. We also characterized how difficulty of learning-based admission control changes with system parameters. Complementing the lower bound, we propose a fixed-horizon sequential hypothesis-testing policy, π^{test} whose regret matches the lower bound up to constants in both regimes, as well as an anytime variant that incurs an additional factor in the "always accept" parameter regime. Monte Carlo simulations of our algorithm support the theoretical findings. More broadly, our results show that sequential hypothesis testing is an effective framework for online decision-making even under partial observation settings. Future directions include extending the framework to more general reward structures and service-time distributions, studying heterogeneous-server systems, and exploring connections with pure-exploration problems such as best-arm identification. Designing statistically efficient algorithms for partially observed stochastic control problems remains an important direction for future research.

AI Assistance Statement: Generative AI tools were used to assist with language editing, code development, and the exploration of proof strategies. All results presented in this paper were independently verified by the authors, who take full responsibility for the contents of this paper.

References

Saghar Adler, Mehrdad Moharrami, and Vijay Subramanian. Learning to admit optimally in an M/M/k/k+N queueing system with unknown service rate. *arXiv preprint arXiv:2202.02419*, 2022.

Rajeev Agrawal. Sample mean based index policies by o(log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(12): 1249–1259, 1989.

Venkat Anantharam, Pravin Pratap Varaiya, and Jean C. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards. *IEEE Transactions on Automatic Control*, 32:968–976, 1987.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Vivek Borkar and Pravin Varaiya. Adaptive control of Markov chains, I: Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957, 1979.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Asaf Cohen, Vijay Subramanian, and Yili Zhang. Learning-based optimal admission control in a single-server queuing system. *Stochastic Systems*, 14(1):69–107, 2024.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1761–1769, 2017.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2006.

Donald A Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences*, 1967.

John Duchi. *Lecture Notes on Statistics and Information Theory*. Course at Stanford University, 2023.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Jason Gaitonde and Éva Tardos. The price of anarchy of strategic queuing systems. *J. ACM*, 70(3), May 2023.

Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management 5(2):79-141.*, 2003.

Aditya Gopalan and Shie Mannor. Thompson Sampling for Learning Parameterized Markov Decision Processes. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 861–898, 2015.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

Bruce Hajek. *Random Processes for Engineers*. Cambridge University Press, 2015.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

Mor Harchol-Balter. *Performance modeling and design of computer systems: Queueing theory in action*. Cambridge University Press, 2013.

Frank Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

Subhashini Krishnasamy, Ari Arapostathis, Ramesh Johari, and Sanjay Shakkottai. On learning the cμ rule in single and parallel server networks. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 153–154, 2018.

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, 2020.

H. W. Kuhn. *Extensive Games and the Problem of Information*. Princeton University Press, 2016.

P. Kumar and Woei Lin. Optimal adaptive controllers for unknown Markov chains. *IEEE Transactions on Automatic Control*, 27(4):765–774, 1982.

P. R. Kumar and A. Becker. A new family of optimal adaptive controllers for Markov chains. *IEEE Transactions on Automatic Control*, 27:137–146, 1982.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2017.

P. Mandl. Estimation and control in Markov chains. *Advances in Applied Probability*, 6(1), 1974.

Peter Marbach, Atilla Eryilmaz, and Asu Ozdaglar. Asynchronous CSMA policies in multihop wireless networks with primary interference constraints. *IEEE Transactions on Information Theory*, 57(6):3644–3676, 2011.

P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

Herbert Robbins and David Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, 1974.

R Srikant and Lei Ying. *Communication networks: An optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2008.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

A EXTENDED RELATED WORK

Following up on the seminal work on adaptive control by Mandl [1974] on the optimality of the CE policy, Borkar and Varaiya [1979] revealed a fundamental limitation of CE control: once the identifiability condition is relaxed, the MLE need not converge to the true parameter, but instead to an *observationally equivalent* parameter—one that induces the exact same closed-loop transition probability. They further showed that the CE policy can prematurely converge to an incorrect parameter and thereafter rely solely on a suboptimal control strategy. Purely exploitative strategies are therefore generally insufficient, and additional exploration is required.

Broadly, two mechanisms have been used to engender this exploration. The first is *implicit* exploration, where the decision rule itself balances exploration and exploitation. One prominent approach is "optimism in the face of uncertainty": Kumar and Becker [1982] and Kumar and Lin [1982] incorporated this by biasing the likelihood function toward parameters with lower optimal costs, while Agrawal [1995] and Auer et al. [2002] later formalized this principle through the upper confidence bound (UCB) algorithm. Another approach is based on Thompson or posterior sampling [Gopalan and Mannor, 2015]. The second mechanism is *forced*, or explicit, exploration in which the controller intentionally takes informative actions irrespective of the current estimate, as in Agrawal et al. [1989].

This line of work in adaptive control was complemented by Lai and Robbins [1985], who developed an information-theoretic framework for instance-dependent regret lower bounds in sequential decision-making under uncertainty. Using a change-of-measure argument, they showed that in the single-parameter multi-armed bandit problem, any uniformly efficient policy must sample a suboptimal arm at least logarithmically often, with the logarithmic coefficient determined by the Kullback–Leibler (KL) divergence quantity between the suboptimal arm's true distribution and the closest alternative under which that arm becomes optimal. Burnetas and Katehakis [1996] generalized this to multi-parameter and non-parametric bandits, where the coefficient instead depends on the minimum KL divergence required to make a suboptimal arm optimal. Anantharam et al. [1987] extended these results to Markovian bandits, and Agrawal et al. [1989] and Graves and Lai [1997] extended them further to controlled Markov chains. In the MDP setting, Gopalan and Mannor [2015] developed a Thompson sampling algorithm for general parametrized MDPs achieving logarithmic regret, with the leading constant dependent on problem's information complexity.

Closest to our setting, Lattimore and Szepesvari [2017] showed that in structured bandit problems, optimism- and posterior-sampling- based algorithms can be arbitrarily suboptimal relative to the instance-dependent asymptotic lower bound—motivating our departure from index-based and posterior-sampling policies in favor of a sequential hypothesis-testing approach. Combes et al. [2017] generalized this observation to general structured bandit problems, deriving an instance-specific regret lower bound that specifies the minimum exploration rate required for each suboptimal arm; their algorithm OSSB, achieves this by directly tracking the optimal exploration rates.

Adaptive control in queues differs starkly from the classical bandits or MDP formulations in two structural ways: rewards are delayed and trajectory-dependent, and exploration induces a carryover cost [Gaitonde and Tardos, 2023]—for example, admitting a customer raises the cost of subsequent decisions until the queue is empty. Despite these differences, our lower bound follows the information-theoretic principle of Lai and Robbins [1985], by characterizing the rate at which information about the service time distribution can be accumulated under any uniformly efficient policy.

A growing line of work investigates learning-based admission and scheduling control in queuing systems with unknown parameters under full state observation, with performance measured via cumulative regret against a genie-aided policy. For scheduling in multiclass parallel-server systems with unknown service rates, Krishnasamy et al. [2018] analyzed the empirical $c\mu$ rule. Under conditions sufficient for stability and geometric ergodicity, they showed that the empirical rule achieves $O(1)$ holding-cost regret via forced exploration. Krishnasamy et al. [2020] later adopted a multi-armed bandit perspective for discrete-time multiclass queuing systems, proposing UCB- and Thompson sampling-based algorithms with forced exploration that achieve polylogarithmic queue regret post-stabilization. For admission control, Cohen et al. [2024] studied an $M/M/1$ queue with unknown arrival and service rates under the Naor [1969] reward structure, which induces a threshold-based admission scheme. Observing the continuous queue-length path, they proposed a batched algorithm alternating between forced-exploration and exploitation phases, achieving $O(1)$ regret when the optimal threshold is strictly positive, and $O(\log^{1+\epsilon}(N))$ regret otherwise (for any $\epsilon > 0$).

Crucially, in all of these existing works, service completions are directly observed or can inferred. Our setting differs fundamentally in its information structure: departures are never directly observed, and the only signal available is the occupancy transition recorded at each arrival epoch. Under this partial-observation model, a direct empirical estimator is not available. While Adler et al. [2022] addressed this using an MLE-based algorithm; we instead propose a sequential hypothesis-testing scheme arising from the information theoretic lower bound.

B STANDARD RESULTS (PROOF OMITTED)

Theorem B.1 (Bretagnolle–Huber inequality)**.** *[Lattimore and Szepesvári, 2020, Theorem 14.2],[Tsybakov, 2008, Section 2.3],[Duchi, 2023, Section 2.2] Consider two probability distribution P and Q on same measurable space and let A be any measurable event. Then,*

$$Q(A) - P(A) \leq TV(P,Q) \leq 1 - \frac{1}{2}e^{-D_{KL}(P\|Q)}.$$

Equivalently, $P(A) + Q(A^C) \geq \frac{1}{2}e^{-D_{KL}(P\|Q)}$.

Theorem B.2 (Ville's inequality)**.** *[Ville, 1939] Let $\{L_n : n \geq 0\}$ be a non-negative supermartingale. Then, for any $\lambda > 0$,*

$$\mathbb{P}(\exists n \geq 1 : L_n \geq \lambda) \leq \frac{\mathbb{E}[L_0]}{\lambda}$$

Theorem B.3 (Doob's Optional Stopping Theorem)**.** *[Hajek, 2015, Corollary 10.17] Suppose $(X_n : n \geq 0)$ is a martingale relative to $(\Omega, \mathcal{F}, P)$ such that:*

(i) there is a constant c such that $\mathbb{E}[|X_{n+1} - X_n| \mid \mathcal{F}_n] \leq c$ for $n \geq 0$, and

(ii) T is a stopping time such that $\mathbb{E}[T] < \infty$.

Then $E[X_T] = E[X_0]$.

Theorem B.4 (Bernstein-type bound for MDS)**.** *[Wainwright, 2019, Theorem 2.19] Let $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a martingale difference sequence, and suppose that*

$$\mathbb{E}[e^{\lambda D_i} \mid \mathcal{F}_{i-1}] \leq e^{\lambda^2 \nu_i^2/2}$$

almost surely for any $|\lambda| < 1/\alpha_i$. Then the following hold:

(a) The sum $\sum_{i=1}^{n} D_i$ is sub-exponential with parameters $\left(\sqrt{\sum_{i=1}^{n} \nu_k^2}, \alpha_*\right)$ where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$.

(b) The sum satisfies the concentration inequality

$$P\left(\left|\sum_{k=1}^{n} D_k\right| \geq t\right) \leq \begin{cases} 2\exp\left(-\dfrac{t^2}{2\sum_{k=1}^{n} \nu_k^2}\right), & \text{if } 0 \leq t \leq \dfrac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*} \\ 2\exp\left(-\dfrac{t}{2\alpha_*}\right), & \text{if } t > \dfrac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*}. \end{cases} \tag{B.1}$$

Theorem B.5 (Convergence Theorem)**.** *[Levin et al., 2017, Theorem 4.9] Suppose P is irreducible and aperiodic with stationary distribution w. Then the chain converges geometrically fast to stationarity in total variation: there exist constants $\phi \in (0,1)$ and $G > 0$ such that*

$$\max_x \|P^t(x,\cdot) - w\|_{\mathrm{TV}} \leq G\phi^t.$$

C PROOFS OF RESULTS USED IN SECTION 2

Lemma C.1 (Arrival-occupancy counting)**.** *Following the notation introduced in Section 1.1 and let $\mathcal{F}_t$ denote the natural filtration generated by the history $\mathcal{H}_t$ at arrival t. Then, we have the following:*

$$\sum_{t=0}^{n-1} N_t = \sum_{t=0}^{n-1} A_t C_t(n-1), \qquad \textit{almost surely.} \tag{C.1}$$

$$\overline{C_t(n)} := \mathbb{E}_{\mu}^{\pi}[C_t(n)|\mathcal{F}_t] \leq \frac{\lambda}{\mu}. \tag{C.2}$$

Proof. For $0 \leq t < l \leq n-1$, define the indicator, $I_{t,l} = \mathbf{1}\{\text{arrival } t \text{ is accepted and still in service at arrival } l\}$, and observe that

$$N_l = N_0 + \sum_{t=0}^{l-1} I_{t,l} = \sum_{t=0}^{l-1} I_{t,l},$$

since $N_0 = 0$ by assumption. We now sum over $l = 0$ to $l = n-1$ and interchange the order of summations to get

$$\sum_{l=0}^{n-1} N_l = \sum_{l=0}^{n-1}\sum_{t=0}^{l-1} I_{t,l} = \sum_{t=0}^{n-1}\sum_{l=t+1}^{n-1} I_{t,l} = \sum_{t=0}^{n-1} A_t C_t(n-1).$$

The last equality follow because if $A_t = 0$, then $I_{t,l} = 0$ for all $l > t$. So $I_{t,l}$ can be non-zero only if $A_t = 1$. When $A_t = 1$, $\sum_{l=t+1}^{n-1} I_{t,l}$ counts the number of arrivals in $\{t+1, t+2, \ldots, n-1\}$ during which arrival t is still in service. This is exactly $C_t(n-1)$ by definition.

