Online Learning and Optimization for Queues with Unknown Arrival Rate and Service Distribution

Xinyun Chen
School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, chenxinyun@cuhk.edu.cn

Guiyu Hong
College of Bussiness, Shanghai University of Finance and Economics, Shanghai, China, hongguiyu@sufe.edu.cn

Yunan Liu
Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695-7906, yliu48@ncsu.edu

We investigate an optimization problem in a queueing system where the service provider selects the optimal service fee p and service capacity μ to maximize the cumulative expected profit (the service revenue minus the capacity cost and delay penalty). The conventional *predict-then-optimize* (PTO) approach takes two steps: first, it estimates the model parameters (e.g., arrival rate and service-time distribution) from data; second, it optimizes a model taking these parameters as input. A major drawback of PTO is that its solution accuracy can often be highly sensitive to the parameter estimation errors because PTO is unable to effectively account for how these errors (step 1) will impact the solution quality of the downstream optimization (step 2). To remedy this issue, we develop an online learning framework that automatically incorporates the aforementioned parameter estimation errors in the optimization process; it is an end-to-end approach that can learn the optimal solution without needing to set up the parameter estimation as a separate step as in PTO. Effectiveness of our online learning approach is substantiated by (i) theoretical results including the algorithm convergence and analysis of the *regret* ("cost" to pay over time for the algorithm to learn the optimal policy), and (ii) engineering confirmation via simulation experiments of a variety of representative examples. We also provide careful comparisons between PTO and our online learning method.

Key words: online learning in queues; service systems; capacity planning; staffing; pricing in service systems

1. Introduction

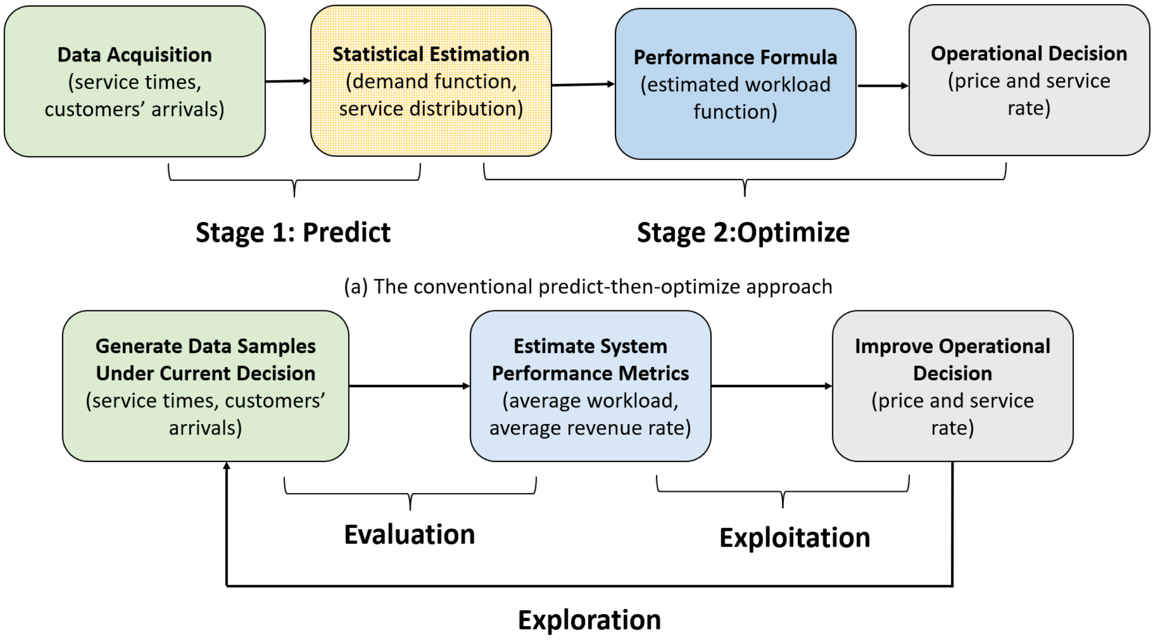
The conventional performance analysis and optimization in queueing systems require the precise knowledge of certain distributional information of the arrival process and service times. For example, consider the $M/GI/1$ queue having Poisson arrivals and general service times, the expected steady-state workload $W(\lambda,\mu,c_s^2)$ is a function of the arrival rate λ, service rate μ and second moment or *squared coefficient of variation* (SCV) $c_s^2 \equiv \mathrm{Var}(S)/\mathbb{E}[S]^2$ of the service time S. In particular, according to the famous Pollaczek–Khinchine (PK) formula (Pollaczek 1930), we have

$$\mathbb{E}[W(\lambda,\mu,c_s^2)] = \frac{\rho}{1-\rho}\frac{1+c_s^2}{2}, \qquad \text{with} \quad \rho \equiv \frac{\lambda}{\mu}. \tag{1}$$

One can never overstate the power of the PK formula because it has such a nice structure that insightfully ties the system performance to all model primitives λ, μ and c_s^2. Indeed, the PK formula has been predominantly used in practice and largely extended to several more general settings such as the $GI/GI/1$ queue with non-Poisson arrivals (Abate et al. 1993) and $M/GI/n$ queue with multiple servers (Cosmetatos 1976).

To optimize desired queueing performance, it is natural to follow the *predict-then-optimize* (PTO) approach, where "predict" means the estimation of required model parameters (e.g., λ, μ and c_s^2) from data (e.g., arrival times and service times) and "optimize" means the optimization of certain queueing decisions using formulas such as (1) with the predicted parameters treated as the true parameters. See panel (a) in Figure 1 for a flow chart of PTO. A potential issue of PTO is that the required queueing formulas can be highly sensitive to the estimation errors of the input parameters (e.g., λ and μ), especially when the system's congestion is critical. For example, when $c_s = \mu = 1$ and $\lambda = 0.99$, the PK formula (1) yields that $\mathbb{E}[W(\lambda,\mu,c_s^2)] = 99$. But a 0.5% increase of the demand rate λ will yield $\mathbb{E}[W(\lambda,\mu,c_s^2)] = 197$, resulting in a 99% relative error in the predicted workload. Consequently, the practical effectiveness of PTO heavily relies on the accuracy of the prediction step to provide near-perfect estimates of the input parameters. Without such precision, solution methods based on these convenient formulas may prove counterproductive or even fail to deliver the desired outcomes.

The performance shortcomings of PTO, particularly in heavy-traffic conditions, stem from its inability to adequately account for parameter estimation errors and the substantial impact these errors have on the quality of the resulting "optimized" decision variables. To help remedy this issue, we propose *an online learning framework that automatically incorporates the aforementioned parameter estimation errors in the solution prescription process; it is an end-to-end approach that can learn the optimal solution more directly from data, so that we no longer need to set up the parameter estimation as a separate stage as in PTO.* In this paper, we solve a pricing and capacity


Data Acquisition
(service times, customers' arrivals)
Statistical Estimation
(demand function, service distribution)
Performance Formula
(estimated workload function)
Operational Decision
(price and service rate)
Stage 1: Predict
Stage 2:Optimize
Generate Data Samples Under Current Decision
(service times, customers' arrivals)
Estimate System Performance Metrics
(average workload, average revenue rate)
Improve Operational Decision
(price and service rate)
Evaluation
Exploitation
Exploration

(a) The conventional predict-then-optimize approach

(b) The end-to-end online learning approach

Figure 1 Schematic presentations for (a) the two-step conventional *predict-then-optimize* scheme and (b) the end-to-end *online learning* scheme.

sizing problem in an $M/GI/1$ queue, where the service provider seeks the optimal service fee p and service rate μ so as to maximize the long-term profit, which is the revenue minus the staffing cost and the queueing penalty, namely,

$$\max_{\mu,p}\ \mathcal{P}(\mu,p) \equiv \lambda(p)p - h_0\mathbb{E}[W] - c(\mu), \tag{2}$$

where W is the system's steady-state workload, $c(\mu)$ is the cost for providing service capacity μ and h_0 is a holding cost per job per unit of time. Problems in this framework have a long history, see for example Kumar and Randhawa (2010), Lee and Ward (2014), Lee and Ward (2019), Maglaras and Zeevi (2003), Nair et al. (2016), Kim and Randhawa (2018), Chen et al. (2024) and the references therein. The major distinction is that in the present paper, we assume that neither the arrival rate $\lambda(p)$ (as a function of p) or the service-time distribution is explicitly available to the service provider. (As showed In Section 6.1.1, we will see that the PTO approach for solving Problem (2) indeed suffer from unaccountable estimation errors in the model parameters.)

Our online learning approach operates in successive cycles in which the service provider's operational decisions are being continuously evolved using newly generated data. Data here include customers' arrival and service times under the policy presently in use. See panel (b) in Figure 1 for an illustration of the online learning approach. In each iteration k, the service provider evaluates the current decision (μ_k, p_k) based on the newly generated data. Then, the decision is updated to

(μ_{k+1}, p_{k+1}) according to the evaluation result (the exploitation step). In the next iteration, the service provider continues to operate the system under (μ_{k+1}, p_{k+1}) to generate more data (the exploration step). We call this algorithm *Learning in Queue with Unknown Arrival Rate* (LiQUAR).

1.1. Advantages and challenges.

First, the conventional queueing control problem builds heavily on formulas such as (1) and requires the precise knowledge of certain distributional information which may not always be readily available. For example, the acquisition of an accurate estimate of the function $\lambda(p)$ across the entire spectrum of the price p is not straightforward and can be both time consuming and costly. In contrast, the online learning approach eliminates the need for such prior information, excelling at "learning from scratch". Second, unlike the two-step PTO procedure, the online learning approach is an integrated method that inherently accounts for estimation errors in observed data during the decision-making process. This allows it to utilize data more effectively, leading to improved decisions that are more robust and effective. In contrast to PTO's "static" learning, where prediction and optimization are distinctly separate steps, LiQUAR employs a reactive learning approach, characterized by its continuous and dynamic interaction with data.

On the other hand, the development of online learning methodologies in queue systems is far from a straightforward extension of their use in other fields, as it must address the unique characteristics of queueing dynamics. First, when the control policy is updated at the beginning of a cycle, the previously established near steady-state dynamics are disrupted, and the system enters a transient phase. The dynamics during this period are endogenously influenced by the updated control policy, giving rise to the so-called *regret of nonstationarity.* Second, the convergence of decision iterations depends heavily on the statistical efficiency of the evaluation step and the specific properties of the queueing data. This introduces new challenges due to the distinctive nature of queueing dynamics. Unlike standard online learning settings (e.g. stochastic bandits), queueing data such as waiting times and queue lengths are often *biased, unbounded, and temporally correlated.* These unique features of queueing models present significant obstacles to the design and analysis of online learning methodologies, necessitating novel approaches that account for these complexities. Finally, our algorithm operates without requiring knowledge of the arrival rate function or the service-time distribution. This makes our research problem more challenging because we cannot take advantage of the detailed structure of the underlying model. Therefore, we are motivated to develop a conceptually simple, model-free learning framework in order to address the above-mentioned challenges.

1.2. Contributions and organization

Our paper makes the following contributions.

- We are the first to develop an online learning scheme for the $M/GI/1$ queue with unknown demand function and service-time distribution. The effectiveness of our algorithm stems from its well-integrated queueing features, encompassing both the overall algorithm design and the optimization of hyperparameters. For our online learning algorithm, we establish a regret bound of $O(\sqrt{T}\log(T))$. In comparison with the standard $O(\sqrt{T})$ regret for model-free *stochastic gradient descent* (SGD) methods assuming unbiased and independent reward samples, our regret analysis exhibits an extra $\log(T)$ term which rises from the nonstationary queueing dynamics due to the policy updates. For the $M/M/1$ model, we derived a more detailed regret bound expressed explicitly as a function of the traffic intensity.
- At the heart of our regret analysis is to properly link the estimation errors from queueing data to the algorithm's hyperparameters and regret bound. For this purpose, we develop new results that establish useful statistical properties of data samples generated by a $M/G/1$ queue. Besides serving as building blocks for our regret analysis in the present paper, these results are of independent research interest and may be used to analyze the estimation errors of data in sequential decision making in ergodic queues. Hence, the theoretic analysis and construction of the gradient estimator may be extended to other queueing models which share similar ergodicity properties.
- Supplementing the theoretical results, we evaluate the practical effectiveness of our method by conducting comprehensive numerical experiments. In particular, our numerical results confirm that the online learning algorithm is efficient and robust to several model and algorithm parameters such as service distributions and updating step sizes; we also generalize our algorithm to the $GI/GI/1$ model. Next, we conduct a systematic analysis and experiments to compare LiQUAR to (i) PTO and (ii) gradient-based reinforcement learning methods.

Organization of the paper. In Section 2, we review the related literature. In Section 3, we introduce the model and its assumptions. In Section 4, we present LiQUAR and describe how the queueing data is processed in our algorithm. In Section 5, we conduct the convergence and regret analysis for LiQUAR. The key steps of our analysis form a quantitative explanation of how estimation errors in queueing data propagate through our algorithm flow and how they influence the quality of the LiQUAR solutions. We analyze the total regret by separately treating *regret of non-stationarity* - the part of regret stemming from transient system dynamics, *regret of suboptimality* - the part aroused by the errors due to suboptimal decsions, and *regret of finite difference* - the part originating from the need of estimation of gradient. In Section 5.3, we report a regret bound explicitly expressed as a function of the traffic intensity for $M/M/1$. In Section 6, we conduct numerical experiments to confirm the effectiveness and robustness of LiQUAR. In Sections 7 and 8, We compare LiQUAR to PTO and gradient-based reinforcement learning methods. We provide concluding remarks in Section 9. Technical proofs and supplementary results are given in the e-Companion.

2. Related Literature

The present paper is related to the following four streams of literature.

Pricing and capacity sizing in queues. There is rich literature on pricing and capacity sizing for service systems under various settings. Maglaras and Zeevi (2003) studies a static pricing and capacity sizing problem in a processor sharing queue motivated by internet applications; Kumar and Randhawa (2010) considers a static pricing problem for a single-server system with nonlinear delay cost; Nair et al. (2016) studies static capacity sizing problem for $M/M/1$ and $M/M/k$ systems with network effect among customers; Kim and Randhawa (2018) considers a dynamic pricing problem in an $M/M/1$ queue. The specific problem that we consider here is related to Lee and Ward (2014), which considers joint static pricing and capacity sizing for $GI/GI/1$ queues with known demand. Later, they further extend their results to the $GI/GI/1+G$ model with customer abandonment in Lee and Ward (2019). Although the present work is motivated by the pricing and capacity sizing problem for service systems, unlike the above-cited works, we assume no knowledge of the demand rate and service distribution.

Demand Learning. Broder and Rusmevichientong (2012) considers a dynamic pricing problem for a single product with an unknown parametric demand curve and establishes an optimal minimax regret in the order of $O(\sqrt{T})$. Keskin and Zeevi (2014) investigates a pricing problem for a set of products with an unknown parameter of the underlying demand curve. Besbes and Zeevi (2015) studies demand learning using a linear curve as a local approximation of the demand curve and establishes a minimax regret in the order of $O(\sqrt{T})$. Later, Cheung et al. (2017) solves a dynamic pricing and demand learning problem with limited price experiments. We draw distinctions from these papers by studying a pricing and capacity sizing problem with demand learning in a queueing setting where our algorithm design and analysis need to take into account unique features of the queueing systems.

Machine learning in queueing systems Our paper is related to a small but booming literature on machine earning in queueing systems. Dai and Gluzman (2021) studies an actor-critic algorithm for queueing networks. Liu et al. (2019) and Shah et al. (2020) develop reinforcement learning techniques to treat the unboundedness of the state space of queueing systems. Raeis et al. (2021) applies deep deterministic policy gradient algorithms to a service rate control problem. Murthy et al. (2024) considers natural policy gradient for control problems in communication networks. Comte et al. (2023) develops a policy gradient method with known score functions of queueing systems. Krishnasamy et al. (2021) develops bandit methods for scheduling problems in a multi-server queue with unknown service rates. Zhong et al. (2024) proposes an online learning method to study a scheduling problem for a multiclass $M_t/M/N+M$ system with unknown service rates and abandonment rates. Chen et al. (2024) studies the joint pricing and capacity sizing problem for $GI/GI/1$ with known demand. See Walton and Xu (2021) for a review of the role of information and learning in queueing systems. Recent research has explored the application of deep learning methods to predict queueing performance: Baron et al. (2023) proposes a deep-learning-based steady-state predictor for the $GI/GI/1$ queue; Garyfallos et al. (2024a,b) develop recurrent neural network models to predict transient performance in nonstationary queues. Our paper is most closely related to Jia et al. (2024) which studies a price-based revenue management problem in an $M/M/c$ queue with unknown demand and discrete price space, under a multi-armed bandit framework. Later, Jia et al. (2022) extends the results in Jia et al. (2024) to the problem setting with a continuous price space and considers linear demand functions. Similar to Jia et al. (2024, 2022), we also study a queueing control problem with unknown demand and continuous decision variables. The major distinction is that in addition to maximizing the service profit as by Jia et al. (2024), the present paper also includes a *queueing penalty* in our optimization problem as a measurement of the quality of service (Kumar and Randhawa 2010, Lee and Ward 2014, 2019). However, this introduces new technical challenges in algorithm design and regret analysis, such as addressing the bias and autocorrelation inherent in queueing data. Besides, the present paper considers more general service distributions and demand functions.

Online learning with continuous decision-making has also been explored in inventory systems. For instance, Huh et al. (2009) proposed an SGD-based algorithm to optimize base-stock policies for inventory systems with positive lead times. Later, Zhang et al. (2020) developed a simulation-based algorithm for the same problem, achieving an optimal regret bound of $O(T^{1/2})$. More recently, Yuan et al. (2021) integrated stochastic gradient descent with bandit algorithms to address convexity challenges and optimize (s,S) policies. While our approach also employs gradient-based methods, the fundamental differences between queueing system dynamics and inventory models lead to distinct algorithm designs, particularly in the construction of gradient estimators, as well as in the theoretical analysis.

3. Model and Assumptions

We study an $M/GI/1$ queueing system having customer arrivals according to a Poisson process (the M), *independent and identically distributed* (I.I.D.) service times following a general distribution (the GI), and a single server that provides service following the *first-in-first-out* (FIFO) discipline. Each customer upon joining the queue is charged by the service provider a fee $p > 0$. The demand arrival rate (per time unit) depends on the service fee p and is denoted as $\lambda(p)$. To maintain a service rate μ, the service provider continuously incurs a staffing cost at a rate $c(\mu)$ per time unit.

For $\mu \in [\underline{\mu}, \bar{\mu}]$ and $p \in [\underline{p}, \bar{p}]$, we have $\lambda(p) \in [\underline{\lambda}, \bar{\lambda}] \equiv [\lambda(\bar{p}), \lambda(\underline{p})]$, and the service provider's goal is to determine the optimal service fee p^* and service capacity μ^* with the objective of maximizing the steady-state expected profit, or equivalently minimizing the objective function $f(\mu, p)$ as follows

$$\min_{(\mu,p)\in\mathcal{B}} f(\mu,p) \equiv h_0 \mathbb{E}[W_\infty(\mu,p)] + c(\mu) - p\lambda(p), \qquad \mathcal{B} \equiv [\underline{\mu}, \bar{\mu}] \times [\underline{p}, \bar{p}]. \tag{3}$$

Here $W_\infty(\mu, p)$ is the stationary workload process observed in continuous time under control parameter (μ, p).

Remark 1 (Static Price and Service Capacity as Benchmark). In this paper, we focus on optimizing a static service fee p and service rate μ, whose optimal values can be explicitly characterized using the P–K formula when model parameters are known. This choice allows us to better focus on studying the effect of parameter uncertainty on the performance of learning algorithms in queues, while keeping the benchmark analytically tractable. Although the theoretically optimal policy over a finite horizon would be state- and time-dependent, analyzing such dynamic policies is substantially more complex and beyond the scope of this work. Moreover, static policies are widely adopted in practice, admit closed-form analysis, and are known to achieve robust performance in many settings (Kumar and Randhawa 2010, Elmachtoub and Shi 2023, Bergquist and Elmachtoub 2023).

In detail, under control parameter (μ, p), customers arrive according to a Poisson process with rate $\lambda(p)$. Let V_n be an I.I.D. sequence corresponding to customers' *workloads* under unit service rate (under service rate μ, customer n has service time V_n/μ). We have $\mathbb{E}[V_n] = 1$ so that the mean service time is $1/\mu$ under service rate μ. Denote by $N(t)$ the number of arrivals by time t. The total amount of workload brought by customers at time t is denoted by $J(t) = \sum_{k=1}^{N(t)} V_k$. Then the workload process $W(t)$ follows the *stochastic differential equation* (SDE)

$$dW(t) = dJ(t) - \mu \mathbf{1}(W(t) > 0)\, dt.$$

In particular, given the initial value of $W(0)$, we have

$$W(t) = R(t) - 0 \wedge \min_{0\le s\le t} R(s), \quad R(t) \equiv W(0) + J(t) - \mu t.$$

The difference $(W(t)-R(t))/\mu$ is the total idle time of the server by time t. It is known in the literature (Asmussen 2003, Corollary 3.3, Chapter X) that under the stability condition $\lambda(p)<\mu$, the workload process $W(t)$ has a unique stationary distribution and we denote by $W_\infty(\mu,p)$ the stationary workload under parameter (μ,p).

We impose the following assumptions on the $M/GI/1$ system throughout the paper.

Assumption 1. (***Demand rate, staffing cost, and uniform stability***)

(*a*) *The arrival rate $\lambda(p)$ is continuously differentiable in the third order and non-increasing in p. Besides,*

$$C_1<\lambda'(p)<C_2,$$

where

$$C_1\equiv 2\max\left(g(\bar{\mu})\frac{\lambda''(p)}{\lambda'(p)},g(\underline{\mu})\frac{\lambda''(p)}{\lambda'(p)}\right)\lambda(p)-\frac{4\lambda(p)(\underline{\mu}-\lambda(p))}{h_0C},\quad C_2\equiv-\max\left(\sqrt{\frac{0\vee(-\lambda''(p)(\bar{\mu}-\lambda(p)))}{2}}\ ,\ \frac{p\lambda''(p)}{2}\right),$$

$g(\mu)=\frac{\mu}{\mu-\lambda(p)}-\frac{p(\mu-\lambda(p))}{h_0C}$ *and* $C=(1+c_s^2)/2$.

(*b*) *The staffing cost $c(\mu)$ is continuously differentiable in the third order, non-decreasing and convex in μ.*

(*c*) *The lower bounds $\underline{p}$ and $\underline{\mu}$ satisfy that $\lambda(\underline{p})<\underline{\mu}$ so that the system is uniformly stable for all feasible choices of (μ,p).*

Although Condition (a) looks complicated, it essentially requires that the derivative of $\lambda(p)$ be not too large or too small. Condition (a) will be used to ensure that the objective function $f(\mu,p)$ is convex in the convergence analysis of our gradient-based online learning algorithm in Section 5.1. The two inequalities hold for a variety of commonly used demand functions, including both convex functions and concave functions. Examples include (1) linear demand $\lambda(p)=a-bp$ with $0<b<4\underline{\lambda}(\underline{\mu}-\bar{\lambda})/h_0C$; (2) quadratic demand $\lambda(p)=c-ap^2$ with $a,c>0$, and $\frac{\bar{\mu}-c}{3\underline{p}^2}<a<\left(\frac{3(\underline{\mu}-\bar{\lambda})\underline{p}}{h_0C}-\frac{\mu}{\underline{\mu}-\bar{\lambda}}\right)\frac{\underline{\lambda}}{\bar{p}^2}$; (3) exponential demand $\lambda(p)=\exp(a-bp)$ with $0<b<2/\bar{p}$; (4) logit demand $\lambda(p)=M_0\exp(a-bp)/(1+\exp(a-bp))$ with $a-b\bar{p}<\log(1/2)$ and $0<b<2/\bar{p}$. See Section EC.4 for detailed discussions.

