NONPREEMPTIVE SCHEDULING WHILE LEARNING CONTEXT-DEPENDENT SERVICE RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nonpreemptive contextual queueing bandits in a single-server system. Each job is represented by a d-dimensional context vector; in each round, a job may arrive with its context drawn from an unknown distribution $\mathcal{D}$, and its departure probability is determined by a logistic model of that context vector with an unknown parameter θ^*. The server learns from service outcomes while deciding which waiting job to serve and whether to idle, aiming to minimize queue-length regret, the gap between its expected terminal queue length and the minimum achievable by an admissible policy. Once selected, a job must be served until completion, and we refer to this as the nonpreemptive setting. A central challenge is that, even with full model knowledge, the optimal policy cannot in general be characterized by a simple myopic rule, since the optimal action can change with the remaining horizon at the same queue state. Nevertheless, when the model and horizon are known, the optimal action can be obtained through a finite-horizon Bellman recursion. Motivated by this, we propose Learn–Clear–Plan (LCP), which estimates the system and uses the resulting Bellman recursion to make horizon-dependent decisions. LCP achieves $\widetilde{O}(\sqrt{d/T})$ queue-length regret, while a lower-bound construction gives $\Omega(\min\{1/\sqrt{d}, \sqrt{d/T}\})$ regret for every learning policy on some instance, establishing optimality up to polylogarithmic factors when $T \geq d^2$. When the horizon is unknown, no horizon-independent policy achieves vanishing regret against the finite-horizon optimum. We therefore use SEPT, the policy that serves a waiting job with the highest probability of departure, as a fixed reference, and suggest an estimated-SEPT algorithm that achieves a tracking error of $\widetilde{O}(\sqrt{d/t})$ without knowing the model.

1 INTRODUCTION

Queueing systems are central to cloud computing, online service platforms, communication networks, and other applications in which heterogeneous jobs compete for limited service capacity (Neely, 2010). In many such systems, a job's completion probability depends on observable features, while the corresponding service model is initially unknown. A scheduler must therefore learn context-dependent service rates from departure feedback while simultaneously controlling congestion. Queueing bandits formalize this learning-while-scheduling problem, and queue-length regret measures the excess number of unfinished jobs relative to an optimal policy with full knowledge of the service model (Krishnasamy et al., 2021; Stahlbuhk et al., 2021).

Contextual queueing bandits extend this framework to job-specific features. Under the assumption that the contexts of arriving jobs are drawn independently from a fixed distribution, Bae et al. (2026) introduce a preemptive contextual queueing bandit model, in which the scheduler may reconsider the entire queue and select a different job–server pair in the next round after an unsuccessful service attempt. They obtain $\widetilde{O}(T^{-1/4})$ queue-length regret in this setting. For the same stochastic-context setting, Bae & Lee (2026) subsequently propose the CQB-η-2 algorithm and improve the upper bound to $\widetilde{O}(T^{-1/2})$. They further show that every learning algorithm incurs $\Omega(T^{-1/2})$ regret on some instance, showing that this dependence on T is optimal up to logarithmic factors.

This preemptive abstraction facilitates modeling bandit feedback, characterizing optimal policies, and analyzing regret, but it excludes systems in which an ongoing computation, repair, or com-

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

Table 1: Summary of main results. Known/unknown indicates whether the server knows the horizon; IA/WC denote the idling-allowed/work-conserving settings, and SEPT denotes the shortest-expected-processing-time policy. Logarithmic factors and fixed model constants are suppressed.

Setting	Benchmark / Objective	Result	Note
Known T, IA	Regret against optimal IA	$\widetilde{O}(\sqrt{d/T})$	Algorithm 1, Theorem 1
Known T, IA/WC	Regret against optimum	$\Omega(\min\{d^{-1/2}, \sqrt{d/T}\})$	Theorem 2
Unknown, IA/WC	Vanishing regret against finite-horizon optimum	Impossible: $\limsup_{t\to\infty} R_t^S(\pi) > 0$	Theorem 3
Unknown, WC	Tracking true SEPT	$\widetilde{O}(\sqrt{d/t})$	Algorithm 2, Theorem 4

munication session cannot be interrupted, or in which preemption and migration are prohibitively costly. To capture such service commitments, we study a contextual queueing bandit model with a *nonpreemptive setting*, where once a job starts service, the server must continue processing it until completion. This change substantially complicates optimal scheduling. In the preemptive setting, an optimal policy is *work-conserving*, i.e., not idle, and selects an available job–server pair with the highest departure probability (Bae et al., 2026). However, under nonpreemption, the optimal action can depend on the current queue state, remaining horizon, and arriving job's distribution even in the single-server case (see Section 3 for a detailed discussion). So, the optimal policy need not admit such a simple myopic characterization as in the preemptive setting, making learning and planning fundamentally intertwined. To study these difficulties in their most basic form, we focus first on the single-server setting. Our main results for known and unknown horizons are summarized in Table 1.

Known horizon: Learn–Clear–Plan algorithm and its optimality For a known horizon T, we compare with the optimal finite-horizon policy that may idle. Our Learn–Clear–Plan (LCP) algorithm estimates the departure probability function of jobs and their departure probability distribution, waits until the queue empties, and then follows an empirical finite-horizon Bellman policy (the policy that follows from the Bellman recursion, which is calculated by using the estimated function and distribution). Under some conditions related to job arrivals, Theorem 1 shows that it achieves $\widetilde{O}(\sqrt{d/T})$ as a regret upper bound. In particular, unlike the preemptive results of Bae et al. (2026); Bae & Lee (2026), this guarantee does not require a lower-eigenvalue bound on the context covariance matrix. We complement this result with the lower bounds in Theorem 2: for fixed model and traffic parameters, every learning policy incurs $\Omega(\min\{d^{-1/2}, \sqrt{d/T}\})$ regret against the optimum policy on some instance, regardless of whether idling is permitted or not. Hence, when $T \geq d^2$, our upper bound under fixed horizon is optimal in its dependence on d and T up to logarithms.

Unknown horizon: Impossibility and SEPT tracking. When the horizon is unknown, Theorem 3 shows that there is no horizon-free policy such that the t-th round queue-length regret converges to 0 as $t \to \infty$, even when the service model is known. We therefore use SEPT, the policy that serves a waiting job with the highest departure probability whenever the server is free, as a fixed horizon-independent reference policy. Our estimated-SEPT algorithm updates its logistic estimator only between busy periods and follows the resulting priority rule during each busy period. By Theorem 4, its tracking error satisfies $\mathcal{T}_t^{\mathrm{SEPT}} = \widetilde{O}(\sqrt{d/t})$ for fixed model and traffic parameters. Thus, our known-horizon result concerns regret relative to a finite-horizon optimum, whereas our unknown-horizon result concerns tracking a fixed reference policy.

2 Problem Setting

We study the nonpreemptive contextual queueing bandit problem in a discrete-time system consisting of a single queue and a single server. Each job is represented by a context vector in $\mathcal{X} \subseteq \mathbb{R}^d$, which determines its service-completion probability. We study both a *known-horizon* setting, where the server knows the evaluation horizon $T \in \mathbb{N}$ and may design its policy accordingly, and an *unknown-horizon* setting, where the server does not know the evaluation horizon and must use a horizon-independent policy (we call this an *anytime* policy). Let Q_t denote the number of unfinished jobs present at the beginning of round t, and call it the *queue length* at round t. We assume

that the system is initially empty, so $Q_1 = 0$. A key feature of our model is the *nonpreemptive* setting. This means that once the server selects a job, it must continue serving that job in every subsequent round until the job departs or the horizon ends. We call a job that is currently in service a *job in service*, and call all other unfinished jobs *waiting jobs*. The server is called *free* when no job is in service and the queue is not empty, and when it is free, it may start one waiting job or remain *idle*, meaning that it performs no service in that round. If the queue is empty, it idles necessarily.

Each round consists of a *service stage* (*departure process*) followed by an *arrival stage* (*arrival process*). If there is a job in service at the beginning of round t, the server must serve that job again. Otherwise, it may start one waiting job of its choice or idle. Here, the server can observe whether the served job is completed. Let $D_t \in \{0, 1\}$ be the departure indicator in round t, with $D_t = 0$ if the server idles. If a job with context x is served in round t, its departure indicator has distribution $\mathrm{Bern}(p(x))$, where $p(x) := \mu(x^\top \theta^*)$ is called the *success probability function* and $\mu(z) := (1 + e^{-z})^{-1}$. Here, $\theta^* \in \mathbb{R}^d$ is called the *parameter vector*. Conditional on the contexts, departure indicators are independent across jobs and rounds. Equivalently, if $G(x)$ is the number of service attempts required to complete a job with context x, then $G(x) \sim \mathrm{Geom}(p(x))$ and $\mathbb{E}[G(x)] = 1/p(x)$ holds. After the service stage, a new job arrives with indicator $A_t \sim \mathrm{Bern}(\lambda)$. Conditional on $A_t = 1$, draw $X_t \sim \mathcal{D}$ from a context distribution $\mathcal{D}$ independently. The arriving job becomes available for service in round $t + 1$, and we denote X_t as the context of the job that arrived in round t. Then the queue length evolves as $Q_{t+1} = (Q_t - D_t)^+ + A_t$, where $(q)^+ := \max\{0, q\}$.

Admissible policies may randomize, although the algorithms proposed in this paper are deterministic conditional on the observed history. We consider two problem settings. In the *idling-allowed* (IA) setting, policies are allowed to idle whenever the server is free. In the *work-conserving* (WC) setting, policies must start a waiting job whenever the server is free and the queue is nonempty. We refer to policies admissible in these settings as IA and WC policies, respectively. Thus, every WC policy is also an IA policy. For a horizon $T \in \mathbb{N}$, denote $J_T^\pi := \mathbb{E}_\pi[Q_{T+1}]$ as the expected final queue length when the server follows policy π. Let V_T^{IA} and V_T^{WC} denote the minimum expected final queue lengths over all IA and WC policies, respectively, with knowledge of T, θ^*, and $\mathcal{D}$. We define the corresponding regrets as $R_T^{\mathrm{IA}}(\pi) := J_T^\pi - V_T^{\mathrm{IA}}$, $R_T^{\mathrm{WC}}(\pi) := J_T^\pi - V_T^{\mathrm{WC}}$.

Assumption 1. The sequence $(A_t, X_t)_{t \geq 1}$ is i.i.d., with A_t independent of X_t. The server knows $\lambda \in (0, 1)$, $S > 0$, and $0 < p_- < p_+ < 1$, while θ^* and $\mathcal{D}$ are unknown. Moreover, $\|X\|_2 \leq 1$ almost surely, $\|\theta^*\|_2 \leq S$, $p(X) \in [p_-, p_+]$ almost surely, and $\rho := \lambda \mathbb{E}_{X \sim \mathcal{D}}[1/p(X)]] < 1$ holds.

In Assumption 1, except for the final condition $\rho < 1$, these are nothing but the standard boundedness and stochastic-context assumptions for logistic bandit models (Bae et al., 2026; Bae & Lee, 2026). The condition $\rho < 1$ plays a crucial role in our analysis, as it is a key ingredient in establishing the workload properties in Lemma 1, which are repeatedly used throughout the paper. See Section 3.2 and Appendix C for its interpretation and detailed use. It also serves as a replacement for the traffic-slack condition used in the prior work.

3 Main Difficulties and New Techniques

In this section, we highlight the main difficulty arising from the nonpreemptive setting and introduce a new tool, which we call workload, to address this difficulty.

3.1 Difficulties of Nonpreemptive Setting

A central difficulty in nonpreemptive contextual queueing bandits is the loss of the simple optimal scheduling rule available in the preemptive setting. In the preemptive model of Bae et al. (2026); Bae & Lee (2026), an optimal policy is work-conserving and selects a waiting job with the highest departure probability when the server is free. Under nonpreemption, this myopic rule need not be optimal, as the following counterexamples show. Suppose that a free server has one waiting job with departure probability p_s. With one round remaining, serving it immediately is trivially optimal. However, with two rounds remaining, the advantage in expected departures of idling first and then serving the best available job, compared with serving immediately, is at least the following value.

$$p_s^2 - p_s(1 + \lambda) + \lambda \left(\mathbb{E}[\max\{p_s, p(X)\}] - p_s \mathbb{E}[p(X)] \right). \tag{1}$$

A detailed proof is included in Appendix G.1.

We can construct a two-context instance satisfying Assumption 1 such that the quantity (1) is positive, as shown in Appendix G.1. This counterexample shows that a work-conserving policy is not always optimal. Furthermore, even in the WC setting, choosing a job with the highest departure probability is not always optimal, as shown in the following proposition.

Proposition 1. *There is a stable three-context instance with $\lambda = \frac{1}{2}$ and departure probabilities $p_A = \frac{7}{20}$, $p_B = \frac{13}{20}$, and $p_F = \frac{19}{20}$ such that when the server is free and exactly one A job and one B job are waiting, serving B first is the unique optimal action with one round remaining. With fourteen rounds remaining, serving A first is the unique optimal action. This instance satisfies Assumption 1.*

The proof of this proposition will be given in Appendix G.2. Hence, in the nonpreemptive setting, there is no universal horizon-independent priority rule, and we need to know the remaining horizon, the current queue, and the future-arrival distribution to decide which action is optimal in each round. Thus, this explains that determining an optimal action requires solving the finite-horizon Bellman recursion at each round.

3.2 A New Technique: Workload-Based Arrival Analysis

In the nonpreemptive setting, the optimal policy depends on the queue state and remaining horizon rather than a simple myopic rule, making departure-based regret analysis difficult. From this viewpoint, we shift our attention to the arrival process, whose law is much simpler and independent of the scheduling policy. For each arriving job with context X_t, independently draw its required service time $G(X_t) \sim \mathrm{Geom}(p(X_t))$, and define the *workload* W_t as the sum of the remaining required service times of all jobs present at the beginning of round t. Under any WC policy, the following recursion holds.

$$W_{t+1} = (W_t - 1)^+ + A_t G(X_t) \tag{2}$$

Thus, under a common coupling of arrivals, arrived jobs, and their required service times, the workload is independent of the service policy and order, satisfies $Q_t \le W_t$, and vanishes iff $Q_t = 0$. Thus, any two WC queues $Q_t^{(1)}, Q_t^{(2)}$ coupled as above share the same workload process $(W_t)_{t\ge 1}$, become empty at the same times, and satisfy $|Q_t^{(1)} - Q_t^{(2)}| \le W_t$ for all t. Furthermore, with this definition, $\rho = \lambda \mathbb{E}[1/p(X)] = \lambda \mathbb{E}_{X\sim\mathcal{D}, G(X)\sim\mathrm{Geom}(p(X))}[G(X)]$ is precisely the expected amount of workload brought into the system by arrivals in each round.

We now introduce the following basic workload properties. The constants in the following are formally defined in Appendix C, and the properties are proved in Appendices C.1 and C.3.

Lemma 1. *Under Assumption 1, there exist positive constants $r \in (0,1)$, $\psi(r) \in (0,1)$, $M_G(r)$, K_r and $C_W(r)$ such that the WC workload process started from $W_1 = 0$ satisfies $\mathbb{E}e^{rW_t} \le K_r$ and $\mathbb{E}W_t \le C_W(r)$ for every $t \ge 1$. Moreover, for $u \ge 1$, let $\tau_0 := \inf\{j \ge 0 : W_{u+j} = 0\}$. Then,*

$$\mathbb{E}\left[W_{u+h}\mathbf{1}\{\tau_0 > h\} \,|\, W_u\right] \le \frac{e^{rW_u}}{r}\psi(r)^h \tag{3}$$

for every $h \ge 0$. Hence, $\mathbb{E}[W_{u+h}\mathbf{1}\{\tau_0 > h\}] \le (K_r/r)\psi(r)^h$. We refer to these two inequalities as the workload tail bounds. Lastly, $\mathbb{E}[W_{t+1} - W_t \mid W_t = w] = \rho - 1 < 0$ for every $w \ge 1$, so the workload recursion admits a stationary distribution W that stochastically dominates W_t for every $t \ge 1$.

Before concluding this subsection, we introduce the notion of a *busy period*, which will be used later in our SEPT analysis. Under a WC policy, a *busy period* is defined as the sequence of consecutive service rounds that begins when the server first serves a job after the system has become empty and ends in the first round whose service stage leaves the system empty, before the arrival at the end of that round. Let τ_{bp} denote its length.

4 Known Horizon: Upper and Lower Bounds of Regret

We now consider the known-horizon IA setting and present the *Learn-Clear-Plan* (LCP) algorithm, which achieves $R_T^{\mathrm{IA}} = \widetilde{O}(\sqrt{d/(\lambda T)})$. Before presenting LCP, we define the estimators used by the algorithm and provide high-probability upper bounds, and introduce the concept of *Bellman policy*.

4.1 Estimating Departure Probabilities and Their Distribution

Index jobs by their arrival order. Let X_i be the context of job i and Y_i the departure indicator from its first service attempt. Then $(X_i, Y_i)_{i\geq 1}$ is an i.i.d. sequence with $Y_i \mid X_i \sim \mathrm{Bern}(p(X_i))$. Using the first n pairs, we estimate θ^* by $\widehat{\theta}_n$ through empirical risk minimization, and use this estimate to construct a predictor $\widehat{p}_n$ of p. For $y \in \{0,1\}$, define $\ell_\theta(x,y) := \log(1+e^{x^\top\theta}) - yx^\top\theta$, and let

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\|\theta\|_2 \leq S} \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i), \qquad \widehat{p}_n(x) := \mathrm{proj}_{[p_-,p_+]}\Big(\mu(x^\top\widehat{\theta}_n)\Big). \tag{4}$$

We next estimate the distribution of future jobs' departure probabilities. Using the contexts from pairs $(X_i, Y_i)_{i=n+1}^{2n}$ and $\widehat{p}_n$, we define the empirical distribution $\widehat{F}_n := \frac{1}{n}\sum_{i=n+1}^{2n} \delta_{\widehat{p}_n(X_i)}$. This distribution estimates $F_{\widehat{p}_n} := \mathrm{Law}(\widehat{p}_n(X))$, where $X \sim \mathcal{D}$ is independent of the training block.

To quantify the estimation errors, for any predictor $q : \mathbb{R}^d \to [p_-, p_+]$, define $\mathcal{E}(q) := \mathbb{E}_{X\sim\mathcal{D}}|q(X) - p(X)|$ and denote W_1 as the 1-Wasserstein distance. Let $\bar{S} := \max\{1, S\}$. For $n \geq 5$, define $\epsilon_n(\delta) := \min(1, \sqrt{(8e^{\bar{S}}\left(d\log(32\bar{S}n) + \log(1/\delta)\right) + 1)/2n})$ and set $\epsilon_n(\delta) = 1$ for $n < 5$. We further define $\eta_n(\delta) := (p_+ - p_-)\sqrt{\log(2/\delta)/2n}$.

Lemma 2. *For every $n \geq 1$ and $\delta \in (0,1)$, each of the following bounds holds with probability at least $1-\delta$.*

$$\mathcal{E}(\widehat{p}_n) \leq \epsilon_n(\delta), \qquad W_1(\widehat{F}_n, F_{\widehat{p}_n}) \leq \eta_n(\delta). \tag{5}$$

The proofs of the two inequalities are provided in Appendices D.1 and D.3. We refer to them as concentration inequalities.

4.2 Bellman Policy for the Known-Horizon IA Setting

We now explain how to characterize an optimal policy in the known-horizon IA setting through a Bellman recursion when p and $\mathcal{D}$ are known. We first represent the current state mathematically and then present the corresponding finite-horizon Bellman recursion. We first consider a future-job success-probability distribution F with finite support. Let $r_1, \ldots, r_m$ be the distinct values consisting of the support of F together with the success probabilities of all currently present jobs, and let $\nu_j = F(\{r_j\})$; thus $\nu_j = 0$ for a value that appears only among the currently present jobs. Represent the queue at the beginning of a round by a state (c, a), where $c \in \mathbb{Z}_{\geq 0}^m$, c_j is the number of waiting jobs with success probability r_j, and $a \in \{0, 1, \ldots, m\}$ identifies the job type in service; $a = 0$ means that the server is free. The job in service is excluded from c and is represented by a. Let $V_h(c,a)$ be the minimum expected terminal queue length with h rounds remaining and starting state (c,a), so that $V_0(c,a) = \sum_j c_j + \mathbf{1}\{a \neq 0\}$, the number of unfinished jobs. Writing e_j for the j-th standard basis vector, define the arrival operator, which averages over the arrival at the end of the current round, by $(\mathcal{A}f)(c,a) = (1-\lambda)f(c,a) + \lambda\sum_{j=1}^m \nu_j f(c+e_j, a)$.

If $a \neq 0$, nonpreemption forces the continuation of service, giving

$$V_h(c,a) = r_a(\mathcal{A}V_{h-1})(c,0) + (1-r_a)(\mathcal{A}V_{h-1})(c,a). \tag{6}$$

If the server is free, it may idle or start a waiting job:

$$\begin{aligned} &V_h(c,0) \\ &= \min\Big\{(\mathcal{A}V_{h-1})(c,0),\ \min_{j:c_j>0}\big[r_j(\mathcal{A}V_{h-1})(c-e_j,0) + (1-r_j)(\mathcal{A}V_{h-1})(c-e_j,j)\big]\Big\}. \end{aligned} \tag{7}$$

Here, the first term in (7) corresponds to idling, and each term in the inner minimum corresponds to starting a type-j job. The inner minimum is omitted when $c = 0$. Each term is the expected optimal continuation value resulting from the corresponding action. Hence, if the first term attains the minimum, idling is optimal, while if the term indexed by j reaches the minimum, starting a type-j job is optimal. Repeating such minimizing choices whenever the server is free yields an optimal finite-horizon policy. For a general F, an analogous recursion represents the waiting jobs by a multiset of success probabilities and replaces the arrival sum by an integral with respect to F.

We call a policy that makes these minimizing choices at every free-server state a *Bellman policy* for the corresponding model and horizon. The *true Bellman policy* based on (p, F_p) uses $p(x)$ as a success probability function and $F_p = \mathrm{Law}(p(X))$, $X \sim \mathcal{D}$, for future-arrivals success-probability distribution. Thus, a true Bellman policy is an optimal IA policy. In particular, starting from an empty queue with T rounds remaining, it attains V_T^{IA}. Suppose now that the true model is (p, F_p), but the server has only an estimated success-probability function q and an estimated future-job distribution $\widehat{F}$. The *estimated Bellman policy* based on $(q, \widehat{F})$ treats $(q, \widehat{F})$ as the true model and follows the minimizing actions of the corresponding finite-horizon Bellman recursion.

4.3 The Learn-Clear-Plan Algorithm

We now introduce the *Learn–Clear–Plan* (LCP) algorithm, which first estimates the unknown model, clears the queue, and then follows the estimated Bellman policy based on that estimated model.