Let job t arrive at time u_t and has service time S_t, then:,

$$C_t(n) := \min\Big(n - t, \#\big\{\text{arrivals during } [u_t,\, u_t + S_t)\big\}\Big).$$

On taking expectation:

$$\begin{aligned}
\overline{C_t(n)} &= \mathbb{E}\left[\sum_{j=1}^{n-t} j \frac{(\lambda S_t)^j}{j!} e^{-\lambda S_t}\right] + \mathbb{E}\left[(n-t)\sum_{j=n-t+1}^{\infty} \frac{(\lambda S_t)^j}{j!} e^{-\lambda S_t}\right] \\
&= \sum_{j=1}^{n-t} j \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\lambda+\mu}\right) + (n-t)\sum_{j=n-t+1}^{\infty} \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\lambda+\mu}\right) \\
&= \sum_{j=1}^{n-t} j \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\lambda+\mu}\right) + (n-t)\left(\frac{\lambda}{\lambda+\mu}\right)^{n-t+1} \\
&= \frac{\lambda}{\mu}\left(1 - (n-t+1)\left(\frac{\lambda}{\lambda+\mu}\right)^{n-t}\right) + \frac{\lambda}{\mu}\left((n-t)\left(\frac{\lambda}{\lambda+\mu}\right)^{n-t+1}\right) + (n-t)\left(\frac{\lambda}{\lambda+\mu}\right)^{n-t+1} \\
&= \frac{\lambda}{\mu}\left(1 - \left(\frac{\lambda}{\lambda+\mu}\right)^{n-t}\right) \leq \frac{\lambda}{\mu}
\end{aligned}$$

□

C.1 PROOF OF LEMMA 2.3

Proof. Let $\Lambda = \Lambda^n_{\text{Leb}} \times \kappa$ be a reference dominating measure over the sample space of realizations $h_n = \{(a_t)_{t=0}^{n-1}, (t_t)_{t=1}^{n}, (n_t)_{t=0}^{n}\}$, where Λ^n_{Leb} is the standard Lebesgue measure on $\mathbb{R}^n_+$ governing the inter-arrival times, and κ is the counting measure governing the discrete action and state spaces. Let p^π_μ denote the joint Radon-Nikodym derivative of $\mathbb{P}^\pi_\mu$ with respect to Λ. Since π is a behavioral policy, the distribution of action A_t depends only on the realized history $\mathcal{H}_t$. By Assumption 1.2, we have

$$p^\pi_\mu(h_n) = \left(\prod_{t=1}^{n} \lambda e^{-\lambda t_t}\right) \prod_{t=0}^{n-1} p^\pi(a_t \mid h_t)\, P_\mu(n_{t+1} \mid n_t, a_t, t_{t+1}), \tag{C.3}$$

where $p^\pi(a_t \mid h_t) := \mathbb{P}^\pi_\mu(A_t = a_t \mid \mathcal{H}_t = h_t)$ is the probability of choosing action a_t given history h_t under π (independent of service rate μ) and $P_\mu(n_{t+1} \mid n_t, a_t, t_{t+1}) := \mathbb{P}^\pi_\mu(N_{t+1} = n_{t+1} \mid N_t = n_t, A_t = a_t, T_{t+1} = t_{t+1})$ is the state transition probability under service rate μ (independent of policy π). By memoryless property of exponential distribution, each of the $n_t + a_t$ servers independently completes the service during inter-arrival time t_{t+1} with probability $1 - e^{-\mu t_{t+1}}$. Thus, conditional on N_t, A_t, and T_{t+1}, the remaining number of busy servers N_{t+1} follows a binomial distribution:

$$P_\mu(N_{t+1} = n_{t+1} \mid n_t, a_t, t_{t+1}) = \binom{n_t + a_t}{n_{t+1}} (e^{-\mu t_{t+1}})^{n_{t+1}} (1 - e^{-\mu t_{t+1}})^{n_t + a_t - n_{t+1}}.$$

To evaluate the relative entropy, we construct the log-likelihood ratio between the parameters μ_0 and μ_1. Because the inter-arrival time density terms and the policy action probabilities $p^\pi(a_t \mid h_t)$ are independent of the service rate μ, they cancel identically in the ratio, yielding

$$\log \frac{p^\pi_{\mu_0}(\mathcal{H}_n)}{p^\pi_{\mu_1}(\mathcal{H}_n)} = \sum_{t=0}^{n-1} \log \frac{P_{\mu_0}(N_{t+1} \mid N_t, A_t, T_{t+1})}{P_{\mu_1}(N_{t+1} \mid N_t, A_t, T_{t+1})}. \tag{C.4}$$

Taking the expectation of this log-likelihood ratio under the probability measure $\mathbb{P}^\pi_{\mu_0}$, and using the standard identity for the KL divergence between two binomial distributions, $D_{\mathrm{KL}}\big(\mathrm{Bin}(m, q_0) \parallel \mathrm{Bin}(m, q_1)\big) = m\, d(q_0 \parallel q_1)$, we obtain

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}^\pi_{\mu_0} \parallel \mathbb{P}^\pi_{\mu_1}) &= \sum_{t=0}^{n-1} \mathbb{E}^\pi_{\mu_0}\left[(N_t + A_t)\, d(e^{-\mu_0 T_{t+1}} \parallel e^{-\mu_1 T_{t+1}})\right] \\
&\overset{(i)}{=} \mathbb{E}^\pi_{\mu_0}\left[\sum_{t=0}^{n-1}(N_t + A_t)\right] D^\lambda(\mu_0 \parallel \mu_1) \\
&\overset{(ii)}{=} \mathbb{E}^\pi_{\mu_0}\left[\sum_{t=0}^{n-1}(A_t(1 + C_t(n-1)))\right] D^\lambda(\mu_0 \parallel \mu_1) \\
&\overset{(iii)}{=} \mathbb{E}^\pi_{\mu_0}\left[\sum_{t=0}^{n-1}(A_t(1 + \overline{C_t(n-1)}))\right] D^\lambda(\mu_0 \parallel \mu_1) \\
&\overset{(iv)}{\leq} \left(1 + \frac{\lambda}{\mu_0}\right) \mathbb{E}^\pi_{\mu_0}\left[\sum_{t=0}^{n-1} A_t\right] D^\lambda(\mu_0 \parallel \mu_1).
\end{aligned}$$

In the above display, (i) follows from Assumption 1.2, (ii) follows from (C.1), (iii) follows from tower property by taking conditional expectation given $\sigma(\mathcal{H}_t)$ and (iv) follows from (C.2). The term $(1 + \lambda/\mu_0)D^\lambda(\mu_0 \parallel \mu_1)$ quantifies the expected information generated by each accepted arrival to distinguish between the two service rates based on future observations of ongoing service at arrivals epochs. □

C.2 PROOF OF LEMMA 2.4

Proof. By definition of the inter-arrival KL divergence,

$$\begin{aligned}
D^\lambda(\mu \parallel \nu) &= \mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}[d(e^{-\mu\tau} \parallel e^{-\nu\tau})]. \\
&= \underbrace{\mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}\left[e^{-\mu\tau} \log \frac{e^{-\mu\tau}}{e^{-\nu\tau}}\right]}_{I_1} + \underbrace{\mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}\left[(1 - e^{-\mu\tau}) \log \frac{(1 - e^{-\mu\tau})}{(1 - e^{-\nu\tau})}\right]}_{I_2}
\end{aligned}$$

Calculating the two terms separately gives us:

$$I_1 = \mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}\left[e^{-\mu\tau}(-\mu\tau - (-\nu\tau))\right] = \mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}\left[(\nu - \mu)\tau e^{-\mu\tau}\right] = (\nu - \mu)\int_0^\infty t e^{-\mu t} \lambda e^{-\lambda t} dt = (\nu - \mu)\frac{\lambda}{(\lambda + \mu)^2},$$

and

$$I_2 = \mathbb{E}_{\tau \sim \mathrm{Exp}(\lambda)}\left[(1 - e^{-\mu\tau}) \log\left(\frac{1 - e^{-\mu\tau}}{1 - e^{-\nu\tau}}\right)\right] = \int_0^\infty (1 - e^{-\mu t}) \log\left(\frac{1 - e^{-\mu t}}{1 - e^{-\nu t}}\right) \lambda e^{-\lambda t} dt.$$

On differentiating with respect to ν we get:

$$\frac{\partial}{\partial \nu} D^\lambda(\mu \parallel \nu) = \frac{\lambda}{(\lambda + \mu)^2} + \int_0^\infty (1 - e^{-\mu t}) \frac{t\lambda e^{-\lambda t}}{(1 - e^{-\nu t})} dt$$

Observe that $\frac{\lambda}{(\lambda+\mu)^2} > 0$ and $\forall t > 0$, $(1 - e^{-\mu t})\frac{t\lambda e^{-\lambda t}}{(1 - e^{-\nu t})} > 0$. Thus, $\frac{\partial}{\partial \nu} D^\lambda(\mu \parallel \nu) > 0$, $\forall \nu > \mu$. Hence, for $\mu < \mu^*$, $D^\lambda(\mu \parallel \nu)$ is strictly increasing for $\nu > \mu$. The infimum of $D^\lambda(\mu \parallel \nu)$ in the feasible set $\nu \geq \mu^*$ is attained at the end point, that is, at μ^*. □

D ACHIEVABILITY RESULTS IN ACCEPT-OPTIMAL REGIME

D.1 PROOF OF LEMMA 3.3

Proof. Let $\mathcal{F}_t := \sigma(\mathcal{H}_t, A_t)$ denote the natural filtration generated by the observed history $\mathcal{H}_t = \{N_0, A_0, T_1, \ldots, N_{t-1}, A_{t-1}, T_t, N_t\}$ and the action A_t at arrival t. For any $\nu < \mu^*$, define the process $\{M_t(\nu) : t \geq 0\}$ as :

$$M_0(\nu) = 1 \quad \text{and} \quad \forall t > 0, \quad M_t(\nu) := \frac{L_t(\nu)}{L_t(\mu^*)}.$$

For service rate $\mu \geq \mu^*$, observe that $M_t(\nu)$ is non-negative and $\mathcal{F}_t$-measurable as likelihood depends only on $(N_i, A_i, T_{i+1}, N_{i+1})$ for $i = 0, \ldots, t-1$, all of which are $\mathcal{F}_t$-measurable. Moreover, $M_t(\mu)$ is integrable for each t. It follow from (7) that

$$\mathbb{E}_\mu^{\pi^{\text{test}}}\left[\frac{L_{t+1}(\nu)}{L_{t+1}(\mu^*)}\middle|\mathcal{F}_t\right] = \frac{L_t(\nu)}{L_t(\mu^*)}\,\mathbb{E}_\mu^{\pi^{\text{test}}}\left[\frac{P_\nu(N_{t+1} \mid N_t, A_t, T_{t+1})}{P_{\mu^*}(N_{t+1} \mid N_t, A_t, T_{t+1})}\middle|\mathcal{F}_t\right].$$

Under rate $\mu' > 0$, the conditional distribution of N_{t+1} given $(\mathcal{H}_t, A_t, T_{t+1})$ is $\mathrm{Bin}(N_t + A_t, e^{-\mu' T_{t+1}})$, that is,

$$P_{\mu'}(N_{t+1} = n_{t+1} \mid N_t = n_t, A_t = a_t, T_{t+1} = t_{t+1}) = \binom{n_t + a_t}{n_{t+1}}(e^{-\mu' t_{t+1}})^{n_{t+1}}(1 - e^{-\mu' t_{t+1}})^{n_t + a_t - n_{t+1}}.$$

Since the number of servers k is finite, n_{t+1} and $n_t + a_t$ bounded above by k and $k+1$ respectively. As $e^{-\mu T}$ is strictly decreasing in μ for a fixed T, we obtain the following:

$$\nu < \mu^* < \mu \quad \Longrightarrow \quad e^{-\nu T} > e^{-\mu^* T} > e^{-\mu T}. \tag{D.1}$$

To simplify expression, introduce $a := e^{-\nu t_{t+1}}$, $b := e^{-\mu^* t_{t+1}}$, $c := e^{-\mu t_{t+1}}$, and observe that

$$\frac{P_\nu(N_{t+1} \mid N_t, A_t, T_{t+1})}{P_{\mu^*}(N_{t+1} \mid N_t, A_t, T_{t+1})} = \left(\frac{e^{-\nu t_{t+1}}}{e^{-\mu^* t_{t+1}}}\right)^{n_{t+1}}\left(\frac{1 - e^{-\nu t_{t+1}}}{1 - e^{-\mu^* t_{t+1}}}\right)^{n_t + a_t - n_{t+1}} = \left(\frac{a}{b}\right)^{n_{t+1}}\left(\frac{1-a}{1-b}\right)^{n_t + a_t - n_{t+1}}.$$

Taking expectation under service rate μ, we obtain

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^{\text{test}}}\left[\frac{P_\nu(N_{t+1} \mid N_t, A_t, T_{t+1})}{P_{\mu^*}(N_{t+1} \mid N_t, A_t, T_{t+1})}\middle|\mathcal{F}_t\right] &= \sum_{n_{t+1}=0}^{n_t+a_t}\binom{n_t + a_t}{n_{t+1}}c^{n_{t+1}}(1-c)^{n_t+a_t-n_{t+1}}\left(\frac{a}{b}\right)^{n_{t+1}}\left(\frac{1-a}{1-b}\right)^{n_t+a_t-n_{t+1}} \\
&= \left(\frac{ac}{b} + \frac{(1-a)(1-c)}{(1-b)}\right)^{n_t+a_t} \\
&\leq 1.
\end{aligned} \tag{D.2}$$

To see why the last inequality in (D.2) is true, observe that

$$1 - \frac{ac}{b} - \frac{(1-a)(1-c)}{1-b} = \frac{(a-b)(b-c)}{b(1-b)} \geq 0,$$

since $0 < c \leq b \leq a \leq 1$. For $\nu < \mu^* < \mu$ from (D.1) and (D.2), we get

$$\mathbb{E}_\mu^{\pi^{\text{test}}}\left[M_{t+1}(\nu) \mid \mathcal{F}_t\right] = \mathbb{E}_\mu^{\pi^{\text{test}}}\left[\frac{L_{t+1}(\nu)}{L_{t+1}(\mu^*)}\middle|\mathcal{F}_t\right] \leq \frac{L_t(\nu)}{L_t(\mu^*)} = M_t(\nu) \quad \text{a.s.} \tag{D.3}$$

Thus, under $\mu \geq \mu^*$, $\{M_t(\nu)\}_{t\geq 0}$ is a nonnegative supermartingale for every $\nu < \mu^*$. Finally, since $W_t = (1/\mu^*)\int_0^{\mu^*} M_t(\nu)d\nu$, we have

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[W_{t+1}|\mathcal{F}_t] = \mathbb{E}_\mu^{\pi^{\text{test}}}\left[\frac{1}{\mu^*}\int_0^{\mu^*} M_{t+1}(\nu)d\nu\middle|\mathcal{F}_t\right] \overset{(i)}{=} \frac{1}{\mu^*}\int_0^{\mu^*}\mathbb{E}_\mu^{\pi^{\text{test}}}\left[M_{t+1}(\nu)|\mathcal{F}_t\right]d\nu \overset{(ii)}{\leq} \frac{1}{\mu^*}\int_0^{\mu^*} M_t(\nu)d\nu = W_t$$

almost surely. Here, (i) follows from an application of Tonelli's theorem, and (ii) uses (D.3). Hence, $\{W_t\}_{t\geq 0}$ is a nonnegative supermartingale under the null, that is, when $\mu \geq \mu^*$. $\square$

E PROOF OF LEMMA 3.7

This section establishes the key technical result needed for the proof of Theorem 3.4. We establish an upper bound on the expected stopping time – defined in (9)– under the alternative hypothesis. We first recall the statement of Lemma 3.7.