Condition (c) of Assumption 1 is commonly used in the literature of SGD methods for queueing models to ensure that the steady-state mean waiting time $\mathbb{E}[W_\infty(\mu,p)]$ is differentiable with respect to model parameters. See Chong and Ramadge (1993), Fu (1990), L'Ecuyer et al. (1994), L'Ecuyer and Glynn (1994), and also Theorem 3.2 of Glasserman (1992). In Section EC.5.2, we present an initial attempt to relax Assumption 1(c).

We do not require full knowledge of service and inter-arrival time distributions. But in order to bound the estimation error of the queueing data, we require the individual workload to be light-tailed. Specifically, we make the following assumptions on V_n.

Assumption 2. (***Light-tailed individual workload***) *There exists a sufficiently small constant* $\eta > 0$ *such that*

$$\mathbb{E}[\exp(\eta V_n)] < \infty.$$

In addition, there exist constants $0 < \theta < \eta/2\bar{\mu}$ *and* $\gamma_0 > 0$ *such that*

$$\phi_V(\theta) < \log\left(1 + \underline{\mu}\theta/\bar{\lambda}\right) - \gamma_0, \tag{4}$$

where $\phi_V(\theta) \equiv \log \mathbb{E}[\exp(\theta V_n)]$ *is the cumulant generating functions of* V_n*.*

Note that $\phi_V'(0) = 1$ as $\mathbb{E}[V_n] = 1$. Suppose ϕ_V is smooth around 0, then we have $\phi_V(\theta) = \theta + o(\theta)$ by Taylor's expansion. On the other hand, as $\underline{\mu} > \bar{\lambda}$ under Assumption 1, there exists $a > 0$ such that $\log\left(1 + \underline{\mu}\theta/\bar{\lambda}\right) = (1+a)\theta + o(\theta)$. This implies that, we can choose θ small enough such that $\log\left(1 + \underline{\mu}\theta/\bar{\lambda}\right) - \phi_V(\theta) > \frac{a\theta}{2}$ and then we set $\gamma_0 = \frac{a\theta}{2}$. Hence, a sufficient condition that warrants (4) is to require that ϕ_V be smooth around 0, which is true for many distributions of V considered in common queueing models. Assumption 2 will be used in our proofs to establish ergodicity result.

4. Our Online Learning Algorithm

We first explain the main ideas in the design of LiQUAR and provide the algorithm outline in Section 4.1. The key step in our algorithm design is to construct a data-based gradient estimator, which is explained with details in Section 4.2. As a unique feature of service systems, there is a delay in data observation of individual workloads, i.e., they are revealed only after service completion. We also explain how to deal with this issue in Section 4.2. The design of algorithm hyperparameters in LiQUAR will be specified later in Section 5 based on the regret analysis results. In the rest of the paper, we use bold symbols for vectors and matrices.

4.1. Algorithm outline

The basic structure of LiQUAR follows the online learning scheme as illustrated in Figure 1. It interacts with the queueing system in continuous time and improves pricing and staffing policies iteratively. In each iteration $k \in \{1, 2, ...\}$, LiQUAR operates the queueing system according to control parameters $\bar{\boldsymbol{x}}_k \equiv (\bar{\mu}_k, \bar{p}_k)$ for a certain time period, and collects data generated by the queueing system during the period. At the end of an iteration, LiQUAR estimates the gradient of the objective function $\nabla f(\bar{\boldsymbol{x}}_k)$ based on the collected data and accordingly updates the control parameters. The updated control parameters will be used in the next iteration.

We use the *finite difference* (FD) method (Broadie et al. 2011) to construct our gradient estimator. Our main purpose is to make LiQUAR model-free and applicable to the settings where the demand function $\lambda(p)$ is unknown. To obtain the FD estimator of $\nabla f(\bar{\boldsymbol{x}}_k)$, LiQUAR splits total time of iteration k into two equally divided intervals (i.e., cycles) each with T_k time units. We index the two cycles by $2k-1$ and $2k$, in which the system is respectively operated under control parameters

$$\boldsymbol{x}_{2k-1} \equiv \bar{\boldsymbol{x}}_k - \delta_k \cdot \boldsymbol{Z}_k/2 \equiv (\mu_{2k-1}, p_{2k-1}) \quad \text{and} \quad \boldsymbol{x}_{2k} \equiv \bar{\boldsymbol{x}}_k + \delta_k \cdot \boldsymbol{Z}_k/2 \equiv (\mu_{2k}, p_{2k}), \tag{5}$$

where δ_k is a positive and small number and $\boldsymbol{Z}_k \in \mathbb{R}^2$ is a random vector independent of system dynamics such that $\mathbb{E}[\boldsymbol{Z}_k] = (1,1)^\top$. Using data collected in the two cycles, LiQUAR obtains estimates of the system performance $\hat{f}(\boldsymbol{x}_{2k})$ and $\hat{f}(\boldsymbol{x}_{2k-1})$, which in turn yield the FD approximation for the gradient $\nabla f(\bar{\boldsymbol{x}}_k)$:

$$\boldsymbol{H}_k \equiv \frac{\hat{f}(\boldsymbol{x}_{2k}) - \hat{f}(\boldsymbol{x}_{2k-1})}{\delta_k}.$$

Then, LiQUAR updates the control parameter according to a SGD recursion as $\bar{\boldsymbol{x}}_{k+1} = \Pi_{\mathcal{B}}(\bar{\boldsymbol{x}}_k - \eta_k \boldsymbol{H}_k)$, where $\Pi_{\mathcal{B}}$ is the operator that projects $\bar{\boldsymbol{x}}_k - \eta_k \boldsymbol{H}_k$ to $\mathcal{B}$. We give the outline of LiQUAR below.

Outline of LiQUAR:

0. Input: hyper-parameters $\{T_k, \eta_k, \delta_k\}$ for $k = 1, 2, ...$, initial policy $\bar{\boldsymbol{x}}_1 = (\bar{\mu}_1, \bar{p}_1)$.
   For $k = 1, 2, ..., L$,
1. Obtain $\boldsymbol{x}_l$ according to (5) for $l = 2k-1$ and $2k$. In cycle l, operate the system with policy $\boldsymbol{x}_l$ for T_k units of time.
2. Compute $\hat{f}(\boldsymbol{x}_{2k-1})$ and $\hat{f}(\boldsymbol{x}_{2k})$ from the queueing data to build an estimator $\boldsymbol{H}_k$ for $\nabla f(\mu_k, p_k)$.
3. Update $\bar{\boldsymbol{x}}_{k+1} = \Pi_{\mathcal{B}}(\bar{\boldsymbol{x}}_k - \eta_k \boldsymbol{H}_k)$.

Next, we explain in details how the gradient estimator $\boldsymbol{H}_k$, along with $\hat{f}(\boldsymbol{x}_{2k-1})$ and $\hat{f}(\boldsymbol{x}_{2k})$, are computed from the queueing data in Step 2.

4.2. Computing Gradient Estimator from Queueing Data

We first introduce some notation to describe the system dynamics under LiQUAR and the queueing data generated by LiQUAR. For $l \in \{2k-1, 2k\}$, let $W_l(t)$ be the present workload at time $t \in [0, T_k]$ in cycle l. By definition, we have $W_{l+1}(0) = W_l(T_k)$ for all $l \geq 1$. We assume that the system starts empty, i.e., $W_1(0) = 0$. At the beginning of each cycle l, the control parameter is updated to (μ_l, p_l). The customers arrive in cycle l according to a Poisson process $N_l(t)$ with rate $\lambda(p_l)$, $0 \leq t \leq T_k$. Let $\{V_i^l : i = 1, 2, ..., N_l\}$ be a sequence of I.I.D. random variables denoting customers' individual workloads, where $N_l = N_l(T_k)$ is the total number of customer arrival in cycle l. Then, the dynamics of the workload process $W_l(t)$ is described by the SDE:

$$W_l(t) = W_l(0) + \sum_{i=1}^{N_l(t)} V_i^l - \mu_l \int_0^t \mathbf{1}(W_l(s) > 0) ds. \tag{6}$$

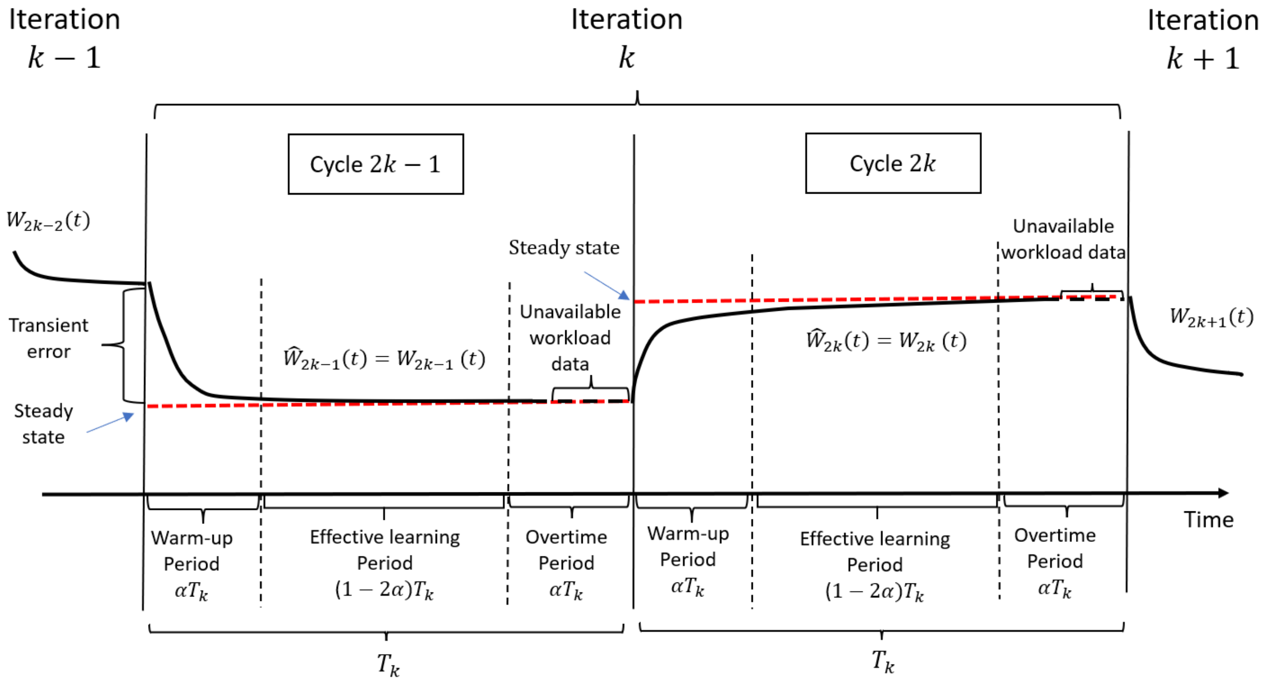

Iteration $k-1$
Iteration k
Iteration $k+1$
Cycle $2k-1$
Cycle $2k$
$W_{2k-2}(t)$
Transient error
Steady state
Steady state
Unavailable workload data
$\widehat{W}_{2k-1}(t) = W_{2k-1}(t)$
Unavailable workload data
$\widehat{W}_{2k}(t) = W_{2k}(t)$
$W_{2k+1}(t)$
Time
Warm-up Period αT_k
Effective learning Period $(1-2\alpha)T_k$
Overtime Period αT_k
Warm-up Period αT_k
Effective learning Period $(1-2\alpha)T_k$
Overtime Period αT_k
T_k
T_k

Figure 2 The system dynamics under LiQUAR.

If the system dynamics is available continuously in time (i.e. $W_l(t)$ was known for all $t \in [0, T_k]$ and $l = 2k-1, 2k$), then a natural estimator for $f(\mu_l, p_l)$ would be

$$\hat{f}(\mu_l, p_l) = \frac{-pN_l}{T_k} + \frac{h_0}{T_k}\int_0^{T_k} W_l(t)dt + c(\mu_l).$$

4.2.1. Retrieving workload data from service and arrival times. We assume that LiQUAR can observe each customer's arrivals in real time, but can only recover the individual workload at the service completion time. This assumption is consistent with real practice in many service systems. For example, in call center, hospital, etc., customer's individual workload is realized only after the service is completed. Hence, the workload process $W_l(t)$ is not immediately observable at t.

In LiQUAR, we approximate $W_l(t)$ by $\hat{W}_l(t)$ which we elaborate below. For given $l \geq 1$ and $t \in [0, T_k]$, if all customers arriving by time t can finish service by the end of cycle l, then all of their service times are realized, so we can recover $W_l(t)$ from the arrival times and service times of these customers using (6). Since customers are served under FIFO, it is straightforward to see that this happens if and only if $W_l(t) \leq \mu_l(T_k - t)$, i.e., the workload at time t is completely processed by T_k. Hence, we define the approximate workload as

$$\hat{W}_l(t) = \begin{cases} W_l(t), & \text{if } W_l(t) \leq \mu_l(T_k - t) \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

As illustrated in Figure 2, to reduce approximation error incurred by delayed observations of service times, we discard the $\hat{W}_l(t)$ data for $t \in ((1-\alpha)T_k, T_k]$; we call the subinterval $((1-\alpha)T_k, T_k]$ the overtime period in cycle l. The following Proposition 1 ensures that the approximation error $|\hat{W}_l(t) - W_l(t)|$ incurred by delayed observation vanishes exponentially fast as length of the overtime period increases. This result will be used in Section 5 to bound the estimation errors of the FD gradient estimator H_k.

Proposition 1 **(Bound on Error of Delayed Observation)**. *Under Assumptions 1 and 2, there exist some constants M and $\theta_0 > 0$ such that, for all $l \geq 1$ and $0 \leq t \leq T_k$,*

$$\mathbb{E}[|\hat{W}_l(t) - W_l(t)|] \leq \exp(-\theta_0 \underline{\mu}/2 \cdot (T_k - t))M.$$

Roughly speaking, M is the moment bound under the busiest traffic intensity, and θ_0 is a small number depending on θ in Assumption 2. The existence of them are shown in the Lemma EC.9 in Section EC.2.2.

4.2.2. Computing the gradient estimator. As illustrated in Figure 2, we also discard the data at the beginning of each cycle (i.e., $\hat{W}_l(t)$ for $t \in [0, \alpha T_k]$ in cycle l) in order to reduce the bias due to transient queueing dynamics incurred by the changes of the control parameters. We call $[0, \alpha T_k]$ the warm-up period of cycle l. Thus, we give the following system performance estimator under control x_l, $l \in \{2k-1, 2k\}$:

$$\hat{f}^G(\mu_l, p_l) = \frac{-pN_l}{T_k} + \frac{h_0}{(1-2\alpha)T_k} \int_{\alpha T_k}^{(1-\alpha)T_k} \hat{W}_l(t)dt + c(\mu_l), \tag{8}$$

and the corresponding FD gradient estimator

$$\boldsymbol{H}_k = \frac{\boldsymbol{Z}_k \cdot (\hat{f}^G(\mu_{2k}, p_{2k}) - \hat{f}^G(\mu_{2k-1}, p_{2k-1}))}{\delta_k}. \tag{9}$$

Unlike standard zero-order methods used in offline optimization problems, our data is generated through online interactions with the real system. Consequently, we must carefully tune our algorithm parameters to control the variance of H_k, as techniques like common random numbers can not be used to achieve variance reduction for $\hat{f}^G(\mu_{2k}, p_{2k}) - \hat{f}^G(\mu_{2k-1}, p_{2k-1})$.

The psuedo code of LiQUAR is given in Algorithm 1. To complete the design of LiQUAR algorithm, we still need to specify the hyperparameters T_k, η_k, δ_k for $k \geq 1$. We seek to optimize these hyperparameters in Section 5 to achieve minimized regret bound.

Remark 2 (Integrating queueing features into LiQUAR). For LiQUAR to effectively address our queueing control problem, its design should be well informed by the features and structures inherent to the underlying queueing system. First, we utilize workload data to construct the

```
Algorithm 1: LiQUAR
  Input: number of iterations $L$;
  parameters $0 < \alpha < 1$, and $T_k$, $\eta_k$, $\delta_k$ for $k = 1, 2, .., L$;
  initial value $\bar{x}_1 = (\bar{\mu}_1, \bar{p}_1)$, $W_1(0) = 0$;
  for $k = 1, 2, ..., L$ do
      Randomly draw $\boldsymbol{Z}_k \in \{(0,2), (2,0)\}$;
      Run Cycle $2k-1$: Run the system for $T_k$ units of time under control parameter
          $$\boldsymbol{x}_{2k-1} = \bar{\boldsymbol{x}}_k - \delta_k \boldsymbol{Z}_k/2 = (\bar{\mu}_k, \bar{p}_k) - \delta_k \boldsymbol{Z}_k/2,$$
      Run Cycle $2k$: Run the system for $T_k$ units of time under control parameter
          $$\boldsymbol{x}_{2k} = \bar{\boldsymbol{x}}_k + \delta_k \boldsymbol{Z}_k/2 = (\bar{\mu}_k, \bar{p}_k) + \delta_k \boldsymbol{Z}_k/2,$$
      Compute FD gradient estimator:
          $$\boldsymbol{H}_k = \frac{\boldsymbol{Z}_k}{\delta_k}\left[\frac{h_0}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\left(\hat{W}_{2k}(t) - \hat{W}_{2k-1}(t)\right)dt - \frac{p_{2k}N_{2k} - p_{2k-1}N_{2k-1}}{T_k} + c(\mu_{2k}) - c(\mu_{2k-1})\right]$$
          where $\hat{W}_l(\cdot)$ is an approximate of $W_l(\cdot)$ as specified in (7).
      Update $\bar{\boldsymbol{x}}_{k+1} = \Pi_{\mathcal{B}}(\bar{\boldsymbol{x}}_k - \eta_k \boldsymbol{H}_k)$.
  end
```

gradient estimator, moving beyond traditional reliance on arrival and service time data. This shift is particularly advantageous in heavy-traffic scenarios where queueing formulas are highly sensitive to input estimates such as demand and service distributions. To calculate the gradient estimator which involves integrating the workload process with delayed observations, we leverage the fact that the workload process is almost surely piecewise linear. This helps simplify the computation of the gradient estimator. Second, to mitigate transient biases and errors caused by delayed observations in the queueing data, we exclude data from a designated warm-up interval at the start of each cycle and an over-time interval at its end. The lengths of these intervals are determined by the exponential ergodic rate of the $M/GI/1$ queue. Third, the efficiency of learning algorithms, particularly gradient-based methods, critically depends on the choice of hyperparameters. Leveraging the regret analysis in Section 5, which extensively utilizes queueing properties such as transient bias and autocorrelation in the data, we carefully determine optimal hyperparameters to minimize regret.

5. Convergence Rate and Regret Analysis

In Section 5.1, we establish the rate of convergence for our decision variables (μ_k, p_k) under LiQUAR (Theorem 1). Besides, our analysis illustrate how the estimation errors in the queueing data will propagate to the iteration of (μ_k, p_k) and thus affect the quality of decision making. We follow three steps: First, we quantify the bias and mean square error of the estimated system performance $\hat{f}^G(\mu_l, p_l)$ computed from the queueing data via (8) (Proposition 2). To bound the estimation errors, we need to deal with the transient bias and stochastic variability in the queueing data. Next, using these estimation error bounds, we can determine the accuracy of the FD gradient estimator H_k in terms of the algorithm hyperparameters (Proposition 3). Finally, following the convergence analysis framework of SGD algorithms, we obtain the convergence rate of LiQUAR in terms of the algorithm hyperparameters (Theorem 1). The above three steps together form a quantitative explanation of how the errors are passed on from the queueing data to the learned decisions (whereas there is no such steps in PTO so its performance is much more sensitive to the errors in the data). In addition, the convergence result enables us to obtain the optimal choice of hyperparameters if the goal is to approximate $\boldsymbol{x}^* = (\mu^*, p^*)$ accurately with minimum number of iterations, which is often preferred in simulation-based offline learning settings.

In Section 5.2, we investigate the cost performance of dynamic pricing and capacity sizing decisions made by LiQUAR, via analyzing the total regret, which is the gap between the amount of cost produced by LiQUAR and that by the optimal control $\boldsymbol{x}^*$. Utilizing the convergence rate established by Theorem 1 and a separate analysis on the transient behavior of the system dynamics under LiQUAR (Proposition 4), we obtain a theoretic bound for the total regret of LiQUAR in terms of the algorithm hyperparameters. By simple optimization, we obtain an optimal choice of hyperparameters which leads to a total regret bound of order $O(\sqrt{T}\log(T))$ (Theorem 2), where T is the total amount of time in which the system is operated by LiQUAR .

5.1. Convergence Rate of Decision Variables

As $\bar{\boldsymbol{x}}_k$ evolves according to an SGD iteration in Algorithm 1, its convergence depends largely on how accurate the gradient is approximated by the FD estimator H_k. In the theoretical analysis, the accuracy of $\boldsymbol{H}_k$ is measured by the following two quantities:

$$B_k \equiv \mathbb{E}\left[\|\mathbb{E}[\boldsymbol{H}_k - \nabla f(\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]\|^2\right]^{1/2} \quad \text{and} \quad \mathcal{V}_k \equiv \mathbb{E}[\|\boldsymbol{H}_k\|^2],$$

where $\mathcal{F}_k$ is the σ-algebra including all events in the first $2(k-2)$ cycles and $\|\cdot\|$ is Euclidean norm in $\mathbb{R}^2$. Intuitively, B_k measures the bias of the gradient estimator $\boldsymbol{H}_k$ and $\mathcal{V}_k$ measures its variability.

According to (9), the gradient estimator $\boldsymbol{H}_k$ is computed using the estimated system performance $\hat{f}^G(\mu_{2k}, p_{2k})$ and $\hat{f}^G(\mu_{2k-1}, p_{2k-1})$. So, the accuracy of $\boldsymbol{H}_k$ essentially depends on the estimation errors of the system performance, i.e., how close is $\hat{f}^G(\mu_l, p_l)$ to $f(\mu_l, p_l)$. Note that the control parameters (μ_l, p_l) for $l \in \{2k-1, 2k\}$ are random and dependent on the events in the first $2(k-2)$ cycles. Accordingly, we need to analyze the estimation error of $\hat{f}^G(\mu_l, p_l)$ conditional on the past events, which is also consistent with our definition of B_k. For this purpose, we denote by $\mathcal{G}_l$ the σ-algebra including all events in the first $l-1$ cycles and write $\mathbb{E}_l[\cdot] \equiv \mathbb{E}[\cdot|\mathcal{G}_l]$. The following Proposition 2 establishes bounds on the conditional bias and mean square error of $\hat{f}^G(\mu_l, p_l)$, in terms of the initial workload $W_l(0)$ and the hyperparameter T_k.

Proposition 2 **(Estimation Errors of System Performance)**. *Under Assumptions 1 and 2, for any $T_k > 0$, the bias and mean square error of $\hat{f}^G(\mu_l, p_l)$, conditional on $\mathcal{G}_l$, have the following bounds:*

1. *Bias*

$$\left|\mathbb{E}_l\left[\hat{f}^G(\mu_l, p_l) - f(\mu_l, p_l)\right]\right| \leq \frac{2\exp(-\theta_1 \alpha T_k)}{(1-2\alpha)\theta_1 T_k} \cdot M(M + W_l(0))(\exp(\theta_0 W_l(0)) + M).$$

2. *Mean square error*

$$\mathbb{E}_l[(\hat{f}^G(\mu_l, p_l) - f(\mu_l, p_l))^2] \leq K_M T_k^{-1}(W_l^2(0) + 1)\exp(\theta_0 W_l(0)),$$

where $\theta_1 \equiv \min(\gamma, \theta_0 \underline{\mu}/2)$, and γ and K_M are two positive constants that are independent of $l, T_k, W_l(0), \mu_l$ and p_l.