Algorithm 1 Learn–Clear–Plan

Require: horizon T, planning-window length $H < T$
$L \leftarrow T - H$, $n \leftarrow \lfloor \lambda L/8 \rfloor$
$Z \leftarrow ()$, PLAN $\leftarrow 0$
for $t = 1, \ldots, T$ **do**
  if $t > L$, $Q_t = 0$, and PLAN $= 0$ **then**
    if $|Z| \geq 2n$ and $n \geq 1$ **then**
      compute $\widehat{p}_n$ from the first n pairs in Z
      $\widehat{F}_n \leftarrow \frac{1}{n} \sum_{i=n+1}^{2n} \delta_{\widehat{p}_n(X_i)}$
    else
      $\widehat{p}_n(x) \leftarrow (p_- + p_+)/2$, $\widehat{F}_n \leftarrow \delta_{(p_- + p_+)/2}$
    end if
    PLAN $\leftarrow 1$
  end if
  if PLAN $= 1$ **then**
    take an optimal action from the Bellman recursion with $T - t + 1$ rounds under $(\widehat{p}_n, \widehat{F}_n)$
  else
    continue the job in service; otherwise serve FCFS, or idle if $Q_t = 0$
  end if
  observe D_t
  if job i begins service in round t **then**
    set $Y_i \leftarrow D_t$ and append (X_i, Y_i) to Z
  end if
  observe A_t and, if $A_t = 1$, the arriving context X_t
end for

Learn. Choose a planning-window length $H < T$, let $L := T - H$, and set $n := \lfloor \lambda L/8 \rfloor$. During the first L rounds, whenever the server is free, it starts the earliest-arriving waiting job; we refer to this rule as *FCFS policy*, meaning that first-come, first-served. When job i with context X_i first begins service, the server retains only the departure indicator Y_i from its first service attempt. If at least $2n$ samples are collected, the first n pairs (X_i, Y_i) are used to construct $\widehat{p}_n$, while the contexts of the next n jobs are used to construct $\widehat{F}_n$. Separating the samples into two blocks ensures that, conditional on $\widehat{p}_n$, the contexts in the second block remain i.i.d.

Clear. If the queue remains nonempty after round L, the server continues serving jobs according to FCFS without idling until it reaches a round that begins with an empty queue. This provides an empty initial state for the subsequent planning phase.

Plan. Once the queue has been cleared, the server follows the *estimated-model Bellman policy* based on $(\widehat{p}_n, \widehat{F}_n)$ for the current queue state and the remaining number of rounds. Importantly, the current queue need not consist only of jobs whose estimated success probabilities lie in the support of $\widehat{F}_n$. The distribution $\widehat{F}_n$ describes only the success probabilities of future arrivals assumed in the Bellman calculation; the jobs already present in the queue are part of the given current state. Thus,

if a job with context X arrives in the actual system with $\widehat{p}_n(X)$ outside the support of $\widehat{F}_n$ during the Plan phase, it is simply included in the current state as an additional type with zero future-arrival mass. This leaves $\widehat{F}_n$ unchanged while allowing the job to be considered among the available actions in subsequent Bellman calculations.

4.4 Upper bound of regret

Theorem 1. *Under Assumption 1, let $1 \le H < T$, $L = T - H$, $n = \lfloor \lambda L/8 \rfloor \ge 1$, and $\delta_T = T^{-4}$. Let π^{LCP} be the policy in Algorithm 1. Then*

$$R_T^{\mathrm{IA}}(\pi^{\mathrm{LCP}}) \le \frac{K_r}{r}\psi(r)^H + H\left(e^{-\lambda L/8} + K_r e^{-r\lambda L/4} + 2\delta_T\right) + \frac{2H^2}{p_-}\epsilon_n(\delta_T) + \frac{(H-1)H(H+1)(H+2)}{12}\eta_n(\delta_T). \tag{8}$$

Here, $r, \psi(r), K_r$ are constants as given in Lemma 1. For $H = \lceil \log^2(eT) \rceil$ and sufficiently large T, $R_T^{\mathrm{IA}}(\pi^{\mathrm{LCP}}) = \widetilde{O}(\sqrt{d/(\lambda T)})$ for fixed model and traffic parameters. The total number of Bellman recursion terms evaluated over all decision epochs in the Plan phase is at most $\exp(O(\log^3 T))$.

Sketch of Proof. Consider the event that Clear occurs before the end of the horizon, at least $2n$ jobs have begun service by round L, and both bounds in Lemma 2 hold. The complementary event contributes the first two terms of Equation (8) by concentration inequalities and the workload tail bound in Lemma 1. In this event, Plan starts from an empty queue with $h \le H$ rounds remaining. Let $q = \widehat{p}_n$ and let $\widehat{\pi}$ be the estimated Bellman policy based on $(q, \widehat{F}_n)$. For an empty-start h-round system, let $J_h^r(\pi)$ denote the expected terminal queue length under the success probability function r, and let V_h^r be its optimal value among all IA policies. We decompose regret for this event as follows.

$$J_h^p(\widehat{\pi}) - V_h^p = \left(J_h^p(\widehat{\pi}) - J_h^q(\widehat{\pi})\right) + \left(J_h^q(\widehat{\pi}) - V_h^q\right) + \left(V_h^q - V_h^p\right).$$

The first and third terms compare models that differ in their success-probability functions, whereas the middle term captures the suboptimality under (q, F_q) of the Bellman policy computed using $(q, \widehat{F}_n)$. We now invoke Equations (32) and (33) to bound these terms. The first inequality below bounds the difference in the expected terminal queue length when the same policy is used under p and q in terms of $\mathcal{E}(q)$, while the second bounds the loss from using the Bellman policy computed with $\widehat{F}_n$ instead of F_q, with q fixed, in terms of $W_1(F_q, \widehat{F}_n)$:

$$|J_h^p(\pi) - J_h^q(\pi)| \le \frac{h^2}{p_-}\mathcal{E}(q), \qquad J_h^q(\widehat{\pi}) - V_h^q \le \frac{(h-1)h(h+1)(h+2)}{12}W_1(F_q, \widehat{F}_n). \tag{9}$$

Applying the first inequality to $\widehat{\pi}$ and to each of the optimal policies under p and q bounds the first and third terms by $2H^2\epsilon_n(\delta_T)/p_-$, which yields the third term in Theorem 1. The second inequality, together with Lemma 2, bounds the middle term by the final term in Equation (8). The complete proof appears in Appendix E.

4.5 Lower Bounds of Regret

We next establish a worst-case regret lower bound for every learning policy, including randomized policies. We construct a family of hard instances with dummy, baseline, and candidate jobs that share the same context distribution $\mathcal{D}$ and model constants λ, S, p_-, p_+, and differ only in θ^*. These instances are statistically hard to distinguish and are chosen so that, if the final round begins with the server free and exactly one baseline and one candidate waiting, either job selection is suboptimal on some instance, with a departure-probability (regret) gap of $\Omega(\min\{1/\sqrt{d}, \sqrt{d/T}\})$.

For WC policies, we define an event H that, followed by the arrival of a candidate, produces this final decision state. Stationary workload domination bounds the probability of this event away from zero, and standard testing arguments imply that, on some instance, the learner takes a suboptimal final action with sufficiently high probability, yielding the desired regret lower bound. The IA proof follows the same strategy, but requires an additional contradiction argument because the event H requires the server to start a waiting job when it becomes free, while an IA policy may instead choose to idle. The complete proofs appear in Appendix F.

Theorem 2. *For every $d \geq 3$, there exist a context distribution $\mathcal{D}$ on $\mathbb{R}^d$ and constants $\lambda \in (0,1)$, $S > 0$, and $0 < p_- < p_+ < 1$ with the following property. For every $T \geq 4$, every setting $\mathsf{S} \in \{\mathrm{IA}, \mathrm{WC}\}$, and every learning policy π in that setting, including randomized policies, there exists a parameter vector θ^* such that $(\mathcal{D}, \theta^*)$ satisfies Assumption 1 and*

$$R_T^{\mathsf{S}}(\pi) \geq c_{\mathsf{S}} \min\left(1/\sqrt{d}, \sqrt{d/T}\right), \tag{10}$$

even if $\mathcal{D}$ is known to the server. The constants $c_{\mathrm{IA}}, c_{\mathrm{WC}} > 0$ and the model constants can be chosen independently of d and T.

For $T \geq d^2$, the IA lower bound is $\Omega(\sqrt{d/T})$, matching the upper LCP bound up to polylogarithmic factors for the fixed model and the traffic parameters. Thus, LCP achieves the optimal worst-case regret rate up to these factors in the known-horizon IA setting.

5 Unknown Horizon: Impossibility and SEPT Tracking

5.1 Nonexistence of an Anytime Policy with Vanishing Regret

A natural question in the unknown-horizon setting is whether an anytime IA or WC policy π can satisfy $R_t^{\mathrm{IA}}(\pi) \to 0$ (resp. $R_t^{\mathrm{WC}}(\pi) \to 0$) as $t \to \infty$. However, the two counterexamples in Equation (1) and Proposition 1 show that this is impossible for both settings, even when θ^* and $\mathcal{D}$ are known.

Theorem 3. *For each setting $\mathsf{S} \in \{\mathrm{IA}, \mathrm{WC}\}$, there exists a fixed instance satisfying Assumption 1, with two contexts for IA and three for WC, such that every anytime S policy π, even with full knowledge of θ^* and $\mathcal{D}$, satisfies*

$$\limsup_{t \to \infty} R_t^{\mathsf{S}}(\pi) > 0. \tag{11}$$

We present only the proof idea for the WC setting here. The IA case is proved similarly using the queue state in Equation (1), for which the optimal action depends on the remaining horizon. For WC, we use the state in Proposition 1, where the server is free with exactly one A job and one B job waiting. Using workload domination, we show that this state occurs at the beginning of the round s, for every $s \geq 5$, with probability bounded away from zero. In this event, starting B is uniquely optimal if the evaluation horizon is s, whereas starting A is uniquely optimal if the evaluation horizon is $s + 13$. Thus, neither choosing A nor choosing B is optimal for both evaluation horizons, and each suboptimal choice incurs a fixed positive regret gap. Therefore, for any randomized action distribution over A and B, at least one of the regrets at horizons s and $s + 13$ is bounded below by a positive constant. This contradicts the convergence of regret to zero. The complete WC proof appears in Appendix G.3, while the complete IA proof appears in Appendix G.1 within the same appendix section.

Thus, in both IA and WC settings, no anytime policy, including randomized policies, can achieve vanishing regret against the finite-horizon optimum, i.e., $R_t^{\mathsf{S}}(\pi) \to 0$ as $t \to \infty$. Moreover, in the free-server states of these counterexamples, no action is optimal for all horizons: each choice is suboptimal for some horizon.

5.2 SEPT Policy and Estimated SEPT Algorithm

Because the optimal action at a given state can vary with the evaluation horizon, and no anytime policy achieves vanishing regret against the finite-horizon optimum, we adopt the *shortest-expected-processing-time* (SEPT) policy as an anytime reference policy. SEPT is the WC policy that, whenever the server is free, starts a waiting job with the highest success probability, breaking ties by arrival order. Since we cannot follow the true SEPT policy because p is unknown, our goal is to construct an anytime policy whose expected terminal queue length approaches that of true SEPT as the evaluation horizon grows, without assuming that SEPT is optimal in the present nonpreemptive setting. To this end, we introduce the *estimated-SEPT* algorithm, a WC policy that updates an estimate $\widehat{p}$ only at the end of each busy period, keeps it fixed throughout the next busy period, and uses it in place of p to apply the same priority rule as SEPT. Updating $\widehat{p}$ only between busy periods ensures that the predictor used within a busy period depends only on observations preceding that busy period

Algorithm 2 Busy-period estimated SEPT

$Z \leftarrow (), \widehat{p}(x) \leftarrow (p_- + p_+)/2$
for $t = 1, 2, \ldots$ **do**
  if a job is in service **then**
    continue serving that job
  else if $Q_t > 0$ **then**
    serve the earliest arrival in $\arg\max_{j \text{ waiting}} \widehat{p}(X_j)$
  else
    idle
  end if
  observe D_t; if job i starts service, set $Y_i \leftarrow D_t$ and append (X_i, Y_i) to Z
  if the queue is empty after service and before the arrival, and $|Z| \geq 1$ **then**
    recompute $\widehat{p}$ using Equation (4) from all pairs in Z
  end if
  observe A_t and, if $A_t = 1$, the arriving context X_t
end for

and is therefore independent of that busy period's arriving job contexts. Let J_t^{alg} and J_t^{SEPT} denote the expected final queue lengths after t rounds under estimated-SEPT and true SEPT, respectively, and define $\mathcal{T}_t^{\mathrm{SEPT}} := |J_t^{\mathrm{alg}} - J_t^{\mathrm{SEPT}}|$. For $t \geq 6$, put $u = \lfloor t/2 \rfloor$, $m_t = \lfloor \lambda u/2 \rfloor$, and $\delta_t = t^{-4}$. If $m_t \geq 5$, set $\bar{\epsilon}_t = \min\left(1, \sqrt{(8e^{\bar{S}}\left(d\log(32\bar{S}t) + \log(\pi^2 t^2/6\delta_t)\right) + 1)/2m_t}\right)$, $b_t = e^{-\lambda u/8} + \delta_t$.

Theorem 4. *Under Assumption 1, if $m_t \geq 5$, then Algorithm 2 satisfies*

$$\mathcal{T}_t^{\mathrm{SEPT}} \leq \Phi_r(\bar{\epsilon}_t) + \frac{M_G(r)}{r(1-\psi(r))}\psi(r)^u + \frac{\sqrt{K_r b_t}}{r}. \tag{12}$$

Here, $\Phi_r(\epsilon)$ is the function defined on $\epsilon \in (0, 1]$ and satisfies $\Phi_r(\epsilon) = O(\epsilon(1 + \log^3(1/\epsilon)))$, and $r, \psi(r), M_G(r), K_r$ are constants as given in Lemma 1. Consequently, for fixed model and traffic parameters, $\mathcal{T}_t^{\mathrm{SEPT}} = \widetilde{O}(\sqrt{d/(\lambda t)})$. If $m_t < 5$, the uniform bound $\mathcal{T}_t^{\mathrm{SEPT}} \leq C_W(r)$ holds.

The proof uses ideas similar to those in the LCP analysis. Under common arrivals, contexts, and service times, estimated-SEPT and true SEPT have identical workload processes and busy period boundaries, so only the last busy period can affect the terminal queue-length difference. We decompose the tracking error according to its starting round. For recent busy periods with small prediction error, Lemma 15 together with the workload tail bound gives the first term in Equation (12); old busy periods are controlled directly by the workload tail bound, giving the second. Finally, Lemma 13 controls the contribution of recent busy periods with large prediction error by the third term. The complete proof appears in Appendix H.

6 Conclusion and Future Work

We studied contextual queueing bandits under nonpreemptive service. For a known horizon, our LCP algorithm achieves $\widetilde{O}(\sqrt{d/(\lambda T)})$ regret against the IA optimum, matching our lower bound up to logarithmic factors for $T \geq d^2$. For an unknown horizon, we show that no anytime policy achieves vanishing regret against the finite-horizon IA or WC optimum, and that our estimated-SEPT algorithm tracks true SEPT at rate $\widetilde{O}(\sqrt{d/(\lambda t)})$.

Extending these results to multiple servers is challenging because the workload contributed by an arriving job depends on which server will eventually process it in the future. Thus, unlike in the single-server setting, the workload is policy-dependent, and our common-workload and common-empty-time coupling arguments no longer apply directly. Nevertheless, the techniques developed in this work, together with the multi-server analysis of Bae & Lee (2026), suggest that under appropriate assumptions it may be possible to obtain upper and lower regret bounds of a similar form in the multi-server nonpreemptive setting. Establishing such results is an interesting direction for future work.

REFERENCES

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

Seoungbin Bae and Dabeen Lee. Algorithm for contextual queueing bandits with rate-optimal queue length regret. *arXiv preprint arXiv:2606.09668*, 2026.

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

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

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

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

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

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

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

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

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

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

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

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

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

Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pp. 1085–1093. PMLR, 2017.

Nadav Merlis, Hugo Richard, Flore Sentenac, Corentin Odic, Mathieu Molina, and Vianney Perchet. On preemption and learning in stochastic scheduling. In *International Conference on Machine Learning*, pp. 24478–24516. PMLR, 2023.

Michael Neely. Stochastic network optimization with application to communication and queueing systems, 2010.

Rhonda Righter and J George Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, 3(3):323–333, 1989.

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

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

Richard R Weber, Pravin Varaiya, and Jean Walrand. Scheduling jobs with stochastically ordered processing times on parallel machines to minimize expected flowtime. *Journal of Applied Probability*, 23(3):841–847, 1986.

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

Yueyang Zhong, John R Birge, and Amy R Ward. Learning to schedule in multiclass many-server queues with abandonment. *Operations Research*, 73(6):3085–3103, 2025.

A Appendix Overview

The appendix is organized as follows.

1. Appendix B provides additional discussion of related work.
2. Appendix C proves the workload bounds in Lemma 1 and establishes additional workload and busy-period properties used throughout the subsequent analyses.
3. Appendix D proves the departure-probability and distribution estimation bounds in Lemma 2, together with a simultaneous prediction bound.
4. Appendix E gives the complete proof of the known-horizon upper bound in Theorem 1, together with the supporting lemmas used in the analysis.
5. Appendix F constructs the hard family of instances and proves the regret lower bounds in Theorem 2 for both the WC and IA settings.
6. Appendix G proves Theorem 3, namely, that for both $\mathsf{S} \in \{\mathrm{IA}, \mathrm{WC}\}$, there is no anytime S policy π such that $R_t^{\mathsf{S}}(\pi) \to 0$ as $t \to \infty$, even when the model is fully known.
7. Finally, Appendix H gives the complete proof of the tracking guarantee for the estimated-SEPT algorithm in Theorem 4, together with the supporting lemmas used in the analysis.

B Additional Related Work

Queueing bandits and learning while scheduling. Queueing bandits study learning-while-scheduling problems in which unknown service rates must be learned while controlling queue lengths (Krishnasamy et al., 2016; 2021). Queue-length regret has subsequently been studied under several queueing architectures and information structures (Stahlbuhk et al., 2021; Liang & Modiano, 2018). Related work considers job dispatching with unknown service rates (Choudhury et al., 2021), decentralized learning in queueing systems (Sentenac et al., 2021; Freund et al., 2022), integrated online learning and queue control (Hsu et al., 2022), and learning-enhanced MaxWeight-type scheduling in multi-server systems (Yang et al., 2023). Other work combines Lyapunov-drift scheduling with adversarial bandit feedback (Huang et al., 2024). Most of these models do not simultaneously incorporate job-specific contextual service probabilities and nonpreemptive service commitments, which are the two central features of our setting.

Contextual queueing and logistic bandits. Context-aware queueing models allow service outcomes to depend on job or queue features. Kim & Oh (2024) study a multi-class multi-server queueing-matching problem with feature-dependent multinomial-logit service preferences. The closest models to ours are the contextual queueing bandits of Bae et al. (2026); Bae & Lee (2026). In the stochastic-context setting most directly comparable to ours, jobs arrive with random contexts, and in each round the scheduler selects a job–server pair whose departure probability is determined by a logistic model. The objective is to minimize the expected queue length at a terminal horizon, or equivalently to control queue-length regret relative to a known-model oracle. Their setting is preemptive: after an unsuccessful service attempt, the scheduler may select a different job–server pair in the next round. Consequently, when the service model is known, an optimal policy is work-conserving and selects, in each nonempty round, an available job–server pair with the largest departure probability.

Our nonpreemptive model differs in that, once a job–server pair is selected, the same pair must be served until the job departs. Hence the current decision constrains future actions, and the finite-horizon optimal policy need not admit the same myopic characterization. In addition, the stochastic-context guarantees of Bae et al. (2026); Bae & Lee (2026) use a pointwise traffic-slackness condition and a positive lower bound on a feature second-moment matrix, whereas our main single-server analysis uses the workload condition $\lambda \mathbb{E}[1/p(X)] < 1$ and does not require such a covariance lower bound.

Our logistic service model is also related to generalized linear and logistic bandits (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020; Abeille et al., 2021).

Nonpreemptive stochastic scheduling and learning. Nonpreemptive scheduling has been extensively studied when service-time distributions are known. For fixed collections of jobs, SEPT and related priority rules are optimal under suitable stochastic-order assumptions for classical objectives such as expected flow time (Weber et al., 1986), while other priority rules can maximize the number of successful completions before a due date under additional structural conditions (Righter & Shanthikumar, 1989). More recently, learning has been incorporated into stochastic scheduling. Merlis et al. (2023) study single-machine scheduling when job-type duration distributions are initially unknown, considering both nonpreemptive and preemptive settings and comparing their learning costs. Zhong et al. (2025) study scheduling with unknown model primitives in multiclass many-server queues with abandonment.

Our setting differs from these works in several respects. Jobs arrive stochastically over time and carry individual contextual features, their service-completion probabilities are governed by an unknown logistic model, and the objective is the expected queue length at a finite terminal horizon. Moreover, under nonpreemption, each selected job occupies the server until completion, so the effect of a scheduling decision persists across multiple rounds and interacts with future arrivals. Consequently, neither the classical SEPT optimality results nor existing learning-to-schedule guarantees directly characterize the finite-horizon benchmark studied here.

C Workload properties and stationary workload

This section proves the properties stated in Lemma 1 and establishes several additional properties of the stationary workload associated with the workload recursion. These results will be used later as key ingredients in the regret lower-bound analyses.

We begin by defining some constants introduced in Lemma 1. For $r > 0$ such that $(1 - p_-)e^r < 1$, define

$$M_G(r) := \mathbb{E}e^{rG(X)} = \mathbb{E}\frac{p(X)e^r}{1-(1-p(X))e^r}, \qquad \psi(r) := e^{-r}\big(1 - \lambda + \lambda M_G(r)\big).$$

Since $\psi(0) = 1$ and $\psi'(0) = \rho - 1 < 0$, there exists $r > 0$ satisfying both $\psi(r) < 1$ and $(1 - p_-)e^r < 1$. Fix any such r and define

$$K_r := \frac{1 - \lambda + \lambda M_G(r)}{1 - \psi(r)}, \qquad C_W(r) := \frac{1}{r}\log K_r.$$

These five constants: $r \in (0, 1), \psi(r) \in (0, 1), M_G(r), K_r$ and $C_W(r)$ are what we introduced in Lemma 1.

For convenience, we divide the conclusions of Lemma 1 into three items for the purposes of this appendix. We refer to the bounds $\mathbb{E}e^{rW_t} \le K_r$ and $\mathbb{E}W_t \le C_W(r)$ for every $t \ge 1$ as *Item 1: uniformly bounded expected workload*. We refer to the conditional workload bound in Equation (3), together with its unconditional consequence $\mathbb{E}[W_{u+h}\mathbf{1}\{\tau_0 > h\}] \le (K_r/r)\psi(r)^h$, as *Item 2: workload tail bound*. Finally, we refer to the existence of a stationary workload distribution W that stochastically dominates W_t for every $t \ge 1$ as *Item 3: stationary workload domination*.