Lemma 3.7. Suppose $\mu < \mu^*$. Let τ be the stopping time defined in 9 with respect to the natural filtration $\{\sigma(\mathcal{H}_t, A_t)\}_{t\geq 0}$. Then, the expected stopping time satisfies:

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\tau] \leq \frac{\log(1/\alpha)}{\gamma_*}(1+o(1))$$

where the constant γ_* is given by:

$$\gamma_* = D^{\lambda}(\mu \parallel \mu^*)\left(1 - B\left(k, \frac{\lambda}{\mu}\right)\right)\left(1 + \frac{\lambda}{\mu}\right)$$

Proof Strategy: The proof proceeds in four steps. First, we introduce an auxiliary process $\widetilde{W}_t(\mu)$ and an associated stopping time $\widetilde{\tau}(\mu)$ that stochastically dominates τ, so that it suffices to bound $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$ (Lemma E.1). Second, we decompose $\widetilde{W}_t(\mu)$ via its Doob decomposition into a martingale part $M_t(\mu)$ and a predictable part $Y_t(\mu)$ (Proposition E.2), and show $Y_t(\mu)$ grows at least linearly (Lemma E.4). Third, in Lemma E.5 we verify the hypotheses of Doob's Optional Stopping Theorem for $\{M_t(\mu)\}$ at $\widetilde{\tau}(\mu)$, which requires showing $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] < \infty$; we do this by combining the martingale strong law of large numbers (which shows the linear drift asymptotically dominates the martingale fluctuations almost surely) with a Bernstein-type concentration bound (which quantifies this domination at a rate sufficient for summability). Fourth, we combine the optional stopping result with the drift lower bound and an overshoot bound to obtain the stated bound on $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$; optimizing the auxiliary parameter δ at the end (Lemma 3.7). The rest of this section makes this argument precise.

Define the auxiliary process as:

$$\widetilde{W}_0(\mu) = 0, \quad \text{and} \quad \forall t > 0,\ \widetilde{W}_t(\mu) := \frac{1}{\delta}\int_{\mu-\delta}^{\mu} \log\frac{L_t(\nu)}{L_t(\mu^*)}d\nu, \tag{E.1}$$

and the associated stopping time as:

$$\widetilde{\tau}(\mu) := \inf\left\{t \geq 1 : \widetilde{W}_t(\mu) \geq \log\left(\frac{\mu^*}{\delta\alpha}\right)\right\}. \tag{E.2}$$

Both $\widetilde{W}_t(\mu)$ and $\widetilde{\tau}(\mu)$ depend on the auxiliary parameter δ, which we suppress from the notation until Section E.2, where it is fixed to an explicit value depending on μ and α.

E.1 SUPPORTING TECHNICAL LEMMAS

Lemma E.1. *Suppose $\mu < \mu^*$. Then for all $\delta \in (0, \mu^*)$,*

$$W_t \geq \frac{\delta}{\mu^*}\exp(\widetilde{W}_t(\mu)),$$

where, the process W_t is defined in (8) *and $\widetilde{W}_t(\mu)$ is defined in* (E.1). *Consequently, the stopping time $\widetilde{\tau}(\mu)$ (defined in* (E.2)*) satisfies $\tau \leq \widetilde{\tau}(\mu)$ and so, $\mathbb{E}_{\mu}^{\pi^{test}}[\tau] \leq \mathbb{E}_{\mu}^{\pi^{test}}[\widetilde{\tau}(\mu)]$.*

Proof. The proof is a consequence of the nonnegativity of likelihood ratios and the convexity of exponential function as shown below:

$$\begin{aligned} W_t := \frac{1}{\mu^*}\int_0^{\mu^*}\frac{L_t(\nu)}{L_t(\mu^*)}d\nu &\geq \frac{1}{\mu^*}\int_{\mu-\delta}^{\mu}\frac{L_t(\nu)}{L_t(\mu^*)}d\nu \\ &= \frac{1}{\mu^*}\int_{\mu-\delta}^{\mu}\exp\left(\log\left(\frac{L_t(\nu)}{L_t(\mu^*)}\right)\right)d\nu \end{aligned}$$

$$
\begin{aligned}
&= \frac{\delta}{\mu^*} \int_{\mu-\delta}^{\mu} \frac{1}{\delta} \exp\left(\log\left(\frac{L_t(\nu)}{L_t(\mu^*)}\right)\right) d\nu \\
&\overset{(i)}{\geq} \frac{\delta}{\mu^*} \exp\left(\frac{1}{\delta} \int_{\mu-\delta}^{\mu} \log\left(\frac{L_t(\nu)}{L_t(\mu^*)}\right) d\nu\right) \\
&= \frac{\delta}{\mu^*} \exp(\widetilde{W}_t(\mu)),
\end{aligned}
$$

where (i) follows from Jensen's inequality applied to the convex function $x \mapsto e^x$ with respect to the uniform probability measure $\frac{1}{\delta} d\nu$ on $[\mu - \delta, \mu]$. This lower bound directly compares the two stopping times: on the event $\{\widetilde{W}_t(\mu) \geq \log(\mu^*/(\delta\alpha))\}$, the bound above gives

$$
W_t \;\geq\; \frac{\delta}{\mu^*} \exp(\widetilde{W}_t(\mu)) \;\geq\; \frac{\delta}{\mu^*} \cdot \frac{\mu^*}{\delta\alpha} \;=\; \frac{1}{\alpha}.
$$

Since, every crossing of $\widetilde{W}_t(\mu)$ above $\log(\mu^*/(\delta\alpha))$, necessarily implies a crossing of W_t above $1/\alpha$, and τ is the first time $W_t \geq 1/\alpha$, it follows that

$$
\tau \leq \widetilde{\tau}(\mu) \qquad \text{almost surely,}
$$

and therefore $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\tau] \leq \mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$. ☐

Define the natural filtration generated by past actions and observations up to arrival t:

$$
\mathcal{F}_t := \sigma(\mathcal{H}_t, A_t) = \sigma(N_0, A_0, T_1, \ldots, N_{t-1}, A_{t-1}, T_t, N_t, A_t). \tag{E.3}
$$

Proposition E.2 (Doob Decomposition)**.** *Let $\widetilde{W}_t(\mu)$ be the process defined in* (E.1) *and let $\{\mathcal{F}_t\}_{t\geq 0}$ be the filtration defined in* (E.3)*. There exists an almost surely unique decomposition*

$$
\widetilde{W}_t(\mu) = M_t(\mu) + Y_t(\mu), \quad \forall t \geq 0,
$$

where $\{M_t(\mu)\}_{t\geq 0}$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-martingale under μ and $\{Y_t(\mu)\}_{t\geq 0}$ is a predictable process. The explicit expression is given by:

$$
\begin{aligned}
M_t(\mu) &:= \widetilde{W}_t(\mu) - \sum_{i=0}^{t-1} \mathbb{E}_{\mu}^{\pi^{test}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) \mid \mathcal{F}_i], \quad M_0(\mu) = 0, \\
\textit{and } Y_t(\mu) &:= \sum_{i=0}^{t-1} \mathbb{E}_{\mu}^{\pi^{test}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i], \quad Y_0(\mu) = 0.
\end{aligned}
$$

Proof. The process $\{\widetilde{W}_t(\mu)\}_{t\geq 0}$ is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ and satisfies $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[|\widetilde{W}_t(\mu)|] < \infty$ for every $t \geq 0$. By (7), the log-likelihood ratio comparing ν to μ^* at arrival t is:

$$
\begin{aligned}
\log \frac{L_t(\nu)}{L_t(\mu^*)} &:= \sum_{i=0}^{t-1} \log \frac{P_\nu(N_{i+1} \mid N_i, A_i, T_{i+1})}{P_{\mu^*}(N_{i+1} \mid N_i, A_i, T_{i+1})} \\
&= \sum_{i=0}^{t-1} (\mu^* - \nu)\, T_{i+1} N_{i+1} + (N_i + A_i - N_{i+1}) \log\left(\frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}}\right).
\end{aligned} \tag{E.4}
$$

Each summand in (E.4) depends only on $(N_i, A_i, T_{i+1}, N_{i+1})$ for $i = 0, \ldots, t-1$, all of which are $\mathcal{F}_t$-measurable. Hence $\log L_t(\nu)/L_t(\mu^*)$ is $\mathcal{F}_t$-measurable for every fixed $\nu \in [\mu - \delta, \mu]$, and since $\widetilde{W}_t(\mu)$, is a Lebesgue integral over a bounded interval of $\mathcal{F}_t$-measurable integrands, it is $\mathcal{F}_t$-measurable.

Thus,

$$
\mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[\left|\log \frac{L_t(\nu)}{L_t(\mu^*)}\right|\right] \overset{(i)}{\leq} \mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[\sum_{i=0}^{t-1} (\mu^* - \nu) T_{i+1} N_{i+1}\right] + \mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[\sum_{i=0}^{t-1} (N_i + A_i - N_{i+1}) \left|\log \frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}}\right|\right]
$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} (\mu^* - \nu)\frac{kt}{\lambda} + (k+1)t\left|\log\frac{\nu}{\mu^*}\right| \\
&\overset{(iii)}{\leq} (\mu^* - \mu + \delta)\frac{kt}{\lambda} + (k+1)t\log\frac{\mu^*}{\mu - \delta}.
\end{aligned}
\tag{E.5}
$$

In the above display (i) follows from the decomposition in (E.4) For (ii), we bound the two terms separately. The first term is controlled using $N_{i+1} \leq k$ and $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[T_{i+1}] = 1/\lambda$. For the second term, we use the inequality

$$
\left|\log\frac{1-e^{-\nu T}}{1-e^{-\mu^* T}}\right| \leq \left|\log\frac{\nu}{\mu^*}\right|, \qquad T > 0,
$$

together with the almost-sure bound $N_i + A_i - N_{i+1} \leq k+1$. (iii) holds because $\nu \in [\mu - \delta, \mu]$.

Finally, integrating over $\nu \in [\mu - \delta, \mu]$ and applying Fubini's theorem gives

$$
\mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[|\widetilde{W}_t(\mu)|\right] \leq K(\mu, \mu^*, \delta, k, \lambda)\, t < \infty,
\tag{E.6}
$$

where $K(\mu, \mu^*, \delta, k, \lambda) := (\mu^* - \mu + \delta)\, k/\lambda + (k+1)\log(\mu^*/(\mu - \delta))$ is a finite constant. Since $\{\widetilde{W}_t(\mu)\}_{t \geq 0}$ is adapted and integrable with respect to $\{\mathcal{F}_t\}_{t \geq 0}$ it admits an almost surely unique Doob decomposition [Hajek, 2015, Example 10.7], given explicitly by

$$
\begin{aligned}
\widetilde{W}_t(\mu) &= M_t(\mu) + Y_t(\mu), \quad \forall t \geq 0, \\
\text{where, } M_t(\mu) &:= \widetilde{W}_t(\mu) - \sum_{i=0}^{t-1}\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) \mid \mathcal{F}_i], \quad M_0(\mu) = 0, \\
\text{and } Y_t(\mu) &:= \sum_{i=0}^{t-1}\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i], \quad Y_0(\mu) = 0.
\end{aligned}
$$

□

Lemma E.3. *Let $\mu > 0$ and suppose $\delta \in (0, \mu/2)$. Then*

$$
\frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^{\lambda}(\mu\|\nu)\, d\nu \leq \frac{\lambda\delta^2}{3(\mu - 2\delta)^3} + \frac{0.65\,\lambda\delta^2}{3(\mu-\delta)^3}.
$$

Proof. Evaluating the δ-strip integral of the inter-arrival KL divergence yields

$$
\int_{\mu-\delta}^{\mu} D^{\lambda}(\mu||\nu)d\nu = \int_{\mu-\delta}^{\mu}\int_0^{\infty}\lambda e^{-\lambda t} d(e^{-\mu t}||e^{-\nu t})dt\, d\nu \overset{(i)}{=} \int_0^{\infty}\lambda e^{-\lambda t}\left(\int_{\mu-\delta}^{\mu} d(e^{-\mu t}||e^{-\nu t})d\nu\right)dt,
$$

where (i) follows from Fubini's theorem.
Let p and q be two Bernoulli distributions with $0 < p < q < 1$, and let $f = q - p$. Then by Taylor's theorem along with the mean-value theorem we get,

$$
\begin{aligned}
d(p \parallel q) &= d(p \parallel p) + \left(-\frac{pf}{1+\alpha f} + \frac{(1-p)f}{1-p-\alpha f}\right)\Bigg|_{\alpha=0} + \frac{f^2}{2}\left(\frac{p}{(p+\tilde{\alpha}f)^2} + \frac{1-p}{(1-p-\tilde{\alpha}f)^2}\right) \\
&= \frac{f^2}{2}\left(\frac{p}{(p+\tilde{\alpha}f)^2} + \frac{1-p}{(1-p-\tilde{\alpha}f)^2}\right),
\end{aligned}
$$

for some $\tilde{\alpha} \in [0,1]$. Maximizing each term separately over $\tilde{\alpha} \in [0,1]$ we have

$$
d(p \parallel p + f) \leq \frac{f^2}{2}\left(\frac{1}{p} + \frac{1-p}{(1-q)^2}\right).
\tag{E.7}
$$

Using (E.7), we have for $\nu \in [\mu - \delta, \mu]$

$$
d(e^{-\mu t} \parallel e^{-\nu t}) \leq \frac{1}{2}(e^{-\mu t} - e^{-\nu t})^2\left(e^{\mu t} + \frac{1-e^{-\mu t}}{(1-e^{-\nu t})^2}\right)
$$

$$
\begin{aligned}
&< \frac{1}{2}(e^{-\mu t} - e^{-\nu t})^2 e^{\mu t} + \frac{(e^{-\mu t} - e^{-\nu t})^2}{(1 - e^{-\nu t})^2} \\
&\overset{(i)}{=} \frac{1}{2}(e^{-\mu t} + e^{(\mu - 2\nu)t} - 2e^{-\nu t}) + \frac{1}{2}(e^{-2\mu t} + e^{-2\nu t} - 2e^{-(\mu+\nu)t}) \sum_{n=0}^{\infty}(n+1)e^{-n\nu t} \\
&= \frac{1}{2}(e^{-\mu t} + e^{(\mu - 2\nu)t} - 2e^{-\nu t}) \\
&\quad + \frac{1}{2}\sum_{n=0}^{\infty}(n+1)e^{-(2\mu + n\nu)t} + (n+1)e^{-(n+2)\nu t} - 2(n+1)e^{-(\mu + (n+1)\nu)t},
\end{aligned} \tag{E.8}
$$

where, (i) follows from geometric series expansion.