The proof of Proposition 2 is given in Section EC.1, where the specification of the two constants γ and K_M are given in (EC.9) and (EC.5) respectively. The key step in the proof is to bound the transient bias (from the steady-state distribution) and auto-correlation of the workload process $\{W_l(t) : 0 \leq t \leq T_k\}$, utilizing an ergodicity analysis. This approach can be applied to other queueing models which share similar ergodicity properties, e.g., GI/GI/1 queue and stochastic networks (Blanchet and Chen 2020).

Based on Proposition 2, we establish the following bounds on B_k and $\mathcal{V}_k$ in terms of the algorithm hyperparameters T_k and δ_k.

Proposition 3 **(Bounds for B_k and $\mathcal{V}_k$)**. *Under Assumptions 1 and 2, the bias and variance of the gradient estimator satisfy*

$$B_k = O\left(\delta_k^2 + \delta_k^{-1}\exp(-\theta_1 \alpha T_k)\right), \quad \mathcal{V}_k = O\left(\delta_k^{-2} T_k^{-1} \vee 1\right). \tag{10}$$

Assumption 1 guarantees that the objective function $f(\mu,p)$ in (3) has desired convex structure (see Lemma EC.5 in Section EC.1 for details). Hence, the SGD iteration is guaranteed to converge to its optimal solution x^* as long as the gradient bias B_k and variance $\mathcal{V}_k$ are properly bounded. Utilizing the bounds on B_k and $\mathcal{V}_k$ as given in Proposition 3, we are able to prove the convergence of LiQUAR and obtain an explicit expression of the convergence rate in terms of algorithm hyperparameters.

Theorem 1 (Convergence rate of decision variables) *Suppose Assumption 1 holds. If there exists a constant $\beta\in(0,1]$ such that the following inequalities hold for all k large enough:*

$$\left(1+\frac{1}{k}\right)^{\beta}\le 1+\frac{K_0}{2}\eta_k,\quad B_k\le\frac{K_0}{8}k^{-\beta},\quad \eta_k\mathcal{V}_k=O(k^{-\beta}). \tag{11}$$

Then, we have

$$\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2\right]=O(k^{-\beta}). \tag{12}$$

If, in further, Assumption 2 holds and the algorithm hyperparameters are set as $\eta_k=O(k^{-a})$, $T_k=O(k^b)$, and $\delta_k=O(k^{-c})$ for some constants $a,b,c\in(0,1]$. We have

$$\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2\right]=O\left(k^{\max(-a,-a-b+2c,-2c)}\right). \tag{13}$$

Remark 3 (optimal convergence rate). According to the bound (13), by minimizing the term $\max(-a,a-b+2c,-2c)$, one can obtain an optimal choice of hyperparameters $\eta_k=O(k^{-1}),T_k=O(k)$ and $\delta_k=O(k^{-1/2})$ under which the decision parameter $\boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$ at a fastest rate of $O(L^{-1})$, in terms of the total number of iterations L. Of course, the above convergence rate analysis does not focus on reducing the total system cost generated through the learning process, which is what we will do in Section 5.2.

5.2. Regret Analysis

Having established the convergence of control parameters under Assumption 1, we next investigate the efficacy of LiQUAR as measured by the cumulative regret which measures the gap between the cost under LiQUAR and that under the optimal control. According to the system dynamics described in Section 4.2, under LiQUAR, the expected cost incurred in cycle l is

$$\rho_l\equiv\mathbb{E}\left[h_0\int_0^{T_k}W_l(t)dt+c(\mu_l)T_k-p_lN_l\right], \tag{14}$$

where $k=\lceil l/2\rceil$. The total regret in the first L iterations (each iteration contains two cycles) is

$$R(L)=\sum_{k=1}^{L}\sum_{l=2k-1}^{2k}R_l=\sum_{l=1}^{2L}R_l,\quad \text{with } R_l\equiv\rho_l-T_kf(\mu^*,\rho^*).$$

Our main idea is to separate the total regret $R(L)$ into three parts as

$$\begin{aligned} R(L) = &\sum_{k=1}^{L} \underbrace{\mathbb{E}\left[2T_k(f(\bar{\boldsymbol{x}}_k) - f(\boldsymbol{x}^*))\right]}_{\equiv R_{1k}\text{: regret of suboptimality}} \\ &+ \sum_{k=1}^{L} \underbrace{\mathbb{E}\left[(\rho_{2k-1} - T_k f(\boldsymbol{x}_{2k-1})) + (\rho_{2k} - T_k f(\boldsymbol{x}_{2k}))\right]}_{\equiv R_{2k}\text{: regret of nonstationarity}} \\ &+ \sum_{k=1}^{L} \underbrace{\mathbb{E}\left[T_k(f(\boldsymbol{x}_{2k-1}) + f(\boldsymbol{x}_{2k}) - 2f(\bar{\boldsymbol{x}}_k))\right]}_{\equiv R_{3k}\text{: regret of finite difference}}, \end{aligned} \tag{15}$$

which arise from the errors due to the suboptimal decisions (R_{1k}), the transient system dynamics (R_{2k}), and the estimation of gradient (R_{3k}), respectively. Then we aim to minimize the orders of all three regret terms by selecting the "optimal" algorithm hyperparameters T_k, η_k and δ_k for $k \geq 1$.

Treating R_{1k}, R_{2k}, R_{3k} ***separately.*** Suppose the hyperparameters of LiQUAR are set in the form of $\eta_k = O(k^{-a})$, $T_k = O(k^b)$, and $\delta_k = O(k^{-c})$ for some constants $a, b, c \in (0, 1]$. The first regret term R_{1k} is determined by the convergence rate of control parameter $\bar{x}_k$. By Taylor's expansion, $f(\bar{\boldsymbol{x}}_k) - f(\boldsymbol{x}^*) = O(\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|_2^2)$, and hence, $R_{1k} = O(T_k\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|_2^2)$. Following Theorem 1, we have $R_{1k} = O(k^{\max(b-a, b-2c, -a+2c)})$. By the smoothness condition in Assumption 1, we can check that $R_{3k} = O(T_k\delta_k^2) = O(k^{b-2c})$ (Lemma EC.8 in Section EC.1).

The remaining regret analysis will focus on the regret of nonstationarity R_{2k}. Intuitively, it depends on the rate at which the (transient) queueing dynamics converges to its steady state. Applying the same ergodicity analysis as used in the analysis of estimation errors of system performance, we can find a proper bound on the transient bias after the warm-up period, i.e., for $W_l(t)$ with $t \geq \alpha T_k$. Derivation of a desirable bound on the transient bias in the warm-up period, i.e., for $W_l(t)$ with $t \in [0, \alpha T_k]$, is less straightforward. The main idea is based on the two facts that (1) $W_l(t)$, when t is small, is close to the steady-state workload corresponding to (μ_{l-1}, p_{l-1}) and that (2) the steady-state workload corresponding to (μ_{l-1}, p_{l-1}) is close to that of (μ_l, p_l). We formalize the bound on R_{2k} in Proposition 4 below. The complete proof is given in Section EC.1.6.

PROPOSITION 4 **(Regret of Nonstationarity)**. *Suppose Assumptions 1 and 2 hold. If* $T_k > \log(k)/\gamma$ *and there exists some constant* $\xi \in (0, 1]$ *such that* $\max(\eta_k\sqrt{\mathcal{V}_k}, \delta_k) = O(k^{-\xi})$*. Then,*

$$R_{2k} = O\left(k^{-\xi}\log(k)\right). \tag{16}$$

If, in further, the algorithm hyperparameters are set as $\eta_k = O(k^{-a})$*,* $T_k = O(k^b)$*, and* $\delta_k = O(k^{-c})$ *for some constants* $a, b, c \in (0, 1]$*, we have*

$$R_{2k} = O\left(k^{\max(-a-b/2+c, -a, -c)}\log(k)\right).$$

By summing up the three regret terms, we can conclude that

$$R(L) \leq \sum_{k=1}^{L} C\left(k^{\max(-a-b/2+c,-a,-c)}\log(k) + k^{\max(b-a,b-2c,-a+2c)} + k^{b-2c}\right),$$

for some positive constant C that is large enough. The order of the upper bound on the right hand side reaches its minimum at $(a,b,c)=(1,1/3,1/3)$. The corresponding total regret and time elapsed in the first L iterations are, respectively,

$$R(L) = O(L^{2/3}\log(L)) \quad \text{and} \quad T(L) = O(L^{4/3}).$$

As a consequence, we have $R(T) = O(\sqrt{T}\log(T))$.

Theorem 2 (Regret Upper Bound) *Suppose Assumptions 1 and 2 hold. If we choose* $\eta_k = c_\eta k^{-1}$ *for some* $c_\eta > 2/K_0$*,* $T_k = c_T k^{1/3}$ *for some* $c_T > 0$ *and* $\delta_k = c_\delta k^{1/3}$ *for some* $0 < c_\delta < \sqrt{K_0/32c}$*, where* c *is a smoothness constant given in Lemma EC.4, then the total regret accumulated in the first* L *rounds by LiQUAR*

$$R(L) = O(L^{2/3}\log(L)) = O(\sqrt{T(L)}\log(T(L))).$$

Here $T(L)$ *is the total units of time elapsed in* L *cycles.*

Remark 4 (On the $O(\sqrt{T}\log(T))$ Regret Bound). Consider a hypothetical setting in which we are no longer concerned with the transient behavior of the queueing system, i.e., somehow we can directly observe an unbiased and independent sample of the objective function with uniform bounded variance in each iteration. In this case, we know that the Kiefer-Wolfowitz algorithm and its variate provide an effective approach for model-free stochastic optimization (Broadie et al. 2011). According to Broadie et al. (2011), the convergence rate of Kiefer-Wolfowitz algorithm is $\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 = O(\eta_k/\delta_k^2)$. In addition, the regret of finite difference is $f(\boldsymbol{x}_{2k-1}) + f(\boldsymbol{x}_{2k}) - 2f(\bar{\boldsymbol{x}}_k) = O(\delta_k^2)$. Since $\eta_k/\delta_k^2 + \delta_k^2 \geq 2\sqrt{\eta_k} \geq k^{-1/2}$, we can conclude that the optimal convergence rate in such a hypothetical setting is $O(k^{-1/2})$. This accounts for the $\sqrt{T}$ part of our regret in Theorem 2. Unfortunately, unlike the hypothetical setting, our queueing samples are biased and correlated. Such a complication is due to the nonstationary error at the beginning of cycles which gives rise to the extra $\log(T)$ term in the regret bound; see Proposition 4 for additional discussion of the $\log(T)$ term in our regret. In Section EC.3, we present a theoretic result showing that the lower bound for the regret of suboptimality is $\Omega(\sqrt{T})$.

5.3. LIQUAR in Heavy Traffic

We now evaluate LiQUAR's performance under heavy traffic conditions. To do this, we construct a series of queueing models with traffic intensities that approach the critical threshold of 1, and define the associated profit optimization problems. Our goal is to derive an explicit regret expression as a function of traffic intensity. To achieve this, we will need to consider a simplified model in order for reduced technicalities in our regret analysis. We consider two simplifications: first, we focus on the case of exponential service times; second, we focus on a pricing problem, treating the service rate μ as a constant. We stress that the simplified model still preserve the challenge of dealing with unknown demand (the core aspect of the learning problem).

Consider a sequence of $M/M/1$ systems indexed by the parameter $h > 0$, which is the queueing congestion cost per time unit. They share a common demand curve $\lambda(p)$ and service rate $\mu = 1$. For the h^{th} model, we aim to minimize the objective function below

$$\min_{p \in \mathcal{B}_h} \quad f_h(p) \equiv -p\lambda(p) + \frac{h\lambda(p)}{1-\lambda(p)}, \tag{17}$$

where $\mathcal{B}_h$ will be specified later. When the holding cost h in (17) decreases, the service provider is incentivized to increase service utilization to maximize profit which places the system under the heavy-traffic regime. Below we will formally show that the traffic intensity under the optimal price ρ_h^* converges to 1 as $h \to 0$, specifically, below we will show $1 - \rho_h^* = O(\sqrt{h})$.

We denote by p_h^* as the optimal solution to (17). To explicitly show the relationship between ρ_h^* and h, we impose the following assumptions on the demand curve.

Assumption 3 **(Demand Curve)**. *We assume that the arrival rate function $\lambda(p)$ satisfies the following conditions:*

1. *The demand function is non-increasing and twice-differentiable.*
2. *The function $r(p) = p\lambda(p)$ is strictly concave.*
3. *The demand function is elastic in the feasible regions: $-\frac{\lambda'(p)}{\lambda(p)} \cdot p > 1$ for $p \in \mathcal{B}_h$.*

The third technical condition is commonly used in the literature of revenue management in queues; see for example, Assumption 1 in Maglaras and Zeevi (2003). Essentially, this condition assumes that customers are price sensitive in the feasible region.

Denote p_0 as the price that makes the system critically loaded, i.e., $\lambda(p_0) = 1$. Before giving our regret bound, we first characterize the optimal pricing decision p_h^* as a function of h and relate the traffic intensity ρ_h^* to h.

Proposition 5. *Under Assumption 3, we have the optimal price*

$$p_h^* \equiv \arg\min f_h(p) = p_0 + \sqrt{h} \cdot \sqrt{\frac{1}{(1+p_0\lambda'(p_0))\lambda'(p_0)}} + o(\sqrt{h}),$$

and the corresponding optimal service excess

$$1-\rho_h^* = \sqrt{h}\cdot\sqrt{\frac{-\lambda'(p_0)}{1+p_0\lambda'(p_0)}} + o(\sqrt{h}) = O(\sqrt{h}).$$

To investigate LiQUAR's performance in heavy traffic, we consider $\mathcal{B}_h$ which asymptotically operates the system in heavy traffic as $h\to 0$. Following Proposition 5, we let

$$p_h^* \in \left[\; p_0 + c_1\sqrt{h}, p_0 + c_2\sqrt{h}\;\right] \equiv \mathcal{B}_h,$$

where the constants c_1 and c_2 satisfy $0 < c_1 < c_0 \equiv 1/\sqrt{\lambda'(p_0)(1+p_0\lambda'(p_0))} < c_2$. Note that as $\rho_h^* \to 1$ (or equivalently, $h\to 0$), the queueing system takes a longer time to converge to its steady-state. Therefore, as $h\to 0$, we need to increase the length of the learning cycle. We operate the h^{th} model under LiQUAR with a total time duration of $T^h \equiv T_0/h$ units where T_0 is a positive constant independent of h. We evaluate our performance using the regret

$$R^h(T_0) \equiv R(T^h).$$

In what follows, we report the theoretical regret bound as a function of the traffic intensity.

Theorem 3 (Regret Bound in Heavy Traffic) *For the h^{th} system operated under LiQUAR to minimize* (17)*, when the hyper-parameters are chosen as*

$$T_k^h = c_T h^{-1} k^{1/3}, \quad \delta_k^h = c_\delta \sqrt{h} k^{-1/3}, \quad \eta_k^h = c_\eta \sqrt{h} k^{-1}, \tag{18}$$

where the constants c_T, c_δ and c_η are independent of h, then the regret is given by

$$R^h(T_0) \le C\sqrt{h^{-1}T_0\log T_0} = O(\sqrt{T_0\log T_0}/(1-\rho_h^*)), \tag{19}$$

where C is a positive constant independent of h.

The result in Theorem 3 refines that in Theorem 2 by emphasizing how the regret depends on the system's traffic intensity. In Section 7, we also conduct heavy-traffic analysis for a nonparametric PTO framework; and we compare the performance under both methods in our heavy-traffic regime both theoretically and numerically. In addition, in Section EC.3, we show that under our heavy-traffic scheme, for **any** algorithm, the lower bound for the regret of suboptimality is $\Omega(\sqrt{T_0/h}) = \Omega\left(\sqrt{T_0}/(1-\rho_h^*)\right)$.

6. Numerical Experiments

We provide engineering confirmations of the effectiveness of LiQUAR by conducting a series of numerical experiments. We will use simulated data to visualize the convergence of LiQUAR, estimate the regret curves and benchmark them with our theoretical bounds. In Section 6.1, we evaluate the performance of LiQUAR using an $M/M/1$ base example with logit demand functions. In Section 6.2, we discuss how to fine-tune the algorithm's hyperparameters including T_k and η_k. In Section 6.3, we generalize LiQUAR to $GI/GI/1$ queues with non-Poisson arrivals and evaluate its performance.

6.1. An $M/M/1$ Base Example

Our base model is an $M/M/1$ queues having Poisson arrivals with rate $\lambda(p)$ and exponential service times with rate μ. We consider a logistic demand function (Besbes and Zeevi 2015)

$$\lambda(p) = M_0 \cdot \frac{\exp(a-bp)}{1+\exp(a-bp)}, \tag{20}$$

with $M_0 = 10, a = 4.1, b = 1$ and a linear staffing cost function

$$c(\mu) = c_0\mu. \tag{21}$$

The demand function is shown in the top left panel in Figure 4. Then, the service provider's profit optimization problem (2) reduces to

$$\max_{\mu,p}\left\{p\lambda(p) - h_0\frac{\lambda(p)/\mu}{1-\lambda(p)/\mu} - c_0\mu\right\}. \tag{22}$$

6.1.1. Performance sensitivity to parameter errors without learning We first illustrate how the parameter estimation error impacts the performance. Here we assume the service provider does not know the true value of $\lambda(p)$ but rather make decisions based on an estimated arrival rate $\hat{\lambda}_\epsilon(p) \equiv (1-\epsilon\%)\lambda(p)$, where ϵ is the percentage estimator error. Let $(\hat{\mu}_\epsilon, \hat{p}_\epsilon)$ and (μ^*, p^*) be the solutions under the estimated $\hat{\lambda}_\epsilon$ and the true value of λ. We next compute the relative profit loss due to the misspecification of the demand function $\left(\mathcal{P}(\mu^*, p^*) - \mathcal{P}(\hat{\mu}_\epsilon, \hat{p}_\epsilon)\right)/\mathcal{P}(\mu^*, p^*)$, which is the relative difference between profit under the miscalculated solutions using the believed $\hat{\lambda}_\epsilon$ and the true optimal profit under λ.

Let $\rho^* \equiv \lambda(p^*)/\mu^*$ be the traffic intensity under the true optimal solution. We are able to impact the value of ρ^* by varying the queueing penalty coefficient h_0. We provide an illustration in Figure 3 with $\epsilon = 5$. From the left panel of Figure 3, we can see that as ρ^* increases, the model fidelity becomes more sensitive to the misspecification error in the demand rate and the relative loss of profit grows dramatically as ρ^* goes closer to 1. This effect arises from the fact that the error predicted

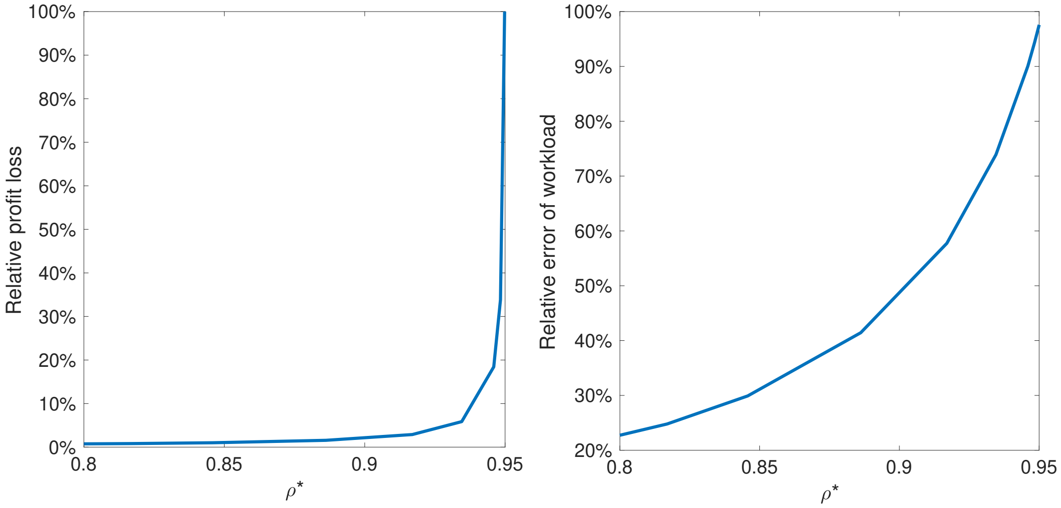


Figure 3 Relative profit loss (left) and workload error (right) for the $M/M/1$ example with $M_0 = 10, a = 4.1$, and $b = 1$ and linear staffing cost $c(\mu) = \mu$.

workload is extremely sensitive to that in the arrival rate and is disproportionally amplified by the PK formula when the system is in heavy traffic (see panel (b) for the relative error of the workload). Later in Section 7, we will conduct a careful comparison to the PTO method where we will compute the PTO regret including profit losses in both the prediction and optimization steps.

6.1.2. Performance of LiQUAR Using the explicit forms of (22), we first numerically obtain the exact optimal solution (μ^*, p^*) and the maximum profit $\mathcal{P}(\mu^*, p^*)$ which will serve as benchmarks for LiQUAR. Taking $h_0 = 1$ and $c_0 = 1$ yields $(\mu^*, p^*) = (8.18, 3.79)$, and the corresponding profit plot is shown in the top right panel of Figure 4. To test the criticality of condition (b) in Assumption 1, we implement LiQUAR when condition (b) does not hold. For this purpose, we set $\mathcal{B} = [6.5, 10] \times [3.5, 7]$, in which the objective (3) is not always convex, let alone the condition (b) of Assumption 1 (top right and middle right panel of Figure 4).