C.1 Proofs of Items 1 and 2 of Lemma 1

Proof. Item 1: Uniformly bounded expected workload.
Let $Z_t = A_t G_t$ be the workload brought by the arrival at the end of round t, with $Z_t = 0$ when $A_t = 0$. Conditional on $A_t = 1$ and X_t, we have $Z_t = G_t \sim \mathrm{Geom}(p(X_t))$. Since the arrival indicator is independent of the context and the latent service requirement,

$$B_r := \mathbb{E}e^{rZ_t} = 1 - \lambda + \lambda \mathbb{E}_{X\sim D} e^{rG(X)} = 1 - \lambda + \lambda M_G(r).$$

Here, $G(X)$ is a random variable that satisfies $G(X) \sim \mathrm{Geom}(p(X))$.
If $W_t = w \ge 1$, the workload is reduced by one during round t and increased by Z_t at the end of the round. Hence

$$\mathbb{E}(e^{rW_{t+1}} \mid W_t = w) = \mathbb{E}e^{r(w-1+Z_t)} = e^{r(w-1)}B_r = e^{rw}\psi(r).$$

If $W_t = 0$, no work is served and $W_{t+1} = Z_t$. Splitting according to whether W_t is zero therefore gives

$$\begin{aligned}\mathbb{E}e^{rW_{t+1}} &= \mathbb{E}\left[e^{rW_t}e^{-r}\mathbb{E}(e^{rZ_t} \mid W_t)\mathbf{1}\{W_t > 0\}\right] + \mathbb{E}\left[\mathbb{E}(e^{rZ_t} \mid W_t)\mathbf{1}\{W_t = 0\}\right] \\ &= \psi(r)\mathbb{E}\left(e^{rW_t}\mathbf{1}\{W_t > 0\}\right) + B_r\mathbb{P}(W_t = 0) \\ &\leq \psi(r)\mathbb{E}e^{rW_t} + B_r.\end{aligned}$$

Starting from $W_1 = 0$ and applying this inequality successively,

$$\begin{aligned}\mathbb{E}e^{rW_t} &\leq \psi(r)^{t-1}\mathbb{E}e^{rW_1} + B_r\sum_{j=0}^{t-2}\psi(r)^j \\ &= \psi(r)^{t-1} + B_r\frac{1-\psi(r)^{t-1}}{1-\psi(r)} \\ &= K_r + \psi(r)^{t-1}(1-K_r) \leq K_r.\end{aligned}$$

Here we used $K_r = B_r/(1-\psi(r))$, and the last inequality holds because $B_r \geq 1$ and $0 < \psi(r) < 1$, so $K_r = B_r/(1-\psi(r)) \geq 1$. Using Jensen's inequality for the function $x \mapsto e^{rx}$ then yields

$$\mathbb{E}W_t \leq \frac{1}{r}\log\mathbb{E}e^{rW_t} \leq \frac{1}{r}\log K_r = C_W(r).$$

Item 2: Workload tail bound.
Fix a positive integer u and let $\tau_0 = \inf\{h \geq 0 : W_{u+h} = 0\}$. On $\{\tau_0 > h\}$ the workload is positive through the first h updates. Since the set of events $\{\tau_0 > h\}$ is equal to $\{\tau_0 > h-1\} \cap \{W_{u+h} > 0\}$, we can get the following recurrence inequality.

$$\begin{aligned}&\mathbb{E}\left(e^{rW_{u+h}}\mathbf{1}\{\tau_0 > h\} \mid W_u\right) \\ &\quad= \mathbb{E}\left[\mathbf{1}\{\tau_0 > h-1\}\mathbb{E}\left(e^{rW_{u+h}}\mathbf{1}\{W_{u+h} > 0\} \mid W_{u+h-1}\right)\middle| W_u\right] \\ &\quad\leq \mathbb{E}\left[\mathbb{E}\left(e^{rW_{u+h}} \mid W_{u+h-1}\right)\mathbf{1}\{\tau_0 > h-1\}\middle| W_u\right] \\ &\quad\leq \psi(r)\mathbb{E}\left(e^{rW_{u+h-1}}\mathbf{1}\{\tau_0 > h-1\} \mid W_u\right)\end{aligned}$$

Applying this inequality successively, we get

$$\mathbb{E}\left(e^{rW_{u+h}}\mathbf{1}\{\tau_0 > h\} \mid W_u\right) \leq \mathbb{E}\left(e^{rW_u}\mathbf{1}\{\tau_0 > 0\} \mid W_u\right)\psi(r)^h \leq e^{rW_u}\psi(r)^h.$$

Because $w \leq e^{rw}/r$ for $w \geq 0$, we obtain

$$\mathbb{E}\left(W_{u+h}\mathbf{1}\{\tau_0 > h\} \mid W_u\right) \leq \frac{e^{rW_u}}{r}\psi(r)^h,$$

which is Equation (3). Finally, by taking the expectation of W_u, we obtain a useful inequality like below.

$$\mathbb{E}\left(W_{u+h}\mathbf{1}\{\tau_0 > h\}\right) = \mathbb{E}\left[\mathbb{E}\left(W_{u+h}\mathbf{1}\{\tau_0 > h\} \mid W_u\right)\right] \leq \frac{\mathbb{E}(e^{rW_u})}{r}\psi(r)^h \leq \frac{K_r}{r}\psi(r)^h \tag{13}$$

□

C.2 Upper bound of long busy period probability

We now show that the probability that a busy period lasts for more than h rounds is upper bounded by a function that decays exponentially in h. This requires a slightly different stopped workload because an arrival in the round that empties the system right before that arrival belongs to the next busy period. Suppose a busy period starts in round s (so in round $s-1$, the queue becomes empty after arrival and a new job arrives immediately later), and let G be the service time of its first job (which arrived at the end of round $s-1$). Set $\widetilde{W}_0 = G$ and, for $h \geq 0$, define

$$\widetilde{W}_{h+1} = \begin{cases}\widetilde{W}_h - 1 + Z_{s+h}, & \widetilde{W}_h \geq 2, \\ 0, & \widetilde{W}_h \leq 1.\end{cases}$$

Thus $\widetilde{W}_h$ is the amount of work belonging to this busy period after h service rounds (in other words, at the beginning of round $s+h$), and it is easy to check that $\widetilde{W}_h$ remains zero after it hits zero for the first time. So if we define $\tau_{\rm bp}$ as the smallest h such that the queue $\widetilde{W}_h$ hits zero, then $\{\tau_{\rm bp} > h\} = \{\widetilde{W}_h > 0\}$ holds. In particular, when $\widetilde{W}_h = 1$, the last unit is served in round $s+h$ and the process is stopped just before the arrival at the end of that round is added.

If $\widetilde{W}_h = w \geq 2$, the same calculation as above gives

$$\mathbb{E}\left(e^{r\widetilde{W}_{h+1}}\mathbf{1}\{\widetilde{W}_{h+1} > 0\} \mid \widetilde{W}_h = w\right) = e^{rw}\psi(r).$$

If $w \leq 1$, the left-hand side is zero. Using the same inductive argument that we used to prove Equation (3), we get

$$\mathbb{E}\left(e^{r\widetilde{W}_h}\mathbf{1}\{\tau_{\rm bp} > h\} \mid G\right) \leq e^{rG}\psi(r)^h. \tag{14}$$

Here, we used $W_0 = G$. Since $e^{r\widetilde{W}_h} \geq 1$ on $\{\tau_{\rm bp} > h\}$,

$$\mathbb{P}(\tau_{\rm bp} > h \mid G) \leq e^{rG}\psi(r)^h.$$

Averaging over G and using $\mathbb{E}e^{rG} = M_G(r)$ proves the desired result.

C.3 Proof of Item 3 of Lemma 1 and Stationary Workload Identities

We now prove the existence of a stationary workload, which we refer to above as Item 3 of Lemma 1, using the assumption $\rho < 1$. We then establish several properties of the stationary workload. All of these results are summarized in the following lemma.

Lemma 3. *The recursion in Equation (2) admits a stationary distribution. Let W denote a random variable with this stationary distribution, and call it the stationary workload. Then the process started from zero and satisfying Equation (2) is stochastically dominated by it. Let $A \sim \mathrm{Bern}(\lambda)$ and, independently of A, let $X \sim \mathcal{D}$ and $G \mid X \sim \mathrm{Geom}(p(X))$. With $Z = AG$,*

$$\mathbb{P}(W > 0) = \rho, \qquad \mathbb{E}W = \frac{\mathbb{E}Z^2 + \rho - 2\rho^2}{2(1-\rho)}. \tag{15}$$

Proof. For every integer $w \geq 1$, workload satisfies the following negative-drift inequality.

$$\begin{aligned}\mathbb{E}[W_{t+1} - W_t \mid W_t = w] &= \mathbb{E}[A_t G(X_t)] - 1 \\ &= \lambda\mathbb{E}_{X\sim\mathcal{D},G(X)\sim\mathrm{Geom}((p(X)))}[G(X)] - 1 \\ &= \lambda\mathbb{E}_{X\sim\mathcal{D}}[1/p(X)] - 1 = \rho - 1 < 0.\end{aligned}$$

From every finite state, a sufficiently long run without arrivals reaches zero. The countable-state Foster–Lyapunov criterion therefore gives positive recurrence. The exponential drift in the proof of Lemma 1 gives a finite stationary exponential moment. Because $w \mapsto (w-1)^+ + Z_t$ is nondecreasing, coupling a zero-started process and a stationary process with the same increments gives stochastic domination.

For the stationary workload W, let $W' = (W-1)^+ + Z$, where Z is independent of W. Since $W' \overset{d}{=} W$ by the definition of the stationary workload, taking expectations of W, W' gives

$$0 = \mathbb{E}(W' - W) = -\mathbb{P}(W > 0) + \mathbb{E}Z,$$

and hence $\mathbb{P}(W > 0) = \mathbb{E}Z = \rho$. Moreover, since $(W-1)^+ = W - \mathbf{1}\{W \geq 1\}$ and $((W-1)^+)^2 = W^2 - 2W + \mathbf{1}\{W \geq 1\}$,

$$\mathbb{E}(W-1)^+ = \mathbb{E}W - \rho, \qquad \mathbb{E}((W-1)^+)^2 = \mathbb{E}W^2 - 2\mathbb{E}W + \rho.$$

Squaring $W' = (W-1)^+ + Z$, taking expectations, and using the independence of W and Z gives

$$\begin{aligned}\mathbb{E}W^2 &= \mathbb{E}((W-1)^+)^2 + 2\mathbb{E}((W-1)^+)\mathbb{E}Z + \mathbb{E}Z^2 \\ &= \mathbb{E}W^2 - 2\mathbb{E}W + \rho + 2\rho(\mathbb{E}W - \rho) + \mathbb{E}Z^2.\end{aligned}$$

Canceling $\mathbb{E}W^2$ from both sides and collecting the terms in $\mathbb{E}W$ gives

$$2(1-\rho)\mathbb{E}W = \mathbb{E}Z^2 + \rho - 2\rho^2.$$

Division by $2(1-\rho) > 0$ proves Equation (15). □

D ESTIMATING DEPARTURE PROBABILITIES

This section proves the two bounds in Lemma 2. These inequalities will serve as key ingredients in the regret analysis.

D.1 PROOF OF FIRST ITEM IN LEMMA 2

Proof. Let $z = x^\top \theta$. The gradient and Hessian of the logistic loss are

$$\nabla \ell_\theta(x, y) = (\mu(z) - y)x, \qquad \nabla^2 \ell_\theta(x, y) = \mu'(z) x x^\top.$$

Because $|\mu(z) - y| \le 1$ and $\|x\|_2 \le 1$, $\|\nabla \ell_\theta(x, y)\|_2 \le 1$. Thus the loss is 1-Lipschitz in θ. The parameter ball has diameter at most $2S \le 2\bar{S}$, so the mean-value theorem gives

$$|\ell_\theta(x, y) - \ell_{\theta'}(x, y)| \le 2\bar{S} \qquad \text{for all } \|\theta\|_2, \|\theta'\|_2 \le S.$$

We next verify exp-concavity. Since $|z| \le S \le \bar{S}$, if $y = 0$, then

$$\frac{\mu'(z)}{\mu(z)^2} = \frac{1 - \mu(z)}{\mu(z)} = e^{-z} \ge e^{-\bar{S}};$$

if $y = 1$, then

$$\frac{\mu'(z)}{(1 - \mu(z))^2} = \frac{\mu(z)}{1 - \mu(z)} = e^{z} \ge e^{-\bar{S}}.$$

In either case,

$$\nabla^2 \ell_\theta(x, y) \succeq e^{-\bar{S}} \nabla \ell_\theta(x, y) \nabla \ell_\theta(x, y)^\top,$$

which is the second-order characterization of $e^{-\bar{S}}$-exp-concavity on the convex parameter ball.

Before applying the excess-risk theorem, observe that θ^* minimizes the population logistic risk over the parameter ball. Indeed, for every θ in this ball, since $Y \sim \mathrm{Bern}(p(X))$ conditional on the context X, the following equality about KL divergence holds.

$$\begin{aligned}
\mathbb{E}\ell_\theta(X, Y) - \mathbb{E}\ell_{\theta^*}(X, Y) &= \left(\mathbb{E}_X\left[\log\left(1 + e^{X^\top \theta}\right)\right] - \mathbb{E}\left[p(X) X^\top \theta\right]\right) \\
&\quad - \left(\mathbb{E}_X\left[\log\left(1 + e^{X^\top \theta^*}\right)\right] - \mathbb{E}\left[p(X) X^\top \theta^*\right]\right) \\
&= \mathbb{E}_X\left[\log\left(\frac{1 + e^{X^\top \theta}}{1 + e^{X^\top \theta^*}}\right) - p(X)\log\left(\frac{e^{X^\top \theta}}{e^{X^\top \theta^*}}\right)\right] \\
&= \mathbb{E}_X\left[p(X)\log\left(\frac{1 + e^{-X^\top \theta}}{1 + e^{-X^\top \theta^*}}\right) + (1 - p(X))\log\left(\frac{1 + e^{X^\top \theta}}{1 + e^{X^\top \theta^*}}\right)\right] \\
&= \mathbb{E}_X\left[p(X)\log\left(\frac{p(X)}{\mu(X^\top \theta)}\right) + (1 - p(X))\log\left(\frac{1 - p(X)}{1 - \mu(X^\top \theta)}\right)\right] \\
&= \mathbb{E}_X D_{\mathrm{KL}}\left(\mathrm{Bern}(p(X)) \,\middle\|\, \mathrm{Bern}(\mu(X^\top \theta))\right) \ge 0
\end{aligned} \tag{16}$$

Since we estimate θ^* by $\widehat{\theta}_n$ by using empirical risk minimization, we can apply Theorem 1 of Mehta (2017). In the notation of that theorem, the feasible set is a d-dimensional closed ball of radius S to which the unknown vector θ^* belongs, so its diameter is $2S \le 2\bar{S}$. For every (x, y), the loss function $\ell_\theta(x, y)$ viewed as a function of the variable θ has Lipschitz constant 1 as shown above. Its exp-concavity parameter is $e^{-\bar{S}}$, and for every $(x, y), \theta, \theta^*$, $|\ell_\theta(x, y) - \ell_{\theta^*}(x, y)| \le |\theta - \theta^*| \le 2\bar{S}$ holds because the loss function has Lipschitz constant 1. The theorem contains the factor $2\bar{S} \vee e^{\bar{S}}$, which equals $e^{\bar{S}}$ because $e^{\bar{S}} \ge 2\bar{S}$ for $\bar{S} \ge 1$. Its logarithmic term becomes $\log(16 \cdot 1 \cdot 2\bar{S} \cdot n) = \log(32\bar{S}n)$. Consequently, for every $n \ge 5$, with probability at least $1 - \delta$,

$$\mathbb{E}\ell_{\widehat{\theta}_n}(X, Y) - \mathbb{E}\ell_{\theta^*}(X, Y) \le \frac{8e^{\bar{S}}\left(d\log(32\bar{S}n) + \log(1/\delta)\right) + 1}{n}. \tag{17}$$

By Equation (16), the left-hand side of Equation (17) equals the expected Bernoulli KL divergence over all $X \sim \mathcal{D}$. Pinsker's inequality gives $|p - q| \le \sqrt{D_{\mathrm{KL}}(\mathrm{Bern}(p) \| \mathrm{Bern}(q))/2}$. Applying Pinsker's inequality pointwise for each X and then using the Cauchy-Schwarz inequality, we get the following inequality.

$$\begin{aligned}
\mathbb{E}_X |p(X) - \mu(X^\top \widehat{\theta}_n)| &\le \mathbb{E}_X \sqrt{\frac{1}{2} D_{\mathrm{KL}} \left(\mathrm{Bern}(p(X)) \,\middle\|\, \mathrm{Bern}(\mu(X^\top \widehat{\theta}_n)) \right)} \\
&\le \sqrt{\mathbb{E}_X \left[\frac{1}{2} D_{\mathrm{KL}} \left(\mathrm{Bern}(p(X)) \,\middle\|\, \mathrm{Bern}(\mu(X^\top \widehat{\theta}_n)) \right) \right] \cdot \mathbb{E}_X[1]} \\
&= \sqrt{\frac{1}{2} \mathbb{E}_X D_{\mathrm{KL}} \left(\mathrm{Bern}(p(X)) \,\middle\|\, \mathrm{Bern}(\mu(X^\top \widehat{\theta}_n)) \right)} \\
&= \sqrt{\frac{1}{2} \left(\mathbb{E} \ell_{\widehat{\theta}_n}(X, Y) - \mathbb{E} \ell_{\theta^*}(X, Y) \right)}
\end{aligned}$$

Projection onto $[p_-, p_+]$ cannot increase the absolute error because $p(X)$ already lies in this interval. So $|p(X) - \widehat{p}(X)| \le |p(X) - \mu(X^\top \widehat{\theta}_n)|$ always holds by the definition of $\widehat{p}(X)$, and this gives $\mathbb{E}_X |p(X) - \widehat{p}(X)| \le \sqrt{\frac{1}{2} \left(\mathbb{E} \ell_{\widehat{\theta}_n}(X, Y) - \mathbb{E} \ell_{\theta^*}(X, Y) \right)}$ by the above inequality. Substituting Equation (17) proves desired result. For $n < 5$, the definition gives $\epsilon_n(\delta) = 1$, while $|\widehat{p}_n(X) - p(X)| \le 1$. So this shows first item of Lemma 2 holds for every $n \ge 1$. □

D.2 A SIMULTANEOUS PREFIX BOUND

Corollary 1. *For $\delta \in (0, 1)$, define $\delta_n = 6\delta/(\pi^2 n^2)$. With probability at least $1 - \delta$, simultaneously for every $n \ge 1$,*

$$\mathcal{E}(\widehat{p}_n) \le \epsilon_n(\delta_n).$$

Proof. For each fixed n, $\mathcal{E}(\widehat{p}_n) \le \epsilon_n(\delta_n)$ fails with probability at most δ_n by first item of Lemma 2. So the probability that there exists a natural number such that $\mathcal{E}(\widehat{p}_n) \le \epsilon_n(\delta_n)$ d $\sum_{n \ge 1} \delta_n = \delta$, a union bound proves the result. No stopping-time conditioning is used: all estimators are defined on the same infinite i.i.d. sequence of job-indexed observations $(X_i, Y_i)_{i \ge 1}$. □

D.3 PROOF OF SECOND ITEM IN LEMMA 2

Conditional on $\widehat{p}_n$, the variables $\widehat{p}_n(X'_1), \ldots, \widehat{p}_n(X'_n)$ are i.i.d. with law $F_{\widehat{p}_n}$ and support in $[p_-, p_+]$. The Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956) gives, with probability at least $1 - \delta$,

$$\sup_z |\widehat{F}_n((-\infty, z]) - F_{\widehat{p}_n}((-\infty, z])| \le \sqrt{\frac{\log(2/\delta)}{2n}}.$$

For probability distributions supported on $[p_-, p_+]$, the one-dimensional identity for W_1 proves second item of Lemma 2 as follows.

$$\begin{aligned}
W_1(\widehat{F}_n, F_{\widehat{p}_n}) &= \int_{p_-}^{p_+} |\widehat{F}_n((-\infty, z]) - F_{\widehat{p}_n}((-\infty, z])| \, dz \\
&\le (p_+ - p_-) \sqrt{\frac{\log(2/\delta)}{2n}},
\end{aligned}$$

Furthermore, couple $\widehat{p}_n(X)$ and $p(X)$ using the same independent $X \sim \mathcal{D}$ and applying the definition of Wasserstein distance, we get following inequality.

$$W_1(F_{\widehat{p}_n}, F_p) = \inf_{\substack{(P, \widehat{P}) \text{ coupling} \\ P \sim F,\ \widehat{P} \sim \widehat{F}}} \mathbb{E}\big[|P - \widehat{P}|\big] \le \mathbb{E}|p(X) - \widehat{p}_n(X)| = \mathcal{E}(\widehat{p}_n).$$

E KNOWN-HORIZON ANALYSIS

We now prove Theorem 1. The proof uses three ingredients; Lemma 4, Lemma 5, and Lemma 6. The first is used to compare the empirical Bellman model with the true system. The second ensures that the learning phase can collect a sufficient amount of data to make an empirical distribution and an estimated probability function, with high probability. The third one compares the minimum expected queue length of a queue that starts with fewer remaining rounds to the minimum expected queue length of the T-round benchmark, and this allows us to calculate the regret upper bound correctly.

The first ingredient is the planning-transfer lemma. If the system assumes its unknown probability function and distribution to be $q : \mathbb{R}^d \to [p_-, p_+]$ and $\widehat{F}$ and uses the Bellman recursion that follows $(q, \widehat{F})$ (for convenience of exposition, we write this Bellman recursion as the empirical Bellman recursion) to find out the optimal decision in each round, this empirical Bellman recursion differs from the true optimal system (the system that makes its decision according to the true Bellman recursion) in two ways. First, for the probability function, it replaces $p(X)$ by $q(X)$ for every job. Second, for the distribution, it uses the distribution $\widehat{F}$ which is an estimated distribution of $F_q = \mathrm{Law}(q(X))$, although the real distribution of departure probabilities of jobs is $F_p = \mathrm{Law}(p(X))$. By focusing on these probability/distribution differences, we can find an upper bound of regret between a queue that follows the empirical Bellman recursion and a queue that follows the true Bellman recursion (i.e. a queue with the smallest expected queue length). Here, we upper bound the prediction error of p, q by the $L^1(D)$ distance of p, q, and the distributional error by the Wasserstein distance between $F_q, \widehat{F}$.

For an h-round system that follows the probability function q and starts empty, let $J_h^p(\pi)$ be the expected terminal queue under policy π, and let V_h^p be its minimum over the relevant policy class in the given setting. Then the following lemma holds for the estimated probability function q and empirical distribution $\widehat{F}$.

Lemma 4 (Planning transfer)**.** *Fix $q : \mathbb{R}^d \to [p_-, p_+]$ before an h-round process starts with an empty queue. Let $F_q = \mathrm{Law}(q(X))$, let $\widehat{F}$ be a distribution on $[p_-, p_+]$, and let $\widehat{\pi}$ be an optimal policy following the Bellman recursion defined by $(q, \widehat{F})$. When $\widehat{\pi}$ is run in the true contextual model,*

$$J_h^p(\widehat{\pi}) - V_h^p \le \frac{2h^2}{p_-}\mathbb{E}|q(X) - p(X)| + \frac{(h-1)h(h+1)(h+2)}{12} W_1(F_q, \widehat{F}). \tag{18}$$

The result holds both in the idling-allowed setting and the work-conserving setting.

The second ingredient ensures that the FCFS learning phase has observed a complete prefix of jobs in arrival order. This matters because the statistical events are defined on the job-indexed sequence $(X_i, Y_i)_{i \ge 1}$, independently of the random time at which the algorithm happens to have served enough jobs.

Lemma 5 (Sample availability)**.** *Let N_L be the number of jobs that have started service by the end of round L, and let $n = \lfloor \lambda L/8 \rfloor$. Under FCFS,*

$$\mathbb{P}(N_L < 2n) \le e^{-\lambda L/8} + K_r e^{-r\lambda L/4}. \tag{19}$$

On $\{N_L \ge 2n\}$, the first $2n$ observations (X_i, Y_i) recorded by the algorithm correspond to the first $2n$ jobs in arrival order.

The final ingredient is a monotonicity property of the finite-horizon empty-state value.