For the contribution of the second term in (E.8) to the inter-arrival KL divergence, we interchange the order of integration and summation by Tonelli's theorem, since each summand is non-negative. Thus,

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \int_0^{\infty} \lambda e^{-\lambda t} \left((n+1)e^{-(2\mu+n\nu)t} + (n+1)e^{-(n+2)\nu t} - 2(n+1)e^{-(\mu+(n+1)\nu)t}\right) dt \\
&= \sum_{n=0}^{\infty}(n+1)\lambda \left(\frac{1}{\lambda + 2\mu + n\nu} + \frac{1}{\lambda + (n+2)\nu} - \frac{2}{\lambda + \mu + (n+1)\nu}\right) \\
&= 2\lambda(\mu - \nu)^2 \sum_{n=0}^{\infty} \frac{n+1}{(\lambda + \mu + (n+1)\nu)(\lambda + 2\mu + n\nu)(\lambda + (n+2)\nu)} \\
&\le 2\lambda(\mu - \nu)^2 \frac{1}{(\mu - \delta)^3} \left(\sum_{n=0}^{\infty} \frac{1}{(n+2)^2}\right) = 2\lambda(\mu - \nu)^2 \frac{1}{(\mu - \delta)^3}\left(\frac{\pi^2}{6} - 1\right) \le 1.3\frac{\lambda(\mu - \nu)^2}{(\mu - \delta)^3}.
\end{aligned}
$$

Assuming that $\delta < \mu/2$, the integral of the first term in (E.8) gives

$$
\begin{aligned}
\int_0^{\infty} \lambda e^{-\lambda t}(e^{-\mu t} + e^{(\mu - 2\nu)t} - 2e^{-\nu t})dt &= \frac{\lambda}{\lambda + \mu} + \frac{\lambda}{\lambda + 2\nu - \mu} - \frac{2\lambda}{\lambda + \nu} \\
&= \frac{2\lambda(\mu - \nu)^2}{(\lambda + 2\nu - \mu)(\lambda + \nu)(\lambda + \mu)} \le \frac{2\lambda(\mu - \nu)^2}{(\mu - 2\delta)^3}.
\end{aligned}
$$

Combining the two bounds yields

$$
\int_{\mu-\delta}^{\mu} D^{\lambda}(\mu||\nu)d\nu \le \int_{\mu-\delta}^{\mu}\left(\int_0^{\infty} \lambda e^{-\lambda t} d(e^{-\mu t}||e^{-\nu t})dt\right) d\nu \le \int_{\mu-\delta}^{\mu}\left(\frac{\lambda(\mu - \nu)^2}{(\mu - 2\delta)^3} + 0.65\frac{\lambda(\mu - \nu)^2}{(\mu - \delta)^3}\right) d\nu.
$$

Therefore, we have

$$
\frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^{\lambda}(\mu||\nu)d\nu \le \frac{1}{3}\frac{\lambda\delta^2}{(\mu - 2\delta)^3} + \frac{0.65}{3}\frac{\lambda\delta^2}{(\mu - \delta)^3}.
$$

□

Lemma E.4. *There exists $\delta^*(\mu) > 0$ such that for every $\delta \in (0, \delta^*(\mu))$, the process $\{\widetilde{W}_t(\mu)\}_{t\ge 0}$ defined in* (E.1) *is a submartingale under μ with respect to the filtration $\{\mathcal{F}_t\}_{t\ge 0}$ defined in* (E.3).

Proof. Let $\widetilde{W}_t(\mu) = M_t(\mu) + Y_t(\mu)$ be the the (a.s.) unique Doob decomposition of $\widetilde{W}_t(\mu)$ under μ with respect to $\{\mathcal{F}_t\}_{t\ge 0}$ defined in Proposition E.2. $\widetilde{W}_t(\mu)$ is an $\{\mathcal{F}_t\}_{t\ge 0}$-submartingale under μ if and only if $Y_t(\mu)$ is almost surely non-decreasing in t [Durrett, 2019, Section 4.2]. From (E.4) we have,

$$
\begin{aligned}
\widetilde{W}_t(\mu) &:= \frac{1}{\delta}\int_{\mu-\delta}^{\mu} \log\left(\frac{L_t(\nu)}{L_t(\mu^*)}\right) d\nu \\
&= \frac{1}{\delta}\int_{\mu-\delta}^{\mu}\left(\sum_{i=0}^{t-1}(\mu^* - \nu)T_{i+1}N_{i+1} + (N_i + A_i - N_{i+1})\log\left(\frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}}\right)\right) d\nu
\end{aligned}
$$

$$= \sum_{i=0}^{t-1} \frac{1}{\delta} \int_{\mu-\delta}^{\mu} \left((\mu^* - \nu) T_{i+1} N_{i+1} + (N_i + A_i - N_{i+1}) \log \left(\frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}} \right) \right) d\nu. \tag{E.9}$$

By definition $Y_t(\mu) \coloneqq \sum_{i=0}^{t-1} \mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i]$ for $t > 0$ and $Y_0 = 0$. The summands of $Y_t(\mu)$ satisfy:

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i] &= \mathbb{E}_\mu^{\pi^{\text{test}}} \left[\frac{1}{\delta} \int_{\mu-\delta}^{\mu} \left((\mu^* - \nu) T_{i+1} N_{i+1} + (N_i + A_i - N_{i+1}) \log \left(\frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}} \right) \right) d\nu \middle| N_i, A_i \right] \\
&= \mathbb{E}_\mu^{\pi^{\text{test}}} \left[(N_i + A_i) \left(\frac{1}{\delta} \int_{\mu-\delta}^{\mu} D^\lambda(\nu \| \mu^*) d\nu \right) \middle| N_i, A_i \right] \\
&= \mathbb{E}_\mu^{\pi^{\text{test}}} \left[(N_i + A_i) \left(\frac{1}{\delta} \int_{\mu-\delta}^{\mu} \left(D^\lambda(\mu \| \mu^*) - D^\lambda(\mu \| \nu) \right) d\nu \right) \middle| N_i, A_i \right] \\
&= (N_i + A_i) \left(D^\lambda(\mu \| \mu^*) - \frac{1}{\delta} \int_{\mu-\delta}^{\mu} D^\lambda(\mu \| \nu) d\nu \right)
\end{aligned}$$

By Lemma E.3,

$$\frac{1}{\delta} \int_{\mu-\delta}^{\mu} D^\lambda(\mu || \nu) d\nu \le \frac{1}{3} \frac{\lambda \delta^2}{(\mu - 2\delta)^3} + \frac{0.65}{3} \frac{\lambda \delta^2}{(\mu - \delta)^3} \xrightarrow[\delta \downarrow 0]{} 0.$$

Since $D^\lambda(\mu \| \mu^*) > 0$ whenever $\mu < \mu^*$, and the right-hand side above vanishes as $\delta \downarrow 0$, there exists $\delta^*(\mu) > 0$ such that for all $\delta \in (0, \delta^*(\mu))$,

$$\frac{1}{\delta} \int_{\mu-\delta}^{\mu} D^\lambda(\mu || \nu) d\nu < D^\lambda(\mu || \mu^*),$$

that is,

$$\eta(\delta, \mu) \ \coloneqq \ D^\lambda(\mu \| \mu^*) - \frac{1}{\delta} \int_{\mu-\delta}^{\mu} D^\lambda(\mu \| \nu) d\nu \ > \ 0. \tag{E.10}$$

Fix any $\delta \in (0, \delta^*(\mu))$ for the remainder of this lemma. Combining (E.10) and $N_i + A_i = \min(N_i + 1, k) \ge 1$ almost surely yields

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i] = (N_i + A_i) \, \eta(\delta, \mu) \ > \ 0,$$

Hence, the predictable part grows at least linearly in t:

$$Y_t(\mu) = \sum_{i=0}^{t-1} \mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_{i+1}(\mu) - \widetilde{W}_i(\mu) | \mathcal{F}_i] \ge t \, \eta(\delta, \mu).$$

This implies that $\{\widetilde{W}_t(\mu)\}_{t \ge 0}$ is an $\mathcal{F}_t$-submartingale under $\mu < \mu^*$ for every $\delta \in (0, \delta^*(\mu))$.

$\square$

Lemma E.5. *Let $\delta \in (0, \delta^*(\mu))$ be fixed as in Lemma E.4, let $\widetilde{\tau}(\mu)$ be the stopping time defined in* (E.2) *and let the $\{M_t(\mu), \mathcal{F}_t\}_{t \ge 0}$ be the martingale defined in Proposition E.2. Then $\{M_t(\mu)\}$ satisfies the hypotheses of Doob's Optional Stopping Theorem (recalled in Theorem B.3) at $\widetilde{\tau}(\mu)$; that is,*

1. ***(Bounded increments)*** *There exists a constant $c_0 < \infty$ such that*

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[|M_{t+1}(\mu) - M_t(\mu)| \mid \mathcal{F}_t] \le c_0$$

almost surely on $\{\widetilde{\tau}(\mu) > t\}$ for every $t \ge 0$.

2. ***(Finite expected stopping time)*** $\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] < \infty$.

Consequently, Doob's Optional Stopping Theorem applies to $\{M_t(\mu)\}$ at $\widetilde{\tau}(\mu)$ and

$$\mathbb{E}_\mu^{\pi^{\text{test}}} \left[M_{\widetilde{\tau}(\mu)}(\mu) \right] = \mathbb{E}_\mu^{\pi^{\text{test}}} \left[M_0(\mu) \right] = 0.$$

Proof. This proof will proceed in 3 steps. Step 0 provides a qualitative description. Step 1 and 2 verify the two hypothesis required for applying Doob's OST.

Step 0: Martingale difference and the strong law. Define $D_t := M_t(\mu) - M_{t-1}(\mu)$. By the definition of $M_t(\mu)$,

$$D_t = \widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu) - \mathbb{E}_\mu^{\pi^{\text{test}}}\big[\widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu) \mid \mathcal{F}_{t-1}\big],$$

and hence

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[D_t^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}_\mu^{\pi^{\text{test}}}\big[(\widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu))^2 \mid \mathcal{F}_{t-1}\big] - \big(\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu) \mid \mathcal{F}_{t-1}]\big)^2. \tag{E.11}$$

By the per-step log-likelihood ratio bound established in (E.5) we have,

$$\begin{aligned}\frac{1}{\delta}\int_{\mu-\delta}^{\mu} (\mu^* - \nu)\,\mathbb{E}_\mu^{\pi^{\text{test}}}[T_{i+1}N_{i+1} \mid N_i, A_i] + \mathbb{E}_\mu^{\pi^{\text{test}}}\left[(N_i + A_i - N_{i+1})\log\left(\frac{1 - e^{-\nu T_{i+1}}}{1 - e^{-\mu^* T_{i+1}}}\right) \,\middle|\, N_i, A_i\right] d\nu \\ \le (\mu^* - (\mu - \delta))\frac{k}{\lambda} + k\log\left(\frac{\mu^*}{\mu - \delta}\right).\end{aligned} \tag{E.12}$$

Combining (E.9), (E.11) and (E.12) we get,

$$\begin{aligned}\mathbb{E}_\mu^{\pi^{\text{test}}}[D_t^2 \mid \mathcal{F}_{t-1}] \le \frac{2k^2}{\lambda^2}(\mu^* - (\mu - \delta))^2 + k^2\log^2\left(\frac{\mu^*}{\mu - \delta}\right) + \frac{2k^2}{\lambda}(\mu^* - (\mu - \delta))\log\left(\frac{\mu^*}{\mu - \delta}\right) \\ - \left(D^\lambda(\mu\|\mu^*) - \frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^\lambda(\mu\|\nu)\,d\nu\right)^2 \;<\; \infty.\end{aligned} \tag{E.13}$$

Since $\sup_t \mathbb{E}_\mu^{\pi^{\text{test}}}[D_t^2 \mid \mathcal{F}_{t-1}] < \infty$, the martingale strong law of large numbers [Hall and Heyde, 1980, Theorem 2.15] applies, giving

$$M_t(\mu)/t \to 0 \quad \text{almost surely under } \mathbb{P}_\mu^{\pi^{\text{test}}}.$$

By Lemma E.4 the predictable part satisfies

$$Y_t(\mu) \ge \eta(\delta, \mu)\,t,$$

where $\eta(\delta, \mu) > 0$. Combined with the almost sure convergence $M_t/t \to 0$, it follows that that the linear drift $Y_t(\mu)$ eventually dominates the martingale fluctuation $M_t(\mu)$ almost surely. Consequently, the event $\{\widetilde{W}_t(\mu) < \log(\mu^*/\delta\alpha)\}$ requires $M_t(\mu)$ to fall a linear-in-t amount below its almost-sure limiting trajectory, since $Y_t(\mu) \ge \eta(\delta, \mu)t$. This is precisely a large-deviation event, so its probability must vanish as $t \to \infty$. Step 2 below makes this quantitative via a Bernstein-type concentration bound, yielding a rate of decay fast enough to guarantee summability of $\mathbb{P}_\mu^{\pi^{\text{test}}}(\widetilde{\tau}(\mu) > t)$ over t, and hence $\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] < \infty$.