Then we implement LiQUAR without exploiting the specific knowledge of the exponential service distribution or the form of $\lambda(p)$. In light of Theorem 2, we set the hyperparameters $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$ and $\alpha = 0.1$. From Figure 4, we observe that the pair (μ_k, p_k), despite some stochastic fluctuations, converges to the optimal decision rapidly. The regret is estimated by averaging 100 sample paths and showed in the bottom left panel of Figure 4. To better relate the regret curve to its theoretical bounds as established in Theorem 2, we also draw the logarithm of regret as a function of the logarithm of the total time; we fit the log-log curve to a straight line (bottom right panel of Figure 4) so that the slope of the line may be used to quantify

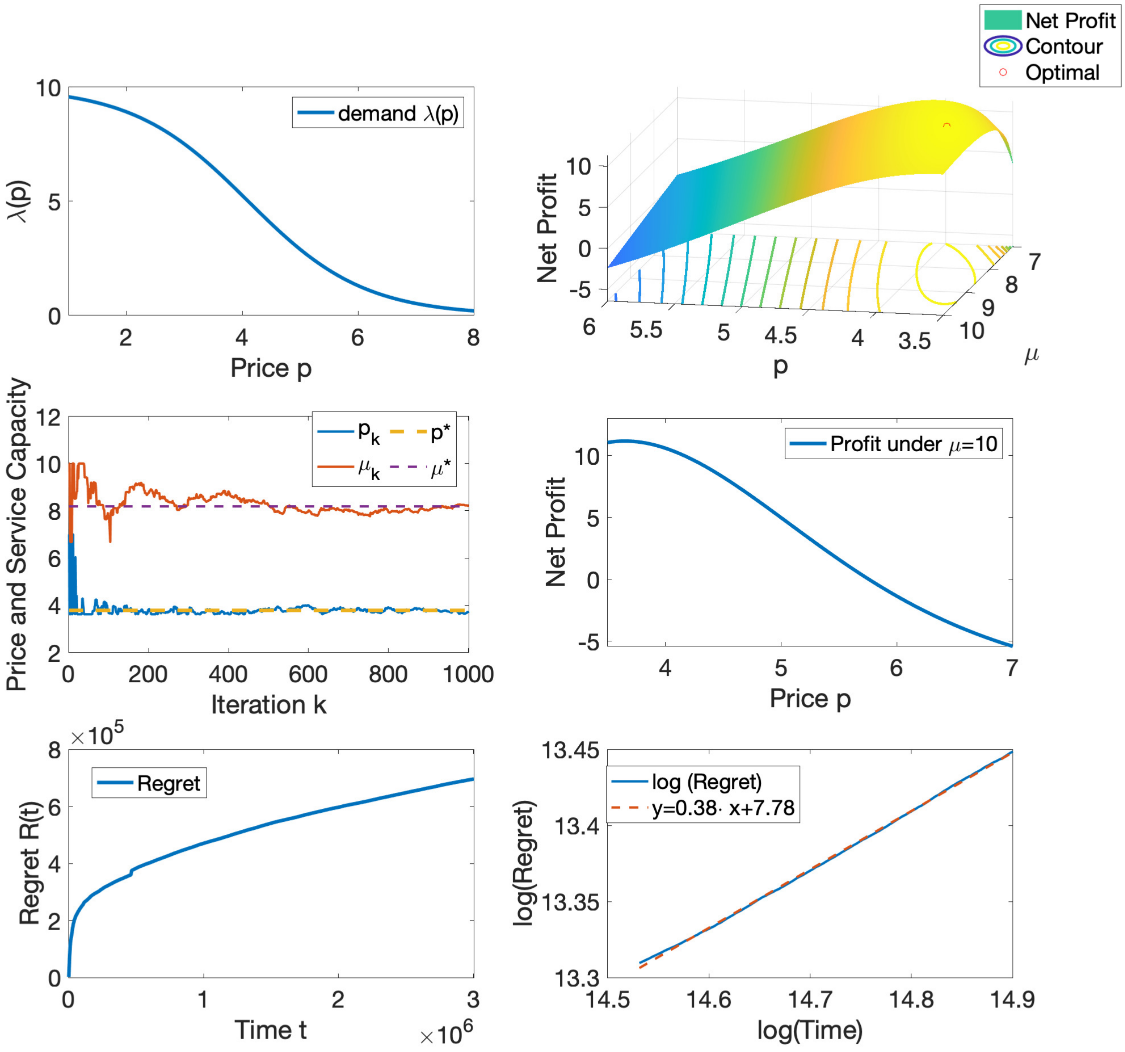


Figure 4 Joint pricing and staffing in the $M/M/1$ logistic demand base example with $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$, $p_0 = 5$, $\mu_0 = 10$ and $\alpha = 0.1$: (i) Demand function $\lambda(p)$ (top left panel); (ii) net profit function (top right panel); (iii) sample trajectories of decision parameters (middle left); (iv) One dimensional net profit function when $\mu = 10$; (v) average regret curve estimated by 100 independent runs (bottom left); (vi) a linear fit to the regret curve in logarithm scale.

the theoretic order of regret: the fitted slope (0.38) is less than its theoretical upper bound (0.5). Such “overperformance” is not too surprising because the theoretic regret bound is established based on a worst-case analysis. In summary, our numerical experiment shows that the technical condition (b) in Assumption 1 does not seem to be too restrictive.

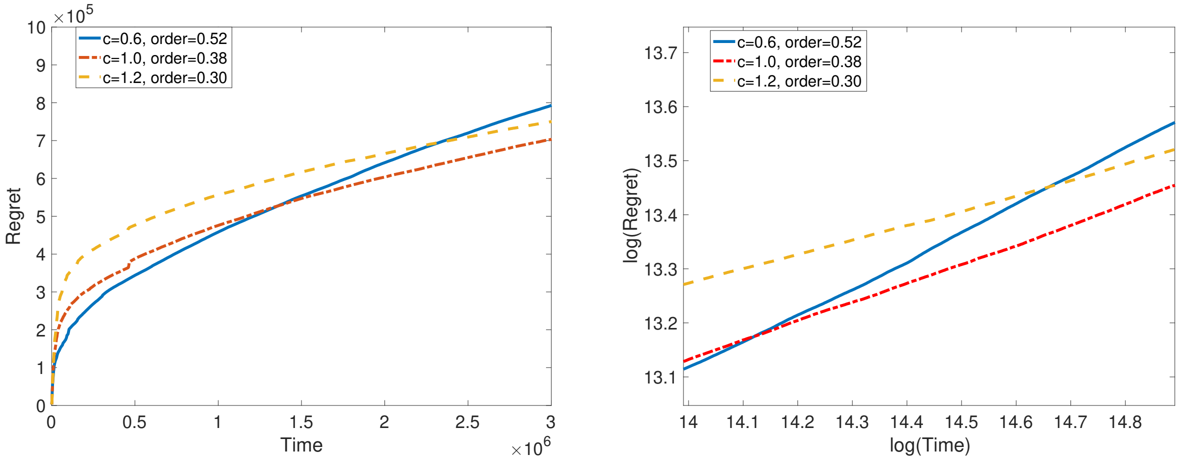


Figure 5 Regret under different $c \in \{0.6, 1.0, 1.2\}$: (i) average regret from 100 independent runs (left panel); (ii) regret curve in logarithm scale, with $T_k = 200k^{1/3}$, $\eta_k = c \cdot 4k^{-1}, \delta_k = \min(0.1, c \cdot 0.5k^{-1/3})$ and $\alpha = 0.1$.

6.2. Tuning the hyperparameters for LiQUAR

Next, we test the performance of LiQUAR on the base $M/M/1$ example under different hyperparameters. We also provide some general guidelines on the choices of hyperparameters when applying LiQUAR in practice.

6.2.1. Step lengths η_k and δ_k. In the first experiment, we tune the step length η_k and δ_k jointly within the following form:

$$\eta_k = c \cdot 4k^{-1}, \quad \text{and} \quad \delta_k = \min(0.1, c \cdot 0.5k^{-1/3}). \tag{23}$$

To understand the rationale of this form, note that these parameters give critical control to the variance of the gradient estimator. We aim to keep the variance of the term $\eta_k H_k$ at the same level in the gradient descent update

$$\boldsymbol{x}_{k+1} = \Pi_{\mathcal{B}}(\boldsymbol{x}_k - \eta_k \boldsymbol{H}_k), \qquad \text{with} \quad \eta_k \boldsymbol{H}_k = \eta_k \frac{\hat{f}(\boldsymbol{x}_k + \delta_k/2 \cdot \boldsymbol{Z}_k) - \hat{f}(\boldsymbol{x}_k - \delta_k/2 \cdot \boldsymbol{Z}_k)}{\delta_k}.$$

In this experiment, we let $c \in \{0.6, 1.0, 1.2\}$ and fix $T_k = 200k^{1/3}$ and $\alpha = 0.1$. For each case, the regret curve is estimated by 100 independent runs for $L = 1000$ iterations. The regret and its linear fit are reported in Figure 5. As shown in the right panel of Figure 5, the regret of LiQUAR has slopes of the linear regret fit close to 0.5 in all three cases. Comparing the two curves with $c = 0.6$ and $c = 1.2$ (left panel of Figure 5), we find that the larger value of c immediately accumulates a large regret in the early stages but performs better in the later iterations. This observation may be explained by the trade-off between the level of exploration and learning rate of LiQUAR. In particular, a larger value of c leads to larger values of η_k and δ_k, which allows more aggressive exploration and higher learning rate.

Although the tuning of c will not affect the convergence of asymptotic regret of the algorithm, it may be critical to decision making in a finite-time period. For example, a myopic decision maker who prefers good system performance in a short term should consider small values of c, while a far-sighted decision maker who values more the long-term performance should adopt a larger c.

6.2.2. Cycle length T_k. In this experiment, we test the impact of T_k on the performance of LiQUAR. We again use the $M/M/1$ base example. The step-length hyperparameters are set to $\eta_k = 4k^{-1}$ and $\delta_k = \min(0.1, 0.5k^{-1/3})$. We choose different values of T_k in the form of

$$T_k = T \cdot k^{1/3}, \ T \in \{40, 200, 360\}.$$

For different values of T, iteration numbers L_T are chosen to maintain equal total running times for LiQUAR. In particular, we choose $L_T = \left\lceil 1000 \cdot (200/T)^{3/4} \right\rceil$. Results of all above-mentioned cases are reported in Figure 6.

The right panel of Figure 6 shows that the slope of the linear fits all below 0.5. According to the three regret curves in the left panel, we can see how different values of T impact the exploration-exploitation trade-off: a larger value of T, e.g., $T = 360$, yields a higher regret in the early iterations but ensures a flatter curve in the later iterations. This is essentially due to the trade-off between the learning cost and the quality of the gradient estimator. A larger cycle length T_k guarantees a high-quality gradient estimators as more data are generated and used in each iteration which help reduce the gradient estimator's transient bias and variance. On the other hand, it demands that the system be operated for a longer time under suboptimal control policies, especially in the early iterations. The above analysis provides the following guidance for choosing T in practice: A smaller

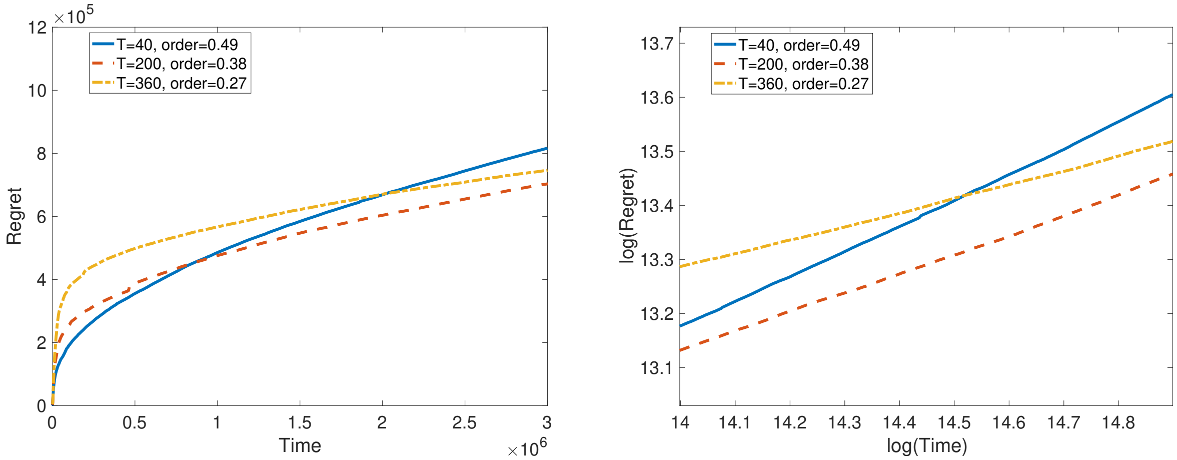


Figure 6 Regret under different $T \in \{40, 200, 360\}$: (i) average regret from 100 independent runs (left panel); (ii) regret curve in logarithm scale, with $T_k = T \cdot k^{1/3}$, $\eta_k = 4k^{-1}$, $\delta_k = \min(0.1, 0.5k^{-1/3})$ and $\alpha = 0.1$.

T is preferred if the service provider's goal is to make the most efficient use of the data in order to make timely adjustment on the control policy. This guarantees good performance in short term (the philosophy here is similar to that of the temporal-difference method with a small updating cycle). But if the decision maker is more patient and aims for good long-term performance, he/she should select a larger T which ensures that the decision update is indeed meaningful with sufficient data (this idea is similar to the Monte-Carlo method with batch updates).

6.3. Queues with non-Poisson arrivals

In this section, we consider the more general $GI/GI/1$ model having arrivals according to a renewal process. Similar to the service times, we model the interarrival times using scaled random variables $U_1/\lambda(p), U_2/\lambda(p), \ldots$ for a given p, with $U_1, U_2, \ldots$ being a sequence of I.I.D. random variables with $\mathbb{E}[U_n] = 1$.

The PTO framework is not applicable here because $\mathbb{E}[W_\infty]$ does not have a closed-form solution in the $GI/GI/1$ setting. This provides additional motivations for our online learning approach. On the other hand, generalizing the theoretical regret analysis rigorously from $M/GI/1$ to $GI/GI/1$ is by no means a straightforward extension. A key step in our analysis is to give a proper bound for the bias of the gradient estimator. When the arrival process is Poisson, the memoryless property ensures that N_l/T_k in (8) is an unbiased estimator for the arrival rate. For renewal arrivals, the arrival rate bias has an order $O(1/T_k) = O(k^{-1/3})$ (see for example Lorden's inequality (Asmussen 2003, Section V, Proposition 6.2)), which contributes to the bias of the FD with an order of $O(1/T_k\delta_k) = O(1)$. This contradicts Theorem 1 which requires $B_k = O(k^{-1})$. This part of the analysis requires additional investigations (in order to establish a more delicate bias bound). We leave the careful regret analysis of $GI/GI/1$ to future research.

Nevertheless, from the engineering perspective, the increased bias due to the GI arrival process may not be too significant (note that the theoretical bias bound is obtained from a worst-case analysis). We next conduct some preliminary numerical experiments to test the performance of LiQUAR under GI arrivals. We consider an $E_2/M/1$ queue example having Erlang-2 interarrival times with mean $1/\lambda(p)$ and exponential service times with rate μ to illustrate the performance of LiQUAR in $GI/GI/1$'s case. We continue to consider the logit demand function (20) with $M = 10, a = 4.1, b = 1$ and linear staffing cost function (21). Unlike the $M/GI/1$ case where the PK formula provides a closed-form formula for the steady-state waiting time, here we numerically compute the optimal solution (μ^*, p^*) by using matrix geometric method (note that the state process of $E_2/M/1$ is quasi-birth-and-death process). Letting $h_0 = c_0 = 1$ yields the optimal decision $(\mu^*, p^*) = (7.78, 3.75)$. We implement LiQUAR with hyperparameters $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$, and $\alpha = 0.1$. Figure 7, as an analog to Figure 4, shows that the refined LiQUAR continues to be

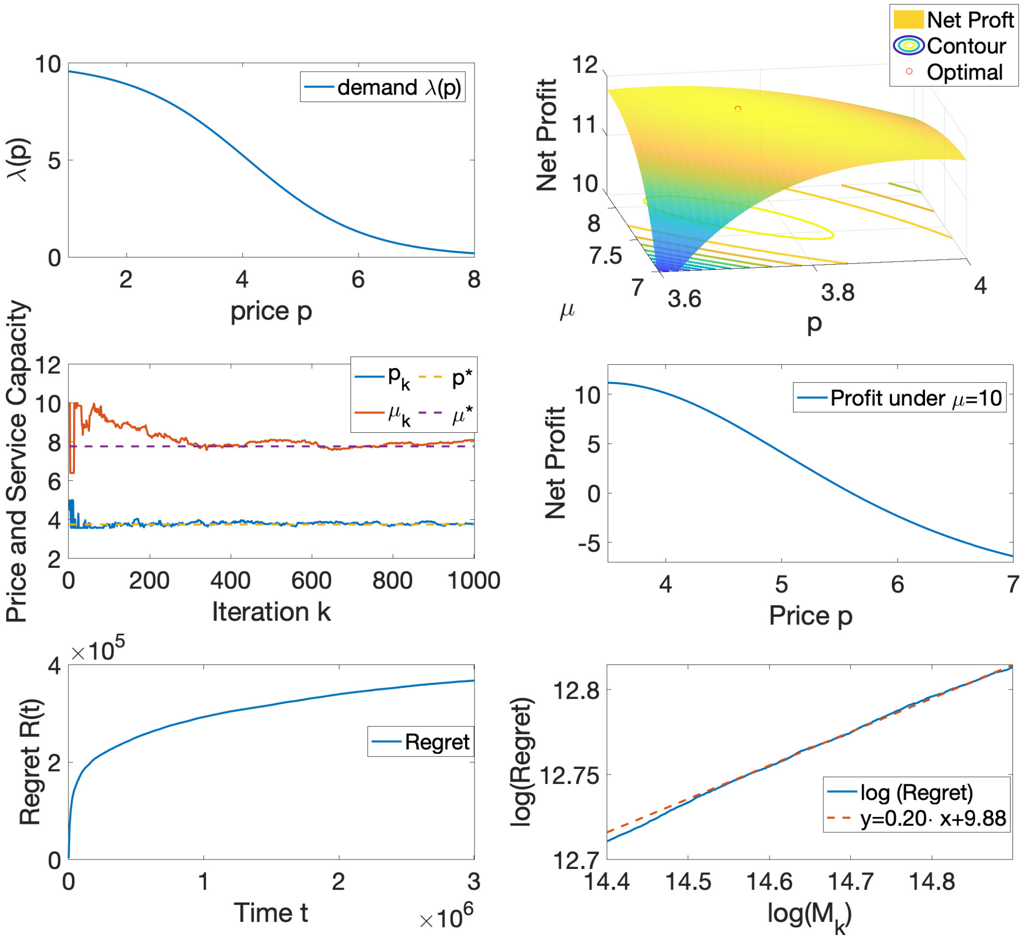


Figure 7 Joint pricing and staffing in the $E_2/M/1$ queue with $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$, $\alpha = 0.1$, $p_0 = 5$ and $\mu_0 = 10$.

effective, exhibiting a rapid converge to the optimal decision and a slowly growing regret curve (bottom left panel of Figure 7). Despite of the good performance of the above $E_2/M/1$ example, we acknowledge that this is only a preliminary step, and the full investigation of the $GI/GI/1$ case requires careful theoretical analysis and comprehensive numerical studies.

7. LiQUAR vs. PTO

In this section, we contrast the performance of LiQUAR to that of the conventional PTO method. In principle, a PTO algorithm undergoes two phases: (i) "prediction" of the model (e.g., estimation of the demand function and service distribution) and (ii) "optimization" of the decision variables (e.g., setting the optimal service price and capacity). Taking the demand function $\lambda(\cdot)$ as an example, PTO relies on the "prediction" phase to provide a good estimate $\widehat{\lambda}(p)$, which will next be fed to the "optimization" phase for generating desired control policies. In case no historical data is available so that the "prediction" completely relies on the newly generated data, one needs to learn the unknown demand curve $\lambda(p)$ by significantly experimenting the decision parameters in real time in order to generate sufficient demand data that can be used to obtain an accurate $\widehat{\lambda}(p)$.

We begin by establishing theoretical results to compare the performance of LiQUAR and PTO within a heavy-traffic framework. These findings are then supplemented by numerical experiments to give engineering confirmations.

7.1. LiQUAR vs. PTO in heavy traffic

In this section, we compare the regret bounds of LiQUAR and PTO when the system is in heavy traffic. Consider the sequence of h-indexed systems described in Section 5.3, we now intend to use PTO to find the optimal value of (17) within $\mathcal{B}_h$ and measure its performance by computing the regret $R^h(T_0)$ at time $T^h = T_0/h$. We consider the following PTO algorithm (Besbes and Zeevi 2009):

- **Input:** Total running time T^h, number of testing points κ^h, time of prediction t_0^h
- **Step 1. Prediction:**
  a. Organize $\mathcal{B}_h$ into κ^h evenly spaced grids and distribute the testing points in all grids.
  b. For each $i = 1, \cdots, \kappa^h$, operate the system at i^{th} under the testing point p_i for t_0^h/κ^h units of time, and approximate the demand curve $\hat{\lambda}(p_i)$ by the time-averaged arrival rate.
- **Step 2. Optimization:**
  a. Calculate $\hat{f}(p_i)$ using $\hat{\lambda}(p_i)$ in the PK formula.
  b. Operate the system under $\hat{p}^* = \arg\max_i \hat{f}(p_i)$ for the rest of time horizon.

We next give a regret bound for the above PTO method under the heavy-traffic learning scheme.

Proposition 6 **(PTO in heavy traffic)**. *Under Assumption 3, in the h^{th} system, PTO with hyperparameters κ^h, t_0^h yields the regret bound:*

$$R^h(T_0) \le C\sqrt{h}t_0^h + CT_0 \cdot \frac{\sqrt{\kappa^h \log T_0/h}}{h\sqrt{t_0^h}} + C\frac{\sqrt{h}T_0}{(\kappa^h)^2}. \tag{24}$$

In addition, if we select $t_0^h = O\left(\frac{T_0^{5/7}(\log T^h)^{2/7}}{h}\right)$ and $\kappa = O\left(\frac{T_0^{1/7}}{\log T^h}\right)$, then the PTO regret bound can be minimized as below:

$$R^h(T_0) \le C \cdot \frac{T_0^{5/7}\log(T_0/h)^{2/7}}{\sqrt{h}} = \tilde{O}\left(\frac{T_0^{5/7}}{1-\rho_h^*}\right),$$

with C being some constant independent with h and $1-\rho_h^*$.

Remark 5 (LiQUAR vs. PTO in heavy traffic). Compared to the original regret bounds presented in Besbes and Zeevi (2009), we have re-optimized the hyperparameters and derived reduced regret bounds under Assumption 1, enabling a fair comparison between LiQUAR and

PTO. According to Theorem 3 and Proposition 6, the regret bounds for both LiQUAR and PTO share a dependence on the traffic intensity in the order of $1/(1-\rho_h^*)$. However, LiQUAR exhibits a slower growth rate with respect to the time horizon, scaling as $\sqrt{T_0}$. This implies that over the long run, LiQUAR achieves a smaller regret bound than PTO. Furthermore, as the two terms involving T_0 and ρ^* interact multiplicatively in the regret bound, the factor $1/(1-\rho^*)$ amplifies LiQUAR's advantage over PTO in heavy-traffic conditions, i.e., as $\rho^* \to 1$. This trend is further validated by the numerical results illustrated in Figure 8.

Next, we numerically investigate the performance of LiQUAR and PTO for a one-dimensional pricing problem under heavy traffic. We consider an $M/M/1$ system having exponential demand function

$$\lambda(p) = \exp(a - bp),$$

with $a = 1 + \log 2,\ b = 1$ and exponential service time distributions. Following the settings in Theorem 3, we consider a sequence of objective functions in (22) indexed by h. We keep $\mu = 1$ held fixed and allow $h \in \{0.1, 0.01, 0.005, 0.001\}$ to account for different values of the traffic intensity. As $h \to 0$, the feasible region $\mathcal{B}_h$ takes the form

$$\mathcal{B}_h = \left[\ p_0 + c_1\sqrt{h}, p_0 + c_2\sqrt{h}\ \right],$$

with $c_1 = 0.6 \cdot c_0, c_2 = 5c_0$ and $c_0 = 1/\sqrt{\lambda'(p_0)(1 + p^*\lambda'(p_0))} = 1.20$.

Then, we apply LiQUAR and PTO in all the instances with different h. For LiQUAR, following Theorem 3, we choose $\eta_k^h = 4\sqrt{h}k^{-1}, \delta_k^h = 2\sqrt{h}k^{-1/3}$ with $T_k^h = h^{-1}k^{1/3}$ for 500 iterations. To make a fair comparison, we pick an equal runtime for LiQUAR and PTO with $T_0^h = \sum_{k=1}^{500} T_k^h$ and $T_0 = h \cdot T_0^h$ for all h. PTO's hyperparameters are chosen as $t_0^h = t \cdot \frac{T_0^{5/7}\log(T_0/h)^{2/7}}{\sqrt{h}}$ and $\kappa^h = h \cdot (t_0/\log T^h)^{1/5}$ with $t \in \{0.1, 0.2, 0.5\}$.

In Figure 8, we report the scaled regret curves of both methods under different holding costs h where each regret curve is estimated by the average of 100 independent replications. To understand how the regret is influenced by the heavy-traffic scaling factor h, we scale time by $h \cdot T^h = T_0$ and scale the regret by $\sqrt{h} \cdot R^h(T_0)$ (as in Theorem 3 and Proposition 6). From Figure 8, we confirm that LiQUAR significantly outperforms PTO in all heavy-traffic scenarios.