Lemma 6 (Empty-state horizon monotonicity)**.** *The minimum terminal cost over policies that may idle is nondecreasing when jobs are added to the initial state. In particular, for systems that start empty,*

$$V_h^{\mathrm{IA}} \le V_{h+1}^{\mathrm{IA}}. \tag{20}$$

The proofs of the above three lemmas will appear after the proof of Theorem 1.

E.1 Proof of Theorem 1

Now, we start to prove Theorem 1.

Proof. Index the jobs in arrival order and let $(X_i, Y_i)_{i\geq 1}$ denote their contexts and first-round departure indicators. This sequence is i.i.d. and does not depend on when the jobs begin service. Define

$$\mathcal{G}_1 = \{N_L \geq 2n\}, \tag{21}$$
$$\mathcal{G}_2 = \{\mathcal{E}(\widehat{p}_n) \leq \epsilon_n(\delta_T)\}, \tag{22}$$
$$\mathcal{G}_3 = \{W_1(\widehat{F}_n, F_{\widehat{p}_n}) \leq \eta_n(\delta_T)\}. \tag{23}$$

The estimator in $\mathcal{G}_2$ uses the first n observations, whereas $\mathcal{G}_3$ uses the estimated probability function and contexts of jobs $n+1, \ldots, 2n$. Neither event is conditioned on $\mathcal{G}_1$. Hence, Lemmas 2 and 5 and a union bound give

$$\mathbb{P}((\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3)^c) \leq \mathbb{P}(\mathcal{G}_1^c) + \mathbb{P}(\mathcal{G}_2^c) + \mathbb{P}(\mathcal{G}_3^c) \leq e^{-\lambda L/8} + K_r e^{-r\lambda L/4} + 2\delta_T. \tag{24}$$

Let $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$ and define the first empty round in the planning window by

$$\tau = \min\{s \in \{L+1, \ldots, T\} : Q_s = 0\},$$

with $\tau = \infty$ when the set is empty (i.e. after starting round L, the first round in which the queue hits zero at the end of that round is round T, or the queue never becomes empty at the end of the remaining rounds). On $\{\tau \leq T\}$, define $h = T - \tau + 1 \leq H$. Conditional on the history at the beginning of round τ, the queue is empty. Thus, on $\mathcal{G} \cap \{\tau \leq T\}$, the arrivals, contexts, and departure indicators from round τ onward are independent of the first $2n$ job-indexed observations used for estimation, since those $2n$ jobs have already completed before the start of round τ.

On $\mathcal{G} \cap \{\tau \leq T\}$, set $q = \widehat{p}_n$. After the queue hits zero at the beginning of round τ, the final queue length Q_{T+1} is only related to the arrival, departure and job selections in rounds $\tau, ..., T$, and all the previous ones are completely independent of it. Since the Bellman policy used after the start of τ is optimal under $(q, \widehat{F}_n)$, applying Lemma 4 and then Lemma 6 shows that the conditional terminal cost $J_h^p(\widehat{\pi})$ is at most

$$V_h^{\mathrm{IA}} + \frac{2h^2}{p_-}\epsilon_n(\delta_T) + \frac{(h-1)h(h+1)(h+2)}{12}\eta_n(\delta_T) \leq V_T^{\mathrm{IA}} + \frac{2H^2}{p_-}\epsilon_n(\delta_T) + \frac{(H-1)H(H+1)(H+2)}{12}\eta_n(\delta_T). \tag{25}$$

On $\mathcal{G}^c \cap \{\tau \leq T\}$, the queue is empty at the start of round τ, and hence there are only $h \leq H$ rounds remaining. During these remaining rounds, at most H arrivals can occur, so $Q_{T+1} \leq H$ regardless of the planning model.

It remains to control the event $\{\tau = \infty\}$. Before planning, Algorithm 1 is work conserving and, in the event $\{\tau = \infty\}$, the queue remains non-empty at the beginning of rounds $L+1, ..., T$; our queue must follow a work-conserving policy starting from round $L+1$ to round T. Recall $Q_s = 0$ if and only if $W_s = 0$ in a work-conserving policy. If we define $\tau_0 = \inf\{j \geq 0 : W_{L+1+j} = 0\}$, on $\{\tau = \infty, W_{T+1} > 0\}$, the workload has not reached zero in any of the first H updates, and hence $\tau_0 > H$. On $\{\tau = \infty, W_{T+1} = 0\}$, the terminal queue is zero. Consequently, Equation (3) gives

$$\begin{aligned}
\mathbb{E}\left(Q_{T+1}\mathbf{1}\{\tau = \infty\}\right) &\leq \mathbb{E}\left(W_{T+1}\mathbf{1}\{\tau = \infty\}\right) \\
&= \mathbb{E}\left(W_{T+1}\mathbf{1}\{\tau = \infty, W_{T+1} > 0\}\right) \\
&\leq \mathbb{E}\left(W_{T+1}\mathbf{1}\{\tau_0 > H\}\right) \\
&\leq \frac{\psi(r)^H}{r}\mathbb{E}e^{rW_{L+1}} \leq \frac{K_r}{r}\psi(r)^H.
\end{aligned} \tag{26}$$

For clarity, let

$$\Gamma_H = \frac{2H^2}{p_-}\epsilon_n(\delta_T) + \frac{(H-1)H(H+1)(H+2)}{12}\eta_n(\delta_T).$$

For an event E, define the event-restricted terminal cost

$$J_T^\pi(E) := \mathbb{E}\left[Q_{T+1}^\pi \mathbf{1}_E\right].$$

Then we have

$$J_T^{\pi^{\mathrm{LCP}}} = J_T^{\pi^{\mathrm{LCP}}}(\mathcal{G} \cap \{\tau \le T\}) + J_T^{\pi^{\mathrm{LCP}}}(\mathcal{G}^c \cap \{\tau \le T\}) + J_T^{\pi^{\mathrm{LCP}}}(\{\tau = \infty\}).$$

We bound these three terms separately by the properties we've shown above.

For each event $E \in \mathcal{G} \cap \{\tau \le T\}$, we define $J_h^p(\widehat{\pi}, E)$ (here $h = T - \tau + 1$) as the conditional terminal cost of the queue when the first $1, ..., \tau - 1$ rounds have happened as E. In other words, $J_h^p(\widehat{\pi}, E) = \mathbb{E}\left(Q_{T+1}^{\pi^{\mathrm{LCP}}} \mid E\right)$. The queue has been cleared by the beginning of round τ, so the planning phase starts from an empty queue with at most H rounds remaining. So by Equation (25), $J_h^p(\widehat{\pi}, E) \le V_T^{\mathrm{IA}} + \Gamma_H$ holds. Hence,

$$\begin{aligned}
J_T^{\pi^{\mathrm{LCP}}}(\mathcal{G} \cap \{\tau \le T\}) &= \mathbb{E}\left[\mathbb{E}\left(Q_{T+1}^{\pi^{\mathrm{LCP}}} \mid E\right) \mathbf{1}(E \in \mathcal{G} \cap \{\tau \le T\})\right] \\
&= \mathbb{E}\left[J_h^p(\widehat{\pi}, E) \mathbf{1}(E \in \mathcal{G} \cap \{\tau \le T\})\right] \\
&\le \mathbb{E}\left[\left(V_T^{\mathrm{IA}} + \Gamma_H\right) \mathbf{1}(E \in \mathcal{G} \cap \{\tau \le T\})\right] \\
&= \left(V_T^{\mathrm{IA}} + \Gamma_H\right) \mathbb{P}(\mathcal{G} \cap \{\tau \le T\}).
\end{aligned}$$

For $\mathcal{G}^c \cap \{\tau \le T\}$, as we've shown that $Q_{T+1}^{\pi^{\mathrm{LCP}}} \le H$ holds regardless of the planning model, we can get the following inequality directly.

$$J_T^{\pi^{\mathrm{LCP}}}(\mathcal{G}^c \cap \{\tau \le T\}) \le H\mathbb{P}(\mathcal{G}^c \cap \{\tau \le T\}).$$

Finally, on $\{\tau = \infty\}$, the following inequality holds by Equation (26).

$$J_T^{\pi^{\mathrm{LCP}}}(\{\tau = \infty\}) \le \frac{K_r}{r}\psi(r)^H.$$

Combining the three bounds gives

$$\begin{aligned}
J_T^{\pi^{\mathrm{LCP}}} &\le \left(V_T^{\mathrm{IA}} + \Gamma_H\right) \mathbb{P}(\mathcal{G} \cap \{\tau \le T\}) + H\mathbb{P}(\mathcal{G}^c \cap \{\tau \le T\}) + \frac{K_r}{r}\psi(r)^H. \\
&\le V_T^{\mathrm{IA}} + \Gamma_H + H\mathbb{P}(\mathcal{G}^c) + \frac{K_r}{r}\psi(r)^H.
\end{aligned}$$

Subtracting V_T^{IA} and using $V_T^{\mathrm{IA}} \ge 0$ yields

$$\begin{aligned}
R_T^{\mathrm{IA}}(\pi^{\mathrm{LCP}}) &= J_T^{\pi^{\mathrm{LCP}}} - V_T^{\mathrm{IA}} \\
&\le \Gamma_H + H\mathbb{P}(\mathcal{G}^c) + \frac{K_r}{r}\psi(r)^H.
\end{aligned}$$

Substituting Equation (24) proves Equation (8).

For $H = \lceil \log^2(eT) \rceil$ and sufficiently large T, $L \ge T/2$ and $n \ge \lambda T/32$. Hence $\epsilon_n(\delta_T) = \widetilde{O}(\sqrt{d/(\lambda T)})$ and $\eta_n(\delta_T) = \widetilde{O}((\lambda T)^{-1/2})$. The factors H^2 and $(H-1)H(H+1)(H+2)$ are polylogarithmic. Since $\psi(r) < 1$, the workload and sampling-tail terms are smaller than any inverse polynomial in T. This proves the fact that regret is $\widetilde{O}(\sqrt{d/(\lambda T)})$. An upper bound on the total number of computations required to evaluate the Bellman recursions during the planning stage is given in Appendix E.2. □

E.2 Total Bellman Computation Cost during the Plan Phase

We now bound the total number of Bellman terms evaluated during the planning stage. Suppose that the planning stage starts with h rounds remaining and with empirical distribution $\widehat{F}_n$, where $h \le H = T - L$. During planning, the algorithm recomputes the empirical Bellman recursion from

the current state. For $s = 1, \dots, h$, let B_s be the number of elementary Bellman terms evaluated when there are s rounds remaining and the server is free. We bound $\sum_{s=1}^{h} B_s$.

It is useful to view one evaluation of the Bellman recursion as a finite directed acyclic computation graph (DAG). A node at depth k represents a reachable state with $s - k$ rounds remaining, and directed edges point from a state to the one-step successor states appearing on the right-hand side of the Bellman recursion (here, Bellman recursion means the recursion formulas of $V_k(c, a), (\mathcal{A}V_k)(c, a)$). The terminal nodes, corresponding to states of 0 rounds remaining, have values given directly by the terminal queue length and therefore do not require any Bellman update. Thus B_s is bounded by the number of outgoing edges from nonterminal nodes in this computation graph.

Since the planning stage starts from an empty queue, at the beginning of a round with s rounds remaining, at most $h - s$ jobs have arrived during the planning stage. Thus the system contains at most $h - s$ jobs. Also, at most $h - s$ fitted departure-probability labels outside the original support of $\widehat{F}_n$ can have appeared as present-only labels. These labels are added to the finite label set used in the Bellman recursion and are assigned zero future-arrival mass, so the distribution $\widehat{F}_n$ itself does not change. Therefore, if m denotes the size of the finite label set used in this Bellman computation, then

$$m \le n + h - s.$$

In this Bellman computation, future arrivals are assumed to have departure-probability labels in this fixed finite label set. Hence every reachable state can be represented as (c, a), where $c \in \mathbb{Z}^m_{\ge 0}$ is the count vector and $a \in \{0, 1, \dots, m\}$ denotes the label of the job in service, with $a = 0$ representing a free server. After k further rounds of the recursion, $k = 0, \dots, s - 1$, the total number of jobs present is at most $h - s + k$. Thus the number of possible count vectors at that depth is at most

$$\#\{c \in \mathbb{Z}^m_{\ge 0} : \|c\|_1 \le h - s + k\} = \binom{m + h - s + k}{m}.$$

For each count vector, the service label a has at most $m + 1$ possible values. Therefore the number of nonterminal reachable states in this Bellman computation is at most

$$(m + 1) \sum_{k=0}^{s-1} \binom{m + h - s + k}{m} \le (m + 1) \binom{m + h}{m + 1},$$

where the last inequality follows from the hockey-stick identity.

For each nonterminal reachable state, evaluating its value requires at most $m + 1$ values of the form $\mathcal{A}V$ at the next step, and each such value depends on at most $m + 1$ next-stage values of V. Thus, in the corresponding computation graph, each nonterminal node has outdegree at most $(m + 1)^2$. Hence,

$$B_s \le (m + 1)^3 \binom{m + h}{m + 1}.$$

Using $m \le n + h - s$, we obtain

$$B_s \le (n + h - s + 1)^3 \binom{n + 2h - s}{h - 1} \le (n + h + 1)^3 \binom{n + 2h - s}{h - 1}$$

Consequently,

$$\begin{aligned} \sum_{s=1}^{h} B_s &\le (n + h + 1)^3 \sum_{s=1}^{h} \binom{n + 2h - s}{h - 1} \\ &\le (n + h + 1)^3 \binom{n + 2h}{h}. \end{aligned} \tag{27}$$

Since $h \le H$, the total number of elementary Bellman terms evaluated during the whole planning stage is at most

$$(n + H + 1)^3 \binom{n + 2H}{H}.$$

When $n = \Theta(T)$ and $H = \Theta(\log^2 T)$, the logarithm of this bound is $O(H \log T) = O(\log^3 T)$.

E.3 Proof of Lemma 4

We first compare two Bellman recursions that may differ in two ways: the two systems may assign different departure probabilities to the same present jobs, and the departure-probability distributions of future arrivals may be different. This slightly more general comparison will be used later with suitable specializations.

Let $V_k^F(s)$ be the minimum expected terminal queue length from state s with k rounds remaining, when the departure probability of each future arrival is drawn from F. Equivalently, $V_k^F(s)$ is the value function obtained from the Bellman recursion for the model that uses the departure probabilities recorded in the current state for present jobs and draws the departure probability of each future arrival from F. Let $Q_k^F(s,a)$ be the corresponding one-step value obtained by taking an admissible action a in state s and then acting optimally. Here admissibility includes the nonpreemptive constraint: if a job is currently in service, the only admissible action is to continue serving that job. For a policy π and an empty initial state, write $J_h^F(\pi)$ for its expected terminal queue length, and write V_h^F for the corresponding minimum over the relevant policy class.

We compare two states s and $\widetilde{s}$ whose currently present jobs are paired, and whose jobs in service are paired as well. A paired present job may carry different departure-probability labels in the two states: we denote its label by p_j in s and by $\widetilde{p}_j$ in $\widetilde{s}$. Define

$$\Delta(s,\widetilde{s}) := \sum_{j \text{ present}} |p_j - \widetilde{p}_j|.$$

Lemma 7. *Let F and $\widetilde{F}$ be distributions of the departure probabilities of future jobs, with $W_1(F,\widetilde{F}) = \eta$. For states s and $\widetilde{s}$ described above and every $k \geq 0$,*

$$|V_k^F(s) - V_k^{\widetilde{F}}(\widetilde{s})| \leq \frac{k(k+1)}{2}\Delta(s,\widetilde{s}) + \frac{(k-1)k(k+1)}{6}\eta. \tag{28}$$

The bound holds both when policies may idle and when they must be work conserving.

Proof. For convenience of explanation, we define $a_k := \frac{k(k+1)}{2}, g_k := \frac{(k-1)k(k+1)}{6}$. We will prove our lemma by induction on k. For $k = 0$, both terminal values equal the common number of jobs in the two states, so Equation (28) holds because $a_0 = g_0 = 0$. Assume it holds with $k-1$ rounds remaining.

Take any action that is admissible in both states. If the action serves a job which is in both of $s, \widetilde{s}$, let p and $\widetilde{p}$ be its departure probabilities and put $\delta = |p - \widetilde{p}|$. Use a common uniform random variable $U \sim \mathrm{Unif}[0,1]$ to generate the two departure indicators. If the selected job has departure probabilities p and $\widetilde{p}$ in the systems started with $s, \widetilde{s}$ respectively, set

$$D = \mathbf{1}\{U \leq p\}, \qquad \widetilde{D} = \mathbf{1}\{U \leq \widetilde{p}\}.$$

Then $D \sim \mathrm{Bern}(p)$, $\widetilde{D} \sim \mathrm{Bern}(\widetilde{p})$, and

$$\mathbb{P}(D \neq \widetilde{D}) = |p - \widetilde{p}| = \delta.$$

Use the same arrival indicator $A \sim \mathrm{Bern}(\lambda)$ in both systems. Conditional on $A = 1$, choose an optimal coupling $(P, \widetilde{P})$ of F and $\widetilde{F}$ such that

$$\mathbb{E}|P - \widetilde{P}| = W_1(F,\widetilde{F}) = \eta.$$

Existence of this coupling comes from the definition of Wasserstein distance. The two arriving jobs are then paired and assigned departure-probability labels P and $\widetilde{P}$, respectively. If $A = 0$, no new job is added to either system.

First, we see the case when the two departure indicators agree, i.e. both 0 or both 1. In both cases, the successor states then contain the same set of jobs, so we can use the induction hypothesis in both cases because they have the same set of jobs. If both indicators are 1, its contribution δ is removed from $\Delta(s,\widetilde{s})$; if both indicators are 0, that contribution is retained. A new arrival adds $|P - \widetilde{P}|$ in both cases. In either case, the expected sum of the probability differences in the successor states

is at most $\Delta(s,\widetilde{s})+\eta$. The induction hypothesis therefore bounds the expected difference of their continuation values by

$$a_{k-1}\big(\Delta(s,\widetilde{s})+\eta\big)+g_{k-1}\eta = a_{k-1}\Delta(s,\widetilde{s})+(a_{k-1}+g_{k-1})\eta. \tag{29}$$

We now see the case when the two departure indicators are different. WLOG let $p\le\widetilde{p}$, so in the queue started with state s the departure indicator was 1, but in the queue started with state s the departure indicator was 0. On their disagreement event, one successor state has exactly one more job than the other. Over the remaining $k-1$ rounds, the terminal queue difference caused by adding or deleting one job in the starting queue is at most k: the initial job-count difference is one, and the two systems can differ by at most one departure in each remaining round. The disagreement event has probability δ, so its contribution is at most $k\delta$. Combining this contribution with Equation (29) and using $\delta\le\Delta(s,\widetilde{s})$ gives

$$\begin{aligned}|Q_k^F(s,a)-Q_k^{\widetilde{F}}(\widetilde{s},a)| &\le (a_{k-1}\Delta(s,\widetilde{s})+(a_{k-1}+g_{k-1})\eta)+k\Delta(s,\widetilde{s}) &(30)\\ &= a_k\Delta(s,\widetilde{s})+g_k\eta, &(31)\end{aligned}$$

where $a_k=a_{k-1}+k$ and $g_k=g_{k-1}+a_{k-1}$. If the common action is to idle, no departure indicator is generated, and at the beginning of the next round the two queues have the same set of jobs. So by using the induction hypothesis again, it's not difficult to show the same bound holds without the $k\delta$ term.

Let a^* minimize $Q_k^F(s,a)$ and let $\widetilde{a}^*$ minimize $Q_k^{\widetilde{F}}(\widetilde{s},a)$ (so that $Q_k^F(s,a^*)=V_k^F(s), Q_k^{\widetilde{F}}(\widetilde{s},a)=V_k^{\widetilde{F}}(\widetilde{s})$). Of course these $a^*,\widetilde{a}^*$ denote the job in service when the server is not free in the starting state s. Applying Equation (31) to $\widetilde{a}^*$ gives

$$V_k^F(s)-V_k^{\widetilde{F}}(\widetilde{s})\le Q_k^F(s,\widetilde{a}^*)-Q_k^{\widetilde{F}}(\widetilde{s},\widetilde{a}^*)\le a_k\Delta(s,\widetilde{s})+g_k\eta.$$

Applying it to a^* with the two models interchanged gives the reverse inequality. This proves Equation (28). □

Setting $s=\widetilde{s}$ (so all the probabilities are equal in the initial queue) in Lemma 7 gives

$$|V_k^F(s)-V_k^{\widetilde{F}}(s)|\le g_k\eta.$$

The common-action bound Equation (31), again with $s=\widetilde{s}$, gives $|Q_k^F(s,a)-Q_k^{\widetilde{F}}(s,a)|\le g_k\eta$ for every admissible action a. Let $\widetilde{\pi}_k(s)$ be an optimal action with k rounds remaining under $\widetilde{F}$ (so we can think that $\widetilde{\pi}_k(s)$ is the optimal decision according to the Bellman recursion defined by the starting queue s, the given success probability for each job, and the future success-probability distribution $\widetilde{F}$). Of course it denotes the job in service when the server is not free in the starting state s. Since $Q_k^{\widetilde{F}}(s,\widetilde{\pi}_k(s))=V_k^{\widetilde{F}}(s)$,

$$\begin{aligned}Q_k^F(s,\widetilde{\pi}_k(s))-V_k^F(s) &= Q_k^F(s,\widetilde{\pi}_k(s))-Q_k^{\widetilde{F}}(s,\widetilde{\pi}_k(s))\\ &\quad+V_k^{\widetilde{F}}(s)-V_k^F(s)\\ &\le 2g_k\eta.\end{aligned}$$

Let S_k be the state with k rounds remaining when $\widetilde{\pi}$ (recall that this policy selects its decision according to the Bellman recursion, which is defined by the given success probability for each job, and the future success-probability distribution $\widetilde{F}$) is run in the model with future-job distribution F. By the definition of Q_k^F,

$$Q_k^F(S_k,\widetilde{\pi}_k(S_k))=\mathbb{E}\left(V_{k-1}^F(S_{k-1})\mid S_k\right).$$

So $\mathbb{E}\left(Q_k^F(S_k,\widetilde{\pi}_k(S_k))\right)=\mathbb{E}\left[\mathbb{E}\left(V_{k-1}^F(S_{k-1})\mid S_k\right)\right]=\mathbb{E}\left(V_{k-1}^F(S_{k-1})\right)$ holds. And also by the definition and assumptions of Lemma 4, $S_h=\varnothing$, $\mathbb{E}\left(V_h^F(\varnothing)\right)=V_h^F$ and $J_h^F(\widetilde{\pi})=\mathbb{E}(S_0)=$

$\mathbb{E}\left(V_0^F(S_0)\right)$ hold. Taking expectations and summing over $k = h, h-1, \ldots, 1$ telescopes:

$$\begin{aligned} J_h^F(\widetilde{\pi}) - V_h^F &= \sum_{k=1}^{h} \left(\mathbb{E}\left(V_{k-1}^F(S_{k-1})\right) - \mathbb{E}\left(V_k^F(S_k)\right)\right) \\ &= \sum_{k=1}^{h} \left(\mathbb{E}\left(Q_k^F(S_k, \widetilde{\pi}_k(S_k))\right) - \mathbb{E}\left(V_k^F(S_k)\right)\right) \\ &= \sum_{k=1}^{h} \mathbb{E}\left(Q_k^F(S_k, \widetilde{\pi}_k(S_k)) - V_k^F(S_k)\right) \\ &\leq 2\eta \sum_{k=1}^{h} g_k = 2\eta \binom{h+2}{4} = \frac{(h-1)h(h+1)(h+2)}{12}\eta. \end{aligned} \tag{32}$$

We next control the change from fitted to true departure probabilities for the same contexts.