Step 1: Bounded increments. On the event $\{\widetilde{\tau}(\mu) > t\}$, it follows from (E.9) and (E.12) that,

$$\begin{aligned}\mathbb{E}_\mu^{\pi^{\text{test}}}\big[|M_{t+1}(\mu) - M_t(\mu)| \mid \mathcal{F}_t\big] &= \mathbb{E}_\mu^{\pi^{\text{test}}}\left[\left|\widetilde{W}_{t+1}(\mu) - \widetilde{W}_t(\mu) - \mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{W}_{t+1}(\mu) - \widetilde{W}_t(\mu) \mid \mathcal{F}_t]\right| \,\middle|\, \mathcal{F}_t\right] \\ &\le 2\,\mathbb{E}_\mu^{\pi^{\text{test}}}\big[\widetilde{W}_{t+1}(\mu) - \widetilde{W}_t(\mu) \big| \mathcal{F}_t\big] \\ &\le (\mu^* - (\mu - \delta))\frac{2k}{\lambda} + 2k\log\left(\frac{\mu^*}{\mu - \delta}\right) =: c_0 < \infty.\end{aligned}$$

Thus, the conditional expectation of the absolute increments of $\{M_t(\mu)\}_{t\ge0}$ is almost surely bounded by c_0, uniformly in t.
Step 2: Finite expected stopping time. Since $\widetilde{\tau}(\mu)$ is a non-negative integer-valued random variable,

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] = \sum_{t=0}^{\infty} \mathbb{P}_\mu^{\pi^{\text{test}}}(\widetilde{\tau}(\mu) > t).$$

For any $t > 0$, by definition of $\widetilde{\tau}(\mu)$ in (E.2) it follows that,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}(\widetilde{\tau}(\mu) > t) = \mathbb{P}_\mu^{\pi^{\text{test}}}\Big(\bigcap_{i=0}^{t}\{\widetilde{W}_i(\mu) < \log(\mu^*/\delta\alpha)\}\Big) \le \mathbb{P}_\mu^{\pi^{\text{test}}}\big(\widetilde{W}_t(\mu) < \log(\mu^*/\delta\alpha)\big) = \mathbb{P}_\mu^{\pi^{\text{test}}}\big(M_t(\mu) + Y_t(\mu) < \log(\mu^*/\delta\alpha)\big).$$

By Lemma E.4, since $Y_t(\mu) \geq \eta(\delta, \mu)\, t$, we obtain

$$\mathbb{P}_\mu^{\pi^{\text{test}}}(\widetilde{\tau}(\mu) > t) \leq \mathbb{P}_\mu^{\pi^{\text{test}}}\big(M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\big).$$

Sub-exponential tails for the martingale differences. The following holds almost surely:

$$\begin{aligned}|D_t| &= \Big|\widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu) - \mathbb{E}_\mu^{\pi^{\text{test}}}\big[\widetilde{W}_t(\mu) - \widetilde{W}_{t-1}(\mu) \,\big|\, \mathcal{F}_{t-1}\big]\Big| \\ &\leq \underbrace{(\mu^* - (\mu - \delta))\, k\, T_t}_{\text{linear-in-}T_t\text{ term}} \;+\; \underbrace{2k \log\left(\frac{\mu^*}{\mu - \delta}\right) + (\mu^* - (\mu - \delta))\frac{k}{\lambda}}_{\text{constant term}} \;=:\; D_t^U.\end{aligned}$$

Define

$$a := (\mu^* - (\mu - \delta))\, k, \qquad C := 2k \log\left(\frac{\mu^*}{\mu - \delta}\right) + (\mu^* - (\mu - \delta))\frac{k}{\lambda},$$

we have $D_t^U = aT_t + C$. Since $T_t \sim \mathrm{Exp}(\lambda)$, it follows that

$$\mathbb{P}_\mu^{\pi^{\text{test}}}(|D_t| > x) \;\leq\; \mathbb{P}_\mu^{\pi^{\text{test}}}(D_t^U > x) \;=\; \mathbb{P}_\mu^{\pi^{\text{test}}}\left(T_t > \frac{x - C}{a}\right) \;=\; \exp\left(-\frac{\lambda(x - C)}{a}\right) \;=\; c_1 e^{-c_2 x},$$

where $c_1 := e^{\lambda C/a}$ and $c_2 := \lambda/a$.

By the standard equivalence between sub-exponential tails and sub-exponential MGF bounds [Wainwright, 2019, Theorem 2.6], the tail bound $\mathbb{P}_\mu^{\pi^{\text{test}}}(|D_t| > x) \leq c_1 e^{-c_2 x}$ implies that there exist (ν_t, α_t) depending only on (c_1, c_2), such that

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[e^{\theta D_t} \mid \mathcal{F}_{t-1}] \leq e^{\theta^2 \nu_t^2/2} \quad \text{a.s. for all } |\theta| < 1/\alpha_t.$$

Set $\nu_*^2 := \sum_{i=1}^t \nu_i^2$ and $\alpha_* := \max_{1 \leq i \leq t} \alpha_i$. Then $M_t(\mu) = \sum_{i=1}^t D_i$ is sub-exponential with parameters (ν_*, α_*). By the Bernstein-type tail bound (recalled in Theorem B.4), for all $s \geq 0$,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}(M_t(\mu) \leq -s) \leq \begin{cases} 2\exp\left(-\dfrac{s^2}{2\nu_*^2}\right), & 0 \leq s \leq \nu_*^2/\alpha_*, \\ 2\exp\left(-\dfrac{s}{2\alpha_*}\right), & s > \nu_*^2/\alpha_*. \end{cases}$$

Since $\log(\mu^*/\delta\alpha)$ is a fixed constant, there exists T_0 such that $\log(\mu^*/\delta\alpha) \leq \frac{\eta(\delta,\mu)}{2} t$ for all $t \geq T_0$. Hence for $t \geq T_0$,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}\big(M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\big) \leq \mathbb{P}_\mu^{\pi^{\text{test}}}\left(M_t(\mu) \leq -\frac{\eta(\delta, \mu)}{2} t\right).$$

Set $s = \eta(\delta, \mu)\, t/2$ and $C' := \max_{1 \leq i \leq t} \nu_i^2$, so that $\nu_*^2 \leq C' t$. In the quadratic regime,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}\left(M_t(\mu) \leq -\frac{\eta(\delta, \mu)}{2} t\right) \leq 2\exp\left(-\frac{(\eta(\delta, \mu)^2/4)t^2}{2C' t}\right) = 2\exp\left(-\frac{\eta(\delta, \mu)^2}{8C'} t\right),$$

and in the linear regime,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}\left(M_t(\mu) \leq -\frac{\eta(\delta, \mu)}{2} t\right) \leq 2\exp\left(-\frac{\eta(\delta, \mu)}{4\alpha_*} t\right).$$

In either case, taking $c' := \min\left\{\dfrac{\eta(\delta, \mu)^2}{8C'}, \dfrac{\eta(\delta, \mu)}{4\alpha_*}\right\}$, we obtain, for all sufficiently large t,

$$\mathbb{P}_\mu^{\pi^{\text{test}}}\big(M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\big) \leq 2e^{-c' t}.$$

This geometric-type decay is summable: for $t \geq T_0$,

$$\sum_{t=T_0}^{\infty} \mathbb{P}_\mu^{\pi^{\text{test}}}\big(M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\big) \leq \sum_{t=T_0}^{\infty} 2e^{-c' t} = \frac{2e^{-c' T_0}}{1 - e^{-c'}} < \infty,$$

and the finitely many remaining terms $t < T_0$ contribute a finite sum trivially. Recalling that $\{\widetilde{\tau}(\mu) > t\} \subset \{M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\}$, it follows that

$$\sum_{t=0}^{\infty} \mathbb{P}_{\mu}^{\pi^{\text{test}}}(\widetilde{\tau}(\mu) > t) \leq \sum_{t=0}^{\infty} \mathbb{P}_{\mu}^{\pi^{\text{test}}}\big(M_t(\mu) < \log(\mu^*/\delta\alpha) - \eta(\delta, \mu)\, t\big) < \infty,$$

i.e., $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] < \infty$.

Combining Steps 1 and 2, we conclude that $\{M_t(\mu)\}$ satisfies the hypotheses of Doob's Optional Stopping Theorem at $\widetilde{\tau}(\mu)$. Thus,

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[M_{\widetilde{\tau}(\mu)}(\mu)\big] = \mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[M_0(\mu)\big] = 0.$$

□

We have now established all the intermediate results required for the proof of Lemma 3.7. We now complete the proof in the remainder of this section.

E.2 PROOF OF LEMMA 3.7

Fix any $\delta \in (0, \delta^*(\mu))$, so that the conclusions of Lemmas E.1–E.5 all hold; we optimize the choice of δ at the end of the proof. Lemma E.1 implies that $\tau \leq \widetilde{\tau}(\mu)$. Therefore, it suffices to establish an upper bound on $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$.

By Lemma E.5, Doob's Optional Stopping Theorem gives

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[M_{\widetilde{\tau}(\mu)}(\mu)\big] = \mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[M_0(\mu)\big] = 0.$$

Recall the decomposition $\widetilde{W}_t(\mu) = M_t(\mu) + Y_t(\mu)$. Applying optional stopping to the sum defining $Y_{\widetilde{\tau}(\mu)}(\mu)$, together with the per-step drift

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[\widetilde{W}_{t+1}(\mu) - \widetilde{W}_t(\mu) \,\big|\, \mathcal{F}_t\big] = (N_t + A_t)\, \eta(\delta, \mu)$$

established in Lemma E.4, we obtain

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu)\big] = \mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[\sum_{t=0}^{\widetilde{\tau}(\mu)-1} \mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[\widetilde{W}_{t+1}(\mu) - \widetilde{W}_t(\mu) \,\big|\, \mathcal{F}_t\big]\right] = \mathbb{E}_{\mu}^{\pi^{\text{test}}}\left[\sum_{t=0}^{\widetilde{\tau}(\mu)-1} (N_t + A_t)\right] \eta(\delta, \mu). \tag{E.14}$$

Since $N_t + A_t = \min(N_t + 1, k)$, the stationary average is

$$\mathbb{E}_w[N_t + A_t] = (1 + \rho)\big(1 - B(k, \rho)\big), \qquad \rho := \lambda/\mu,$$

where w is the stationary distribution of the state occupancy and $B(k, \rho)$ is the Erlang–B blocking probability, as defined in Definition 3.5. Combining this with mixing-time bound from Lemma 3.6 and (E.14) gives

$$(1 + \rho)\big(1 - B(k, \rho)\big)\, \eta(\delta, \mu)\, \mathbb{E}[\widetilde{\tau}(\mu)] \;\leq\; \mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu)\big] + \frac{Gk\, \eta(\delta, \mu)}{1 - \phi}, \tag{E.15}$$

Define $\gamma(\delta) := (1 + \rho)\big(1 - B(k, \rho)\big)\, \eta(\delta, \mu)$.

Since $\widetilde{\tau}(\mu)$ is the first time the process $\{\widetilde{W}_t(\mu)\}_{t \geq 0}$ crosses the level $\log(\mu^*/\delta\alpha)$, let $B_{\widetilde{\tau}(\mu)}$ denote the overshoot of $\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu)$ above this level, so that

$$B_{\widetilde{\tau}(\mu)} \leq \big(\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu) - \widetilde{W}_{\widetilde{\tau}(\mu)-1}(\mu)\big)_+,$$

and hence

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu)\big] \leq \log\left(\frac{\mu^*}{\delta\alpha}\right) + \mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[B_{\widetilde{\tau}(\mu)}\big].$$

It remains to bound $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[B_{\widetilde{\tau}(\mu)}]$. Since the martingale difference sequence is square-integrable, Jensen's inequality together with the tower property give

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[(\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu) - \widetilde{W}_{\widetilde{\tau}(\mu)-1}(\mu))_+\big] \leq \sqrt{\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[(\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu) - \widetilde{W}_{\widetilde{\tau}(\mu)-1}(\mu))^2\big]}$$

$$= \sqrt{\mathbb{E}_{\mu}^{\pi^{\text{test}}}\Big[\mathbb{E}_{\mu}^{\pi^{\text{test}}}\big[(\widetilde{W}_{\widetilde{\tau}(\mu)}(\mu) - \widetilde{W}_{\widetilde{\tau}(\mu)-1}(\mu))^2 \mid \mathcal{F}_{\widetilde{\tau}(\mu)-1}\big]\Big]},$$

and applying the bound from (E.13) to the inner conditional expectation shows that this quantity is at most $K(\delta) < \infty$, where

$$K(\delta)^2 = \frac{2k^2}{\lambda^2}(\mu^* - (\mu - \delta))^2 + k^2 \log^2\left(\frac{\mu^*}{\mu - \delta}\right) + \frac{2k^2}{\lambda}(\mu^* - (\mu - \delta)) \log\left(\frac{\mu^*}{\mu - \delta}\right).$$

Substituting into (E.15), we obtain

$$\gamma(\delta)\, \mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] \le \log\left(\frac{\mu^*}{\delta\alpha}\right) + K(\delta) + \frac{Gk\,\eta(\delta,\mu)}{1-\phi}. \tag{E.16}$$

Choice of δ. The bound (E.16) holds for every fixed $\delta \in (0, \delta^*(\mu))$; we now choose δ as a function of μ and α. As $\delta \downarrow 0$, $\eta(\delta,\mu) \to D^{\lambda}(\mu\|\mu^*)$ and correspondingly $\gamma(\delta) \to \gamma_* := (1+\rho)\big(1 - B(k,\rho)\big)\, D^{\lambda}(\mu\|\mu^*)$; write $\eta(\delta,\mu) = D^{\lambda}(\mu\|\mu^*) - \epsilon(\delta)$ with

$$\epsilon(\delta) := \frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^{\lambda}(\mu\|\nu)\, d\nu = O(\delta^2)$$

by Lemma E.3. We choose δ small enough, that (i) $\eta(\delta,\mu) > 0$ (i.e. $\delta < \delta^*(\mu)$, so that Lemmas E.4 and E.5 apply) and, (ii) the term $\log(1/\delta)$ in $\log(\mu^*/(\delta\alpha)) = \log(1/\alpha) + \log(\mu^*/\delta)$ is asymptotically negligible relative to the leading term $\log(1/\alpha)$. Concretely, take

$$\delta = \delta(\alpha) := \min\left(\frac{\delta^*(\mu)}{2},\ \frac{\mu}{2},\ \frac{1}{\log(1/\alpha)}\right).$$