7.2. LiQUAR vs. objective-informed PTO

In this section, we compare LiQUAR to an advanced parametric PTO framework, where the prediction step incorporates information from the downstream objective function. We refer to this approach as *objective-informed PTO* (oiPTO). In oiPTO, it is assumed that the decision-maker knows the parametric form of the demand function, $\lambda(\cdot;\boldsymbol{\beta})$, with parameters $\boldsymbol{\beta}$ that are initially

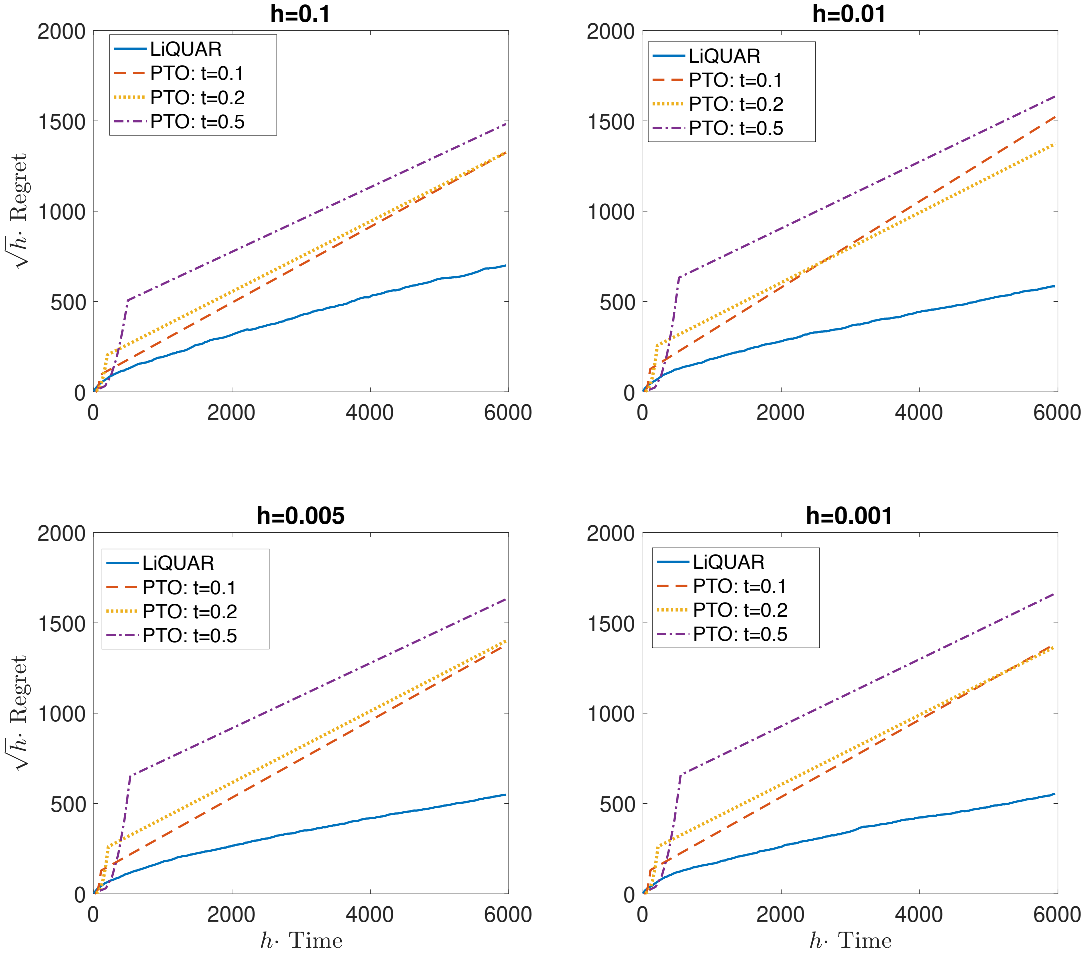


Figure 8 PTO vs. LiQUAR; $\eta_k^h = 4\sqrt{h}k^{-1}, \delta_k^h = 2\sqrt{h}k^{-1/3}$ with $T_k^h = h^{-1}k^{1/3}$; $t_0^h = t\cdot\frac{T_0^{5/7}\log(T_0/h)^{2/7}}{\sqrt{h}}$ and $\kappa = t_0^{1/5}\cdot h^{1/5}$ with $t\in\{0.1,0.2,0.5\}$

unknown. During the prediction phase, oiPTO estimates $\boldsymbol{\beta}$ by leveraging the structure of the downstream objective function (see (25)). In the optimization phase, oiPTO uses the demand function with the estimated parameters, denoted as $\boldsymbol{\beta}_{oi}$, to compute the optimal decisions.

Specifically, let $\theta\in(0,1)$ represent the exploration ratio and T denote the total time (or learning budget). The oiPTO approach divides the total time horizon T into two phases. In the first phase, corresponding to the interval $[0,\theta T]$, the focus is on learning the parameters of the demand function. In the second phase, covering the interval $[\theta T, T]$, the system operates using decisions optimized based on the estimated parameters. The details of these two steps are provided below:

- **Prediction:** Suppose m parameters of the demand function are to estimated, and we uniformly select $(p_1,\mu_1),\cdots,(p_m,\mu_m)\in\mathcal{B}$ as experimentation decisions. We sequentially operate the system under each of the experimentation decision for $\theta T/m$ units of time. Then, based on the arrival and workload data, the estimated parameter is given by

$$\boldsymbol{\beta}_{oi} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{m} (f(p_k, \mu_k; \boldsymbol{\beta}) - \hat{f}(p_k, \mu_k))^2, \tag{25}$$

where $f(p,\mu;\boldsymbol{\beta}) = -p\lambda(p;\boldsymbol{\beta}) + h \cdot \frac{\lambda(p;\boldsymbol{\beta})}{\mu - \lambda(p;\boldsymbol{\beta})} + c(\mu)$ and $\hat{f}(p,\mu)$ is the time average cost estimation of $f(p,\mu)$.

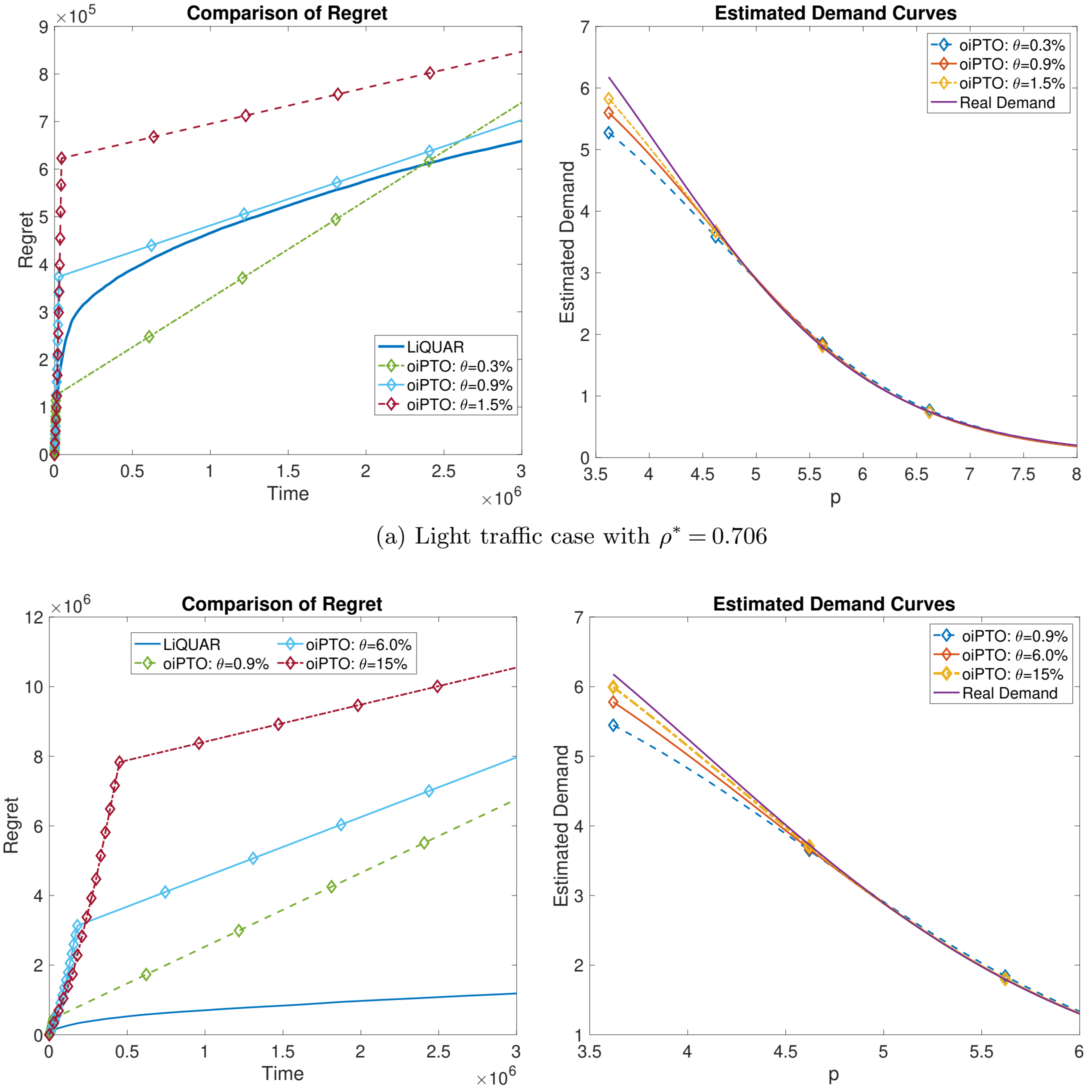
- **Optimization:** Next, we obtain the oiPTO-optimal policy $\hat{x}^*$ by maximizing our objective function with $\lambda(\cdot;\boldsymbol{\beta})$ replaced by $\lambda(\cdot;\boldsymbol{\beta}_{oi})$. Then we implement this policy for the rest of time horizon.

Experiment settings and results. We consider our base logit example in Section 6.1 having demand function (20) with $M_0 = 10, a = 4.1, b = 1$ and exponential service times. Throughout this experiment, we fix the staffing cost $c(\mu) = \mu$. To understand the impact of the system's congestion level on performance of oiPTO and LiQUAR, we consider two scenarios specified by the optimal traffic intensity ρ^*: (i) A light-traffic case with $\rho^* = 0.709$ ($h_0 = 1$) and (ii) A heavy-traffic case with $\rho^* = 0.987$ ($h = 0.001$).

For LiQUAR, we consistently select the hyperparameters $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, initial values $(\mu_0, p_0) = (10, 7)$ and $T_k = 200k^{1/3}$ for $L = 1000$ iterations with a total running time $T = 2\sum_{k=1}^{L} 200k^{1/3}$. For oiPTO, we use the same total time T and consider several values of the exploration ratio $\theta \in \{0.3\%, 0.9\%, 1.5\%, 6\%, 15\%\}$ to account for different levels of exploration efforts.

In Figure 9, we present the regret results for LiQUAR and oiPTO, showcasing the three oiPTO curves with the lowest regrets. The left-hand panels illustrate that the exploration ratio θ has a significant impact on oiPTO's performance. The regret for oiPTO exhibits a piecewise linear pattern: during the prediction phase, regret grows rapidly due to periodic exploration across all experimentation variables; in the optimization phase, regret continues to increase linearly, albeit at a slower rate, as the system operates based on the oiPTO-optimized solution, which remains suboptimal. A larger (smaller) θ leads to higher (lower) regret during the prediction phase but results in a more (less) accurate model. This improved accuracy generates decisions that are closer to optimal, resulting in a slower (faster) regret growth during the optimization phase.

Furthermore, comparing case (a) to case (b), we observe that while oiPTO incorporates downstream objective information, LiQUAR consistently outperforms oiPTO by achieving a lower regret. This advantage is especially pronounced in heavy-traffic scenarios. The primary reason is that LiQUAR employs an integrated learning approach, continuously refining its decision-making through direct interaction with the environment. In contrast, oiPTO follows a static learning strategy, investing fixed efforts into parameter prediction and relying entirely on these predictions in the optimization phase. The limitations of oiPTO become more apparent in heavy traffic, where the nonlinear structure of workload amplifies the cost of suboptimal decisions.



(a) Light traffic case with $\rho^* = 0.706$

(b) Heavy traffic case with $\rho^* = 0.987$

Figure 9 oiPTO vs. LiQUAR: (i) low traffic scenario $\rho^* = 0.705$; (ii) high traffic scenario $\rho^* = 0.987$. Hyperparameters for LiQUAR are $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3}), T_k = 200k^{1/3}$ in both scenarios. All regret curves are estimated by averaging 1,000 independent simulation runs.

8. LiQUAR vs. Reinforcement Learning

In this section, we compare LiQUAR with reinforcement learning (RL) methods. While the machine learning literature offers a wide range of RL approaches, we focus on the policy gradient type methods (PG methods) for comparison due to the following reasons: (i) both LiQUAR and PG methods rely on gradient-based optimization, making them conceptually aligned; and (ii) our problem involves an infinite state space (e.g., queue length or workload) and a continuous action space, where the PG methods demonstrate particular advantages over other RL techniques.

Problem Settings and Algorithms. Because an RL method is underpinned by its corresponding Markov decision process (MDP), and setting up an MDP requires the model to be Markovian, we now restrict our attention to the $M/M/1$ queue. Specifically, we consider the following discrete-time MDP with the objective of maximizing its long-run average reward. Our MDP has

- *Time steps:* $t = 1, 2, \dots$.
- *States:* Queue length at beginning of period t, denoted by S_t.
- *Actions:* Choices of price and service rate at each time step $A_t = (p, \mu)$.
- *Rewards:* The net profit gained in time slot t, denoted by R_t.

Under the above setting, we write the Bellman equation as below:

$$q_\pi(s,a) + h_\pi = \mathbb{E}_{s' \sim P(s,a), a' \sim \pi}[R(s,a) + q(s',a')],$$

with $q_\pi(s,a)$ is the q-function of policy π and h_π is the long-run average revenue under π.

Following (Sutton and Barto 2018, Section 13.6), we apply the Gaussian parameterization for our actions. Specifically, we draw $p \sim N(\bar{p}, \sigma_p^2)$ and $\mu \sim N(\bar{\mu}, \sigma_\mu^2)$ independently. Let $\theta \equiv (\bar{p}, \bar{\mu}, \sigma_p^2, \sigma_\mu^2)$ and we denote π_θ as the Gaussian density with parameter θ. According to the policy gradient theorem (Sutton and Barto 2018, p.339), the gradient on the policy function can be represented as $\nabla_\theta h_{\pi_\theta} = \mathbb{E}[\nabla \log \pi_\theta(A_t|S_t) \cdot q(S_t, A_t)]$. We consider two types of PG algorithms, i.e., a base policy gradient algorithm (PG-base) and a policy gradient algorithm with score-aware gradient (PG-SAGE) introduced in Comte et al. (2023). See EC.8 for more details about the two algorithms.

Experiment settings and results. We now compare PG algorithms with LiQUAR using our base example as described in Section 6.1. Specifically, we consider an $M/M/1$ queue having logit demand function (20) with $M_0 = 10$, $a = 4.1$, $b = 1$ and exponential service times with holding cost $c(\mu) = \mu$. For LiQUAR, the hyperparameter are $\eta_k = 4k^{-1}$, $T_k = 200k^{1/3}$, and $\delta_k = \min(0.1, 0.5k^{-1/3})$. For PG type algorithms, we consider picking both the constant step sizes $\eta \in \{1, 0.1, 0.01, 0.001, 0.0001\}$ and the decreasing step sizes $\eta_k \in \{4k^{-1}, 2k^{-1}, k^{-1}, 0.1k^{-1}, 0.01k^{-1}\}$. The cycle lengths are $T \in \{10, 100, 300\}$. In addition, we keep the total running time of LiQUAR and PG equal in order for a fair comparison.

We report the regret curves in Figure 10. For the clarity of the figure, we report the curves with the lowest regret for each PG algorithm. From Figure 10, we find that LiQUAR is more effective than PG in a wide range of hyper-parameter choices.

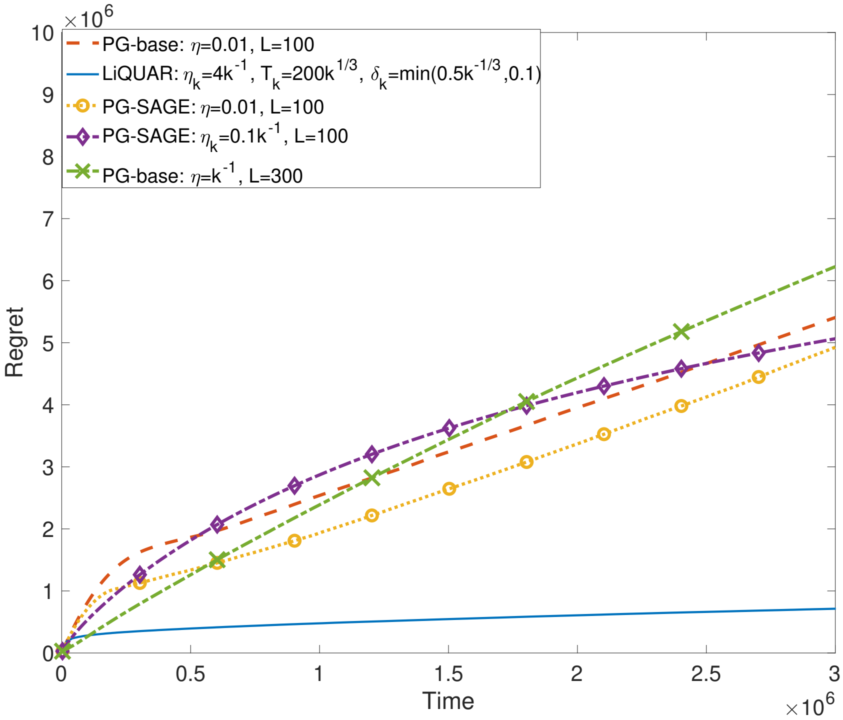


Figure 10 Comparison of the regret of LiQUAR with PG type algorithms in our base example. The hyperparameter choices are: (i) LiQUAR : $\eta_k = 4k^{-1}, T_k = 200k^{1/3}, \delta_k = \min(0.5k^{1/3}, 0.1)$; (ii) PG: $\eta_k \in \{1, 0.1, 0.01, 0.001, 4k^{-1}, 2k^{-1}, k^{-1}, 0.1k^{-1}, 0.01k^{-1}\}, L \in \{1, 10, 100, 300\}$, Episode length $T = 3000/L$. For clarity of the figure, we plot the lowest regret curve for PG algorithms. All regret curves are estimated from 100 independent simulation replications.

Remark 6 (LiQUAR vs. PG). In the PG algorithms, the gradient estimators rely on accurately learning the $q_\pi(s,a)$ function (as outlined in the policy gradient theorem), which is a two-dimensional function. Inaccuracies in this estimation can result in significant variance in the gradient calculations. In contrast, LiQUAR only requires learning the values of individual actions, substantially reducing the complexity and effort required for learning. Furthermore, by the design of LiQUAR, the tuning of its hyperparameters can leverage domain-specific knowledge of the queueing system, e.g., the transient bias and auto-correlation between queueing data; also see Remark 2. In comparison, tuning hyperparameters in the PG algorithms is considerably more challenging, as RL methods are generally considered black-box approaches.

9. Conclusions

In this paper we develop an online learning framework, dubbed LiQUAR, designed for dynamic pricing and staffing in an $M/GI/1$ queue with unknown arrival rate function and service distribution. LiQUAR's main appeal is its "model-free" attribute. Unlike the conventional "predict-then-optimize" approach where precise estimations of the demand function and service distribution must be conducted (as a separate step) before the decisions may be optimized, LiQUAR is an integrated method that recursively evolves the control policy to optimality by effectively using the newly generated queueing data (e.g., arrival times and service times). LiQUAR's main advantage is its solution robustness; its algorithm design is able to automatically relate the parameter estimation errors to the fidelity of the optimized solutions. Comparing to the conventional method, this advantage becomes more significant when the system is in heavy traffic.

Effectiveness of LiQUAR is substantiated by (i) theoretical results including the algorithm convergence and regret analysis, and (ii) engineering confirmation via simulation experiments of a variety of representative queueing models. Theoretical analysis of the regret bound in the present paper may shed lights on the design of efficient online learning algorithms (e.g., bounding gradient estimation error and controlling proper learning rate) for more general queueing systems. In addition, the analysis on the statistical properties for our gradient estimator has independent interests and may contribute to the general literature of stochastic gradient decent. We also extend LiQUAR to the more general $GI/GI/1$ model and confirm its good performance by conducting numerical studies.

There are several venues for future research. One dimension is to extend the method to queueing models under more general settings such as non-Poisson arrivals, customer abandonment and multiple servers, which will make the framework more practical for service systems such as call centers and healthcare. Another interesting direction is to theoretically relax the assumption of uniform stability by developing a "smarter" algorithm that automatically explores and then settles on control policies that guarantee stable system performance. Another practical direction is to investigate how to learn the optimal dynamic policy, where for a finite horizon T the optimal service rate μ and price p are inherently state- and time-dependent as given by a continuous-time optimal control problem.

References

Abate J, Choudhury GL, Whitt W (1993) Calculation of the GI/G/1 waiting-time distribution and its cumulants from Pollaczek's formulas. *Archiv für Elektronik und Ubertragungstechnik* 47(5/6):311–321.

Abate J, Whitt W (1988a) The correlation functions of RBM and M/M/1. *Communications in Statistics. Stochastic Models* 4(2):315–359.

Abate J, Whitt W (1988b) Transient behavior of the M/M/1 queue via Laplace transforms. *Advances in Applied Probability* 20(1):145–178.

Asmussen S (2003) *Applied Probability and Queues*, volume 2 (Springer).

Baron O, Krass D, Senderovich A, Sherzer E (2023) Supervised ML for solving the GI/GI/1 queue. *INFORMS Journal on Computing* 2015(586-594).

Bergquist J, Elmachtoub AN (2023) Static pricing guarantees for queueing systems. *arXiv preprint arXiv:2305.09168* .

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations research* 57(6):1407–1420.

Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* 61(4):723–739.

Blanchet J, Chen X (2020) Rates of convergence to stationarity for reflected Brownian motion. *Mathematics of Operations Research* 45(2):660–681.

Broadie M, Cicek D, Zeevi A (2011) General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Operations Research* 59(5):1211–1224.

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Operations Research* 60(4):965–980.

Chen X, Liu Y, Hong G (2024) An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research* 72(6):2677–2697, URL `http://dx.doi.org/10.1287/opre.2020.0612`.

Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Operations Research* 65(6):1722–1731.

Chong EKP, Ramadge PJ (1993) Optimization of queues using an infnitesimal perturbation analysis-based stochastic algorithm with general update times. *SIAM Journal on Control and Optimization* 31:698–732.

Comte C, Jonckheere M, Sanders J, Senen-Cerda A (2023) Score-aware policy-gradient methods and performance guarantees using local lyapunov conditions: Applications to product-form stochastic networks and queueing systems. *arXiv preprint arXiv:2312.02804* .

Cosmetatos GP (1976) Some approximate equilibrium results for the multi-server queue (M/G/r). *Journal of the Operational Research Society* 27(3):615–620.

Dai JG, Gluzman M (2021) Queueing network controls via deep reinforcement learning. *Stochastic Systems* 12(1):30–67.

Elmachtoub AN, Shi J (2023) The power of static pricing for reusable resources. *arXiv preprint arXiv:2302.11723* .

Fu MC (1990) Convergence of a stochastic approximation algorithm for the GI/G/1 queue using infinitesimal perturbation analysis. *Journal of Optimization Theory and Applications* 65:149–160.

Garyfallos S, Liu Y, Barlet-Ros P, Cabellos-Aparicio A (2024a) Neuralinq: A neural network method for the transient performance analysis in non-Markovian queues. *Working Paper.*

Garyfallos S, Liu Y, Barlet-Ros P, Cabellos-Aparicio A (2024b) Service level prediction in non-Markovian nonstationary queues: A simulation-based deep learning approach. *Winter Simulation Conference (WSC)* 2015(586-594).

Glasserman P (1992) Stationary waiting time derivatives. *Queueing Systems* 12:369–390.

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* 34(2):397–416.

Jia H, Shi C, Shen S (2022) Online learning and pricing with reusable resources: Linear bandits with sub-exponential rewards. *International Conference on Machine Learning*, 10135–10160.