Lemma 8. *For $p, q \in [p_-, p_+]$, there is a coupling of $G_p \sim \mathrm{Geom}(p)$ and $G_q \sim \mathrm{Geom}(q)$ such that*

$$\mathbb{P}(G_p \neq G_q) = \frac{|p-q|}{\max(p,q)} \leq \frac{|p-q|}{p_-}.$$

Proof. Let $U_1, U_2, \ldots$ be i.i.d. uniform random variables and set $G_p = \inf\{g : U_g \leq p\}$ and $G_q = \inf\{g : U_g \leq q\}$. If $p \leq q$, then at the first time g such that $U_g \leq q$ holds, the service times agree exactly when $U_g \leq p$, so this event has conditional probability p/q. Thus $\mathbb{P}(G_p \neq G_q) = 1 - p/q = (q-p)/q$. The other case is symmetric. □

For this part, $J_h^q(\pi)$ and V_h^q denote the cost of a fixed policy π and the optimal value when the same contexts are used but every departure probability $p(X)$ is replaced by $q(X)$. Fix q before an h-round process starts empty, and use the same arrivals and contexts in the true and fitted systems (i.e. systems using the same distribution of context vectors, but different departure probability functions). Couple the two service times of each arriving job using Lemma 8. Define $\mathcal{I}_h$ as the set of jobs arriving during these h rounds, and define $\mathcal{H}$ as the set of events such that there is at least one job such that it started to be served in both queues during the h rounds, but the service times of that job in the two queues were different. Then, conditional on their contexts and $\mathcal{I}_h$, each job $X_i \in \mathcal{I}_h$ has service time disagreement probability at most $\frac{|p(X_i)-q(X_i)|}{p_-}$. So a union bound gives

$$\mathbb{P}(\mathcal{H} \mid \mathcal{I}_h) \leq \frac{1}{p_-} \sum_{i \in \mathcal{I}_h} |p(X_i) - q(X_i)|.$$

Because q is fixed before the process begins and each arriving context has distribution $\mathcal{D}$,

$$\begin{aligned} \mathbb{E} \sum_{i \in \mathcal{I}_h} |p(X_i) - q(X_i)| &= \mathbb{E}\left[|\mathcal{I}_h| \cdot \mathbb{E}_{X \sim \mathcal{D}} |p(X) - q(X)|\right] \\ &= \lambda h \, \mathbb{E}_{X \sim \mathcal{D}} |p(X) - q(X)| \leq h\mathcal{E}(q). \end{aligned}$$

Thus the probability that there exists a job that is served in both queues during the h rounds and the service time of that job is different in the two queues is at most $h\mathcal{E}(q)/p_-$. If every pair of service times agrees, the two systems have identical histories and take identical actions under the same nonanticipating policy. On the disagreement event, each terminal queue contains at most the h jobs that could have arrived, so the final queue length difference is at most h. Therefore, for every fixed policy π,

$$|J_h^p(\pi) - J_h^q(\pi)| \leq \frac{h^2}{p_-}\mathcal{E}(q). \tag{33}$$

Let π_p and π_q be optimal in the true and fitted systems, respectively (so $J_h^p(\pi_p) = V_h^p, J_h^q(\pi_q) = V_h^q$). Applying Equation (33) to these two policies gives

$$V_h^p - V_h^q \leq J_h^p(\pi_q) - J_h^q(\pi_q) \leq \frac{h^2}{p_-}\mathcal{E}(q),$$

$$V_h^q - V_h^p \leq J_h^q(\pi_p) - J_h^p(\pi_p) \leq \frac{h^2}{p_-}\mathcal{E}(q).$$

Hence $|V_h^p - V_h^q| \le h^2 \mathcal{E}(q)/p_-$.

Finally, compare the true model that follows (p, F_p) with the intermediate model (q, F_q) and then with the empirical model $(q, \widehat{F})$:

$$J_h^p(\widehat{\pi}) - V_h^p = (J_h^p(\widehat{\pi}) - J_h^q(\widehat{\pi})) + (J_h^q(\widehat{\pi}) - V_h^q) + (V_h^q - V_h^p).$$

The first and third terms are bounded by Equation (33); Equation (32) bounds the middle term with $\eta = W_1(F_q, \widehat{F})$. Adding the three bounds gives Equation (18).

E.4 Proof of Lemma 5

Proof. Let $M_L = \sum_{t=1}^{L} A_t$ be the number of arrivals through round L. Since $M_L \sim \mathrm{Binomial}(L, \lambda)$ and $\mathbb{E}(M_L) = \lambda L$, the multiplicative Chernoff bound with relative deviation $1/2$ gives $\mathbb{P}(M_L < \lambda L/2) \le e^{-\lambda L/8}$. Also, $2n \le \lambda L/4$ by definition. Therefore, on $\{M_L \ge \lambda L/2, N_L < 2n\}$, the number of jobs that have arrived but not begun service by the end of round L is $M_L - N_L > \lambda L/2 - 2n \ge \lambda L/4$. So more than $\lambda L/4$ jobs have arrived but not begun service by the end of round L. Each such job contributes at least one unit to W_{L+1}, so $W_{L+1} > \lambda L/4$. Markov's inequality applied to $e^{rW_{L+1}}$ and Lemma 1 give

$$\begin{aligned} \mathbb{P}(M_L \ge \lambda L/2, N_L < 2n) &\le \mathbb{P}(W_{L+1} > \lambda L/4) \\ &= \mathbb{P}(e^{rW_{L+1}} > e^{r\lambda L/4}) \\ &\le \frac{\mathbb{E}(e^{rW_{L+1}})}{e^{r\lambda L/4}} \le K_r e^{-r\lambda L/4}. \end{aligned} \tag{34}$$

So by a union bound, we can get the desired inequality as follows.

$$\begin{aligned} \mathbb{P}(N_L < 2n) &= \mathbb{P}(M_L \ge \lambda L/2, N_L < 2n) + \mathbb{P}(M_L < \lambda L/2, N_L < 2n) \\ &\le \mathbb{P}(M_L \ge \lambda L/2, N_L < 2n) + \mathbb{P}(M_L < \lambda L/2) \\ &\le K_r e^{-r\lambda L/4} + e^{-\lambda L/8} \end{aligned} \tag{35}$$

Under FCFS, jobs begin service in arrival order. Hence, whenever $N_L \ge 2n$, the first $2n$ recorded observations are $(X_i, Y_i)_{i=1}^{2n}$. □

E.5 Proof of Lemma 6

Proof. Let s be an initial queue and let s^+ be the queue that contains all jobs of s together with some additional jobs. Fix a policy π^+ for s^+. We construct a policy π for s by maintaining a virtual copy of the larger system, and show that by this algorithm its expected terminal queue length is not bigger than that of π^+ started from s^+.

Couple the two systems by the same future arrivals and by the same service times for every common job. Whenever π^+ selects a common job, the smaller system selects the same job; whenever π^+ selects one of the additional jobs, the smaller system idles. Idling is admissible in the policy class considered here. By induction over the number of rounds, we can directly prove that every job in the smaller system is also present in the virtual larger system, and the algorithm always works well in every round. Hence the terminal queue under the constructed policy is not bigger than the terminal queue under π^+. This shows that, if we take an infimum of the expected queue length (at the end) of the queue started with s among all possible policies, it is not bigger than the expected queue length of the queue started with s^+ and using policy ϕ^+. Now, taking an infimum on both sides of the inequality over all possible policies π^+, we can prove monotonicity in the initial state.

Now start an $(h+1)$-round system with an empty queue. No job is available for service in its first round, and the server must idle. After the arrival at the end of that round, the queue state S is either empty or contains one waiting job. Since what we've proved above implies $V_h^{\mathrm{IA}}(S) \ge V_h^{\mathrm{IA}}(\varnothing) = V_h^{\mathrm{IA}}$, the Bellman recursion and the initial-state monotonicity give

$$V_{h+1}^{\mathrm{IA}} = \mathbb{E}V_h^{\mathrm{IA}}(S) \ge \mathbb{E}V_h^{\mathrm{IA}}(\varnothing) = V_h^{\mathrm{IA}}.$$

□

F LOWER BOUNDS

This section constructs the common family of instances and proves Theorem 2. For work-conserving policies, the last round reduces to choosing the faster of two waiting jobs. For policies that may idle, we first show that any policy with small regret reaches the same decision state with probability bounded away from zero.

F.1 THE HARD FAMILY

We first define the distribution $\mathcal{D}$, job context vectors, probability constants (λ, p_-, p_+), and parameter vector θ^* that will be used to make a lower bound of regret as follows.

1. Job types and distribution $\mathcal{D}$:
   Let $m = d - 2$, and let there be d types of jobs in our distribution: a dummy job C_D, a baseline job C_0, and m candidate jobs $C_1, \dots, C_m$. Conditional on an arrival, let's define each job's arrival probability
   $$\mathbb{P}(C_D) = \nu_D, \quad \mathbb{P}(C_0) = \nu_0, \quad \mathbb{P}(C_j) = \frac{\nu_C}{m}.$$
   where ν_D, ν_0, ν_C are three nonnegative real numbers that sum to 1. So $\mathcal{D}$ is the distribution that consists of jobs $C_D, C_0, C_1, \dots, C_m$ with arrival probability as above.
2. Job context vectors and probability constants:
   For the d-dimensional standard basis set $(e_j)_{j=1}^d$, define
   $$x_D = e_1, \qquad x_0 = \frac{e_2}{\sqrt{2}}, \qquad x_{C_j} = \frac{e_2 + e_{j+2}}{\sqrt{2}}.$$
   Choose fixed z_D, z_0, a_0, S $(a_0, S > 0)$ such that $z_D > z_0 + a_0$ and $S^2 > z_D^2 + 2z_0^2$, and set
   $$c_S = \sqrt{\frac{S^2 - z_D^2 - 2z_0^2}{2}}, \qquad \ell_0 = \min_{|z - z_0| \le a_0} \mu'(z) > 0.$$
   For example, we can choose $z_D = \frac{S}{2}, z_0 = -\frac{S}{2}, a_0 = 0.99S$.

   Now, we choose fixed bounds $p_- \le \mu(z_0 - a_0)$ and $p_+ \ge \mu(z_D)$ with $0 < p_- < p_+ < 1$, and a fixed $\lambda > 0$ small enough that the following inequality holds.
   $$\lambda < p_-, \qquad \frac{\lambda(1-\lambda)}{p_- - \lambda} \le \frac{(1 - \mu(z_D))^4}{16} \tag{36}$$
   We can always choose such $\lambda > 0$ since the functions $\lambda \mapsto \lambda, \lambda \mapsto \frac{\lambda(1-\lambda)}{p_- - \lambda}$ (defined on $0 < \lambda < p_-$) go to 0 as $\lambda \to 0^+$.
3. Parameter vector θ^*:
   For fixed constants $\alpha > 0$ that will be determined later, let $a_T = \min\left(a_0, \frac{c_S}{\sqrt{m}}, \alpha\sqrt{\frac{m}{T}}\right)$.
   Next, for each $\omega \in \{-1, +1\}^m$, set the parameter vector corresponding to ω as
   $$\theta_\omega = (z_D, \sqrt{2}z_0, \sqrt{2}\omega_1 a_T, \dots, \sqrt{2}\omega_m a_T).$$
   Then the dummy job has departure probability $p_D = \mu(z_D)$, the baseline job has departure probability $p_0 = \mu(z_0)$, and the candidate job C_j has departure probability $p_{C_j} = \mu(z_0 + \omega_j a_T)$.

The following lemma shows that the above definitions satisfy Assumption 1.

Lemma 9. *The above construction satisfies Assumption 1 for every d, T.*

Proof. Since $a_T \le \frac{c_S}{\sqrt{m}}$ by definition and $w_i^2 = 1$ for all $i \in [m]$,
$$\begin{aligned}\|\theta_\omega\|_2^2 &= z_D^2 + 2z_0^2 + 2(w_1^2 + \dots + w_m^2)a_T^2 \\ &= z_D^2 + 2z_0^2 + 2ma_T^2 \\ &\le z_D^2 + 2z_0^2 + 2c_S^2 = S^2.\end{aligned}$$

Since $a_T \le a_0$ by definition, $z_0 + w_j a_T \in \{z_0 - a_T, z_0 + a_T\} \subset [z_0 - a_0, z_D]$ holds. Also $z_0, z_D \in [z_0 - a_0, z_D]$, so every job's departure probability lies in $[\mu(z_0 - a_0), \mu(z_D)] \subset [p_-, p_+]$. Finally,

$$\rho_\omega = \lambda \mathbb{E}_\omega \frac{1}{p(X)} \le \frac{\lambda}{p_-} < 1.$$

□

F.2 Information in the observed history

Fix an adaptive learning policy π. For each $\omega \in \{-1, +1\}^m$, let P_ω^T denote the probability law induced by running π for T rounds on the instance with parameter θ_ω. This law is defined on the complete interaction history through round T, including all arrival indicators and contexts, the learner's internal randomization and resulting actions, and all observed departure indicators. Thus, for any event $\mathcal{E}$ determined by the interaction through round T, $P_\omega^T(\mathcal{E})$ is the probability that $\mathcal{E}$ occurs under parameter θ_ω and policy π. We suppress the dependence of P_ω^T on the fixed policy π in the notation.
Let $\omega^{(j)}$ be the vector obtained by changing the sign of coordinate j (ω_j) of ω, and let $N_j(T)$ be the number of rounds through T in which a candidate-j job is served.

Lemma 10. *For every adaptive nonpreemptive policy, including policies that may idle,*

$$D_{\mathrm{KL}}(P_\omega^T \| P_{\omega^{(j)}}^T) \le \frac{a_T^2}{2} \mathbb{E}_\omega N_j(T) \le \frac{\lambda \nu_C T a_T^2}{2mp_-}, \tag{37}$$

$$\mathrm{TV}(P_\omega^T, P_{\omega^{(j)}}^T) \le a_T \sqrt{\frac{\lambda \nu_C T}{4mp_-}} \le \frac{\alpha}{2} \sqrt{\frac{\lambda \nu_C}{p_-}}. \tag{38}$$

Proof. Only the departure indicators observed while serving candidate-j jobs have different conditional distributions under ω and $\omega^{(j)}$. The two neighboring logits are $z_0 + \omega_j a_T$ and $z_0 - \omega_j a_T$. If we define $A(z) = \log(1 + e^z)$, then $A'(z) = \mu(z)$ and $D_{\mathrm{KL}}\left(\mathrm{Bern}(\mu(u)) \,\|\, \mathrm{Bern}(\mu(v))\right) = A(v) - A(u) - A'(u)(v - u)$ hold, and the following inequality can be obtained from Taylor's theorem.

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \,\|\, \mathrm{Bern}(\mu(z_0 - \omega_j a_T))\right) \\
&= A(z_0 + \omega_j a_T) - A(z_0 - \omega_j a_T) - 2A'(z_0 - \omega_j a_T)\omega_j a_T \\
&= \frac{1}{2} A''(\varepsilon)(2\omega a_T)^2 \le 2 \left(\sup_z \frac{e^z}{(1 + e^z)^2} \right) a_T^2 \le \frac{a_T^2}{2}
\end{aligned}$$

Here, ε is some real number in $(z_0 - \omega_j a_T, z_0 + \omega_j a_T)$ obtained by second order Taylor's theorem, and we used $A''(\varepsilon) = \frac{e^\varepsilon}{(1+e^\varepsilon)^2} \le \sup_z \frac{e^z}{(1+e^z)^2} = \frac{1}{4}$.

Let $I_t^{(j)}$ indicate that a candidate-j job is served in round t. Let $\mathcal{F}_t^-$ denote the information available immediately before the departure indicator D_t is observed; this includes the complete history through round $t - 1$, as well as the learner's randomization and resulting action in round t. In particular, $I_t^{(j)}$ is $\mathcal{F}_t^-$-measurable.

Conditional on $\mathcal{F}_t^-$, the two neighboring instances have identical departure-indicator laws when $I_t^{(j)} = 0$. When $I_t^{(j)} = 1$, the conditional laws of D_t under ω and $\omega^{(j)}$ are, respectively,

$$\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \quad \text{and} \quad \mathrm{Bern}(\mu(z_0 - \omega_j a_T)).$$

All other conditional kernels, including those governing arrivals, contexts, learner randomization, and actions, are identical under the two instances. Therefore, the KL chain rule gives

$$\begin{aligned}
D_{\mathrm{KL}}(P_\omega^T \| P_{\omega^{(j)}}^T) &= \sum_{t=1}^{T} \mathbb{E}_\omega \left[I_t^{(j)} D_{\mathrm{KL}} \left(\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \,\|\, \mathrm{Bern}(\mu(z_0 - \omega_j a_T)) \right) \right] \\
&= \left(\sum_{t=1}^{T} \mathbb{E}_\omega I_t^{(j)} \right) D_{\mathrm{KL}} \left(\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \,\|\, \mathrm{Bern}(\mu(z_0 - \omega_j a_T)) \right) \\
&= \left(\mathbb{E}_\omega \sum_{t=1}^{T} I_t^{(j)} \right) D_{\mathrm{KL}} \left(\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \,\|\, \mathrm{Bern}(\mu(z_0 - \omega_j a_T)) \right) \\
&= E_\omega N_j(T) \cdot D_{\mathrm{KL}} \left(\mathrm{Bern}(\mu(z_0 + \omega_j a_T)) \,\|\, \mathrm{Bern}(\mu(z_0 - \omega_j a_T)) \right) \\
&\leq \frac{a_T^2}{2} \mathbb{E}_\omega N_j(T),
\end{aligned}$$

where $N_j(T) = \sum_{t=1}^{T} I_t^{(j)}$. This proves the first inequality in Equation (37).

Pre-sample the geometric service time of every candidate-j job. Regardless of the scheduling decisions, the number of rounds that one specific candidate-j job is served by time T cannot exceed its service time. Therefore $N_j(T) \leq \sum_{t=1}^{T} \mathbf{1}\{A_t = 1,\, X_t = C_j\} G_t$, and the following holds.

$$\begin{aligned}
\mathbb{E}_\omega N_j(T) &\leq \mathbb{E}_\omega \left(\sum_{t=1}^{T} \mathbf{1}\{A_t = 1,\, X_t = C_j\} G_t \right) \\
&= \sum_{t=1}^{T} \mathbb{E}_\omega \left(\mathbf{1}\{A_t = 1,\, X_t = C_j\} G_t \right) \\
&= T \cdot \mathbb{E}_\omega \left(\mathbf{1}\{A_t = 1,\, X_t = C_j\} G_t \right) \\
&= T \cdot \lambda \frac{\nu_C}{m} \mathbb{E}_\omega (G_t \mid X_t = C_j) \\
&= \frac{\lambda \nu_C T}{m p(x_{C_j})} \leq \frac{\lambda \nu_C T}{m p_-}
\end{aligned}$$

This proves the second inequality in Equation (37).

For Equation (38), using Pinsker's inequality $\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \| Q)}$ gives the first bound in Equation (38), and $a_T \leq \alpha \sqrt{m/T}$ gives the second. □

For later use, let $P_\omega^t, P_\omega^{t,\mathrm{pre}}$ denote the law of the history up to the end of round t, and the law of the history after service in round t and before the arrival at the end of that round. This history is a measurable function of the complete observations through round t, so the definition of total variance between two distributions $P_\omega^t, P_{\omega^{(j)}}^t$ gives

$$\mathrm{TV}(P_\omega^{t,\mathrm{pre}}, P_{\omega^{(j)}}^{t,\mathrm{pre}}) \leq \mathrm{TV}(P_\omega^t, P_{\omega^{(j)}}^t). \tag{39}$$

Define

$$\Delta_0 = 1 - \frac{\lambda}{p_-}, \qquad \zeta = \lambda^3 \nu_D \nu_0 \nu_C p_D (1 - p_D), \qquad g_D = (1 - \lambda)^2 p_D (1 - p_D)^2. \tag{40}$$

It's easy to check that these three constants are all positive. Choose the fixed $\alpha > 0$ which is introduced in Appendix F.1 so that following inequality holds.

$$\alpha \leq \sqrt{\frac{p_-}{\lambda \nu_C}} \min \left(\frac{\zeta \Delta_0}{2}, \frac{3\zeta}{16} \right)$$

This condition was chosen to make the inequality $\frac{\alpha}{2} \sqrt{\frac{\lambda \nu_C}{p_-}} \leq \min \left(\frac{\zeta \Delta_0}{4}, \frac{3\zeta}{32} \right)$ hold. So by this condition, $\mathrm{TV}(P_\omega^T, P_{\omega^{(j)}}^T) \leq \frac{\zeta \Delta_0}{4}, \frac{3\zeta}{32}$ holds.

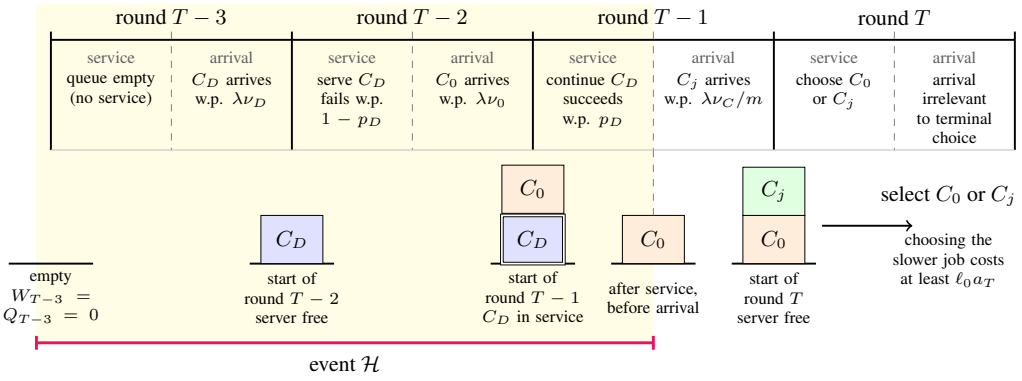


Figure 1: The final four-round construction used in the lower-bound proof. Solid vertical lines mark round boundaries, and dashed lines separate service from the end-of-round arrival. The shaded region corresponds to $\mathcal{H}$; the candidate-j arrival is appended afterward.

F.3 Proof of Theorem 2 for WC case

Proof. Under every work-conserving policy, the workload started with $W_1 = 0$ and satisfies Equation (2) is stochastically dominated by stationary workload(Lemma 1). By Lemma 3, for every ω,

$$\mathbb{P}_\omega(W_{T-3} = 0) \geq \mathbb{P}_\omega(W = 0) = 1 - \rho_\omega \geq \Delta_0 \tag{41}$$

where W denotes the stationary distribution's workload. The above inequality holds because the zero-started workload W_t is stochastically dominated by the stationary workload W for every $t \geq 1$.