With this choice,

$$\log\left(\frac{\mu^*}{\delta(\alpha)\,\alpha}\right) = \log(1/\alpha)\big(1 + o(1)\big).$$

This holds as $K(\delta(\alpha))$ and $\frac{Gk\,\eta(\delta(\alpha),\mu)}{1-\phi}$ remain bounded uniformly in α—since $\delta(\alpha) \le \delta^*(\mu)/2$, and both $K(\delta)$ and $\eta(\delta,\mu)$ are continuous and bounded on $(0, \delta^*(\mu)/2]$—and $\gamma(\delta(\alpha)) \to \gamma_*$ as $\delta(\alpha) \to 0$. Substituting into (E.16) and dividing through by $\gamma(\delta(\alpha))$,

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)] \le \frac{\log(1/\alpha)\big(1+o(1)\big) + O(1)}{\gamma(\delta(\alpha))} = \frac{\log(1/\alpha)}{\gamma_*}\big(1+o(1)\big).$$

Combined with $\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\tau] \le \mathbb{E}_{\mu}^{\pi^{\text{test}}}[\widetilde{\tau}(\mu)]$ from Lemma E.1, we conclude that

$$\mathbb{E}_{\mu}^{\pi^{\text{test}}}[\tau] \le \frac{\log(1/\alpha)}{\gamma_*}\big(1+o(1)\big).$$

This completes the proof. □

F ACHIEVABILITY RESULTS IN REJECT-OPTIMAL REGIME

F.1 PROOF OF LEMMA 3.6

Proof. For each arrival $i \ge 0$, let $\nu_i(\cdot) := \mathbb{P}(N_i \in \cdot)$ denote the distribution of the Erlang–B Markov chain at the i^{th} arrival, where the initial state occupancy N_0 follows an arbitrary distribution ν_0. By Theorem B.5, $\max_x \|P^i(x,\cdot) - w\|_{\text{TV}} \le G\phi^i$, and since total variation distance is convex in its first argument,

$$\|\nu_i - w\|_{\text{TV}} = \|\nu_0 P^i - w\|_{\text{TV}} \le \sum_x \nu_0(x)\, \|P^i(x,\cdot) - w\|_{\text{TV}} \le G\phi^i.$$

In particular, the resulting bound is uniform in the initial distribution, so we need not specify ν_0. Then for any bounded f with $\|f\|_\infty \le m$, by the definition of total variation distance,

$$|\mathbb{E}_{\nu_i}[f] - \mathbb{E}_w[f]| \le m\, \|\nu_i - w\|_{\text{TV}} \le mG\phi^i.$$

Thus,

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{t=0}^{T-1} f(N_t)\right] - \mathbb{E}[T]\mathbb{E}_w[f]\right| &= \left|\sum_{t=1}^{\infty} \mathbb{P}(T=t) \sum_{i=0}^{t-1} (\mathbb{E}_{\nu_i}[f] - \mathbb{E}_w[f])\right| \\ &\leq \left|\sum_{t=1}^{\infty} \mathbb{P}(T=t) \sum_{i=0}^{t-1} mG\phi^i\right| \\ &\leq \mathbb{E}\left[\sum_{i=0}^{T-1} mG\phi^i\right] \leq mG\sum_{i=0}^{\infty}\phi^i \leq \frac{mG}{1-\phi}.\end{aligned}$$

This completes the proof. □

G PROOFS OF RESULTS USED IN SECTION 4

G.1 PROOF OF LEMMA 4.3

Proof. Let T_i denote the first hitting time to state 0 when system starts in state $i \in \{0, 1, \ldots, k\}$ at time 0. Clearly, $T_0 = 0$ almost surely. A one-step analysis yields:

$$\begin{aligned}\mathbb{E}[T_i] &= \frac{1}{\lambda + i\mu} + \frac{\lambda}{\lambda + i\mu}\mathbb{E}[T_{i+1}] + \frac{i\mu}{\lambda + i\mu}\mathbb{E}[T_{i-1}], \qquad 1 \leq i < k, \\ \mathbb{E}[T_k] &= \frac{1}{\lambda + k\mu} + \frac{k\mu}{\lambda + k\mu}\mathbb{E}[T_{k-1}] + \frac{\lambda}{\lambda + k\mu}\mathbb{E}[T_k].\end{aligned}$$

Rearranging the first equation for $1 \leq i < k$ gives

$$\lambda\mathbb{E}[T_{i+1} - T_i] = i\mu\mathbb{E}[T_i - T_{i-1}] - 1.$$

Define $d_i = \mathbb{E}[T_i] - \mathbb{E}[T_{i-1}]$. Then $\{d_i\}_{i=1}^k$ satisfies the recursion

$$\begin{aligned}d_{i+1} &= \frac{i\mu}{\lambda}d_i - \frac{1}{\lambda}, \quad \text{if } 1 \leq i < k, \\ d_k &= \frac{1}{k\mu}.\end{aligned}$$

where the boundary condition follows from the equation for $\mathbb{E}[T_k]$.

Since d_k is know, the recursion can be solved backwards to obtain the unique minimal non-negative solution

$$d_i = \sum_{j=i}^{k} \frac{1}{\mu}\left(\frac{\lambda}{\mu}\right)^{j-i} \frac{(i-1)!}{j!}, \qquad i = 1, 2, \ldots, k.$$

Finally,

$$\mathbb{E}[T_i] = \sum_{l=1}^{i} d_l = \sum_{l=1}^{i}\sum_{j=l}^{k} \frac{1}{\mu}\left(\frac{\lambda}{\mu}\right)^{j-l} \frac{(l-1)!}{j!}, \qquad i = 1, \ldots, k.$$

Hence

$$T_0^h := \max_{i \in \{1,2,\ldots,k\}} \mathbb{E}[T_i] < \infty.$$

since $k < \infty$ and each expectation is finite.

□

Algorithm 3 Fixed-Horizon Implementable Policy π^{alg}

Input: number of servers k, admission threshold μ^*, horizon n, covering width δ, grid size M, grid start point ϵ.
Initialize:
$N_0 \leftarrow 0$ ▷ number of occupied servers.
$A_0 \leftarrow 1$ ▷ admission decision
$\tau \leftarrow n$ ▷ stopping time; default to "never reject before horizon ends".
$\zeta_0(\nu) \leftarrow 0, \ \forall \nu \in [\epsilon, \mu^*]$ ▷ cumulative log-likelihood-ratio process, initialized to zero.
Define $C := e \log(\mu^*/\epsilon)$ ▷ constant controlling the test's significance level.
Construct a deterministic covering grid of $[\varepsilon, \mu^*]$:

$$G_i := \epsilon + (i-1)\frac{\delta}{2}, \qquad i = 1, \ldots, \kappa+1, \text{ where } \kappa := 2\left\lceil \frac{\mu^* - \epsilon}{\delta} \right\rceil.$$

Partition $[\epsilon, \mu^*]$ into κ consecutive bins of width $\delta/2$ using the grid points G_i:

$$I_i := [G_i, G_{i+1}), \quad i = 1, \ldots, \kappa - 1, \text{ and } I_\kappa := [G_\kappa, \mu^*].$$

for each arrival $t \in \{1, \ldots n-1\}$ while $\{t < \tau\}$ **do**
  repeat
    Draw $\nu_1^{(t)}, \ldots, \nu_M^{(t)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\epsilon, \mu^*)$ ▷ fresh i.i.d. draw at every arrival t.
  until every bin I_i, $i \in [\kappa]$, contains at least one $\nu_m^{(t)}$ ▷ rejection sampling: re-draw until coverage holds.
  Observe N_t, T_t and store them.
  For all $\nu_m^{(t)}$ accumulate log-likelihood ratio over all past observations

$$\zeta_t(\nu_m^{(t)}) = \sum_{s=1}^{t} l_{s-1}(\nu_m^{(t)}) - l_{s-1}(\mu^*)$$

  where, the per-step log-likelihood contribution is

$$l_{s-1}(\nu_m^{(t)}) = -\nu_m^{(t)} T_s N_s + (N_{s-1} + A_{s-1} - N_s)\log(1 - e^{-\nu_m^{(t)} T_s}).$$

  Compute the mixture statistic

$$W_t^M := \frac{1}{M} \sum_{m=1}^{M} \frac{L_t(\nu_m^{(t)})}{L_t(\mu^*)} = \frac{1}{M} \sum_{m=1}^{M} \exp(\zeta_t(\nu_m^{(t)})).$$

  if $W_t^M \geq Cn^2$ **then** ▷ reject H_0: the service rate is detected to be below μ^*.
    $\tau \leftarrow t$
    $A_i \leftarrow 0, \ \forall i \in \{t, t+1, \ldots, n-1\}$ ▷ stop admitting for the rest of the horizon.
  else ▷ fail to reject: continue with the greedy admission rule
    if $N_t < k$ **then**
      $A_t \leftarrow 1$ ▷ server available: admit.
    else
      $A_t \leftarrow 0$ ▷ system full: block
    end if
    Store A_t.
  end if
end for

H COMPUTATIONAL TRACTABILITY

Setup: This section follows the same notation introduced in Section 1.1. The continuous mixture statistic defined in (8) is not directly implementable, since it requires evaluating an integral over continuum of service rates in $(0, \mu^*)$. We therefore approximate the integral using a randomized grid (Algorithm 3). At each arrival epoch, a fresh collection of M points

$$\nu_1^{(t)}, \ldots, \nu_M^{(t)}$$

is generated by rejection sampling from $\mathrm{Unif}(\epsilon, \mu^*)$ for some $\epsilon > 0$ until every interval of a deterministic covering grid of width $\delta/2$ contains at least one sampled point. The resulting empirical mixture likelihood ratio is

$$W_t^M := \frac{1}{M} \sum_{m=1}^{M} \frac{L_t(\nu_m^{(t)})}{L_t(\mu^*)}. \tag{H.1}$$

and the associated stopping time is

$$\tau^M := \inf \left\{ t \in \{1, \ldots, n\} : W_t^M \geq b_n \right\}.$$

Here, $b_n := \frac{e^{\eta(n+k)} K}{\alpha}$ and $K = \left\lceil \frac{\log(\mu^*/\epsilon)}{\log(1+\eta)} \right\rceil$. We will choose η and α so that $\frac{e^{\eta(n+k)} K}{\alpha} < Cn^2$ for a constant C independent of n. Therefore, the implementable stopping rule used by the Algorithm 3 is equivalently

$$\tau^M := \inf \left\{ t \in \{1, \ldots, n\} : W_t^M \geq Cn^2 \right\}, \tag{H.2}$$

with $\tau = n$ if the threshold is never crossed, where $C = e \log(\mu^*/\epsilon)$. In the remainder of this section we establish the performance guarantees of policy π^{alg} described in Algorithm 3.

Definition H.1 (Fixed-Horizon Implementable Policy π^{alg})**.** *Consider an Erlang–B system with k identical servers with service rate $\mu > 0$. Let n denote the maximum number of jobs that may arrive, and let τ^M denote the stopping time defined in* (H.2)*, where the empirical mixture likelihood ratio is constructed using the randomized grid procedure described in Algorithm 3.*

Define the policy π^{alg} that takes the following action at the t^{th} arrival:

$$A_t^{\pi^{\mathrm{alg}}} = \mathbf{1}\{t < \tau^M,\, N_t < k\} = \begin{cases} 1, & \text{if } t < \tau^M \text{ and } N_t < k, \\ 0, & \text{otherwise.} \end{cases}$$

That is, the policy admits every arriving job while the randomized-grid sequential test has not rejected the null hypothesis, provided a server is available. Once the stopping time τ^M is reached, the policy permanently rejects all subsequent arrivals for the remainder of the horizon.

H.1 OPTIMAL DECISION IS TO ACCEPT

The level-α guarantee follows immediately when the randomized grid is replaced by a fixed grid across all arrival epochs. Specifically, for every $\nu < \mu^*$ the likelihood ratio process $L_t(\nu)/L_t(\mu^*)$ is a non-negative supermartingale under the null hypothesis $\mu > \mu^*$ (Lemma 3.3). As the randomized grid in Algorithm 3 changes from one arrival epoch to the next, the statistic W_t^M is not itself a supermartingale. Nevertheless, we will show that this process is almost surely dominated by a supermartingale. Applying Ville's inequality to the dominating process then establishes the desired level-α guarantee.

The proof proceeds as follows. First, we compare each randomly sampled point with a nearby point on a fixed deterministic geometric grid (Lemma H.2) Second, we show that this comparison implies a pathwise upper bound on the randomized statistic W_t^M in terms of the maximum likelihood ratio over the deterministic grid. Third, we apply Ville's inequality to each deterministic-grid likelihood ratio and take a union bound over the grid.

Lemma H.2 (Deterministic Multiplicative Grid)**.** *Let ϵ denote the grid start point and fix $\eta > 0$. Construct the geometric grid $\mathcal{G} = \{g_1, \ldots, g_K\}$, where $g_i = \epsilon(1+\eta)^{i-1}$ and $K = \left\lceil \frac{\log(\mu^*/\epsilon)}{\log(1+\eta)} \right\rceil$. For each $\nu \in [\epsilon, \mu^*]$ define $g(\nu) := \max_{1 \leq i \leq K}\{g_i : g_i \leq \nu\}$, that is, $g(\nu)$ is the largest grid point not exceeding ν. Then, for all $t \leq n$,*

$$\frac{L_t(\nu)}{L_t(\mu^*)} \leq e^{\eta(n+k)} \frac{L_t(g(\nu))}{L_t(\mu^*)}$$

Proof. Define $f_t(\nu) := \log \frac{L_t(\nu)}{L_t(\mu^*)}$. Since f_t is concave in ν, for $g(\nu) \le \nu$,

$$f_t(\nu) \;\le\; f_t(g(\nu)) + (\nu - g(\nu))\; f_t'(g(\nu)))$$

Now, the derivative of f_t at $g(\nu) > 0$ is

$$\begin{aligned} f_t'(g(\nu)) &= \sum_{j=0}^{t-1} \left[-N_{j+1}T_{j+1} + (N_j + A_j - N_{j+1}) \frac{T_{j+1}e^{-g(\nu)T_{j+1}}}{1 - e^{-g(\nu)T_{j+1}}} \right] \\ &\overset{(i)}{\le} \sum_{j=0}^{t-1} \frac{N_j + A_j - N_{j+1}}{g(\nu)} \overset{(ii)}{\le} \frac{1}{g(\nu)} \Big(\sum_{j=0}^{t-1} A_j + N_0 - N_t \Big) \overset{(iii)}{\le} \frac{t+k}{g(\nu)} \overset{(iv)}{\le} \frac{n+k}{g(\nu)} \end{aligned} \tag{H.3}$$

where (i) holds by dropping the first (non-positive) term, and by $\frac{xe^{-x}}{1-e^{-x}} = \frac{x}{e^x - 1} \le 1$ for $x > 0$ termwise with $x = g(\nu)T_{j+1}$, (ii) follows from telescoping sum, (iii) and (iv) follows from bounds on N_i, A_i and t respectively.