Jia H, Shi C, Shen S (2024) Online learning and pricing for service systems with reusable resources. *Operations Research* 72(3):1203–1241.

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations research* 62(5):1142–1167.

Kim J, Randhawa RS (2018) The value of dynamic pricing in large queueing systems. *Operations Research* 66(2):409–425.

Krishnasamy S, Sen R, Johari R, Shakkottai S (2021) Learning unknown service rates in queues: A multi-armed bandit approach. *Operations Research* 69(1):315–330.

Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing Service Oper. Management* 12(3):511–526.

L'Ecuyer P, Giroux N, Glynn PW (1994) Stochastic optimization by simulation: Numerical experiments with the M/M/1 queue in steady-state. *Management Science* 40(10):1245–1261.

L'Ecuyer P, Glynn PW (1994) Stochastic optimization by simulation: Convergence proofs for the GI/GI/1 queue in steady state. *Management Science* 40(11):1562–1578.

Lee C, Ward AR (2014) Optimal pricing and capacity sizing for the GI/GI/1 queue. *Operations Research Letters* 42:527–531.

Lee C, Ward AR (2019) Pricing and capacity sizing of a service facility: Customer abandonment effects. *Production and Operations Management* 28(8):2031–2043.

Liu B, Xie Q, Modiano E (2019) Reinforcement learning for optimal control of queueing systems. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 663–670.

Maglaras C, Zeevi A (2003) Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* 49(8):1018–1038.

Murthy Y, Grosof I, Maguluri ST, Srikant R (2024) Performance of npg in countable state-space average-cost rl. *arXiv preprint arXiv:2405.20467* .

Nair J, Wierman A, Zwart B (2016) Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base. *Management Science* 62(6):1830–1841.

Nakayama MK, Shahabuddin P, Sigman K (2004) On finite exponential moments for branching processes and busy periods for queues. *Journal of Applied Probability* 41(A):273–280.

Olivares J, Martin P, Valero E (2018) A simple approximation for the modified bessel function of zero order i0 (x). *Journal of Physics: Conference Series*, volume 1043, 012003 (IOP Publishing).

Pollaczek F (1930) Über eine aufgabe der wahrscheinlichkeitstheorie. I. *Mathematische Zeitschrift* 32(1):64–100.

Raeis M, Tizghadam A, Leon-Garcia A (2021) Queue-learning: A reinforcement learning approach for providing quality of service. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 461–468.

Shah D, Xie Q, Xu Z (2020) Stable reinforcement learning with unbounded state space. Bayen AM, Jadbabaie A, Pappas G, Parrilo PA, Recht B, Tomlin C, Zeilinger M, eds., *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, 581–581.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (The MIT Press), 2nd edition.

Walton N, Xu K (2021) Learning and information in stochastic networks and queues. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, 161–198.

Yuan H, Luo Q, Shi C (2021) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* 67(10):6089–6115, URL `http://dx.doi.org/10.1287/mnsc.2020.3799`.

Zhang H, Chao X, Shi C (2020) Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science* 66(5):1962–1980.

Zhong Y, Birge JR, Ward AR (2024) Learning to schedule in multiclass many-server queues with abandonment. *Operations Research* .

E-Companion

This e-companion provides supplementary materials to the main paper. In Section EC.1, we provide the main proofs of the theorems in the main paper. In Section EC.2, we give the technical proofs in Section EC.1. In Section EC.4, we verify that the Condition (a) of Assumption 1 holds for some commonly used demand functions. In Section EC.5, we conduct additional numerical studies. To facilitate readability, all notations are summarized in Table EC.1 including all model parameters and functions, algorithmic parameters and variables, and constants in the regret analysis.

EC.1. Main Proofs

In this section, we provide the proofs of the main theorems and propositions. Proofs of technical lemmas are given in the Section EC.2.

EC.1.1. Proof of Proposition 1

First, we introduce a technical lemma to uniformly bound the moments of workload under arbitrary control policies.

Lemma EC.1 **(Uniform Moment Bounds)**. *Under Assumptions 1 and 2, there exist some constants $\theta_0>0$ and $M>1$ such that, for any sequence of control parameters $\{(\mu_l,p_l):l\geq 1\}$,*

$$\mathbb{E}[W_l(t)^m]\leq M,\quad \mathbb{E}[W_l(t)^m\exp(2\theta_0 W_l(t))]\leq M,$$

for all $m\in\{0,1,2\}$, $l\geq 1$ and $0\leq t\leq T_k$ with $k=\lceil l/2\rceil$.

Then, following (7),

$$\begin{aligned}\mathbb{E}\left[|\hat{W}_l(t)-W_l(t)|\right]&=\mathbb{E}\left[W_l(t)\cdot\mathbf{1}\left(W_l(t)>\mu_l(T_k-t)\right)\right]\leq\mathbb{E}\left[W_l(t)^2\right]^{1/2}\mathbb{P}\left(W_l(t)>\mu_l(T_k-t)\right)^{1/2}\\&\leq\mathbb{E}\left[W_l(t)^2\right]^{1/2}\cdot\exp\left(-\frac{1}{2}\theta_0\mu_l(T_k-t)\right)\mathbb{E}\left[\exp(\theta_0 W_l(t))\right]^{1/2}\leq\exp\left(-\frac{1}{2}\theta_0\underline{\mu}(T_k-t)\right)M,\end{aligned}$$

where the last inequality follows from Lemma EC.1. □

EC.1.2. Proof of Proposition 2

For each cycle l, the difference between the estimated system performance $\hat{f}^G(\mu_l,p_l)$ and its true value is

$$\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)=\frac{-p_l(N_l-\lambda(p_l)T_k)}{T_k}+\frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}[\underbrace{\hat{W}_l(t)-W_l(t)}_{\text{delayed observation}}+\underbrace{W_l(t)-w_l}_{\text{transient error}}]\,dt,$$

where $w_l=\mathbb{E}[W_\infty(\mu_l,p_l)]$ is the steady-state mean workload. To bound the moments of this difference, which correspond to the bias and MSE of $\hat{f}^G(\mu_l,p_l)$, we construct a stationary workload process $\bar{W}_l(t)$ for $0 \le t \le T_k$. At $t=0$, the initial value $\bar{W}^l(0)$ is independently drawn from the stationary distribution $W_\infty(\mu_l, p_l)$ and $\bar{W}_l(t)$ is *synchronously coupled* with $W_l(t)$ in the sense that they share the same sequence of arrivals and individual workload on $[0, T_k]$.

Bound on the Bias. The *bias* of $\hat{f}^G(\mu_l, p_l)$ can be decomposed as

$$\begin{aligned}
&\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l) - f(\mu_l,p_l)\right] \\
&= \frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\left(\mathbb{E}_l\left[\hat{W}_l(t)\right] - \mathbb{E}_l[\bar{W}_l(t)]\right)dt \le \frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l\left[|\hat{W}_l(t) - \bar{W}_l(t)|\right]dt. \\
&\le \frac{1}{(1-2\alpha)T_k}\left(\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l[|\hat{W}_l(t) - W_l(t)|]dt + \int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l[|W_l(t) - \bar{W}_l(t)|]dt\right). \qquad \text{(EC.1)}
\end{aligned}$$

The first term in (EC.1) is the error caused by delayed observation. Following the same analysis as in Section EC.1.1,

$$\mathbb{E}_l\left[|\hat{W}_l(t) - W_l(t)|\right] \le \mathbb{E}_l[W_l(t)^2]^{1/2}\cdot\exp(-a\mu_l(T_k - t))\mathbb{E}_l[\exp(2aW_l(t))]^{1/2},$$

for $a = \theta_0/2$. It is easy to check that $W_l(t) \le W_l(0) + \bar{W}_l(t)$. Conditional on $\mathcal{G}_l$, for all $0 \le t \le T_k$, $\bar{W}_l(t)$ is the stationary workload with parameter (μ_l, p_l). Following the proof of Lemma EC.1, $\bar{W}_l(t)$ is stochastic bounded by the stationary workload with parameter $(\underline{\mu}, \underline{p})$. Therefore,

$$\begin{aligned}
&\mathbb{E}_l\left[|\hat{W}_l(t) - W_l(t)|\right] \le \mathbb{E}_l[W_l(t)^2]^{1/2}\cdot\exp(-\theta_0\mu_l(T_k-t)/2)\mathbb{E}_l[\exp(\theta_0 W_l(t))]^{1/2} \\
&\le \exp(-\theta_0\mu_l(T_k-t)/2)(W_l(0)^2 + 2W_l(0)\mathbb{E}_l[\bar{W}_l(t)] + \mathbb{E}_l[\bar{W}_l(t)^2])^{1/2}\exp(\theta_0 W_l(0))\mathbb{E}_l[\exp(\theta_0\bar{W}_l(t))]^{1/2} \\
&\le \exp(-\theta_0\mu_l(T_k-t)/2)(W_l(0)^2 + 2MW_l(0) + M^2)^{1/2}\exp(\theta_0 W_l(0))M^{1/2} \\
&\le \exp(-\theta_0\mu_l(T_k-t)/2)M(M + W_l(0))\exp(\theta_0 W_l(0)). \qquad \text{(EC.2)}
\end{aligned}$$

The last inequality holds as $M \ge 1$. The second term in (EC.1) will be bounded using the following lemma on convergence rate of two synchronously coupled workload processes.

LEMMA EC.2 **(Ergodicity Convergence)**. *Suppose Assumptions 1 and 2 hold. Two workload processes $W(t)$ and $\bar{W}(t)$ with equal control parameters $(\mu, p) \in \mathcal{B}$ are synchronously coupled with initial states $(W(0), \bar{W}(0))$. Then, there exists $\gamma > 0$ independent of (μ, p), such that*

$$\mathbb{E}\left[|W(t) - \bar{W}(t)|^m \mid W(0), \bar{W}(0)\right] \le e^{-\gamma t}(e^{\theta_0 W(0)} + e^{\theta_0 \bar{W}(0)})|W(0) - \bar{W}(0)|^m.$$

Using this lemma, we can compute

$$\begin{aligned}
\mathbb{E}_l[|W_l(t) - \bar{W}_l(t)|] &\le \exp(-\gamma t)\mathbb{E}_l\left[|W_l(0) - \bar{W}_l(0)|(\exp(\theta_0 W_l(0)) + \exp(\theta_0\bar{W}_l(0)))\right] \\
&\le \exp(-\gamma t)\left(W_l(0)\exp(\theta_0 W_l(0)) + MW_l(0) + M\exp(\theta_0 W_l(0)) + M\right) \\
&\le \exp(-\gamma t)(M + W_l(0))(\exp(\theta_0 W_l(0)) + M). \qquad \text{(EC.3)}
\end{aligned}$$

Let $\theta_1 = \min(\gamma, \theta_0\underline{\mu}/2)$. Plugging inequalities (EC.2) and (EC.3) into (EC.1), we obtain the following bound for the bias

$$\left|\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l) - f(\mu_l,p_l)\right]\right| \le \frac{1}{(1-2\alpha)T_k}\cdot\frac{2\exp(-\theta_1\alpha T_k)}{\theta_1}\cdot M(M+W_l(0))(\exp(\theta_0 W_l(0))+M).$$

Bound on the Mean Square Error. The mean square error (MSE) of $\hat{f}^G(\mu_l,p_l)$

$$\mathbb{E}_l[(\hat{f}^G(\mu_l,p_l) - f(\mu_l,p_l))^2] \le 2\mathbb{E}_l[E_1^2] + 2\mathbb{E}_l[E_2^2],$$

with

$$\hat{f}^G(\mu_l,p_l) - f(\mu_l,p_l) = \underbrace{\frac{-p_l(N_l - \lambda(p_l)T_k)}{T_k}}_{E_1} + \underbrace{\frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}(\hat{W}_l(t) - w_l)dt}_{E_2}.$$

Conditional on $\mathcal{G}_l$, the observed number of arrivals N_l is a Poisson r.v. with mean $\lambda(p_l)T_k$. So, $\mathbb{E}_l[E_1^2] = p_l^2\lambda(p_l)T_k^{-1} \le \bar{p}^2\bar{\lambda}T_k^{-1}$.

For E_2, we have

$$\mathbb{E}_l[E_2^2] = \frac{1}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l\left[(\hat{W}_l(t)-w_l)(\hat{W}_l(s)-w_l)\right]dtds.$$

According to (7), $\hat{W}_l(\cdot) \le W_l(\cdot)$ and therefore, for any $0 \le s \le t \le T_k$,

$$\begin{aligned}
&\mathbb{E}_l[(\hat{W}_l(t)-w_l)(\hat{W}_l(s)-w_l)] = \mathbb{E}_l[\hat{W}_l(t)\hat{W}_l(s) - w_l(\hat{W}_l(s)+\hat{W}_l(t)) + w_l^2]\\
&\le \mathbb{E}_l[W_l(t)W_l(s) - w_l(\hat{W}_l(s)+\hat{W}_l(t)) + w_l^2]\\
&\le \mathbb{E}_l[(W_l(t)-w_l)(W_l(s)-w_l)] + \left(\mathbb{E}_l[w_l|W_l(s)-\hat{W}_l(s)|] + \mathbb{E}_l[w_l|W_l(t)-\hat{W}_l(t)|]\right)\\
&\le \underbrace{\mathbb{E}_l[(W_l(t)-w_l)(W_l(s)-w_l)]}_{\text{auto-covariance}} + M\underbrace{\left(\mathbb{E}_l[|W_l(s)-\hat{W}_l(s)|] + \mathbb{E}_l[|W_l(t)-\hat{W}_l(t)|]\right)}_{\text{error caused by delayed observations}}
\end{aligned}$$

To bound the auto-covariance term, we introduce the following lemma.

Lemma EC.3 **(Auto-covariance of $W_l(t)$).** *There exists a constant $K_V > 0$ independent of T_k, l, p_l, μ_l such that, for any $l \ge 1$ and $0 \le s \le t \le T_k$,*

$$\mathbb{E}_l[(W_l(t)-w_l)(W_l(s)-w_l)] \le K_V\left(\exp(-\gamma(t-s)) + \exp(-\gamma s)\right)(W_l(0)^2+1)\exp(\theta_0 W_l(0)). \quad \text{(EC.4)}$$

Following (EC.4), we write

$$\begin{aligned}
&\frac{1}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l[(W_l(t)-w_l)(W_l(s)-w_l)]dtds\\
&\le \frac{2K_V(W_l(0)^2+1)\exp(\theta_0 W_l(0))}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\int_{\alpha T_k}^{t}(\exp(-\gamma(t-s)) + \exp(-\gamma s))dsdt\\
&= \frac{2K_V(W_l(0)^2+1)\exp(\theta_0 W_l(0))}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\gamma^{-1}(1-\exp(-\gamma(t-\alpha T_k)) + \exp(-\gamma\alpha T_k) - \exp(-\gamma t))dt\\
&\le \frac{2K_V(W_l(0)^2+1)\exp(\theta_0 W_l(0))}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}2\gamma^{-1}dt \le \frac{4K_V(W_l(0)^2+1)\exp(\theta_0 W_l(0))}{\gamma(1-2\alpha)T_k}.
\end{aligned}$$

For the error of delayed observation, by Proposition 1, we have

$$
\begin{aligned}
&\frac{1}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\left(\mathbb{E}_l[|W_l(s)-\hat{W}_l(s)|]+\mathbb{E}_l[|W_l(t)-\hat{W}_l(t)|]\right)dsdt\\
\le\ &\frac{M(M+W_l(0))\exp(\theta_0W_l(0))}{(1-2\alpha)^2T_k^2}\int_{\alpha T_k}^{(1-\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\left(\exp(-\frac{\theta_0\mu_l}{2}(T_k-t))+\exp(-\frac{\theta_0\mu_l}{2}(T_k-s))\right)dsdt\\
=\ &\frac{4M(M+W_l(0))\exp(\theta_0W_l(0))}{\theta_0\mu_l(1-2\alpha)T_k}\left(\exp(-\frac{\theta_0\mu_l}{2}\alpha T_k)-\exp(-\frac{\theta_0\mu_l}{2}(1-\alpha)T_k)\right)\\
\le\ &\frac{4M(M+W_l(0))\exp(\theta_0W_l(0))}{\theta_0\mu_l(1-2\alpha)T_k}.
\end{aligned}
$$

As $W_l(0)\le (W_l(0)^2+1)/2$ and $M\ge 1$, we have $M+W_l(0)\le (M+1)(1+W_l(0)^2)$. Then, if we choose

$$K_M=\frac{8(K_V+M^3+M^2)}{(1-2\alpha)\min(\gamma,\theta_0\underline{\mu})}+2\bar{p}^2\bar{\lambda} \tag{EC.5}$$

then, we have

$$\mathbb{E}_l[(\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l))^2]\le 2\mathbb{E}_l[E_1^2]+2\mathbb{E}_l[E_2^2]\le K_MT_k^{-1}(W_l(0)^2+1)\exp(\theta_0W_l(0)).$$

□

EC.1.3. Proof of Proposition 3

According to the following lemma, the FD approximation error is of order $O(\delta_k^2)$.

Lemma EC.4. *Under Assumption 1, there exists a smoothness constant $c>0$ such that for any $\mu_1,\mu_2,\mu\in[\underline{\mu},\bar{\mu}]$ and $p_1,p_2,p\in[\underline{p},\bar{p}]$,*

$$
\begin{aligned}
\left|\frac{f(\mu_1,p)-f(\mu_2,p)}{\mu_1-\mu_2}-\partial_\mu f\left(\frac{\mu_1+\mu_2}{2},p\right)\right|&\le c(\mu_1-\mu_2)^2\\
\left|\frac{f(\mu,p_1)-f(\mu,p_2)}{p_1-p_2}-\partial_p f\left(\mu,\frac{p_1+p_2}{2}\right)\right|&\le c(p_1-p_2)^2.
\end{aligned}
$$

So, to bound B_k, it remains to show that

$$\mathbb{E}[\mathbb{E}[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)|\mathcal{F}_k]^2]^{1/2}=O(\exp(-\theta_1\alpha T_k)).$$

Recall that $\mathcal{F}_k$ is the σ-algebra including all events in the first $2(k-2)$ cycles, so $\mathcal{F}_k\subseteq\mathcal{G}_l$ for $l=2k-1,2k$. By Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)|\mathcal{F}_k\right]^2&=\mathbb{E}\left[\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]\Big|\mathcal{F}_k\right]^2\\
&\le\mathbb{E}\left[\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]^2\Big|\mathcal{F}_k\right].
\end{aligned}
$$

Therefore,

$$\mathbb{E}\left[\mathbb{E}\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)|\mathcal{F}_k\right]^2\right]\le\mathbb{E}\left[\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]^2\right].$$

By Proposition 2, the bias of estimated system performance

$$\left|\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]\right|\leq\frac{2\exp(-\theta_1\alpha T_k)}{(1-2\alpha)\theta_1 T_k}\cdot M(M+W_l(0))(\exp(\theta_0 W_l(0))+M).$$

As $(x+y)^2\leq 2x^2+2y^2$, we have, by Lemma EC.1,

$$\begin{aligned}&\mathbb{E}[\mathbb{E}_l[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)]^2]\\ \leq&\frac{4\exp(-2\theta_1\alpha T_k)}{(1-2\alpha)^2\theta_1^2T_k^2}\left(4M^4\mathbb{E}[\exp(2\theta_0W_l(0))+W_l(0)^2]+4M^2\mathbb{E}[W_l(0)^2\exp(2\theta_0W_l(0))]+4M^6\right)\\ \leq&\frac{4\exp(-2\theta_1\alpha T_k)}{(1-2\alpha)^2\theta_1^2T_k^2}\cdot(8M^5+4M^3+4M^6)=O(\exp(-2\theta_1\alpha T_k)).\end{aligned}$$

Therefore, $B_k=O\left(\delta_k^2+\delta_k^{-1}\exp(-\theta_1\alpha T_k)\right)$. The variance

$$\mathbb{E}[\|H_k\|^2]\leq 3\delta_k^{-2}\sum_{l=2k-1}^{2k}\mathbb{E}[(\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l))^2]+3\delta_k^{-2}\mathbb{E}[(f(\mu_{2k},p_{2k})-f(\mu_{2k-1},p_{2k-1}))^2].$$

By the smoothness condition of the objective function $f(x)$ as given in Assumption 1,

$$3\delta_k^{-2}\mathbb{E}[(f(\mu_{2k},p_{2k})-f(\mu_{2k-1},p_{2k-1}))^2]\leq\max_{(\mu,p)\in\mathcal{B}}\|\nabla f(\mu,p)\|^2=O(1).$$

Following Proposition 2, for $l=2k-1,2k$,

$$\mathbb{E}[(\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l))^2]\leq K_MT_k^{-1}\mathbb{E}[(W_l(0)^2+1)\exp(\theta_0W_l(0))]=O(T_k^{-1}).$$

Therefore, $\mathbb{E}[\|H_k\|^2]=O(\delta_k^{-2}T_k^{-1}\vee 1)$. □

EC.1.4. Proof of Theorem 1

To obtain convergence of the SGD iteration, we first need to establish a desirable convex structure of the objective function (3).

LEMMA EC.5 **(Convexity and Smoothness of $f(\mu,p)$).** *Suppose Assumption 1 holds. Then, there exist finite positive constants $0<K_0\leq 1$ and $K_1>K_0$ such that for all $x=(\mu,p)\in\mathcal{B}$,*

(a) $(\boldsymbol{x}-\boldsymbol{x}^*)^T\nabla f(x)\geq K_0\|\boldsymbol{x}-\boldsymbol{x}^*\|^2$,

(b) $|\partial_\mu^3 f(\boldsymbol{x})|,|\partial_p^3 f(\boldsymbol{x})|\leq K_1$.

We comment that although sufficient conditions for convexity of $\mathbb{E}[W(p,\mu)]$ and the $f(\mu,p)$ is not straightforward to characterize, it is quite clear in one-dimensional settings (when one of the control parameter is fixed). In detail, $\mathbb{E}[W(p)]$ is convex as long as $2(\lambda')^2+(\mu-\lambda)\lambda''>0$, while $\mathbb{E}[W(\mu)]$ is always convex. See the proof of Lemma EC.5.

We only sketch the key ideas in the proof of the convergence result (12) under the convexity structure here; the full proof is given in Appendix EC.2.1. Let $b_k=\mathbb{E}[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2]$. Then, following the SGD recursion and some algebra, we get the following recursion on b_k:

$$b_{k+1}\leq(1-2K_0\eta_k+\eta_kB_k)b_k+\eta_kB_k+\eta_k^2\mathcal{V}_k.$$

Under condition (11), we can show that the recursion coefficient $1-2K_0\eta_k+\eta_kB_k<1$, so b_k eventually converges to 0. With more careful calculation as given in Appendix EC.2.1, we can obtain the convergence rate (12) by induction using the above recursion.

Applying the convergence result (12) to LiQUAR relies on knowing the bounds on B_k and $\mathcal{V}_k$. Given Proposition 3, one can check that, if $\eta_k=O(k^{-a})$, $T_k=O(k^b)$ and $\delta_k=O(k^{-c})$, the bounds for B_k and $\mathcal{V}_k$ as specified in condition (11) holds with $\beta=\max(-a,-a-b+2c,-2c)$. Then, (13) follows immediately from (12).

EC.1.5. Proof of Proposition 4

The regret of nonstationarity

$$R_{2k}=\sum_{l=2k-1}^{2k}\mathbb{E}[\rho_l-T_kf(x_l)]=\sum_{l=2k-1}^{2k}\mathbb{E}\left[h_0\int_0^{T_k}(W_l(t)-w_l)dt-p_l(N_l-T_k\lambda(p_l))\right],$$

where $w_l=\mathbb{E}_l[W_\infty(\mu_l,p_l)]$. Conditional on p_l, N_l is a Poisson random variable with mean $T_k\lambda(p_l)$ and therefore,

$$R_{2k}=h_0\sum_{l=2k-1}^{2k}\mathbb{E}\left[\int_0^{T_k}(W_l(t)-w_l)dt\right].$$

Roughly speaking, R_{2k} depends on how fast $W_l(t)$ converges to its steady state for given (μ_l,p_l). Given the ergodicity convergence result in Lemma EC.2, we can show that $W_l(t)$ becomes close to the steady-state distribution after a warm-up period of length $t_k=O(\log(k))$.