Starting from this event $\{W_{T-3} = 0\}$, consider the following arrivals and departures: First, the server must idle since $W_{T-3} = 0$, and then a dummy arrives at the end of round $T-3$. The server starts to process that dummy job in round $T-2$, but it fails to succeed. Next, a baseline job arrives at the end of round $T-2$. In round $T-1$, the server must do the failed dummy job, and it succeeds with it in round $T-1$. Let $\mathcal{H}$ denote this event before the arrival at the end of round $T-1$. The above Figure 1 illustrates this event. By independence of events,

$$\begin{aligned}\mathbb{P}_\omega(\mathcal{H}) &= \mathbb{P}_\omega(W_{T-3} = 0) \cdot (\lambda\nu_D) \cdot (1 - p_D) \cdot (\lambda\nu_0) \cdot p_D \\ &\geq \Delta_0 \lambda^2 \nu_D \nu_0 p_D (1 - p_D) = \frac{\zeta \Delta_0}{\lambda \nu_C}.\end{aligned} \tag{42}$$

Conditional on $\mathcal{H}$, candidate C_j arrives at the end of round $T-1$ with probability $\lambda\nu_C/m$. The server is free in round T and must choose between C_0 and C_j. If $\omega_j = +1$, C_j has the higher departure probability; if $\omega_j = -1$, C_0 does. In either case, the mean-value theorem gives

$$|p_{C_j} - p_0| = |\mu(z_0 + \omega_j a_T) - \mu(z_0)| \geq \ell_0 a_T.$$

Thus, serving the slower job rather than the faster one increases the expected terminal queue length by $|p_{C_j} - p_0| \geq \ell_0 a_T$.

Fix j and pair an instance ω with $\omega^{(j)}$. Given the history just before the arrival at the end of round $T-1$, let the Markov kernel K_j append a candidate-j arrival and then apply the learner's randomized decision in round T. Write $\overline{P}_{\omega,j} = P_\omega^{T-1,\mathrm{pre}} K_j$, and let $\mathcal{S}_j(\omega)$ be the event that, conditional on $\mathcal{H}$ and on C_j arriving at the end of round $T-1$, the decision selects the optimal job between C_0 and C_j in round T for the parameter vector θ_ω. So if $\omega_j = 1$ then $\mathcal{S}_j(\omega)$ is the event that the decision selects C_j, and if $\omega_j = -1$ then $\mathcal{S}_j(\omega)$ is the event that the decision selects C_0. This directly shows that, within this conditional experiment, $\mathcal{S}_j(\omega)^c = \mathcal{S}_j(\omega^{(j)})$ since $[\omega^{(j)}]_j = -\omega_j$. By these definitions,

data processing for K_j, followed by Equation (39), gives

$$\begin{aligned}
&\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c) + \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega^{(j)})^c) \\
&\quad = \overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c) + \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega)) \\
&\quad = \overline{P}_{\omega,j}(\mathcal{H}) - \overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)) + \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega)) \\
&\quad = \overline{P}_{\omega,j}(\mathcal{H}) - \left(\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)) - \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega))\right) \\
&\quad \geq \mathbb{P}_\omega(\mathcal{H}) - \mathrm{TV}(\overline{P}_{\omega,j}, \overline{P}_{\omega^{(j)},j}) \\
&\quad \geq \mathbb{P}_\omega(\mathcal{H}) - \mathrm{TV}(P_\omega^{T-1,\mathrm{pre}}, P_{\omega^{(j)}}^{T-1,\mathrm{pre}}) \\
&\quad \geq \mathbb{P}_\omega(\mathcal{H}) - \mathrm{TV}(P_\omega^{T-1}, P_{\omega^{(j)}}^{T-1}).
\end{aligned}$$

In more detail, the first inequality follows from the definition of total variation, while the second follows from the data-processing inequality $\mathrm{TV}(PK, QK) \leq \mathrm{TV}(P, Q)$, which holds for any probability measures P and Q on the same measurable space and any common Markov kernel K from that space to another measurable space. We also used the identity $\overline{P}_{\omega,j}(\mathcal{H}) = P_\omega(\mathcal{H})$, which holds because K_j leaves the pre-arrival history unchanged.

In the original experiment, a candidate-j arrival occurs with probability $\lambda\nu_C/m$, independently of the history immediately before that arrival; conditional on this arrival, the action law is the one produced by K_j. Multiplying by the arrival probability and applying Equations (38) and (42), the paired probability of a wrong final action is at least

$$\begin{aligned}
\frac{\lambda\nu_C}{m}\left(\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c) + \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega^{(j)})^c)\right) &\geq \frac{\lambda\nu_C}{m}\left(\mathbb{P}_\omega(\mathcal{H}) - \mathrm{TV}(P_\omega^{T-1}, P_{\omega^{(j)}}^{T-1})\right) \\
&\geq \frac{\lambda\nu_C}{m}\left(\frac{\zeta\Delta_0}{\lambda\nu_C} - \frac{\zeta\Delta_0}{4}\right) \\
&= \frac{\zeta\Delta_0}{m}\left(1 - \frac{\lambda\nu_C}{4}\right) \geq \frac{3\zeta\Delta_0}{4m}.
\end{aligned} \tag{43}$$

Now let $p_{\mathrm{err}}(\omega) = \sum_{j=1}^m \frac{\lambda\nu_C}{m}\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c)$. So this is the probability that, for the system with parameter vector θ_ω, the event $\mathcal{H}$ occurs, a candidate job arrives at the end of round $T-1$, and the system chooses the slower job in the final round. Now, by a double-counting argument, the following equalities and inequality hold. More specifically, the third equality follows by applying the double-counting argument to the vertices and edges of the m-dimensional hypercube, whose vertex set is $\{-1,1\}^m$.

$$\begin{aligned}
\mathbb{E}_{\omega\sim\mathrm{Unif}(\{-1,1\}^m)}[p_{\mathrm{err}}(\omega)] &= \frac{1}{2^m}\sum_{\omega\in\{-1,1\}^m} p_{\mathrm{err}}(\omega) \\
&= \frac{1}{2^m}\sum_{\omega\in\{-1,1\}^m}\sum_{j=1}^m \frac{\lambda\nu_C}{m}\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c) \\
&= \frac{1}{2^m}\sum_{\substack{\{\omega,\omega^{(j)}\}\text{ unordered pair} \\ \omega,\omega^{(j)}\in\{-1,1\}^m}} \frac{\lambda\nu_C}{m}\left(\overline{P}_{\omega,j}(\mathcal{H}, \mathcal{S}_j(\omega)^c) + \overline{P}_{\omega^{(j)},j}(\mathcal{H}, \mathcal{S}_j(\omega^{(j)})^c)\right) \\
&\geq \frac{1}{2^m}\sum_{\substack{\{\omega,\omega^{(j)}\}\text{ unordered pair} \\ \omega,\omega^{(j)}\in\{-1,1\}^m}} \frac{3\zeta\Delta_0}{4m} \\
&= \frac{3\zeta\Delta_0}{4m}\cdot\frac{m2^{m-1}}{2^m} = \frac{3\zeta\Delta_0}{8}
\end{aligned} \tag{44}$$

Therefore $\max_{\omega\in\{-1,1\}^m} p_{\mathrm{err}}(\omega) \geq \mathbb{E}_{\omega\sim\mathrm{Unif}(\{-1,1\}^m)}[p_{\mathrm{err}}(\omega)] \geq \frac{3\zeta\Delta_0}{8}$ holds, so if we choose $\omega_0 \in \arg\max_{\omega\in\{-1,1\}^m} p_{\mathrm{err}}(\omega)$, then $p_{\mathrm{err}}(\omega_0) \geq \frac{3\zeta\Delta_0}{8}$. Changing only the final action to the

better member of $\{C_0, C_j\}$ is a feasible work-conserving comparison policy. Thus the queue-length regret for the instance with parameter vector θ_{ω_0} is at least $|p_0 - p_{C_j}| \cdot p_{\mathrm{err}}(\omega_0) \geq \ell_0 a_T \cdot \frac{3\zeta\Delta_0}{8}$, and this instance has regret at least this value.

So this means that for any fixed policy π, there must be some $\omega \in \{-1, 1\}^m$ such that, if the unknown parameter vector were actually θ_ω, then the regret would be at least $\ell_0 a_T \cdot \frac{3\zeta\Delta_0}{8}$. Since this works for every work-conserving policy π, this shows that whenever we use a work-conserving policy for T rounds, there exist a distribution $\mathcal{D}$ and a parameter vector θ^* such that the regret is bounded below by $\ell_0 a_T \cdot \frac{3\zeta\Delta_0}{8}$.

Finally, define $b_{d,T} = d^{-1/2} \wedge \sqrt{d/T}$. Since $b_{d,T} \leq d^{-1/2} \leq 1/\sqrt{3}$ and $\frac{d}{3} \leq m = d - 2 < d$ hold for $d \geq 3$,

$$a_0 \geq \sqrt{3} a_0 b_{d,T}, \qquad \frac{c_S}{\sqrt{m}} > \frac{c_S}{\sqrt{d}} \geq c_S b_{d,T}.$$

If $T \geq d^2$, then $b_{d,T} = \sqrt{d/T}$ and $\sqrt{m/T} \geq b_{d,T}/\sqrt{3}$. If $T < d^2$, then $b_{d,T} = d^{-1/2}$ and $\sqrt{m/T} > \sqrt{m}/d \geq 1/\sqrt{3d} = b_{d,T}/\sqrt{3}$. Consequently,

$$a_T \geq \min\left(\sqrt{3} a_0, c_S, \frac{\alpha}{\sqrt{3}}\right) b_{d,T}.$$

Therefore, the theorem follows with

$$c_{\mathrm{WC}} = \frac{3\ell_0\zeta\Delta_0}{8} \min\left(\sqrt{3} a_0, c_S, \frac{\alpha}{\sqrt{3}}\right),$$

and this lower bound is $\Omega\left(\min\left\{\frac{1}{\sqrt{d}}, \sqrt{\frac{d}{T}}\right\}\right)$. □

F.4 Two facts for policies that may idle

Lemma 11. *Under $\lambda < p_-$,*

$$\sup_T V_T^{\mathrm{IA}} \leq \sup_T V_T^{\mathrm{WC}} \leq \frac{\lambda(1-\lambda)}{p_- - \lambda}.$$

Proof. Run FCFS. In every round with a job in service, couple its departure indicator with a Bernoulli-p_- indicator so that a departure in the comparison system also occurs in the true system. The true queue is then stochastically dominated by the late-arrival Geo/Geo/1 queue

$$\widetilde{Q}_{t+1} = \left(\widetilde{Q}_t - S_t\right)^+ + A_t, \qquad \mathbb{P}(S_t = 1 \mid \widetilde{Q}_t > 0) = p_-.$$

Its drift outside zero is $\lambda - p_- < 0$, so it has a stationary version and its zero-started process is stochastically dominated by the stationary version's queue. Let $(\widetilde{Q}, S, A)$ have the stationary one-round distribution. Taking expectations in the recursion gives

$$\begin{aligned} 0 &= \mathbb{E}\left[\left((\widetilde{Q} - S)^+ + A\right)\right] - \mathbb{E}\widetilde{Q} \\ &= \mathbb{E}\left[(A - S) \cdot \mathbf{1}\left(\widetilde{Q} > 0\right)\right] + \mathbb{E}\left[A \cdot \mathbf{1}(\widetilde{Q} = 0)\right] \\ &= \mathbb{E}A - \mathbb{E}\left[S \cdot \mathbf{1}\left(\widetilde{Q} > 0\right)\right] \\ &= \lambda - p_- \mathbb{P}(\widetilde{Q} > 0) \end{aligned} \tag{45}$$

and therefore $\mathbb{E}\left[S \cdot \mathbf{1}\left(\widetilde{Q} > 0\right)\right] = \mathbb{E}A = \lambda$ and $\mathbb{P}(\widetilde{Q} > 0) = \lambda/p_-$. Since $\mathbb{E}(S \mid \widetilde{Q}) = \mathbb{P}(S_t = 1 \mid \widetilde{Q}_t > 0) = p_-$ for $\widetilde{Q} > 0$,

$$\begin{aligned}\mathbb{E}(\widetilde{Q}S) &= \mathbb{E}\left[\widetilde{Q}S \cdot \mathbf{1}(\widetilde{Q} > 0)\right] + \mathbb{E}\left[\widetilde{Q}S \cdot \mathbf{1}(\widetilde{Q} = 0)\right] \\ &= \mathbb{E}\left[\widetilde{Q}S \cdot \mathbf{1}(\widetilde{Q} > 0)\right] \\ &= \mathbb{E}\left[\mathbb{E}(S \mid \widetilde{Q})\widetilde{Q} \cdot \mathbf{1}(\widetilde{Q} > 0)\right] \\ &= p_-\mathbb{E}\left[\widetilde{Q} \cdot \mathbf{1}(\widetilde{Q} > 0)\right] = p_-\mathbb{E}\widetilde{Q}\end{aligned} \tag{46}$$

Moreover, A is independent of $(\widetilde{Q}, S)$, $A^2 = A$, and $S^2 = S$. Expanding the square of the stationary recursion gives

$$\begin{aligned}0 &= \mathbb{E}\left[\left((\widetilde{Q} - S)^+ + A\right)^2\right] - \mathbb{E}\widetilde{Q}^2 \\ &= \mathbb{E}\left[S \cdot \mathbf{1}(\widetilde{Q} > 0)\right] + \mathbb{E}A - 2\mathbb{E}\left[\widetilde{Q}S \cdot \mathbf{1}(\widetilde{Q} > 0)\right] + 2\mathbb{E}\left[\widetilde{Q}A \cdot \mathbf{1}(\widetilde{Q} > 0)\right] - 2\mathbb{E}\left[SA \cdot \mathbf{1}(\widetilde{Q} > 0)\right] \\ &= 2\lambda - 2p_-\mathbb{E}\widetilde{Q} + 2\lambda\mathbb{E}\widetilde{Q} - 2\lambda^2 \\ &= -2(p_- - \lambda)\mathbb{E}\widetilde{Q} + 2\lambda(1 - \lambda).\end{aligned}$$

Thus $\mathbb{E}\widetilde{Q} = \frac{\lambda(1-\lambda)}{p_- - \lambda}$, and this shows that the expected final queue length of the FCFS policy is at most $\frac{\lambda(1-\lambda)}{p_- - \lambda}$ by stochastic dominance. FCFS is feasible for both benchmark classes, so the result is proved. □

Lemma 12. *Suppose the server is free, only a dummy job with departure probability p_D is waiting, three rounds remain, and every present or future job has departure probability at most p_D. For every policy that idles in the first round, there is a policy that serves the dummy immediately and has expected terminal queue smaller by at least g_D.*

Proof. We first establish a two-round comparison. Suppose a job with departure probability q is waiting and every other present or future job has departure probability at most q. If this job is served immediately, let M be the expected highest departure probability available in the final round, conditional on its departure in the first round. The expected number of departures is

$$q + (1 - q)q + qM = 2q - q^2 + qM.$$

Now suppose another waiting job has departure probability $p \le q$ and is served first. If it departs, the q job is available in the last round; if it remains, nonpreemption requires its service to continue. The expected number of departures is therefore

$$p + (1 - p)p + pq = 2p - p^2 + pq.$$

Because the p job remains available after a first-round departure of the q job, $M \ge p$. Subtracting the two expressions gives

$$(2q - q^2 + qM) - (2p - p^2 + pq) = (q - p)(2 - q - p) + q(M - p) \ge 0.$$

Serving the q job also dominates idling: after idling, at most one departure can occur, with probability at most q, whereas $2q - q^2 + qM \ge q$.

Return to the three-round state in the lemma and condition on the arrivals and their contexts during the next two rounds. Compare the best continuation after serving the dummy in the first round with the best continuation after idling. If the dummy departs in the first round, the former system begins the second round with the same arrivals and without the dummy. Because policies in this comparison may idle, deleting one job cannot increase the minimum expected terminal queue: the system without the dummy can simulate a policy for the larger system and omit every action assigned to the deleted job. If the dummy does not depart, geometric memorylessness gives it a fresh $\mathrm{Geom}(p_D)$ residual service time. The two systems then begin the second round with the same jobs, except that immediate service is committed to the dummy whereas the idling policy may choose its next action.

The dummy has the highest departure probability, so the two-round comparison above shows that continuing its service is optimal. Immediate service in the first round is therefore no worse for every fixed realization of the arrivals and contexts.

It remains to lower-bound the strict improvement. Condition on no arrival at the ends of the first two rounds; this event has probability $(1-\lambda)^2$. Apart from an arrival at the end of the third round, which contributes equally under both first actions, the dummy is the only job. If it is served immediately, it remains at the terminal time with probability $(1-p_D)^3$. After idling in the first round, even the best continuation can serve it in only two rounds, so it remains with probability $(1-p_D)^2$. The conditional reduction in expected terminal queue is therefore

$$(1-p_D)^2 - (1-p_D)^3 = p_D(1-p_D)^2.$$

The comparison is weakly favorable on every other realization of the arrivals. Averaging thus gives the unconditional improvement

$$(1-\lambda)^2(1-p_D)^2 p_D = g_D,$$

so immediate service reduces the expected terminal queue by at least g_D. □

F.5 Proof of Theorem 2 for IA case

Proof. Define

$$c_0 = \min\left(\frac{(1-p_D)^4}{16a_0}, \frac{3\lambda\nu_D g_D}{32a_0}, \frac{11\ell_0\zeta}{32}\right).$$

Suppose for a contradiction that there was a policy π (which can be idle) that satisfies $\max_\omega R_T^{\mathrm{IA}}(\pi,\omega) < c_0 a_T$. Here, $R_T^{\mathrm{IA}}(\pi,\omega)$ is the regret between policy π and the optimum policy among all idling-allowed policies when the unknown parameter vector was θ_ω.

All departure probabilities are at most p_D by definition. Conditional on $Q_{T-3}^\pi > 0$, the probability that no job departs in rounds $T-3, T-2, T-1, T$ is at least $(1-p_D)^4$, regardless of the policy's actions. Here we used the fact that every job has departure failure probability at least $1-p_D$, and idling can be viewed as having a departure failure probability of 1. On this event the terminal queue is nonempty, and hence

$$J_T^\pi = \mathbb{E}\left[Q_{T+1}^\pi \cdot \mathbf{1}\left(Q_{T-3}^\pi > 0, \text{no departure on last 4 rounds}\right)\right] \geq (1-p_D)^4 \mathbb{P}_\omega(Q_{T-3}^\pi > 0)$$

when our parameter vector was θ_ω. The assumed regret bound gives $J_T^\pi = V_T^{\mathrm{IA}} + R_T^{\mathrm{IA}}(\pi) < V_T^{\mathrm{IA}} + c_0 a_T$. By Lemma 11 and Equation (36), $V_T^{\mathrm{IA}} \leq (1-p_D)^4/16$, and the definition of c_0 gives $c_0 a_T \leq (1-p_D)^4 a_T/16a_0 \leq (1-p_D)^4/16$. Combining these inequalities with the preceding lower bound yields

$$\mathbb{P}_\omega(Q_{T-3}^\pi > 0) < \frac{1}{8}. \tag{47}$$

On $\{Q_{T-3}^\pi = 0\}$, consider the event that a dummy arrives at the end of round $T-3$. Let $\mathcal{I}$ be the event that the learner idles rather than serving this dummy in round $T-2$. By Lemma 12, changing only this idle action gives a feasible policy (call it π^*) whose expected terminal cost is smaller by at least $g_D \mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D, \mathcal{I})$. Since V_T^{IA} is the minimum over all policies that may idle,

$$g_D \mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D, \mathcal{I}) \leq J_T^\pi - J_T^{\pi^*} \leq J_T^\pi - V_T^{\mathrm{IA}} = R_T^{\mathrm{IA}}(\pi) < c_0 a_T$$

and this gives $\mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D, \mathcal{I}) \leq \frac{c_0 a_T}{g_D} \leq \frac{3\lambda\nu_D}{32}$.

Now we find a lower bound of $\mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D)$ to get a lower bound of $\mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D, \mathcal{I}^c)$. The dummy arrival at the end of round $T-3$ is independent of Q_{T-3}^π. So by Equation (47), $\mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D) = \mathbb{P}_\omega(Q_{T-3}^\pi = 0) \cdot \lambda\nu_D > \frac{7\lambda\nu_D}{8}$, and $\mathbb{P}_\omega(Q_{T-3}^\pi = 0, A_{T-3} = 1, X_{T-3} = C_D, \mathcal{I}^c) > \frac{7\lambda\nu_D}{8} - \frac{3\lambda\nu_D}{32} = \frac{25\lambda\nu_D}{32}$. This shows that the probability that the following statements all hold is larger than $25\lambda\nu_D/32$: $Q_{T-3}^\pi = 0$, a dummy job arrives at the end of round $T-3$, and the learner starts the dummy in round $T-2$.

Next intersect this event with the event that the dummy fails to depart in round $T-2$, and a baseline job arrives in round $T-2$, followed by departure of the dummy in round $T-1$. Let $\mathcal{H}$ denote the resulting event before the arrival at the end of round $T-1$. Independence gives

$$\begin{aligned}\mathbb{P}_\omega(\mathcal{H}) &= \mathbb{P}_\omega(Q^\pi_{T-3}=0, A_{T-3}=1, X_{T-3}=C_D, \mathcal{I}^c)\cdot(1-p_D)\cdot\lambda\nu_0\cdot p_D \\ &> \frac{25}{32}\lambda^2\nu_D\nu_0 p_D(1-p_D) = \frac{25\zeta}{32\lambda\nu_C}.\end{aligned} \tag{48}$$

Pair neighboring instances $\omega, \omega^{(j)}$ and use the same kernel K_j and event $\mathcal{S}_j(\omega)$ as in the work-conserving proof. For each ω, $\mathcal{S}_j(\omega)^c$ is a wrong decision and includes selecting the smaller probability job or idling at round T. Data processing through K_j, Equation (39), and Equations (38) and (48) show that their paired wrong-action probability, including the independent candidate arrival, is at least

$$\frac{\lambda\nu_C}{m}\left(\mathbb{P}_\omega(\mathcal{H}) - \mathrm{TV}(P_\omega^{T-1}, P_{\omega^{(j)}}^{T-1})\right) \geq \frac{\lambda\nu_C}{m}\left(\frac{25\zeta}{32\lambda\nu_C} - \frac{3\zeta}{32}\right) = \frac{25\zeta}{32m} - \frac{3\zeta\lambda\nu_C}{32m} \geq \frac{11\zeta}{16m}.$$

Using the double counting method on the m-dimensional hypercube to get an average of $p_{\mathrm{err}}(\omega)$ as shown in the work-conserving proof gives an average wrong-action probability at least $\frac{11\zeta}{32}$. A wrong choice costs $|p_0(\omega) - p_{C_j}(\omega)| \geq \ell_0 a_T$ for the case of selecting a small probability job in round T, and $\max\{p_0(\omega), p_{C_j}(\omega)\} \geq |p_0(\omega) - p_{C_j}(\omega)| \geq \ell_0 a_T$ for the case of idling in round T (here, $p_0(\omega), p_{C_j}(\omega)$ mean p_0, p_{C_j} for the parameter vector θ_ω). Both cases have wrong choice costs of at least $\ell_0 a_T$. Thus there must be some $\omega \in \{-1,1\}^m$ such that $R_T^{\mathrm{IA}}(\pi,\omega) \geq \frac{11\ell_0\zeta a_T}{32} \geq c_0 a_T$, a contradiction. Using the final bound on a_T in the proof of Theorem 2 for WC case, the theorem holds with

$$c_{\mathrm{IA}} = c_0 \min\left(\sqrt{3}a_0, c_S, \frac{\alpha}{\sqrt{3}}\right).$$

□

G Benchmark Structure

We first prove the impossibility results for policies that may idle and for work-conserving policies. We then compare the two finite-horizon benchmark values.