By definition of $g(\nu)$,

$$g(\nu) \le \nu \le (1+\eta) g(\nu) \implies \frac{\nu - g(\nu)}{g(\nu)} \le \eta. \tag{H.4}$$

Combining (H.3) and (H.4) gives

$$\frac{L_t(\nu)}{L_t(\mu^*)} \le e^{\eta(n+k)} \frac{L_t(g(\nu))}{L_t(\mu^*)}$$

□

Theorem H.3. *For $\mu \ge \mu^*$, the policy π^{alg} introduced in Algorithm 3 satisfies*

$$\mathbb{E}_\mu^{\text{alg}}[\text{Reg}(n, \pi^{\text{alg}}; \mu)] \le \Delta(\mu).$$

Proof. Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ be the deterministic grid constructed in Lemma H.2. By Lemma H.2,

$$W_t^M = \frac{1}{M} \sum_{m=1}^{M} \frac{L_t(\nu_m^{(t)})}{L_t(\mu^*)} \le e^{\eta(n+k)} \max_{g \in G} \frac{L_t(g)}{L_t(\mu^*)}.$$

Therefore,

$$\{\tau^M < n\} \subseteq \bigcup_{g \in \mathcal{G}} \left\{ \sup_{t<n} \frac{L_t(g)}{L_t(\mu^*)} \ge \frac{K}{\alpha} \right\}.$$

For each fixed $g \in \mathcal{G}$, the process $L_t(g)/L_t(\mu^*)$ is a nonnegative supermartingale under service rate $\mu > \mu^*$ (Lemma 3.3). Ville's inequality gives

$$\mathbb{P}_\mu^{\text{alg}} \left(\sup_{t<n} \frac{L_t(g)}{L_t(\mu^*)} \ge \frac{K}{\alpha} \right) \le \frac{\alpha}{K}.$$

A union bound over the K grid points yields

$$\mathbb{P}_\mu^{\text{alg}}(\tau^M < n) \le \alpha.$$

Consequently set $\alpha = \frac{1}{n}$ and thus

$$\begin{aligned} \mathbb{E}_\mu^{\text{alg}}[\text{Reg}(n, \pi^{\text{alg}}; \mu)] &\le \Delta(\mu)\, n\, \mathbb{P}_\mu^{\pi^{\text{alg}}}(\tau^M < n) \\ &\le \Delta(\mu)\, n\, \frac{1}{n} = \Delta(\mu). \end{aligned}$$

□

We choose $\eta = 1/n$. With this choice, $e^{\eta(n+k)}$ is uniformly bounded in n, while the geometric covering number $K = O(n)$. Taking $\alpha = 1/n$, the threshold

$$\frac{e^{\eta(n+k)} K}{\alpha} \le Cn^2$$

for a n-independent constant $C := e \log(\mu^*/\epsilon)$. Hence the stopping rule in Algorithm 3 is implemented using the simpler threshold Cn^2.

H.2 OPTIMAL DECISION IS TO REJECT

We now prove the expected stopping-time guarantee for the discretized statistic. The argument follows from the continuous-mixture proof, with two modifications. First, the statistic is now an empirical average over randomly sampled grid points, so we restrict attention to sampled points that fall in a small neighborhood $(\mu-\delta,\mu)$ of the service rate μ. Rejection sampling ensures existence of a grid point in each bin of width $\delta/2$, so the number of points falling in any δ length interval in $[\epsilon,\mu^*]$ is non-zero. Jensen's inequality (similar to Lemma E.1) then relates this local log-average to actual likelihood-ratio statistic. Second, the test threshold is no longer $1/\alpha$; due to type-I error correction for the deterministic grid comparison, the relevant threshold is

$$b_n := \frac{e^{\eta(n+k)}K}{\alpha}.$$

Consequently, compared to Lemma 3.7, the stopping-time bound replaces $\log(1/\alpha)$ with $\log(b_n)$ along with an additional $\log(M)$ due to discretization.

Lemma H.4. *Suppose $\mu<\mu^*$ and $\delta\in(0,\mu/2)$. Let τ^M be the stopping time introduced in Definition H.2 with respect to the natural filtration $\{\sigma(\mathcal{H}_t,A_t)\}_{t\geq 0}$. Then the expected stopping time under policy π^{alg} (Algorithm 3) satisfies*

$$\mathbb{E}_\mu^{\mathrm{alg}}[\tau^M] \leq \frac{\log b_n + \log M}{\eta(\delta,\mu)(1+\lambda/\mu)(1-B(k,\lambda/\mu))}(1+o(1)),$$

where

$$\eta(\delta,\mu) :=\ D^\lambda(\mu\|\mu^*) - \frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^\lambda(\mu\|\nu)d\nu$$

and $o(1)$ denotes a term that vanishes as $n\to\infty$.

Proof. The proof follows from same argument as Lemma 3.7; we highlight only the steps that differ due to the Monte Carlo approximation. At each arrival epoch t, π^{alg} draws $\nu_1^{(t)},\ldots,\nu_M^{(t)}$ independently from $\mathrm{Unif}(\epsilon,\mu^*)$, independently across arrivals and independently of the system history. Let

$$I_\delta(\mu) := (\mu-\delta,\mu), \qquad M_\delta^{(t)} := \#\{m : \nu_m^{(t)} \in I_\delta(\mu)\}.$$

Condition on the event $M_\delta^{(t)}\geq 1$ at every arrival. Equivalently, under the rejection-sampling implementation, $M_\delta^{(t)}$ has the zero-truncated $\mathrm{Bin}(M,p)$ distribution, where $p:=\delta/\mu^*$. Define the discretized mixture statistic and auxiliary stopping time, in direct analogy with (E.1) and (E.2):

$$\widetilde{W}_0^M(\mu) = 0, \quad \text{and} \quad \forall t>0,\ \widetilde{W}_t^M(\mu) := \frac{1}{M_\delta^{(t)}}\sum_{i=1}^{M_\delta^{(t)}} \log\frac{L_t(\nu_i^{(t)})}{L_t(\mu^*)}, \qquad , \tag{H.5}$$

$$\widetilde{\tau}^M(\mu) := \inf\Big\{t\geq 1 : \widetilde{W}_t^M(\mu) \geq \log\Big(\frac{M\ b_n}{M_\delta^{(t)}}\Big)\Big\}. \tag{H.6}$$

The same logic as in Lemma E.1 (restricting average to points in $I_\delta(\mu)$ and applying Jensen's inequality) gives

$$W_t^M \geq \frac{M_\delta^{(t)}}{M}\exp\{\widetilde{W}_t^M(\mu)\}.$$

Therefore, whenever $\widetilde{W}_t^M(\mu)\geq \log\left(\frac{b_n M}{M_\delta^{(t)}}\right)$, we have $W_t^M\geq b_n$ almost surely and hence

$$\tau^M \leq \widetilde{\tau}^M(\mu) \quad \text{a.s..}$$

It remains to upper bound $\mathbb{E}_\mu^{\mathrm{alg}}[\widetilde{\tau}^M(\mu)]$. Conditional on $\mathcal{F}_t$, the sampled points in $I_\delta(\mu)$ are i.i.d $\mathrm{Unif}(\mu-\delta,\mu)$, so $\widetilde{W}_t^M(\mu)$ is an unbiased estimator of $\widetilde{W}_t(\mu)$:

$$\mathbb{E}_\mu^{\mathrm{alg}}\left[\widetilde{W}_t^M(\mu)\mid \mathcal{F}_t,\right] = \widetilde{W}_t(\mu),$$

The same Doob-decomposition argument as in the continuous-mixture proof (Lemma E.5) gives

$$\widetilde{W}_t^M(\mu) = M_t^M(\mu) + Y_t^M(\mu),$$

where $M_t^M(\mu)$ is a martingale and

$$\mathbb{E}_\mu^{\text{alg}}[\widetilde{W}_{t+1}^M(\mu) - \widetilde{W}_t^M(\mu) \mid \mathcal{F}_t] = (N_t + A_t)\eta(\delta, \mu),$$

with

$$\eta(\delta, \mu) = D^\lambda(\mu \| \mu^*) - \frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^\lambda(\mu \| \nu)\, d\nu.$$

By Doob's Optional Stopping Theorem applied to the martingale part,

$$\mathbb{E}_\mu^{\text{alg}}[\widetilde{W}_{\widetilde{\tau}^M(\mu)}^M(\mu)] = \eta(\delta, \mu)\mathbb{E}_\mu^\pi\left[\sum_{t=0}^{\widetilde{\tau}^M(\mu)-1}(N_t + A_t)\right].$$

Using the same mixing argument as before,

$$\gamma(\delta)\mathbb{E}_\mu^{\text{alg}}[\widetilde{\tau}^M(\mu)] \le \mathbb{E}_\mu^{\text{alg}}[\widetilde{W}_{\widetilde{\tau}^M}(\mu)] + \frac{Gk\eta(\delta, \mu)}{1-\phi}$$

where $\gamma(\delta) = (1 + \lambda/\mu)(1 - B(k, \lambda/\mu))\eta(\delta, \mu)$.

The remainder of the proof differs from the continuous case only through the stopping threshold. Since $\widetilde{\tau}^M(\mu)$ is the first time $\widetilde{W}_t^M(\mu) \ge \log\left(\frac{b_n M}{M_\delta^{(t)}}\right)$, we have

$$\mathbb{E}_\mu^{\text{alg}}\left[\widetilde{W}_{\widetilde{\tau}^M(\mu)}^M(\mu)\right] \le \mathbb{E}_\mu^{\text{alg}}\left[\log\left(\frac{b_n M}{M_\delta^{(\widetilde{\tau}^M(\mu))}}\right)\right] + \mathbb{E}_\mu^{\text{alg}}\left[B_{\widetilde{\tau}^M(\mu)}\right], \tag{H.7}$$

where $B_{\widetilde{\tau}^M(\mu)}$ denotes the overshoot from the threshold.

To bound the first term in (H.7), let

$$q := \Pr(M_\delta = 2 \mid M_\delta \ge 1).$$

By the orderliness expansion of the zero-truncated Binomial distribution,

$$q = \frac{M-1}{2}\frac{\delta}{\mu^*} + O(\delta^2), \qquad \Pr(M_\delta \ge 3 \mid M_\delta \ge 1) = O(\delta^2).$$

Hence, up to an event of probability $O(\delta^2)$.

$$M_\delta = \begin{cases} 1, & \text{with probability } 1 - q + O(\delta^2), \\ 2, & \text{with probability } q + O(\delta^2), \end{cases}$$

Therefore using $q = \frac{M-1}{2}\frac{\delta}{\mu^*} + O(\delta^2)$ we have,

$$\begin{aligned}\mathbb{E}\left[\log\left(\frac{b_n M}{M_\delta}\right)\right] &= (1-q)\log(b_n M) + q\log\left(\frac{b_n M}{2}\right) + O(\delta^2) \\ &= \log(b_n M) - q\log 2 + O(\delta^2) \\ &= \log b_n + \log M + O(\delta),\end{aligned}$$

The overshoot is bounded exactly as in Lemma 3.7 by a finite constant $K(\delta)$. Consequently,

$$\gamma(\delta)\,\mathbb{E}_\mu^{\text{alg}}[\widetilde{\tau}^M(\mu)] \le \log b_n + \log M + K(\delta) + \frac{Gk\,\eta(\delta, \mu)}{1-\phi} + O(\delta).$$

Since $b_n \le Cn^2$ for an n-independent constant C, $\log b_n \to \infty$ as $n \to \infty$. Since δ is fixed, the terms $K(\delta)$, $Gk\,\eta(\delta,\mu)/(1-\phi)$, and $O(\delta)$ are all $O(1)$ with respect to n and are absorbed into the $(1+o(1))$ factor below. Thus

$$\mathbb{E}_\mu^{\text{alg}}[\tau^M] \le \mathbb{E}_\mu^{\text{alg}}[\widetilde{\tau}^M(\mu)] \le \frac{\log b_n + \log M}{\eta(\delta,\mu)(1+\lambda/\mu)(1-B(k,\lambda/\mu))}(1+o(1)).$$

where the $K(\delta)$, $Gk\eta(\delta,\mu)/(1-\phi)$, and $O(\delta)$ terms are all absorbed into the $(1+o(1))$ factor. We retain the explicit dependence on M rather than absorbing it into the $(1+o(1))$ in order to highlight the effect of discretization on the stopping time. Since M is independent of n it may equally be absorbed into the $(1+o(1))$ factor. □

Theorem H.5. *For $\mu < \mu^*$, the policy π^{alg} introduced in Algorithm 3 satisfies*

$$\mathbb{E}_\mu^{\pi^{\text{test}}}[\text{Reg}(n,\pi^{\text{alg}};\mu)] \le \frac{\Delta(\mu)(2\log n + \log M)}{\eta(\delta,\mu)(1+\lambda/\mu)}(1+o(1)),$$

where

$$\eta(\delta,\mu) = D^\lambda(\mu\|\mu^*) - \frac{1}{\delta}\int_{\mu-\delta}^{\mu} D^\lambda(\mu\|\nu)\,d\nu$$

and $o(1)$ denotes a term that vanishes as $n \to \infty$.