Lemma EC.6 **(Nonstationary Error after Warm-up)**. *Suppose $T_k>t_k\equiv\log(k)/\gamma$, then*

$$\mathbb{E}\left[\int_{t_k}^{T_k}(W_l(t)-w_l)dt\right]=O(k^{-1}).$$

To obtain a finer bound for small values of t, i.e., in the warm-up period, we follow a similar idea as in Chen et al. (2024) and decompose $\mathbb{E}[W_l(t)-w_l]=\mathbb{E}[W_l(t)-w_{l-1}]+\mathbb{E}[w_{l-1}-w_l]$.

Lemma EC.7 **(Nonstationary Error in Warm-up Period)**. *Suppose $T_k>t_k\equiv\log(k)/\gamma$ for all $k\geq1$. Then, there exists a constant C_0 such that for all $l=2k-1,2k$,*

(a) $\mathbb{E}[|w_l-w_{l-1}|]\leq C_0\mathbb{E}[\|\boldsymbol{x}_l-\boldsymbol{x}_{l-1}\|]$;

(b) $\mathbb{E}\left[\int_0^{t_k}W_l(t)-w_{l-1}dt\right]\leq C_0\mathbb{E}[\|\boldsymbol{x}_l-\boldsymbol{x}_{l-1}\|^2]^{1/2}t_k$.

As a consequence,

$$\mathbb{E}\left[\int_0^{t_k}(W_l(t)-w_l)dt\right]=O\left(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)\log(k)\right).$$

Following Lemma EC.6 and Lemma EC.7, we have

$$R_{2k} = h_0 \sum_{l=2k-1}^{2k} \mathbb{E}\left[\int_0^{t_k} W_l(t) - w_l dt + \int_{t_k}^{T_k} W_l(t) - w_l dt\right] = O(k^{-1}) + O(\max(\eta_k\sqrt{\mathcal{V}_k}, \delta_k)\log(k))$$
$$= O(k^{-1}) + O(k^{-\xi}\log(k)) = O(k^{-\xi}\log(k)).$$

Furthermore, if $\eta_k = O(k^{-a}), T_k = O(k^b)$ and $\delta_k = O(k^{-c})$, then by Proposition 3, $\eta_k\sqrt{\mathcal{V}_k} = O(k^{\max(-a-b/2+c,-a)})$. As a result, $\max(\eta_k\sqrt{\mathcal{V}_k}, \delta_k) = O(k^{\max(-a-b/2+c,-a,-c)})$. Therefore, setting $\xi = \max(-a-b/2+c,-a,-c)$ finishes the proof. □

EC.1.6. Proof of Theorem 2

As discussed in Section 5.2, the bound for regret of suboptimality R_{1k} follows immediately from Theorem 1. The bound for R_{2k} follows from Proposition 4. The bound for R_{3k} follows from the smooth condition in Assumption 1.

LEMMA EC.8 **(Exploration Cost)**. *Under Assumption 1, there exists a constant $K_4 > 0$ such that*

$$R_{3k} \le K_4 T_k \delta_k^2. \tag{EC.6}$$

Now, given that $\eta_k = c_\eta k^{-1}$ with $c_\eta > 2/K_0$, $T_k = c_T k^{1/2}$ with $c_T > 0$ and $\delta_k = c_\delta k^{1/3}$ with $0 < c_\delta < \sqrt{K_0/32c}$, by Proposition 3,

$$B_k \le 2c\delta_k^2 + O(\delta_k^{-1}\exp(-\theta_1\alpha T_k)) = \frac{K_0}{16}k^{-2/3} + o(k^{-2/3}) \le \frac{K_0}{8}k^{-2/3},$$

for k large enough, and $\mathcal{V}_k = O(k^{1/3})$. So condition (11) is satisfied with $\beta = 2/3$ and hence $R_{1k} = O(k^{-1/3})$. On the other hand, conditions in Proposition 4 hold with $\xi = 1/3$ and hence $R_{2k} = O(k^{-1/3}\log(k))$. Finally, $R_{3k} = O(T_k\delta_k^2) = O(k^{-1/3})$. So we can conclude that

$$R(L) = \sum_{k=1}^{L}(R_{1k} + R_{2k} + R_{3k}) = \sum_{k=1}^{L} O(k^{-1/3}\log(k)) = O(L^{2/3}\log(L)).$$

As $T_k = O(k^{1/3})$, we have $T(L) = O(L^{4/3})$, and therefore $R(L) = O(\sqrt{T(L)}\log(T(L))$. □

EC.2. Proofs

EC.2.1. Full Proof of Theorem 1

By the SGD recursion, $\bar{\boldsymbol{x}}_{k+1} = \Pi_{\mathcal{B}}(\bar{\boldsymbol{x}}_k - \eta_k\boldsymbol{H}_k)$. Let $\mathcal{F}_k$ be the filtration up to iteration k, i.e. it includes all events in the first $2(k-1)$ cycles. By Lemma EC.5, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\boldsymbol{x}}_{k+1} - \boldsymbol{x}^*\|^2\right] \le \mathbb{E}[\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^* - \eta_k\boldsymbol{H}_k\|^2] \\
=\ &\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 - 2\eta_k\boldsymbol{H}_k\cdot(\bar{\boldsymbol{x}}_k - x^*) + \eta_k^2\|\boldsymbol{H}_k\|^2\right] \\
=\ &\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 - 2\eta_k\nabla f(\bar{\boldsymbol{x}}_k)\cdot(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)\right] - \mathbb{E}[2\eta_k(\boldsymbol{H}_k - \nabla f(\bar{\boldsymbol{x}}_k))\cdot(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)] + \mathbb{E}[\eta_k^2\|\boldsymbol{H}_k\|^2] \\
\le\ &(1 - 2\eta_k K_0)\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2\right] + \mathbb{E}[2\eta_k(\boldsymbol{H}_k - \nabla f(\bar{\boldsymbol{x}}_k))\cdot(\boldsymbol{x}^* - \bar{\boldsymbol{x}}_k)] + \eta_k^2\mathbb{E}[\|\boldsymbol{H}_k\|^2].
\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}[2\eta_k(\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k))\cdot(\boldsymbol{x}^*-\bar{\boldsymbol{x}}_k)] = &\ \mathbb{E}[\mathbb{E}[2\eta_k(\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k))\cdot(\boldsymbol{x}^*-\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]]\\ = 2\eta_k\mathbb{E}[\mathbb{E}[\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]\cdot(\boldsymbol{x}^*-\bar{\boldsymbol{x}}_k)] \leq &\ 2\eta_k\mathbb{E}[\|\mathbb{E}[\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]\|^2]^{1/2}\mathbb{E}[\|\boldsymbol{x}^*-\bar{\boldsymbol{x}}_k\|^2]^{1/2}\\ \leq \eta_k\mathbb{E}[\|\mathbb{E}[\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]\|^2]^{1/2}(1+\mathbb{E}[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2]).&\end{aligned}$$

The second last inequality follows from Hölder's Inequality, and the last inequality follows from $2a\leq 1+a^2$. Let $b_k=\mathbb{E}[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2]$ and recall that we have defined

$$B_k=\mathbb{E}[\|\mathbb{E}[\boldsymbol{H}_k-\nabla f(\bar{\boldsymbol{x}}_k)|\mathcal{F}_k]\|^2]^{1/2},\quad \mathcal{V}_k=\mathbb{E}[\|\boldsymbol{H}_k\|^2].$$

Then, we obtain the recursion

$$b_{k+1}\leq(1-2K_0\eta_k+\eta_kB_k)b_k+\eta_kB_k+\eta_k^2\mathcal{V}_k. \tag{EC.7}$$

Next, we prove by mathematical induction that there exists a large constant $K_2>0$ such that $b_k\leq K_2k^{-\beta}$ for all $k\geq 1$ using recursion (EC.7). Given that $\eta_k\mathcal{V}_k=O(k^{-\beta})$, we can find a constant $K_3>0$ large enough such that $\eta_k\mathcal{V}_k\leq K_3k^{-\beta}$ for all $k\geq 1$. Then, by the induction assumption that $b_k\leq K_2k^{-\beta}$, we have

$$b_{k+1}\leq(1-2K_0\eta_k+\eta_kB_k)b_k+\eta_kB_k+\eta_k^2\mathcal{V}_k\leq\left(1-2K_0\eta_k+\frac{K_0}{8}\eta_kk^{-\beta}\right)b_k+\frac{K_0}{8}\eta_kk^{-\beta}+K_3\eta_kk^{-\beta}.$$

Note that $k^{-\beta}/(k+1)^{-\beta}=(1+\frac{1}{k})^{\beta}\leq 1+\frac{1}{k}\leq 1+\frac{K_0}{2}\eta_k$. So we have

$$\begin{aligned}b_{k+1}&\leq\left(1-2K_0\eta_k+\frac{K_0}{8}\eta_kk^{-\beta}\right)\left(1+\frac{K_0\eta_k}{2}\right)K_2(k+1)^{-\beta}+\frac{K_0}{8}\eta_kk^{-\beta}+K_3\eta_kk^{-\beta}\\&\leq K_2(k+1)^{-\beta}-\eta_kk^{-\beta}\left(\frac{3K_0K_2}{2}-\frac{K_0K_2}{8}k^{-\beta}-\frac{K_0^2K_2}{16}\eta_kk^{-\beta}-\frac{K_0}{8}-K_3\right).\end{aligned}$$

Then, we have $b_{k+1}\leq K_2(k+1)^{-\beta}$ as long as

$$\frac{3K_0K_2}{2}-\frac{K_0K_2}{8}k^{-\beta}-\frac{K_0^2K_2}{16}\eta_kk^{-\beta}-\frac{K_0}{8}-K_3\geq 0.$$

As the step size $\eta_k\to 0$, $\eta_kK_0\leq 1$ for k large enough. Let $k_0=\max\{k\geq 1:\eta_kK_0>1\}$. Then, if $K_2\geq 8K_3/K_0$, for all $k\geq k_0$,

$$\frac{3K_0K_2}{2}-\frac{K_0K_2}{8}\Delta_k-\frac{K_0^2K_2}{16}\eta_k\Delta_k-\frac{K_0}{8}-K_3\geq\frac{3K_0K_2}{2}-\frac{K_0K_2}{8}-\frac{K_0K_2}{16}-\frac{K_0K_2}{8}-\frac{K_0K_2}{8}=\frac{17K_0K_2}{16}>0.$$

Let

$$K_2=\max\left(k_0^{\beta}(|\bar{\mu}-\underline{\mu}|^2+|\bar{p}-\underline{p}|^2),8K_3/K_0\right).$$

Then we have $\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2\leq K_2k^{-\beta}$ for all $1\leq k\leq k_0$, and we can conclude by induction that, for all $k\geq k_0$,

$$\mathbb{E}[\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}^*\|^2]\leq K_2k^{-\beta}.$$

□

EC.2.2. Proofs of Technical Lemmas

In addition to the uniform moment bounds for $W_l(t)$ as stated in Lemma EC.1, we also need to establish similar bounds for the so-called observed busy period $X_l(t)$, which will be used in the proof of Lemma EC.7. In detail, $X_l(t)$ is the units of time that has elapsed at time point t in cycle l since the last time when the server is idle (probably in a previous cycle). So the value of $X_l(t)$ is uniquely determined by $\{W_l(t)\}$, i.e., $X_l(t)=0$ whenever $W_l(t)=0$ and $dX_l(t)=dt$ whenever $W_l(t)>0$.

Lemma EC.9 **(Complete Version of Lemma EC.1)**. *Under Assumptions 1 and 2, there exist some constants $\theta_0>0$ and $M>1$ such that, for any sequence of control parameters $\{(\mu_l,p_l):l\geq 1\}$,*

$$\mathbb{E}[X_l^m(t)]\leq M,\quad \mathbb{E}[W_l(t)^m]\leq M,\quad \mathbb{E}[W_l(t)^m\exp(2\theta_0 W_l(t))]\leq M,$$

for all $m\in\{0,1,2\}$, $l\geq 1$ and $0\leq t\leq T_k$ with $k=\lceil l/2\rceil$.

Proof of Lemma EC.9 We consider a $M/GI/1$ system under a stationary policy such that $\mu_l\equiv \underline{\mu}$ and $p_l\equiv \underline{p}$ for all $l\geq 1$. We call this system the dominating system and denote its workload process by $W_l^D(t)$. In addition, we set $W_1^D(0)\overset{d}{=}W_\infty(\underline{\mu},\underline{p})$ so that $W_l^D(t)\overset{d}{=}W_\infty(\underline{\mu},\underline{p})$ for all $l\geq 1$ and $t\in[0,T_k]$. Then, the arrival process in the dominating system is an upper envelop process (UEP) for all possible arrival processes corresponding to any control sequence (μ_l,p_l) and the service process in the dominating system is a lower envelope process (LEP) for all possible service processes corresponding to any control sequence. In addition, $W_1(0)=0\leq W_l^D(t)$. So we have

$$W_l(t)\leq_{st} W_l^D(t)\overset{d}{=}W_\infty(\underline{\mu},\underline{p}),\ \text{for all } l\geq 1 \text{ and } t\in[0,T_k].$$

By Theorem 5.2 in the Chapter X of Asmussen (2003), the stationary workload process

$$W_\infty(\underline{\mu},\underline{p})\overset{d}{=}Y_1+...+Y_N.$$

Here N is geometric random variable of mean $1/(1-\bar{\rho})$ and $\bar{\rho}=\lambda(\underline{p})/\underline{\mu}$, and Y_n are I.I.D. random variables independent of N. In addition, the density of Y_n is

$$f_Y(t)=\frac{\mathbb{P}(V_n>t)}{\mathbb{E}[V_n]},\quad t\in[0,\infty).$$

Under Assumption 2, we have

$$\mathbb{P}(Y_n>t)=\int_t^\infty f_Y(s)ds=\int_t^\infty \frac{\mathbb{P}(V_n>s)}{\mathbb{E}[V_n]}ds\leq\int_t^\infty \frac{\exp(-\eta s)\mathbb{E}[\exp(\eta V_n)]}{\mathbb{E}[V_n]}ds=\frac{\mathbb{E}[\exp(\eta V_n)]}{\eta\mathbb{E}[V_n]}\cdot\exp(-\eta t).$$

As a consequence, Y_n has finite moment generating function around the origin. As $W_\infty(\underline{\mu},\underline{p})$ is a geometric compound of Y_n, it also has finite moment generating function around the origin. So we can conclude that, there exists some constants $\theta_0\in(0,\theta/2)$ and $C\geq 1$ such that

$$\mathbb{E}[W_l(t)^m]\leq\mathbb{E}[W_\infty(\underline{\mu},\underline{p})^m]\leq C,\quad \mathbb{E}[W_l(t)^m\exp(2\theta_0 W_l(t))]\leq\mathbb{E}[W_\infty(\underline{\mu},\underline{p})^m\exp(2\theta_0 W_\infty(\underline{\mu},\underline{p}))]\leq C,$$

for $m = 1, 2$.

To deal with the observed busy period, we need to do a time-change. In detail, for each cycle l and control parameter (μ_l, p_l), we "slow down" the clock by $\lambda(p_l)$ times so that the arrival rate is normalized to 1 and mean service time to $\lambda(p_l)/\mu_l$. We denote the time-changed workload and observed busy period by $\tilde{W}_l(t)$ and $\tilde{X}_l(t)$ for $t \in [0, \lambda(p_l)T_k]$. Then, for all $t \in [0, T_k]$,

$$W_l(t) \le \frac{1}{\lambda(\bar{p})} \tilde{W}_l\left(\lambda(p_l)t\right), \quad X_l(t) \le \frac{1}{\lambda(\bar{p})} \tilde{X}_l\left(\lambda(p_l)t\right).$$

We denote by $\tilde{X}_l^D(t)$ the time-changed observed busy period corresponding to the dominating system. Then, since $\lambda(p_l)/\mu_l/ \le \lambda(\underline{p})/\underline{\mu}$ for all possible values of (μ_l, p_l), we can conclude that $\tilde{X}_l(t) \le_{st} \tilde{X}_l^D(t)$. Following Nakayama et al. (2004), $\mathbb{E}[\tilde{X}_l^D(t)] \le \mathbb{E}[X_\infty(1, \underline{\mu}/\lambda(\underline{p}))] < \infty$. Let $M = C \vee \left(\mathbb{E}[X_\infty(1, \underline{\mu}/\lambda(\underline{p}))]/\lambda(\bar{p})\right)$ and we can conclude that $\mathbb{E}[X_l(t)] \le M$. □

Proof of Lemma EC.2 Let $N(t)$ be the arrival process under control parameter (μ, p), which is a Poisson process with rate $\lambda(p)$. Define an auxiliary Lévy process as $R(t) = \sum_{i=1}^{N(t)} V_i - \mu t$. For the workload processes $W(t)$ and $\bar{W}(t)$, define two hitting times τ and $\bar{\tau}$ as

$$\tau \equiv \min_{t\ge 0}\{t : W(0) + R(t) = 0\}, \quad \text{and} \quad \bar{\tau} \equiv \min_{t\ge 0}\{t : \bar{W}(0) + R(t) = 0\}.$$

Following Lemma 2 of Chen et al. (2024), we have

$$|W(t) - \bar{W}(t)| \le |W(0) - \bar{W}(0)| \mathbf{1}\left(t < \tau \vee \bar{\tau}\right). \tag{EC.8}$$

Next, we give a bound for the probability $\mathbb{P}(\tau > t)$ by constructing an exponential supermartingale. Define

$$M(t) = \exp\left(\theta_0(W(0) + R(t)) + \gamma t\right),$$

where θ_0 is defined in Lemma EC.9 and the value of γ will be specified in (EC.9). Let $\{\mathcal{F}_t\}_{t\ge 0}$ be the natural filtration associated to $R(t)$. For any $t, s > 0$,

$$\begin{aligned}\mathbb{E}[M(t+s)|\mathcal{F}_t] &= \mathbb{E}[M(t)\exp(\theta_0(R(t+s) - R(t)) + \gamma s)|\mathcal{F}_t] = M(t)\mathbb{E}[\exp(\theta_0 R(s) + \gamma s)] \\ &= M(t)\mathbb{E}\left[\exp\left(\theta_0 \sum_{i=1}^{N(s)} V_i - \theta_0 \mu s + \gamma s\right)\right] = M(t)\mathbb{E}\left[\mathbb{E}[\exp(\theta_0 V_i)]^{N(s)}\right] e^{-\theta_0 \mu s + \gamma s} \\ &= M(t)\exp\left(s\left(\lambda \mathbb{E}[\exp(\theta_0 V_i)] - \lambda - \mu\theta_0 + \gamma\right)\right).\end{aligned}$$

According to Assumption 2, $\phi(\theta) < \log(1 + \underline{\mu}\theta/\bar{\lambda}) - \gamma_0$ for some $\theta, \gamma_0 > 0$. Besides, the function $h(x) \equiv \phi(x) - \log(1 + \underline{\mu}x/\bar{\lambda})$ is convex on $[0, \theta]$. As $0 < \theta_0 < \theta$, we have

$$h(\theta_0) \le (1 - \theta_0/\theta)h(0) + \frac{\theta_0}{\theta}h(\theta) < -\frac{\theta_0}{\theta}\gamma_0.$$

We choose

$$\gamma = \underline{\lambda}\left(1 - e^{-\frac{\theta_0\gamma_0}{\theta}}\right)\left(1 + \underline{\mu}\theta_0/\bar{\lambda}\right). \tag{EC.9}$$

Then, it satisfies that

$$\begin{aligned}\lambda\mathbb{E}[\exp(\theta_0 V_i)] - \lambda - \mu\theta_0 + \gamma = \lambda\left(e^{\phi(\theta_0)} - (1+\frac{\mu\theta_0}{\lambda}) + \frac{\gamma}{\lambda}\right) &< \lambda\left(e^{-\frac{\theta_0}{\theta}\gamma_0}(1+\underline{\mu}\theta_0/\bar{\lambda}) - (1+\mu\theta_0/\lambda) + \frac{\gamma}{\lambda}\right)\\ &< \lambda\left(-\left(1 - e^{\frac{\theta_0\gamma_0}{\theta}}\right)(1+\underline{\mu}\theta_0/\bar{\lambda}) + \frac{\gamma}{\underline{\lambda}}\right) = 0.\end{aligned}$$

Now, we can conclude that $M(t)$ is an non-negative supermartingale with γ as given by (EC.9). By Fatou's lemma,

$$\begin{aligned}\mathbb{P}(\tau > t|W(0)) &\le e^{-\gamma t}\mathbb{E}[\exp(\gamma\tau)|W(0)] = e^{-\gamma t}\mathbb{E}[\liminf_{n\to\infty} M(\tau\wedge n)|W(0)]\\ &\le e^{-\gamma t}\liminf_{n\to\infty}\mathbb{E}[M(\tau\wedge n)|W(0)] \le e^{-\gamma t}\mathbb{E}[M(0)|W(0)] = e^{-\gamma t}\exp(\theta_0 W(0)).\end{aligned}$$

Similarly, $\mathbb{P}(\bar{\tau} > t|\bar{W}(0)) \le e^{-\gamma t}\exp(\theta_0\bar{W}(0))$. Combining these bounds with (EC.8), we can conclude that

$$\begin{aligned}\mathbb{E}\left[|W(t) - \bar{W}(t)|^m|W(0), \bar{W}(0)\right] &\le |W(0) - \bar{W}(0)|^m\mathbb{P}(\tau\vee\bar{\tau} > t|W(0), \bar{W}(0))\\ &\le |W(0) - \bar{W}(0)|^m\left(\mathbb{P}(\tau > t|W(0)) + \mathbb{P}(\bar{\tau} > t|\bar{W}(0))\right)\\ &\le |W(0) - \bar{W}(0)|^m\left(e^{\theta_0 W(0) + \theta_0\bar{W}(0)}\right)e^{-\gamma t}.\end{aligned}$$

□

Proof of Lemma EC.3 We first analyze the conditional expectation $\mathbb{E}_l[(W_l(t) - w_l)(W_l(s) - w_l)]$ for each given pair of (s,t) such that $0 \le s \le t \le T_k$. To do this, we synchronously couple with $\{W_l(r) : s \le r \le T_k\}$ a stationary workload process $\{\bar{W}_l^s(r) : s \le r \le T_k\}$. In particular, $\bar{W}_l^s(s)$ is independently drawn from the stationary distribution $W_\infty(\mu_l, p_l)$. As a result, $\bar{W}_l^s(r)$ is independent of $W_l(s)$ for all $s \le r \le T_k$, and hence

$$\mathbb{E}_l[W_l(s)(\bar{W}_l^s(t) - w_l)] = \mathbb{E}_l[W_l(s)]\left(\mathbb{E}_l[\bar{W}_l^s(t)] - w_l\right) = 0.$$

Then, we have

$$\mathbb{E}_l[(W_l(t) - w_l)(W_l(s) - w_l)] = \mathbb{E}_l[(W_l(t) - \bar{W}_l^s(t))W_l(s)] - w_l\mathbb{E}_l[W_l(s) - w_l].$$

By Lemma EC.2,

$$\mathbb{E}_l[(W_l(t) - \bar{W}_l^s(t))W_l(s)|W_l(s), \bar{W}_l^s(s)] \le \exp(-\gamma(t-s))(e^{\theta_0 W_l(s)} + e^{\theta_0\bar{W}_l^s(s)})(W_l(s) + \bar{W}_l^s(s))W_l(s).$$