G.1 Proof of Theorem 3 for IA case

Proof. Let there be two kinds of jobs: a slow one and a fast one, which will be denoted as s, f respectively. Define the distribution $\mathcal{D}$ so that a slow job comes with probability $\nu = 10^{-4}$, and a fast job comes with probability $1-\nu$. Choose

$$p_s = 0.01, \qquad p_f = 0.5, \qquad \lambda = 0.0205.$$

as the success probabilities for each job, and the arrival probability. Use contexts $e_1 = (1,0), e_2 = (0,1)$ and parameter vector $\theta^* = (\mathrm{logit}(p_s), \mathrm{logit}(p_f))$. Finally, properly choose fixed constants S, p_-, and p_+ so that $|\theta^*| < S, p_- \leq 0.01, p_+ \geq 0.5$. By this definition, its load is

$$\rho = \lambda\left(\frac{\nu}{p_s} + \frac{1-\nu}{p_f}\right) \approx 0.0412 < 1.$$

For the stationary workload increment $Z = AG$ in Lemma 3,

$$\mathbb{E}Z^2 = \lambda\mathbb{E}_X\left[\mathbb{E}_{g\sim\mathrm{Geom}(p(X))}\left(g^2\right)\right] = \lambda\left(\nu\frac{2-p_s}{p_s^2} + (1-\nu)\frac{2-p_f}{p_f^2}\right) \approx 0.1638.$$

Here, we used the fact that $\mathbb{E}_{g\sim\mathrm{Geom}(p)}\left(g^2\right) = \frac{2-p}{p^2}$ holds. Substituting this value and value of ρ into Equation (15) gives $\mathbb{E}W \approx 0.1051 < 0.2$. The zero-started workload is stochastically dominated by the stationary distribution workload. Moreover, every work-conserving policy satisfies $Q_{t+1} \leq W_{t+1}$, and every such policy is feasible when idling is allowed. Therefore $V_t^{\mathrm{IA}} \leq V_t^{\mathrm{WC}} \leq \mathbb{E}W_{t+1} \leq \mathbb{E}W < 0.2$ for every t.

We will show $\sup_{i\geq t}\{R_i^{\mathrm{IA}}(\pi)\} \geq \min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\}$ holds for all $t \geq 1$. Here $g = p_s^2 - p_s(1+\lambda) + \lambda(1-p_s)\mathbb{E}p(X) \approx 4.1506 \times 10^{-5} > 0$, and this fact shows that $\limsup R_t^{\mathrm{IA}}(\pi)$ is at least $\min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\}$, which is a fixed positive constant.

For simplicity, let $C_0 = \min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\}$ and suppose, for a contradiction, there exists a $t_0 \geq 1$ such that $\sup_{i\geq t_0}\{R_i^{\mathrm{IA}}(\pi)\} < C_0$ holds. So $R_t^{\mathrm{IA}}(\pi) < C_0$ for all $t \geq t_0$. Since $J_t^\pi = V_t^{\mathrm{IA}} + R_t^{\mathrm{IA}}(\pi)$ and $V_t^{\mathrm{IA}} < 0.2$ for all t, we have $\mathbb{E}_\pi Q_{t+1} < 0.2 + C_0 \leq 0.5$ for all $t \geq t_0$. Markov's inequality for the nonnegative integer Q_{t+1} gives $\mathbb{P}_\pi(Q_{t+1} = 0) = 1 - \mathbb{P}_\pi(Q_{t+1} \geq 1) \geq 1 - \mathbb{E}_\pi Q_{t+1} > 1/2$. Now fix $t \geq t_0$. Then the arrival at the end of round $t+1$ is independent of the event $\{Q_{t+1} = 0\}$, which means the queue became empty at the beginning of round $t+1$. Hence, with probability greater than $\lambda\nu/2$, the system is empty at the beginning of round $t+1$ and a slow job arrives at the end of that round. At the beginning of round $t+2$, the server is free and exactly one slow job is waiting.

If the policy idles in this state, changing only this action to service reduces the expected terminal queue with horizon $t+2$ by p_s, since the expected number of departures differs by p_s in the final round. Now evaluate the same decision with horizon $t+3$, so that two rounds remain. If the policy serves immediately, then the server has two choices, idle or work, in the final round $t+3$ if the server is free and the queue is not zero at the beginning of round $t+3$. In that case, choosing work in round $t+3$ makes the expected number of departures maximal, so in these two rounds the maximum of the expected number of departures is

$$p_s(1 + \lambda\mathbb{E}p(X)) + (1-p_s)p_s = 2p_s - p_s^2 + \lambda p_s\mathbb{E}p(X).$$

If it follows the policy that idles in round $t+2$ and serves the fastest job available in the last round $t+3$, the expected number of departures is

$$(1-\lambda)p_s + \lambda\mathbb{E}\max\{p_s, p(X)\} = (1-\lambda)p_s + \lambda\mathbb{E}p(X),$$

where the equality uses $p(X) \geq p_s$ in this instance. Thus idling first has an expected-departure advantage at least

$$g = p_s^2 - p_s(1+\lambda) + \lambda(1-p_s)\mathbb{E}p(X) \approx 4.1506 \times 10^{-5} > 0.$$

So the expectation of queue length decreases by at least g if we change our decision to idle in round $t+2$, rather than serve a slow job.

Let $\mathcal{H}$ denote the event defined above: $Q_{T+1} = 0$ and a slow job arrived at the end of round $t+1$, so the server is free and the slow job is the only waiting job at the beginning of round $t+2$. Let $\mathbf{H}_{t+1}$ denote the random history through the end of round $t+1$, and, for each realization $\mathfrak{h} \in \mathcal{H}$, define

$$a(\mathfrak{h}) := \mathbb{P}_\pi(\pi \text{ serves the slow job in round } t+2 \,|\, \mathbf{H}_{t+1} = \mathfrak{h}).$$

Because π is a horizon-free policy, it uses the same conditional randomized decision rule after the same history $\mathfrak{h}$, regardless of whether the evaluation horizon is $t+2$ or $t+3$ (this is because the server doesn't know the horizon). A comparison policy that changes only this action is feasible because idling is allowed. Therefore, applying the preceding comparison argument conditionally on each history $\mathfrak{h} \in \mathcal{H}$ and then averaging over the common distribution of $\mathbf{H}_{t+1}$ gives

$$R_{t+2}^{\mathrm{IA}}(\pi) + R_{t+3}^{\mathrm{IA}}(\pi) \geq \mathbb{E}\left[\mathbf{1}_{\mathcal{H}}\left((1 - a(\mathbf{H}_{t+1}))p_s + a(\mathbf{H}_{t+1})g\right)\right].$$

Since $(1-a)p_s + ag \geq \min\{p_s, g\}$ for every $a \in [0,1]$, and since $\mathbb{P}(\mathcal{H}) \geq \lambda\nu/2$, it follows that

$$R_{t+2}^{\mathrm{IA}}(\pi) + R_{t+3}^{\mathrm{IA}}(\pi) \geq \mathbb{P}(\mathcal{H})\min\{p_s, g\} \geq \frac{\lambda\nu}{2}\min\{p_s, g\} > 0.$$

But by assumption, since $t \geq t_0$ the following inequality holds.

$$R_{t+2}^{\mathrm{IA}}(\pi) + R_{t+3}^{\mathrm{IA}}(\pi) < 2C_0 = 2\min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\} \leq \frac{\lambda\nu}{2}\min\{p_s, g\}$$

This gives a contradiction; therefore we have proved that $\sup_{i\geq t}\{R_i^{\mathrm{IA}}(\pi)\} \geq \min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\}$ holds. Since $g < p_s$ and $\frac{\lambda\nu}{4}p_s < \frac{1}{4} < 0.3$ hold in the above setting, $\min\{\frac{\lambda\nu}{4}p_s, \frac{\lambda\nu}{4}g, 0.3\} = \frac{\lambda\nu}{4}g$ holds and we can simplify the result as follows: for every $t \geq 1$,

$$\sup_{i\geq t}\{R_i^{\mathrm{IA}}(\pi)\} \geq \frac{\lambda\nu}{4}g \approx 2.1272 \times 10^{-11}.$$

□

G.2 Exact certificate for Proposition 1

Define three jobs as A, B, F, and set their arrival probabilities $\nu_A = 1/4$, $\nu_B = 1/100$, and $\nu_F = 37/50$. Together, we set their departure probabilities as $p_A = 7/20, p_B = 13/20, p_F = 19/20$ and $\lambda = 1/2$. These arrival probabilities sum to one, and give

$$\rho = \frac{1}{2}\left(\frac{\nu_A}{p_A} + \frac{\nu_B}{p_B} + \frac{\nu_F}{p_F}\right) = \frac{6521}{8645} < 1.$$

Let $c = (c_A, c_B, c_F)$ record the waiting job counts, and let $a \in \{0, A, B, F\}$ denote the type in service; $a = 0$ means that the server is free. The job in service is not included in c, and e_i increments the waiting count of type i. For a function f, recall the definition of the arrival operator which is defined in Section 4.2.

$$(\mathcal{A}f)(c, a) = (1 - \lambda)f(c, a) + \lambda \sum_{i\in\{A,B,F\}} \nu_i f(c + e_i, a).$$

Now we will use the Bellman recursion formula which was first shown in Section 4.2. Let $V_h(c, a)$ be the minimum expected terminal queue with h rounds remaining. Its boundary value is $V_0(c, a) = c_A + c_B + c_F + \mathbf{1}\{a \neq 0\}$, which denotes the number of remaining jobs (waiting or in service) in the queue. A type-a job in service forces the recursion

$$V_h(c, a) = p_a(\mathcal{A}V_{h-1})(c, 0) + (1 - p_a)(\mathcal{A}V_{h-1})(c, a). \tag{49}$$

If the server is free and $c \neq 0$, define the value of starting type i for i with $c_i > 0$ by

$$Q_h(i; c) = p_i(\mathcal{A}V_{h-1})(c - e_i, 0) + (1 - p_i)(\mathcal{A}V_{h-1})(c - e_i, i)$$

Work conservation gives

$$V_h(c, 0) = \min_{i:c_i>0} Q_h(i; c). \tag{50}$$

When $c = 0$, $V_h(0, 0) = (\mathcal{A}V_{h-1})(0, 0)$. For each fixed finite horizon h, these recursions visit only finitely many states because the distribution has finite support. Moreover, since all probability values are rational, every Bellman term can be computed exactly using rational arithmetic.

For the remainder of this subsection, abbreviate $Q_h(i) = Q_h(i; e_A + e_B)$. Thus, $Q_h(i)$ denotes the expected terminal queue length when the system starts with h rounds remaining, a free server, and the initial queue A, B, and job i is selected for service in the first round. By calculating the Bellman recursion, we can show that the one-round action values are $Q_1(A) = 43/20$ and $Q_1(B) = 37/20$, so B is uniquely optimal with gap $3/10$. Iterating Equations (49) and (50) to $h = 14$ gives following values of $Q_{14}(A), Q_{14}(B)$. Here, both of $Q_{14}(A), Q_{14}(B)$ are rational.

$$\begin{aligned} Q_{14}(A) \approx 1.6375, \qquad Q_{14}(B) \approx 1.6396, \\ Q_{14}(B) - Q_{14}(A) \approx 0.0021 > 0 \end{aligned} \tag{51}$$

Thus A is uniquely optimal. Taking contexts e_A, e_B, e_F and $\theta^* = (\operatorname{logit} p_A, \operatorname{logit} p_B, \operatorname{logit} p_F)$ gives a three-dimensional logistic representation with finite S. The accompanying script evaluates the same recursion using exact fractions and asserts both action inequalities.

The preceding example further shows that, in a finite-horizon, work-conserving, nonpreemptive system, a policy that selects a job with the highest departure probability whenever the server becomes free need not be optimal. Indeed, when only one round remains, serving B, which has the higher departure probability, maximizes the expected number of departures and hence minimizes the expected terminal queue length. In contrast, when 14 rounds remain, serving A, which has the lower departure probability, results in a larger expected number of departures and hence a smaller expected

terminal queue length. Thus, the optimal decision may depend on the number of remaining rounds and need not coincide with the myopic maximum-departure-probability rule. This is one of the fundamental distinctions between the preemptive and nonpreemptive settings: in the finite-horizon, work-conserving preemptive setting, it is well known that an optimal policy selects a currently available job–server pair with the highest departure probability in every round. See Bae & Lee (2026) for more detail.

G.3 Proof of Theorem 3 for WC case

Proof. Use the instance from Proposition 1. By Equation (41), the zero-started common workload satisfies $\mathbb{P}(W_t = 0) \geq 1 - \rho$ for every t. Fix $s \geq 5$, and starting from the event $W_{s-4} = 0$, consider the following event: an F job arrives at the end of round $s-4$; it remains in service in round $s-3$ and an A job arrives; it remains in service in round $s-2$ and a B job arrives; and it departs in round $s-1$, with no arrival at the end of that round. The F job is the only job present when its service begins, and nonpreemption fixes all three service decisions. At the beginning of round s, the server is free with exactly one A and one B job waiting. Define the above event as $\mathcal{H}_{\mathrm{WC}}$. Independence of the arrivals and departure indicators gives the lower bound of the probability that this kind of event happens.

$$\begin{aligned}\mathbb{P}(\mathcal{H}_{\mathrm{WC}}) &= \mathbb{P}(W_{s-4} = 0) \cdot \lambda\nu_F \cdot (1 - p_F) \cdot \lambda\nu_A \cdot (1 - p_F) \cdot \lambda\nu_B \cdot p_F \cdot (1 - \lambda) \\ &\geq (1 - \rho)\lambda^3(1 - \lambda)\nu_F\nu_A\nu_B p_F(1 - p_F)^2\end{aligned} \tag{52}$$

Here, we define $\zeta_{\mathrm{WC}} = (1 - \rho)\lambda^3(1 - \lambda)\nu_F\nu_A\nu_B p_F(1 - p_F)^2 \approx 6.7469 \times 10^{-8}$.

Let $\mathbf{H}_{s-1}$ denote the random history through the end of round $s-1$. For each realization $\mathfrak{h} \in \mathcal{H}_{\mathrm{WC}}$, define

$$\alpha_s(\mathfrak{h}) := \mathbb{P}_\pi\left(\pi \text{ starts } B \text{ in round } s \,|\, \mathbf{H}_{s-1} = \mathfrak{h}\right).$$

Because π is horizon-free, it uses the same conditional randomized decision rule after the same history $\mathfrak{h}$, regardless of whether the terminal horizon is s or $s + 13$.

With terminal horizon s, one round remains after each history $\mathfrak{h} \in \mathcal{H}_{\mathrm{WC}}$. A comparison policy that changes only this last action from A to B is feasible and work-conserving. Therefore

$$R_s^{\mathrm{WC}}(\pi) \geq \mathbb{E}\left[\mathbf{1}_{\mathcal{H}_{\mathrm{WC}}}\left(1 - \alpha_s(\mathbf{H}_{s-1})\right)(p_B - p_A)\right].$$

Now evaluate the same infinite policy at horizon $s + 13$. Fourteen rounds remain after the same history. If the policy starts B, even its best continuation has value $Q_{14}(B)$, whereas a comparison policy that starts A and then continues optimally has value $Q_{14}(A)$. Hence

$$R_{s+13}^{\mathrm{WC}}(\pi) \geq \mathbb{E}\left[\mathbf{1}_{\mathcal{H}_{\mathrm{WC}}}\alpha_s(\mathbf{H}_{s-1})\big(Q_{14}(B) - Q_{14}(A)\big)\right].$$

Adding the two inequalities and using $(1-\alpha)x + \alpha y \geq \min\{x, y\}$ for every $\alpha \in [0, 1]$ and $x, y > 0$ gives

$$\begin{aligned}R_s^{\mathrm{WC}}(\pi) + R_{s+13}^{\mathrm{WC}}(\pi) &\geq \mathbb{E}\left[\mathbf{1}_{\mathcal{H}_{\mathrm{WC}}}\left(\left(1 - \alpha_s(\mathbf{H}_{s-1})\right)(p_B - p_A) + \alpha_s(\mathbf{H}_{s-1})\left(Q_{14}(B) - Q_{14}(A)\right)\right)\right] \\ &\geq \mathbb{P}(\mathcal{H}_{\mathrm{WC}}) \min\left\{p_B - p_A, Q_{14}(B) - Q_{14}(A)\right\} \\ &\geq \zeta_{\mathrm{WC}} \min\left\{p_B - p_A, Q_{14}(B) - Q_{14}(A)\right\} \approx 1.425 \times 10^{-10}.\end{aligned} \tag{53}$$

So this inequality shows that for $s \geq 5$,

$$\begin{aligned}\sup_{i \geq s} R_s^{\mathrm{WC}}(\pi) &\geq \frac{1}{2}\left(R_s^{\mathrm{WC}}(\pi) + R_{s+13}^{\mathrm{WC}}(\pi)\right) \\ &\geq \frac{1}{2}\zeta_{\mathrm{WC}} \min\left\{p_B - p_A, Q_{14}(B) - Q_{14}(A)\right\} \\ &\approx 7.125 \times 10^{-11}.\end{aligned}$$

□

G.4 A COMPARISON BETWEEN THE BENCHMARK CLASSES

Theorem 5. *For every $T \geq 1$, $0 \leq V_T^{\mathrm{WC}} - V_T^{\mathrm{IA}} \leq C_W(r)$. For any policy π, let $D_T^\pi := \mathbb{E}_\pi \sum_{t=1}^T D_t = \lambda T - J_T^\pi$ be its expected number of departures. In particular, $D_T^{\mathrm{IA}} = \lambda T - V_T^{\mathrm{IA}}$ and $D_T^{\mathrm{WC}} = \lambda T - V_T^{\mathrm{WC}}$. Then*

$$D_T^{\mathrm{WC}} \geq \left(1 - \frac{C_W(r)}{\lambda T}\right)^+ D_T^{\mathrm{IA}}.$$

Moreover, a work-conserving learner with regret $R_T^{\mathrm{WC}}(\pi)$ satisfies

$$D_T^\pi \geq \left(1 - \frac{C_W(r) + R_T^{\mathrm{WC}}(\pi)}{\lambda T}\right)^+ D_T^{\mathrm{IA}}.$$

Proof. Every work-conserving policy is also feasible when idling is allowed, so $V_T^{\mathrm{IA}} \leq V_T^{\mathrm{WC}}$. Under a work-conserving policy, $Q_{T+1} \leq W_{T+1}$ pathwise, and Lemma 1 gives $V_T^{\mathrm{WC}} \leq C_W(r)$. This proves the additive comparison.

Summing $Q_{t+1} = Q_t - D_t + A_t$ from $t = 1$ to T and using $Q_1 = 0$ gives

$$Q_{T+1} = \sum_{t=1}^T A_t - \sum_{t=1}^T D_t.$$

Taking expectations and using $\mathbb{E}A_t = \lambda$ yields $D_T^\pi = \lambda T - J_T^\pi$. Hence $D_T^{\mathrm{WC}} \geq \lambda T - C_W(r)$ and $D_T^{\mathrm{IA}} \leq \lambda T$. If $\lambda T \leq C_W(r)$, the first multiplicative bound has a zero right-hand side. Otherwise,

$$D_T^{\mathrm{WC}} \geq \lambda T - C_W(r) \geq \left(1 - \frac{C_W(r)}{\lambda T}\right) D_T^{\mathrm{IA}}.$$

Finally, $J_T^\pi \leq V_T^{\mathrm{WC}} + R_T^{\mathrm{WC}}(\pi) \leq C_W(r) + R_T^{\mathrm{WC}}(\pi)$, and the same two cases prove the learner bound. □

H ANYTIME SEPT TRACKING

We first state the prediction guarantee and the busy-period bound used in the proof of Theorem 4. We then prove the theorem and establish each supporting result. The two sigma-fields below record the order of service, estimation, and arrivals within a round.

Let $\mathcal{F}_t$ be the σ-algebra generated by all information about the queue, the job in service, the observations, and the learner's randomization available at the beginning of round t. The scheduling decision is $\mathcal{F}_t$-measurable. Let $\mathcal{G}_t$ be the σ-algebra obtained by additionally including D_t and any estimator update made after service in round t, but not (A_t, X_t). Under Assumption 1, (A_t, X_t) is independent of $\mathcal{G}_t$, with $A_t \sim \mathrm{Bern}(\lambda)$ and $X_t \sim \mathcal{D}$ independent of A_t.

The following three lemmas are the main ingredients in the proof of Theorem 4. First, Lemma 13 shows that, with high probability, every sufficiently recent busy period uses a predictor with small prediction error. Recall that $u = \lfloor \frac{t}{2} \rfloor$, $m_t = \lfloor \frac{\lambda u}{2} \rfloor$, $b_t = e^{-\lambda u/8} + \frac{1}{t^4}$.

Lemma 13. *Fix $t \geq 6$ such that $m_t \geq 5$. Then there exists an event B_t with $\mathbb{P}(B_t) \leq b_t$ such that, on B_t^c, every predictor used in a busy period beginning in a round s with $u + 2 \leq s \leq t$ satisfies $\mathcal{E}(\widehat{p}) \leq \bar{\epsilon}_t$.*

For $s \geq 2$, let $\mathcal{S}_s$ denote the event that the system is empty immediately after service in round $s - 1$ and that an arrival occurs at the end of that round. Thus, on $\mathcal{S}_s$, a new busy period begins in round s. The next Lemma 14 shows that, under two work-conserving policies, the contribution to the expected terminal queue-length difference from a busy period beginning in round s and remaining active up to round t decays exponentially in its age $h = t - s + 1$.

Lemma 14. *Let $h = t - s + 1$, and let C_s be the event that a busy period begins in round s and remains active after service in round t; equivalently, if we denote τ_{bp} as the length of the busy period starting at round s conditional on the event $\mathcal{S}_s$, C_s is defined as $\mathcal{S}_s \cap \{\tau_{\mathrm{bp}} > h\}$. Assign each job*

the same service time under two work-conserving ordering rules. If their queue lengths after round t are $Q^{(1)}_{t+1}$ and $Q^{(2)}_{t+1}$, define $Z_{s,t} = Q^{(1)}_{t+1} - Q^{(2)}_{t+1}$. Then

$$\mathbb{E}\big(|Z_{s,t}|\mathbf{1}\{C_s\}\big) \le \mathbb{E}\big(W_{t+1}\mathbf{1}\{C_s\}\big) \le \frac{M_G(r)}{r}\psi(r)^h. \tag{54}$$

We refer to the above inequality as the busy-period tail bound.

Finally, Lemma 15 shows that if a predictor q, fixed before a busy period begins, approximates the true departure-probability function p with small prediction error, then q-SEPT and true SEPT have close expected terminal queue lengths when both policies are evaluated under the true model p. Here, q-SEPT is the priority policy that, whenever the server becomes free, selects a waiting job maximizing $q(X)$, thereby treating q as the departure-probability function. For a priority rule π, $J^p_{h,\mathrm{stop}}(\pi)$ denotes the expected queue length after h rounds when the actual departure probability of a job with context X is $p(X)$ and the jobs are selected according to π, with the comparison stopped as soon as service leaves no unfinished job. Thus, $J^p_{h,\mathrm{stop}}(q\text{-SEPT})$ is the expected terminal queue length when jobs are ranked using q, while their actual departures are governed by p.