Proof. The proof follow the same argument as Theorem 3.4 and is immediate by substituting $b_n = Cn^2$ in Lemma H.4. □

H.3 NOTE ON IMPLEMENTATION

Our simulation uses a computationally efficient variant π^{fast}, which generates the rejection-sampled grid once at the beginning of each simulated sample path, rather than resampling it independently at every arrival as prescribed by π^{alg}. To verify that this approximation does not materially affect the stopping time distribution, we conducted a paired equivalence study over four parameter regimes $\mu \in \{0.2, 0.3, 0.5, 0.7\}$, with $n = 2000$, $M = 1000$ and $\delta = 0.05$ using 200 paired replications for each value of μ. Within each pair, the arrival and service paths were identical; only the randomly generated grids differed. For each value of μ, let $\{(\tau_i^{\text{fast}}, \tau_i^{\text{alg}})\}_{i=1}^{200}$ denote the paired stopping times obtained from the two policies, where each pair shares the same arrival and service sample path. We report the sample mean stopping times

$$\overline{\tau}^{\text{fast}} := \frac{1}{200}\sum_{i=1}^{200}\tau_i^{\text{fast}}, \qquad \overline{\tau}^{\text{alg}} := \frac{1}{200}\sum_{i=1}^{200}\tau_i^{\text{alg}},$$

together with the paired t-test p-value (p_t), the Wilcoxon signed-rank p-value (p_W), and the sample Pearson correlation $\rho := \text{Corr}\Big(\{\tau_i^{\text{fast}}\}_{i=1}^{200}, \{\tau_i^{\text{alg}}\}_{i=1}^{200}\Big)$, computed from the 200 matched stopping-time pairs. As reported in Table 1 and Figure 2, the two policies are statistically indistinguishable. In every regime, both the paired t-test and Wilcoxon signed-rank test yield p-values exceeding 0.33, the mean difference is at most 0.20% of the typical stopping time, and the pathwise correlation between τ^{fast} and τ^{alg} exceeds 0.994. These results indicate that fixing the rejection-sampled grid has a negligible effect on performance. Accordingly, we use π^{fast} in all subsequent experiments.

Table 1: Paired equivalence study: π^{fast} vs. π^{alg} (200 paired replications per regime, $n = 2000$, $M = 1000$, $\delta = 0.05$).

μ	$\overline{\tau}^{\text{fast}}$	$\overline{\tau}^{\text{alg}}$	$\overline{\tau}^{\text{fast}} - \overline{\tau}^{\text{alg}}$	Relative diff. (%)	p_t	p_W	ρ
0.2	32.00	31.96	$+0.0300$	$+0.09\%$	0.3316	0.4669	0.99877
0.3	40.59	40.66	-0.0800	-0.20%	0.3332	0.6354	0.99491
0.5	73.87	73.86	$+0.0050$	$+0.01\%$	0.8912	0.8153	0.99969
0.7	147.46	147.62	-0.1600	-0.11%	0.5355	0.4354	0.99696

$\overline{\tau}^{\text{fast}}$ and $\overline{\tau}^{\text{alg}}$ denote the sample mean stopping times over the 200 paired replications. The relative difference is computed as $\frac{\overline{\tau}^{\text{fast}} - \overline{\tau}^{\text{alg}}}{\overline{\tau}^{\text{alg}}} \times 100\%$. Here p_t is the paired t-test p-value, p_W is the Wilcoxon signed-rank p-value, and ρ is the sample Pearson correlation between the paired stopping times $\{(\tau_i^{\text{fast}}, \tau_i^{\text{alg}})\}_{i=1}^{200}$..

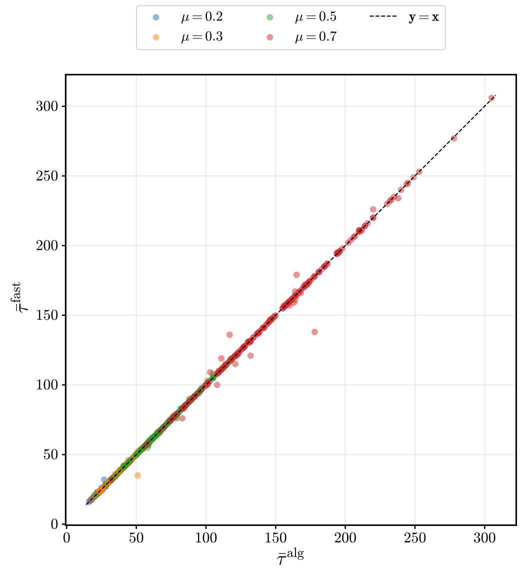


Figure 2: Sample correlation between matched τ values

H.4 SIMULATION-BASED NUMERICAL RESULTS

In this section, we evaluate the performance of the computationally efficient implementation of policy π^{alg} (Algorithm 3) through simulation on synthetic data.

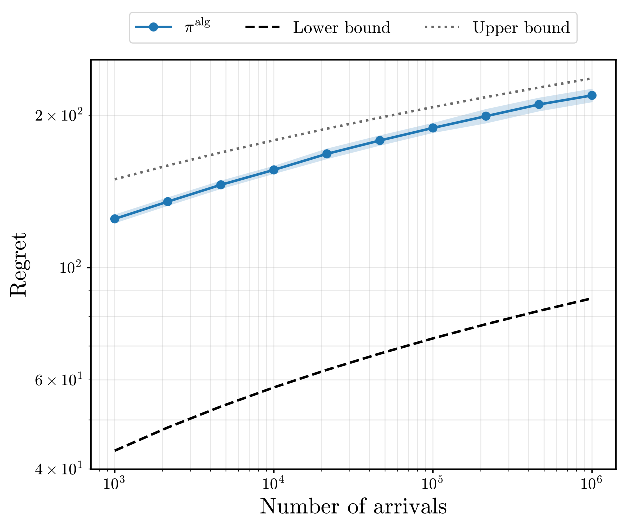
Figure 3a compares the regret of π^{alg} for a fixed service rate, with the instance-dependent asymptotic lower bound (Theorem 2.2) and the corresponding upper bound (Theorem H.5). Both bounds scale as $\Theta(\log n)$, where n denotes the number of arrivals. The gap between the leading constants arises from the additional finite-grid correction term in the upper bound. As shown, the simulated mean regret of π^{alg} lies consistently between the two theoretical bounds—offering empirical evidence for the tightness of the bounds established in Section 3.

Figure 3b compares the regret of π^{alg} as a function of the number of arrivals for three service rates in the reject-optimal regime. As expected, the regret increases monotonically as μ approaches μ^*, consistent with the instance-dependent lower-bound coefficient: instances closer to the threshold are statistically harder to distinguish and therefore require longer expected stopping times before the policy commits to rejecting the null.

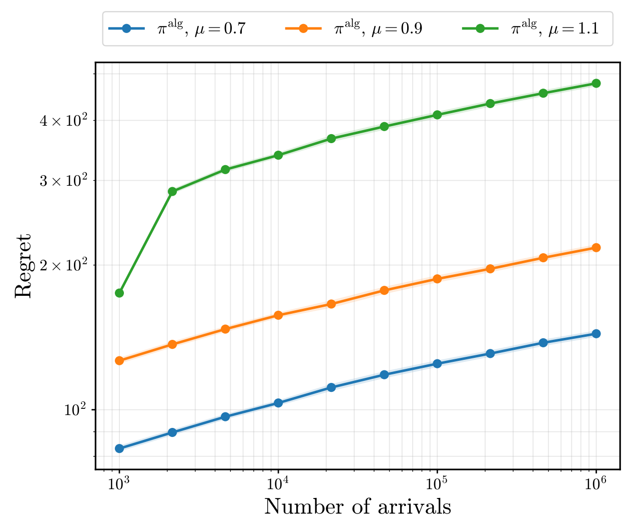
Figure 4 compares the regret as a function of the horizon n for three admission-control policies—π^{alg} (Algorithm 3), π^{pred} and Algorithm 1 of Adler et al. [2022]. The policy π^{pred} is a structurally simpler alternative to π^{alg}. It uses the same stopping threshold as π^{alg}, but instead of forming a mixture of likelihood ratios over a grid, it uses a single point estimate. Specifically, at each arrival t, it computes the maximum likelihood estimate using the history available just before arrival t; hence, it is a predictable estimator. Algorithm 1 of Adler et al. [2022] is qualitatively different from the other two: rather than using a stopping-time test, it is a certainty-equivalent dispatching policy with forced exploration.

Figure 4a compares the regret of the three policies in the reject-optimal regime. Across different values of service rate and horizon, π^{alg} achieves lower regret than π^{pred}—confirming that the mixture-based likelihood yields a statistically more efficient test than a single point-estimate-based procedure. Algorithm 1 of Adler et al. [2022] attains the lowest regret among the three in this regime, which is expected because, rather than committing to a stopping decision, it continuously adapts over the entire horizon.

On the other hand, Figure 4b compares the regret in the accept-optimal regime. Since the stopping-test policies admit whenever there is available capacity, and the probability of false rejection is low, both π^{alg} and π^{pred} achieve zero regret in the simulated runs. In contrast, Algorithm 1 of Adler et al. [2022] is not a stopping-time test and continues to actively explore over the full horizon, regardless of regime. Early in a sample path, the MLE-threshold statistic may be dominated by few unfavorable observations, causing the algorithm to incorrectly infer the regime. It can then take a long time to recover from this “blocked" state, leading to nonzero regret even in the accept-optimal regime.

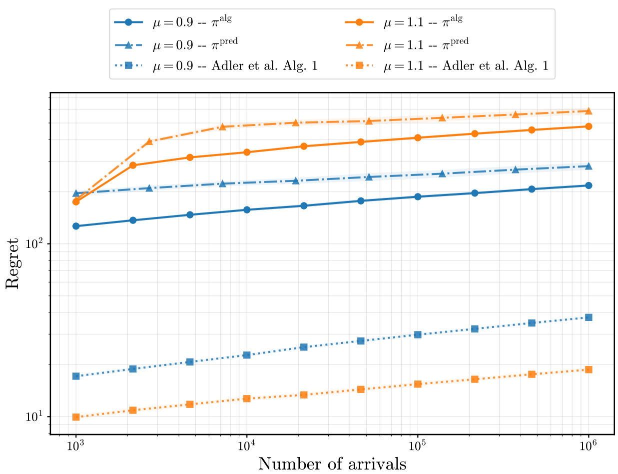


(a) Regret of fixed-horizon implementable policy π^{alg} (Algorithm 3) in reject-optimal regime as a function of the horizon n, at service rate $\mu = 0.9$ (with admission threshold $\mu^* = 1.3$, fixed reward $R = 1$, cost per-unit time $c = 1.3$, number of servers $k = 5$, grid size $M = 1000$, covering width $\delta = 0.05$ and grid start point $\epsilon = 0.001$). The solid curve and markers show mean regret over Monte Carlo simulations with shaded band denoting ± 1 standard error of the mean. The number of replications is reduced from 1000 runs at $n \leq 10^4$ to 500 runs at $n = 10^6$. The dashed curve is the instance-dependent asymptotic lower bound (Theorem 2.2) and the dotted curve is the matching upper bound established for π^{alg} (Theorem H.5), and we show a log-log plot.

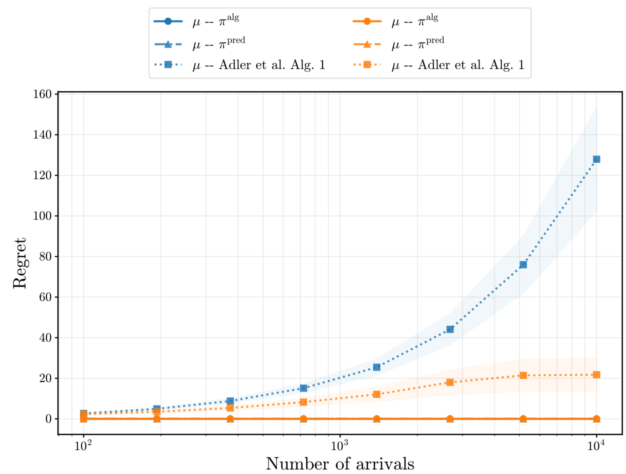


(b) Comparison of regret of fixed-horizon implementable policy π^{alg} (Algorithm 3) in reject-optimal regime for different values of service rate $\mu \in (0.7, 0.9, 1.1)$ as a function of the horizon n (with admission threshold $\mu^* = 1.3$, fixed reward $R = 1$, cost per-unit time $c = 1.3$, number of servers $k = 5$, grid size $M = 1000$, covering width $\delta = 0.05$ and grid start point $\epsilon = 0.001$). The solid curve and markers show mean regret over Monte Carlo simulations with shaded band denoting ± 1 standard error of the mean. The number of replications is reduced from 1000 runs at $n \leq 10^4$ to 500 runs at $n = 10^6$, and we show a log-log plot.

Figure 3: Comparison of performance of policy π^{alg} (Algorithm 3).



(a) Regret versus horizon n for three admission-control policies at service rate $\mu \in \{0.9, 1.1\}$ (with admission threshold $\mu^* = 1.3$, fixed reward $R = 1$, cost per-unit time $c = 1.3$, number of servers $k = 5$, grid size $M = 1000$, covering width $\delta = 0.05$ and grid start point $\epsilon = 0.001$). Circles/solid line: π^{alg} (Algorithm 3). Triangles/dash: π^{pred}, the predictable-MLE test, using the identical stopping threshold as as π^{alg}. Squares/dotted: the dispatching policy of Adler et al. [2022] (Algorithm 1) with forced-exploration probability $1/f(x)$ where $f(x) = \exp(x^{0.6})$ and x counts the number of forced-exploration acceptances at empty system epochs. Shaded bands denote ± 1 standard error of the mean, and we show a log-log plot.



(b) Regret versus horizon n for three admission-control policies at service rate $\mu \in \{1.6, 1.9\}$ (with admission threshold $\mu^* = 1.3$, fixed reward $R = 1$, cost per-unit time $c = 1.3$, number of servers $k = 5$, grid size $M = 1000$, covering width $\delta = 0.05$ and grid start point $\epsilon = 0.001$). Circles/solid line: π^{alg} (Algorithm 3). Triangles/dash: π^{pred}, the predictable-MLE test, using the identical stopping threshold as as π^{alg}. Squares/dotted: the dispatching policy of Adler et al. [2022] (Algorithm 1) with exploration probability $1/f(x)$ where $f(x) = \exp(x^{0.6})$ and x counts the number of forced-exploration acceptances at empty system epochs. Shaded bands denote ± 1 standard error of the mean, and we show a log-linear plot.

Figure 4: Comparison of admission-control policies.