Lemma 15. *Suppose $\mathcal{E}(q) = \mathbb{E}|q(X) - p(X)| = \epsilon$. Run a busy period for h rounds, beginning with the first round in which its initial job is served, and stop both comparison systems as soon as service leaves no unfinished job. Then*

$$\left|J^p_{h,\mathrm{stop}}(q\text{-SEPT}) - J^p_{h,\mathrm{stop}}(p\text{-SEPT})\right| \le \frac{2(h+1)^2}{p_-}\epsilon. \tag{55}$$

We conclude this introductory part preceding the main proofs by introducing the function $\Phi_r(\epsilon)$, which appears in the statement of Theorem 4. For $\epsilon \in (0, 1]$, define

$$\Phi_r(\epsilon) = \sum_{a=0}^{\infty} \min\left(\frac{2(a+2)^2}{p_-}\epsilon, \frac{M_G(r)}{r}\psi(r)^{a+1}\right). \tag{56}$$

Here, $M_G(r)$ and $\psi(r)$ are constants defined at the beginning of Appendix C. Splitting the series at the smallest a for which $\psi(r)^a \le \varepsilon$ and bounding the terms appropriately, we obtain $\Phi_r(\varepsilon) = O\left(\varepsilon\left(1 + \log^3(1/\varepsilon)\right)\right)$, where the constant depends only on r, p_-, $M_G(r)$, and $\psi(r)$. A detailed proof of this asymptotic bound is provided in Appendix H.6.

H.1 Proof of Theorem 4

Proof. Run Algorithm 2 and true SEPT with the same arrivals and contexts, and assign each job the same geometric service time in the two systems. Both policies are work conserving, so Equation (2) gives the same workload path. In particular, the two systems become empty after service in exactly the same rounds and therefore have the same busy-period boundaries (i.e. the starting/ending times of busy periods are all the same in both queues).

For $s \in \{2, \dots, t\}$, let q_s denote the predictor held by the learner immediately after service and any estimator update in round $s-1$. On $\mathcal{S}_s$, this predictor is fixed before the context of the job initiating the new busy period is observed and remains fixed throughout that busy period. Starting with this initial job which arrived at the end of round $s-1$, compare q_s-SEPT and true SEPT using the same subsequent arrivals, contexts, and service times. Stop both comparison systems as soon as service leaves no unfinished job. With $h = t - s + 1$, let $Z_{s,t}$ be the difference between their queue lengths after these h rounds, where both stopped queue lengths are defined to be zero if the busy period has already ended.

By definition, on $\mathcal{S}_s \setminus \mathcal{C}_s$, the busy period that begins in round s ends by the completion of service in round t. The two stopped queue lengths are therefore both zero, and hence $Z_{s,t} = 0$. Consequently, $\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s\}\right) = \mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{C}_s\}\right)$ for every s.

Moreover, only the last busy period occurring during the first t rounds can contribute to the terminal queue-length difference. If no busy period remains active after service in round t, the two coupled systems have the same terminal queue length. It is therefore sufficient to restrict attention to the

event $\bigcup_{s=2}^{t} \mathcal{C}_s$. Since the events $\mathcal{C}_2, \ldots, \mathcal{C}_t$ are pairwise disjoint and $Z_{s,t}$ agrees with the actual terminal queue-length difference on $\mathcal{C}_s$, we obtain

$$\begin{aligned} J_t^{\mathrm{alg}} - J_t^{\mathrm{SEPT}} &= \mathbb{E}\left(\left(Q_{t+1}^{\mathrm{alg}} - Q_{t+1}^{\mathrm{SEPT}}\right)\mathbf{1}\{\bigcup_{s=2}^{t} \mathcal{C}_s\}\right) \\ &= \sum_{s=2}^{t} \mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{C}_s\}\right) = \sum_{s=2}^{t} \mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s\}\right). \end{aligned} \tag{57}$$

A busy period beginning in round $t+1$ need not be included, since its initial job has not yet received service and is present in both terminal queues.

For the statistical comparison, define $E_s = \{\mathcal{E}(q_s) \le \bar{\epsilon}_t\}$ and let $\mathcal{H}_s = \mathcal{G}_{s-1} \vee \sigma(A_{s-1})$. The predictor q_s and the events $\mathcal{S}_s$ and E_s are $\mathcal{H}_s$-measurable. Conditional on $\mathcal{H}_s$, the predictor q_s is fixed. Moreover, on the event $\mathcal{S}_s$, the context of the initial job of the busy period beginning in round s has distribution $\mathcal{D}$, and the subsequent arrival process is independent of $\mathcal{H}_s$.

Recall that $Z_{s,t}$ is defined as the difference between the terminal queue lengths in the stopped h-round comparison of q_s-SEPT and true SEPT, where $h = t - s + 1$. Therefore, conditional on $\mathcal{H}_s$, on the event $\mathcal{S}_s$, its conditional expectation is precisely the difference between the two expected terminal queue lengths appearing in Lemma 15:

$$\mathbb{E}\left(Z_{s,t} \mid \mathcal{H}_s\right) = J^p_{h,\mathrm{stop}}\left(q_s\text{-SEPT}\right) - J^p_{h,\mathrm{stop}}\left(p\text{-SEPT}\right).$$

Consequently, Lemma 15 applies conditionally and gives, on $\mathcal{S}_s$,

$$\left|\mathbb{E}\left(Z_{s,t} \mid \mathcal{H}_s\right)\right| \le \frac{2(h+1)^2}{p_-}\mathcal{E}(q_s).$$

In particular, on $\mathcal{S}_s \cap E_s$,

$$\left|\mathbb{E}\left(Z_{s,t} \mid \mathcal{H}_s\right)\right| \le \frac{2(h+1)^2}{p_-}\bar{\epsilon}_t.$$

Since $\mathcal{S}_s \cap E_s$ is $\mathcal{H}_s$-measurable, the tower property then gives

$$\begin{aligned} \left|\mathbb{E}\big(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s\}\big)\right| &\le \mathbb{E}\left[\mathbf{1}\{\mathcal{S}_s \cap E_s\}\left|\mathbb{E}\big(Z_{s,t} \mid \mathcal{H}_s\big)\right|\right] \\ &\le \frac{2(h+1)^2}{p_-}\bar{\epsilon}_t. \end{aligned}$$

On $\mathcal{S}_s \setminus C_s$, the busy period has ended by round t, so the stopped discrepancy $Z_{s,t}$ is zero. It follows from Lemma 14 that

$$\begin{aligned} \left|\mathbb{E}\big(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s\}\big)\right| &= \left|\mathbb{E}\big(Z_{s,t}\mathbf{1}\{\mathcal{C}_s \cap E_s\}\big)\right| \\ &\le \mathbb{E}\big(|Z_{s,t}|\mathbf{1}\{C_s \cap E_s\}\big) \\ &\le \mathbb{E}\big(|Z_{s,t}|\mathbf{1}\{C_s\}\big) \le \frac{M_G(r)}{r}\psi(r)^h. \end{aligned}$$

Combining the preceding two bounds yields

$$\left|\mathbb{E}\big(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s\}\big)\right| \le \min\left\{\frac{2(h+1)^2}{p_-}\bar{\epsilon}_t, \frac{M_G(r)}{r}\psi(r)^h\right\}. \tag{58}$$

For $s \in \{u+2, \ldots, t\}$, set $a = h - 1 = t - s$. Summing equation 58 over these starting rounds gives

$$\sum_{s=u+2}^{t} \left|\mathbb{E}\big(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s\}\big)\right| \le \sum_{a=0}^{t-u-2} \min\left\{\frac{2(a+2)^2}{p_-}\bar{\epsilon}_t, \frac{M_G(r)}{r}\psi(r)^{a+1}\right\} \le \Phi_r(\bar{\epsilon}_t). \tag{59}$$

By Lemma 13, on B_t^c, every predictor used in a busy period beginning in a round $s \in \{u+2, \ldots, t\}$ satisfies $\mathcal{E}(q_s) \le \bar{\epsilon}_t$. Thus

$$\mathcal{C}_s \cap E_s^c \subseteq \mathcal{S}_s \cap E_s^c \subseteq B_t, \qquad u+2 \le s \le t.$$

Since all events C_s are disjoint, $C_s \cap E_s^c$ $(u+2 \le s \le t)$ are all disjoint in the event B_t, so $\bigcup_{s=u+2}^{t} C_s \cap E_s^c \subset B_t$ holds. Using this fact and the fact that $Z_{s,t} = 0$ in the event $\mathcal{S}_s \setminus \mathcal{C}_s$, we can derive the following inequality.

$$\begin{aligned}\sum_{s=u+2}^{t} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s^c\}\right)\right| &= \sum_{s=u+2}^{t} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{C}_s \cap E_s^c\}\right)\right| \le \sum_{s=u+2}^{t} \mathbb{E}\left(|Z_{s,t}|\,\mathbf{1}\{\mathcal{C}_s \cap E_s^c\}\right) \\ &\le \sum_{s=u+2}^{t} \mathbb{E}\left(W_{t+1}\mathbf{1}\{\mathcal{C}_s \cap E_s^c\}\right) = \mathbb{E}\left(W_{t+1}\mathbf{1}\{\bigcup_{s=u+2}^{t} \mathcal{C}_s \cap E_s^c\}\right) \\ &\le \mathbb{E}\left(W_{t+1}\mathbf{1}\{B_t\}\right)\end{aligned} \tag{60}$$

Moreover, $(rw)^2 \le e^{rw}$ for every $w \ge 0$, so Lemma 1 gives

$$\mathbb{E}W_{t+1}^2 \le \frac{1}{r^2}\mathbb{E}e^{rW_{t+1}} \le \frac{K_r}{r^2}.$$

By using the Cauchy–Schwarz inequality for W_{t+1} and $\mathbf{1}\{B_t\}$, and the inequality in Lemma 13, we get the following upper bound for Equation (60).

$$\mathbb{E}\left(W_{t+1}\mathbf{1}\{B_t\}\right) \le \sqrt{\mathbb{E}W_{t+1}^2\,\mathbb{P}(B_t)} \le \frac{\sqrt{K_r b_t}}{r}$$

Finally, for C_s with $2 \le s \le u+1$, by applying Lemma 14 we can get $\mathbb{E}\left(W_{t+1}\mathbf{1}\{C_s\}\right) \le \frac{M_G(r)}{r}\psi(r)^h$ where $h = t-s+1$. Summing this inequality for $s = 2, \ldots, u+1$, we get the following upper bound.

$$\begin{aligned}\sum_{s=2}^{u+1} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s\}\right)\right| &= \sum_{s=2}^{u+1} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{C}_s\}\right)\right| \le \sum_{s=2}^{u+1} \mathbb{E}\left(|Z_{s,t}|\,\mathbf{1}\{\mathcal{C}_s\}\right) \le \sum_{s=2}^{u+1} \mathbb{E}\left(W_{t+1}\mathbf{1}\{\mathcal{C}_s\}\right) \\ &\le \frac{M_G(r)}{r}\sum_{h=t-u}^{t-1} \psi(r)^h \le \frac{M_G(r)}{r(1-\psi(r))}\psi(r)^{t-u} \le \frac{M_G(r)}{r(1-\psi(r))}\psi(r)^u\end{aligned} \tag{61}$$

Here, we used the fact $t - u \ge u$.

Combining the three preceding estimates Equation (59), Equation (60), Equation (61) with equation 57 gives

$$\begin{aligned}|J_t^{\text{alg}} - J_t^{\text{SEPT}}| &= \left|\sum_{s=2}^{t} \mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s\}\right)\right| \\ &\le \sum_{s=2}^{u+1} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s\}\right)\right| + \sum_{s=u+2}^{t} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s\}\right)\right| + \sum_{s=u+2}^{t} \left|\mathbb{E}\left(Z_{s,t}\mathbf{1}\{\mathcal{S}_s \cap E_s^c\}\right)\right| \\ &\le \frac{M_G(r)}{r(1-\psi(r))}\psi(r)^u + \Phi_r(\bar{\epsilon}_t) + \frac{\sqrt{K_r b_t}}{r}\end{aligned} \tag{62}$$

Therefore we can obtain Equation (12).

For fixed model and traffic parameters,

$$m_t = \Theta(\lambda t) \qquad \text{and} \qquad \bar{\epsilon}_t = \widetilde{O}\left(\sqrt{\frac{d}{\lambda t}}\right).$$

The series calculation in Appendix H.6 therefore gives the stated rate. If $m_t < 5$, both policies are work conserving, so Lemma 1 implies

$$0 \le J_t^{\text{alg}}, J_t^{\text{SEPT}} \le C_W(r),$$

and hence

$$\mathcal{T}_t^{\text{SEPT}} = |J_t^{\text{alg}} - J_t^{\text{SEPT}}| \le C_W(r).$$

□

H.2 Proof of Lemma 13

Proof. Suppose that a busy period begins in round s. Its first job arrived at the end of round $s-1$. The predictor used during this busy period was computed after service in round $s-1$ and before that arrival. Because the queue was empty immediately after service in round $s-1$, every job that arrived by the end of round $s-2$ had already departed, and (X_i, Y_i) had been recorded for each such job. The predictor is therefore fitted to a complete prefix of the i.i.d. job-indexed sequence and is fixed before the context of the initial job is drawn at the end of round $s-1$.

Suppose that $u+2 \le s \le t$. Since $u \le s-2$, the data used by the predictor contain all $M_u = \sum_{j=1}^{u} A_j$ jobs arriving through round u. Since $m_t = \lfloor \lambda u/2 \rfloor$ and $\mathbb{E}M_u = \lambda u$, a multiplicative Chernoff bound with relative deviation $1/2$ gives

$$\mathbb{P}(M_u < m_t) \le \mathbb{P}\left(M_u < \frac{\lambda u}{2}\right) \le e^{-\lambda u/8}.$$

Let Ω_t be the simultaneous prediction event supplied by Corollary 1 with failure probability t^{-4}. Define $B_t = \{M_u < m_t\} \cup \Omega_t^c$. Then

$$\mathbb{P}(B_t) \le \mathbb{P}(\{M_u < m_t\}) + \mathbb{P}(\Omega_t^c) \le e^{-\lambda u/8} + \frac{1}{t^4} = b_t.$$

On B_t^c, consider a predictor used in a busy period beginning in a round $s \in \{u+2, \dots, t\}$. If N denotes the number of observations used to construct this predictor, then $m_t \le N \le s-2 < t$ holds, and Corollary 1 gives $\mathcal{E}(\widehat{p}_N) \le \epsilon_N\left(\frac{6}{\pi^2 N^2 t^4}\right)$. Using $m_t \le N \le t$, we obtain

$$\begin{aligned}
\mathcal{E}(\widehat{p}_N) &\le \min\left(1, \sqrt{\frac{8e^{\bar{S}}\left(d\log(32\bar{S}N) + \log(\pi^2 N^2 t^4/6)\right)+1}{2N}}\right) \\
&\le \min\left(1, \sqrt{\frac{8e^{\bar{S}}\left(d\log(32\bar{S}t) + \log(\pi^2 t^6/6)\right)+1}{2m_t}}\right) \\
&= \bar{\epsilon}_t.
\end{aligned}$$

This proves the claim. □

H.3 Proof of Lemma 14

Proof. Couple the two systems so that their arrival indicators, arriving contexts, and job-specific service times agree. Since both policies are work conserving, their workload processes coincide. Consequently, their busy periods begin and end in the same rounds.

Let G be the service time of the job initiating the busy period in round s, and let $\widetilde{W}_h$ be the stopped workload associated with this busy period after h service rounds. On C_s, $W_{t+1} = \widetilde{W}_h$ and $\tau_{\mathrm{bp}} > h$ hold. Therefore,

$$\mathbb{E}\big(W_{t+1}\mathbf{1}\{C_s\}\big) = \mathbb{P}(\mathcal{S}_s)\mathbb{E}\left(\widetilde{W}_h \mathbf{1}\{\tau_{\mathrm{bp}} > h\} \,\middle|\, \mathcal{S}_s\right).$$

Conditional on $\mathcal{S}_s$ and G, the future workload increments have the same distribution as those in the stopped busy-period process. Using $w \le e^{rw}/r$ and Equation (14), we obtain

$$\mathbb{E}\left(\widetilde{W}_h \mathbf{1}\{\tau_{\mathrm{bp}} > h\} \,\middle|\, \mathcal{S}_s, G\right) \le \frac{1}{r}\mathbb{E}\left(e^{r\widetilde{W}_h}\mathbf{1}\{\tau_{\mathrm{bp}} > h\} \,\middle|\, \mathcal{S}_s, G\right) \le \frac{e^{rG}}{r}\psi(r)^h.$$

The event $\mathcal{S}_s$ is determined before the context of the initial job is observed. Hence the conditional distribution of G given $\mathcal{S}_s$ is the usual mixture $G \mid X \sim \mathrm{Geom}(p(X))$, and

$$\mathbb{E}\big(e^{rG} \mid \mathcal{S}_s\big) = \mathbb{E}_X\left[\frac{p(X)e^r}{1-(1-p(X))e^r}\right] = M_G(r).$$

It follows that

$$\mathbb{E}\big(W_{t+1}\mathbf{1}\{C_s\}\big) \le \mathbb{P}(\mathcal{S}_s)\frac{M_G(r)}{r}\psi(r)^h \le \frac{M_G(r)}{r}\psi(r)^h.$$

Finally, on C_s, both terminal queue lengths lie in $[0, W_{t+1}]$. Therefore,

$$|Z_{s,t}| = |Q_{t+1}^{(1)} - Q_{t+1}^{(2)}| \le W_{t+1}.$$

This proves Equation (54). □

H.4 Sensitivity of SEPT to departure probabilities

Lemma 16. *Fix N jobs, their arrival rounds, a horizon, and a deterministic tie-breaking rule. For $v \in [p_-, p_+]^N$, let $\mathcal{V}(v)$ be the expected terminal queue length when job i has service time $\mathrm{Geom}(v_i)$ and the jobs are served in decreasing order of v_i. Then*

$$|\mathcal{V}(v) - \mathcal{V}(w)| \le \frac{N}{p_-}\|v - w\|_1. \tag{63}$$

The same conclusion holds if the process is stopped when the queue first becomes empty.

Proof. First suppose that the coordinates of v and w induce the same strict order. The two systems then use the same priority rule. Couple their service times coordinatewise as in Lemma 8. A union bound gives

$$\mathbb{P}(\text{some pair of service times differs}) \le \frac{1}{p_-}\sum_{i=1}^{N}|v_i - w_i|.$$

If every pair of service times agrees, the two schedules and their terminal queues agree. Since each terminal queue contains at most N jobs, multiplying the disagreement probability by N proves Equation (63) whenever v and w induce the same strict order.

Next, suppose that $v_i = v_j$. Jobs i and j then have identical service-time distributions. If at most one of them is waiting, their relative priority is irrelevant. If both are waiting, exchanging their labels together with their i.i.d. geometric service times leaves the law of the queue unchanged. Repeating this argument after the first departure, or equivalently proceeding by induction on the number of remaining rounds, shows that either relative order gives the same expected terminal queue length. Thus $\mathcal{V}$ is continuous when two coordinates cross.

For vectors v and w with distinct coordinates, divide the line segment joining them at the finitely many points at which two coordinates become equal. On each resulting subsegment, the coordinate order is fixed, so the preceding coupling argument applies. Summing the resulting bounds proves Equation (63), since the sum of the ℓ_1-lengths of the subsegments equals $\|v - w\|_1$. General v, w follow by continuity. Stopping the process when the queue first becomes empty does not affect the label-exchange argument, so the same proof applies to the stopped process. □

H.5 Proof of Lemma 15

Proof. Condition on the context of the initial job and on all subsequent arrival indicators and contexts in the h-round comparison. Stop the process as soon as service leaves no unfinished job. Let N be the number of jobs described by these variables. Then $N \le h + 1$, and these variables are independent of q, which was fixed before the context of the initial job was drawn. Write

$$p_i = p(X_i), \qquad q_i = q(X_i).$$

Let $J_p(q\text{-SEPT})$ be the conditional expected terminal queue length when job i has departure probability p_i and priority q_i. Compare this system with the intermediate system in which job i has both departure probability and priority q_i. We have

$$J_p(q\text{-SEPT}) - J_p(p\text{-SEPT}) = \big(J_p(q\text{-SEPT}) - J_q(q\text{-SEPT})\big) + \big(\mathcal{V}(q) - \mathcal{V}(p)\big). \tag{64}$$

The priority order is fixed in the first difference. Coordinatewise geometric coupling therefore bounds its absolute value by

$$\frac{N}{p_-}\sum_{i=1}^{N}|p_i-q_i|.$$

The second difference is bounded by the same quantity using Lemma 16. Hence

$$|J_p(q\text{-SEPT})-J_p(p\text{-SEPT})|\le\frac{2N}{p_-}\sum_{i=1}^{N}|p_i-q_i|.$$

Averaging over the arrival variables and contexts gives

$$\mathbb{E}\sum_{i=1}^{N}|p_i-q_i|\le(h+1)\mathcal{E}(q).$$

Since $N\le h+1$, we can show the desired result. □

H.6 Asymptotic dependence of $\Phi_r(\epsilon)$ on ϵ

For $\epsilon\in(0,1]$, let

$$L_\epsilon=\left\lceil\frac{\log(1/\epsilon)}{\log(1/\psi(r))}\right\rceil.$$

By definition, $\psi(r)^{L_\epsilon}\le\epsilon$. Bounding the terms indexed by $a\in\{0,\dots,L_\epsilon\}$ by their statistical bound and the remaining terms by their busy-period tail bound gives

$$\begin{aligned}\Phi_r(\epsilon)&\le\frac{2\epsilon}{p_-}\sum_{a=0}^{L_\epsilon}(a+2)^2+\frac{M_G(r)}{r}\sum_{a=L_\epsilon+1}^{\infty}\psi(r)^{a+1}\\&<\frac{2\epsilon(L_\epsilon+2)(L_\epsilon+3)(2L_\epsilon+5)}{6p_-}+\frac{M_G(r)}{r(1-\psi(r))}\psi(r)^{L_\epsilon+2}\\&\le\frac{2\epsilon(L_\epsilon+3)^3}{3p_-}+\frac{M_G(r)}{r(1-\psi(r))}\epsilon.\end{aligned}$$

Since

$$L_\epsilon\le1+\frac{\log(1/\epsilon)}{\log(1/\psi(r))},$$

we conclude that

$$\Phi_r(\epsilon)=O\left(\epsilon\big(1+\log^3(1/\epsilon)\big)\right),$$

with the dependence stated after Equation (56).

H.7 Average tracking and a uniformly random horizon

From Theorem 4 and $\sum_{t=1}^{T}t^{-1/2}\le2\sqrt{T}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{T}_t^{\mathrm{SEPT}}=\widetilde{O}\left(\sqrt{\frac{d}{\lambda T}}\right).$$

If τ is independent of the queueing process and uniformly distributed on $[T]$, then conditioning on τ gives

$$\mathbb{E}Q_{\tau+1}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}Q_{t+1}.$$

Thus, the same rate compares the learner and SEPT when the evaluation horizon is chosen independently and uniformly from $[T]$.

USE OF GENERATIVE-AI TOOLS

The authors formulated the research problem and used GPT-5.6 Sol to explore possible solutions alongside their own investigation. The model proposed an alternative approach and generated initial theoretical results and proof drafts. The authors checked and revised these drafts, correcting errors and filling gaps in the arguments. In particular, the authors corrected and refined the proofs of the lower bound in Theorem 2 and the impossibility of an anytime policy with vanishing regret under an unknown horizon in Theorem 3. GPT-5.6 Sol was also used for language editing and manuscript polishing. The authors take responsibility for the correctness and presentation of the final work.