As $\bar{W}_l^s(s)$ is independent of $W_l(s)$,

$$\begin{aligned}
&\mathbb{E}_l[(W_l(t)-\bar{W}_l^s(t))W_l(s)|W_l(s)]\\
\le\ &\exp(-\gamma(t-s))\mathbb{E}_l\left[(e^{\theta_0 W_l(s)}+e^{\theta_0\bar{W}_l^s(s)})(W_l(s)+\bar{W}_l^s(s))W_l(s)|W_l(s)\right]\\
=\ &\exp(-\gamma(t-s))(e^{\theta_0 W_l(s)}W_l(s)^2+e^{\theta_0 W_l(s)}W_l(s)\mathbb{E}[\bar{W}_l^s(s)]+W_l(s)^2\mathbb{E}[e^{\theta_0\bar{W}_l^s(s)}]+W_l(s)\mathbb{E}[e^{\theta_0\bar{W}_l^s(s)}\bar{W}_l^s(s)])\\
\le\ &\exp(-\gamma(t-s))(e^{\theta_0 W_l(s)}W_l(s)^2+Me^{\theta_0 W_l(s)}W_l(s)+MW_l(s)^2+MW_l(s)).
\end{aligned}$$

One can check that $W_l(s)\le W_l(0)+\bar{W}_l(s)$, where $\bar{W}_l(s)$ is a stationary workload process synchronously coupled with $W_l(t)$ having an independent drawn initial $\bar{W}_l(0)$. Therefore,

$$\begin{aligned}
\mathbb{E}_l\left[e^{\theta_0 W_l(s)}W_l(s)^2\right]&\le e^{\theta_0 W_l(0)}\mathbb{E}_l\left[(W_l(0)+\bar{W}_l(s))^2e^{\theta_0\bar{W}_l(s)}\right]\\
&=e^{\theta_0 W_l(0)}\left(W_l(0)^2\mathbb{E}_l[e^{\theta_0\bar{W}_l(s)}]+2W_l(0)\mathbb{E}_l\left[\bar{W}_l(s)e^{\theta_0\bar{W}_l(s)}\right]+\mathbb{E}_l\left[W_l(s)^2e^{\theta_0\bar{W}_l(s)}\right]\right)\\
&\le 2Me^{\theta_0 W_l(0)}(1+W_l(0)^2),\\
\mathbb{E}_l\left[e^{\theta_0 W_l(s)}W_l(s)\right]&\le e^{\theta_0 W_l(0)}\mathbb{E}_l\left[W_l(0)e^{\theta_0\bar{W}_l(s)}+\bar{W}_l(s)e^{\theta_0\bar{W}_l(s)}\right]\le e^{\theta_0 W_l(0)}M(1+W_l(0))\\
&\le\frac{3M}{2}e^{\theta_0 W_l(0)}(1+W_l(0)^2),
\end{aligned}$$

where the last inequality holds because the constant $M\ge 1$ and $W_l(0)\le(1+W_l(0)^2)/2$. Note that $W_l(s)^2\le e^{\theta_0 W_l(s)}W_l(s)^2$ and $W_l(s)\le W_l(s)e^{\theta_0 W_l(s)}$, we have

$$\mathbb{E}_l[(W_l(t)-\bar{W}_l^s(t))W_l(s)]\le e^{-\gamma(t-s)}e^{\theta_0 W_l(0)}(1+W_l(0)^2)(2M+5M^2).$$

On the other hand, by Lemma EC.2,

$$\begin{aligned}
|\mathbb{E}_l[W_l(s)-w_l]|&\le\exp(-\gamma s)MW_l(0)(M+W_l(0))\exp(\theta_0 W_l(0))\\
&\le e^{-\gamma s}e^{\theta_0 W_l(0)}M^2(1+W_l(0))^2\le 2M^2e^{-\gamma s}e^{\theta_0 W_l(0)}(1+W_l(0)^2).
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\mathbb{E}_l[(W_l(t)-w_l)(W_l(s)-w_l)]&=\mathbb{E}_l[(W_l(t)-\bar{W}_l^s(t))W_l(s)]-w_l\mathbb{E}_l[W_l(s)-w_l]\\
&\le(e^{-\gamma(t-s)}+e^{-\gamma s})e^{\theta_0 W_l(0)}(1+W_l(0)^2)(2M+5M^2+2M^3).
\end{aligned}$$

and we can conclude (EC.4) with $K_V=2M+5M^2+2M^3$. □

Proof of Lemma EC.4 By the mean value theorem,

$$\begin{aligned}
f(\mu_1,p)&=f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{\mu_1-\mu_2}{2}\partial_\mu f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{(\mu_1-\mu_2)^2}{8}\partial_\mu^2 f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{(\mu_1-\mu_2)^3}{48}\partial_\mu^3 f(\xi_1,p)\\
f(\mu_2,p)&=f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{\mu_2-\mu_1}{2}\partial_\mu f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{(\mu_1-\mu_2)^2}{8}\partial_\mu^2 f\left(\frac{\mu_1+\mu_2}{2},p\right)+\frac{(\mu_2-\mu_1)^3}{48}\partial_\mu^3 f(\xi_2,p),
\end{aligned}$$

where ξ_1 and ξ_2 take values between μ_1 and μ_2. As a consequence, we have

$$\left|\frac{f(\mu_1,p)-f(\mu_2,p)}{\mu_1-\mu_2}-\partial_\mu f\left(\frac{\mu_1+\mu_2}{2},p\right)\right|\le c(\mu_1-\mu_2)^2,$$

with $c=(\max_{(\mu,p)\in\mathcal{B}}|\partial^3_\mu f(\mu,p)|\vee|\partial^3_p f(\mu,p)|)/24$. Following the same argument, we have

$$\left|\frac{f(\mu,p_1)-f(\mu,p_2)}{p_1-p_2}-\partial_\mu f\left(\mu,\frac{p_1+p_2}{2}\right)\right|\le c(p_1-p_2)^2.$$

□

Proof of Lemma EC.5 By Pollaczek-Khinchin formula and PASTA,

$$f(\mu,p)=\frac{h_0(1+c_V^2)}{2}\cdot\frac{\lambda(p)}{\mu-\lambda(p)}+c(\mu)-p\lambda(p).$$

We intend to show that $f(\mu,p)$ is strongly convex in $\mathcal{B}$. For ease of notation, denote $C=\frac{1+c_V^2}{2}$ and

$$g(\mu,\lambda)=\frac{\lambda}{\mu-\lambda}.$$

Write $\lambda(p)$,$\lambda'(p)$ and $\lambda''(p)$ as λ, λ' and λ'' respectively. By direct calculation, we have

$$\partial_\lambda g=\frac{\mu}{(\mu-\lambda)^2},\partial_\mu g=\frac{\lambda}{(\mu-\lambda)^2},\partial^2_{\lambda\lambda}g=\frac{2\mu}{(\mu-\lambda)^3},\partial^2_{\lambda\mu}g=-\frac{\mu+\lambda}{(\mu-\lambda)^3},\partial^2_{\mu\mu}g=\frac{2\lambda}{(\mu-\lambda)^3}.$$

The second-order derivatives are

$$\begin{aligned}\partial_{pp}f&=\frac{h_0C\mu}{(\mu-\lambda)^3}\left(2(\lambda')^2+(\mu-\lambda)\lambda''\right)-p\lambda''-2\lambda'\\ \partial_{p\mu}f&=-\frac{h_0C(\mu+\lambda)}{(\mu-\lambda)^3},\quad \partial_{\mu\mu}f=\frac{2h_0C\lambda}{(\mu-\lambda)^3}+c''(\mu).\end{aligned}$$

By Condition (a) of Assumption 1, we have

$$-p\lambda''-2\lambda'>0\quad\text{and}\quad 2(\lambda')^2+(\mu-\lambda)\lambda''>0\quad\Rightarrow\quad\partial_{pp}f>0.$$

It is easy to check that $\partial_{\mu\mu}f>0$ as $c(\mu)$ is convex. So, to verify the convexity of f, we only need to show that the determinant of Hessian metric $\boldsymbol{H}_f$ is positive in $\mathcal{B}$. By direct calculation,

$$\begin{aligned}|\boldsymbol{H}_f|&=\frac{h_0^2C^2}{(\mu-\lambda)^5}\left(2\mu\lambda\lambda''-(\mu-\lambda)(\lambda')^2\right)+(-p\lambda''-2\lambda')\frac{2h_0C\lambda}{(\mu-\lambda)^3}+c''(\mu)\partial_{pp}f\\ &\ge\frac{h_0^2C^2}{(\mu-\lambda)^5}\left(2\mu\lambda\lambda''-(\mu-\lambda)(\lambda')^2\right)+(-p\lambda''-2\lambda')\frac{2h_0C\lambda}{(\mu-\lambda)^3}\\ &=\frac{h_0C}{(\mu-\lambda)^5}\left[h_0C(2\mu\lambda\lambda''-(\mu-\lambda)(\lambda')^2)+2\lambda(\mu-\lambda)^2(-p\lambda''-2\lambda')\right]\\ &=-\frac{h_0C\lambda'}{(\mu-\lambda)^4}\left[h_0C\lambda'+4\lambda(\mu-\lambda)-2\frac{h_0C\mu-p(\mu-\lambda)^2}{\mu-\lambda}\frac{\lambda''\lambda}{\lambda'}\right].\end{aligned}$$

As $-\lambda' > 0$, we need to prove the term in bracket is positive. Note that the term

$$\frac{h_0 C\mu - p(\mu-\lambda)^2}{\mu-\lambda} = h_0 C + \frac{h_0 C\lambda}{\mu-\lambda} - p(\mu-\lambda)$$

is monotonically decreasing in μ. By Assumption 1, we have, for all $\mu \in [\underline{\mu}, \bar{\mu}]$ and $\lambda \in [\underline{\lambda}, \bar{\lambda}]$,

$$\begin{aligned}
&h_0 C\lambda' + 4\lambda(\mu-\lambda) - 2\frac{h_0 C\mu - p(\mu-\lambda)^2}{\mu-\lambda}\frac{\lambda''\lambda}{\lambda'} \\
\geq& h_0 C\lambda' + 4\lambda(\underline{\mu}-\lambda) - 2\left(h_0 C + \frac{h_0 C\lambda}{\mu-\lambda} - p(\mu-\lambda)\right)\frac{\lambda''\lambda}{\lambda'} \\
\geq& h_0 C\lambda' + 4\lambda(\underline{\mu}-\lambda) - 2h_0 C\frac{\lambda''\lambda}{\lambda'} - 2\max\left\{\left(\frac{h_0 C\lambda}{\underline{\mu}-\lambda} - p(\underline{\mu}-\lambda)\right)\frac{\lambda''\lambda}{\lambda'}, \left(\frac{h_0 C\lambda}{\bar{\mu}-\lambda} - p(\bar{\mu}-\lambda)\right)\frac{\lambda''\lambda}{\lambda'}\right\} \\
>& 0.
\end{aligned}$$

As $\mathcal{B}$ is compact, we can conclude that $f(\mu,p)$ is strongly convex on $\mathcal{B}$. Then by Taylor's expansion, Statement (a) holds for some $1 \geq K_0 > 0$. Statement (b) follows immediately after Assumption 1.

□

Proof of Lemma EC.6 By Lemma EC.2, conditional on μ_l, p_l and $W_l(0)$, we have

$$\begin{aligned}
\mathbb{E}_l[|W_l(t) - \bar{W}_l(t)|] &\leq \exp(-\gamma t)\mathbb{E}_l\left[|W_l(0) - \bar{W}_l(0)|(\exp(\theta_0 W_l(0)) + \exp(\theta_0 \bar{W}_l(0)))\right] \\
&\leq \exp(-\gamma t)\left(W_l(0)\exp(\theta_0 W_l(0)) + MW_l(0) + M\exp(\theta_0 W_l(0)) + M\right) \\
&\leq \exp(-\gamma t)M(M + W_l(0))\exp(\theta_0 W_l(0)).
\end{aligned}$$

As a consequence, for $t \geq t_k$,

$$\begin{aligned}
\mathbb{E}[|W_l(t) - \bar{W}_l(t)|] &\leq \mathbb{E}[\exp(-\gamma t)M(M + W_l(0))\exp(\theta_0 W_l(0))] \\
&= \exp(-\gamma t)\left(M^2\mathbb{E}[\exp(\theta_0 W_l(0))] + M\mathbb{E}[W_l(0)\exp(\theta_0 W_l(0))]\right) \leq \exp(-\gamma t)\cdot(M^2 + M^3)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\int_{t_k}^{T_k}(W_l(t) - w_l)dt\right] &= \int_{t_k}^{T_k}\mathbb{E}[W_l(t) - w_l]dt \ \leq \int_{t_k}^{T_k}\mathbb{E}[|W_l(t) - \bar{W}_l(t)|]dt \\
&\leq \int_{t_k}^{T_k}\exp(-\gamma t)\cdot(M^2 + M^3)dt \ \leq \exp(-\gamma t_k)\cdot\frac{M^2 + M^3}{\gamma} \\
&\leq k^{-1}\cdot\frac{M^2 + M^3}{\gamma} = O(k^{-1}).
\end{aligned}$$

□

Proof of Lemma EC.7 Statement (1) is a direct corollary of Pollaczek–Khinchine formula. The proof of Statement (2) involves coupling workload processes with different parameters. Let us first explain the coupling in detail. Suppose $W^1(t)$ and $W^2(t)$ are two workload processes on $[0,T]$ with parameters (μ_1, λ_1) and (μ_2, λ_2) respectively. Let $W^1(0)$ and $W^2(0)$ be the given initial states. We construct two workload processes $\tilde{W}^1(t)$ and $\tilde{W}^2(t)$ on $[0,\infty)$ with parameters $(\mu_1/\lambda_1, 1)$ and $(\mu_2/\lambda_2, 1)$ such that $\tilde{W}^i(0) = W^i(0)$ for $i = 1, 2$. The two processes $\tilde{W}^1(t)$ and $\tilde{W}^2(t)$ are coupled such that they share the same Poisson arrival process $N(t)$ with rate 1 and the same sequence of individual workload V_n.

Then, we can couple $W^i(t)$ with $\tilde{W}(t)$ via a change of time, i.e. $W^i(t) = \tilde{W}^i(\lambda_i t)$ and obtain

$$\int_0^T W^i(t)dt = \frac{1}{\lambda_i}\int_0^{\lambda_i T} \tilde{W}^i(t)dt, \text{ for } i = 1, 2.$$

Without loss of generality, assuming $\lambda_1 \geq \lambda_2$ and we have

$$\begin{aligned}&\left|\int_0^T W^1(t)dt - \int_0^T W^2(t)dt\right| \\ \leq &\frac{1}{\lambda_1}\left|\int_0^{\lambda_2 T}(\tilde{W}^1(t) - \tilde{W}^2(t))dt\right| + \left|\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right|\int_0^{\lambda_2 T}\tilde{W}^2(t)dt + \frac{1}{\lambda_1}\int_{\lambda_2 T}^{\lambda_1 T}\tilde{W}^1(t)dt. \end{aligned} \tag{EC.10}$$

Following a similar argument as in the proof of Lemma 3 in Chen et al. (2024), we have that

$$|\tilde{W}^1(t) - \tilde{W}^2(t)| \leq \left|\frac{\mu_1}{\lambda_1} - \frac{\mu_2}{\lambda_2}\right|\max(\tilde{X}^1(t), \tilde{X}^2(t)) + |W^1(0) - W^2(0)|,$$

where $\tilde{X}^i(t)$ is the observed busy period at time t, i.e.

$$\tilde{X}^i(t) = t - \sup\{s : 0 \leq s \leq t, \tilde{W}^i(s) = 0\}.$$

To apply (EC.10) to bound $\mathbb{E}[W_l(t) - w_{l-1}]$, we construct a stationary workload process $\bar{W}_{l-1}(t)$ with control parameter (μ_{l-1}, p_{l-1}) synchronously coupled with $W_{l-1}(t)$ since the beginning of cycle $l-1$. In particular, $\bar{W}_{l-1}(0)$ is independently drawn from the stationary distribution of $W_\infty(\mu_{l-1}, p_{l-1})$. We extend the sample path $\bar{W}_{l-1}(t)$ to cycle l, i.e. for $t \geq T_{k(l-1)}$ with $k(l-1) = \lceil (l-1)/2 \rceil$, and couple it with $W_l(t)$ following the procedure described above. Then we have

$$\mathbb{E}\left[\int_0^{t_k}(W_l(t) - w_{l-1})dt\right] \leq \mathbb{E}\left[\left|\int_0^{t_k} W_l(t)dt - \int_0^{t_k}\bar{W}_{l-1}(T_{k(l-1)} + t)dt\right|\right].$$

Without loss of generality, assume $\lambda_l \geq \lambda_{l-1}$. Then following (EC.10), we have

$$\begin{aligned}&\left|\int_0^{t_k} W_l(t)dt - \int_0^{t_k}\bar{W}_{l-1}(T_{k(l-1)} + t)dt\right| \\ \leq\ &\frac{1}{\lambda_l}\left|\int_0^{\lambda_{l-1}t_k}(\tilde{W}_l(t) - \tilde{W}_{l-1}(T_{k(l-1)} + t))dt\right| + \left|\frac{1}{\lambda_l} - \frac{1}{\lambda_{l-1}}\right|\int_0^{\lambda_{l-1}t_k}\tilde{W}_{l-1}(t)dt + \frac{1}{\lambda_l}\int_{\lambda_{l-1}t_k}^{\lambda_l t_k}\tilde{W}_l(t)dt \\ \leq\ &\frac{1}{\lambda_l}\int_0^{\lambda_{l-1}t_k}\left|\tilde{W}_l(t) - \tilde{W}_{l-1}(T_{k(l-1)} + t)\right|dt + \left|\frac{1}{\lambda_l} - \frac{1}{\lambda_{l-1}}\right|\int_0^{\lambda_{l-1}t_k}\tilde{W}_{l-1}(t)dt + \frac{1}{\lambda_l}\int_{\lambda_{l-1}t_k}^{\lambda_l t_k}\tilde{W}_l(t)dt,\end{aligned}$$

where $\tilde{W}_l(\cdot)$ and $\tilde{W}_{l-1}(\cdot)$ are the time-change version of $W_l(\cdot)$ and $\bar{W}_{l-1}(\cdot)$, respectively, such that their Poisson arrival processes are both of rate 1. For the first term, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\tilde{W}_l(t)-\tilde{W}_{l-1}(T_{k(l-1)}+t)\right|\right]\\
\le\ &\mathbb{E}\left[\left|\frac{\mu_l}{\lambda_l}-\frac{\mu_{l-1}}{\lambda_{l-1}}\right|\max(\tilde{X}_l(t),\tilde{X}_{l-1}(T_{k(l-1)}+t))+|W_l(0)-\bar{W}_{l-1}(T_{k(l-1)})|\right]\\
\overset{(a)}{\le}\ &\mathbb{E}\left[\left|\frac{\mu_l}{\lambda_l}-\frac{\mu_{l-1}}{\lambda_{l-1}}\right|\tilde{X}_l^D(t)\right]+\mathbb{E}\left[|W_{l-1}(T_{k(l-1)})-\bar{W}_{l-1}(T_{k(l-1)})|\right]\\
\overset{(b)}{\le}\ &\mathbb{E}\left[\left|\frac{\mu_l}{\lambda_l}-\frac{\mu_{l-1}}{\lambda_{l-1}}\right|\tilde{X}_l^D(t)\right]+O(k^{-1})\\
\le\ &\mathbb{E}\left[\left|\frac{\mu_l}{\lambda_l}-\frac{\mu_{l-1}}{\lambda_{l-1}}\right|^2\right]^{1/2}\mathbb{E}\left[\tilde{X}_l^D(t)^2\right]^{1/2}+O(k^{-1})\\
\overset{(c)}{=}\ &O(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k))+O(k^{-1})=O(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)),
\end{aligned}$$

where $\tilde{X}_l^D(\cdot)$ is the dominant observed busy period defined in the proof of Lemma EC.9. Here inequality (a) follows from the definition of $\tilde{X}_l^D(\cdot)$, inequality (b) from Lemma EC.6 and equality (c) from Lemma EC.9 and the fact that

$$\|\boldsymbol{x}_l-\boldsymbol{x}_{l-1}\|=\begin{cases}\delta_k & \text{for } l=2k\\ \eta_k\|\boldsymbol{H}_{k-1}\| & \text{for } l=2k-1.\end{cases}$$

For the second term,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{\lambda_l}-\frac{1}{\lambda_{l-1}}\right|\int_0^{\lambda_{l-1}t_k}\tilde{W}_{l-1}(t)dt\right]=\mathbb{E}\left[\left|1-\frac{\lambda_{l-1}}{\lambda_l}\right|\int_0^{t_k}W_{l-1}(t)dt\right]\\
\le\ &\frac{1}{\underline{\lambda}}\mathbb{E}\left[(\lambda_l-\lambda_{l-1})^2\right]^{1/2}\mathbb{E}\left[\left(\int_0^{t_k}W_{l-1}(t)dt\right)^2\right]^{1/2}=O(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)t_k).
\end{aligned}$$

Following a similar argument, we have that

$$\mathbb{E}\left[\frac{1}{\lambda_l}\int_{\lambda_{l-1}t_k}^{\lambda_l t_k}\tilde{W}_l(t)dt\right]=\mathbb{E}\left[\int_{\frac{\lambda_{l-1}}{\lambda_l}t_k}^{t_k}W_l(t)dt\right]=O(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)t_k).$$

In summary, we can conclude that there exists a constant $C_0>0$ such that

$$\mathbb{E}\left[\int_0^{t_k}(W_l(t)-w_l)dt\right]\le t_k\mathbb{E}\left[|w_l-w_{l-1}|\right]+\mathbb{E}\left[\left|\int_0^{t_k}(W_l(t)-\bar{W}_{l-1}(T_{k(l-1)}+t))dt\right|\right]\le C_0\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)t_k.$$

As a consequence,

$$\mathbb{E}\left[\int_0^{t_k}(W_l(t)-w_l)dt\right]\le C_0\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)t_k=O\left(\max(\eta_k\sqrt{\mathcal{V}_k},\delta_k)\log(k)\right).$$

□

Proof of Lemma EC.8 By Taylor's expansion and the mean value theorem,

$$R_{3k}=\mathbb{E}[T_k\left(f(\boldsymbol{x}_{2k-1})+f(\boldsymbol{x}_{2k})-2f(\bar{\boldsymbol{x}}_k)\right)]=\mathbb{E}[T_k(f''(\boldsymbol{x}')+f''(\boldsymbol{x}''))\delta_k^2]\le K_4T_k\delta_k^2,$$

where $\boldsymbol{x}',\boldsymbol{x}''\in\mathcal{B}$ and the last inequality follows from Lemma EC.5. □

EC.2.3. Proof of Theorem 3

The proof of Theorem 3 follows a structure similar to that of the proof of Theorem 2. We first need to build bounds on (i) moments; (ii) transient bias of the queueing data; (iii) variance of the queueing data; (iv) and FD approximation error of the gradient in terms of the parameter h which corresponds to Lemmas EC.10 to EC.13. Based on the results, we could bound the bias and variance of our gradient estimator in Lemma EC.14 and the order of strong-convexity coefficient in Lemma EC.15. Then, following the regret decomposition in the main paper, we bound the regret of suboptimality, nonstationary and finite difference in Lemmas EC.16 to EC.18, which complete the proof of Theorem 3.

For M/M/1 queue with unit service rate, the mean stationary workload is equal to mean stationary queueing length (including the customer in service). So, one could estimate the objective function using the observed queue length data, and hence, entirely eliminate the bias of delayed observation. In the following analysis, we use $Q_l^h(t)$ and p_l^h to denote the observed queueing length and control price, respectively, in cycle l when applying LiQUAR to the h-th system.

In addition, when applying LiQUAR to the h-th system, we denote the gradient estimator in iteration k as

$$H_k^h = \frac{1}{2\delta_k^h}\left[-p_{2k-1}^h\frac{N_{2k-1}^h}{T_k} + p_{2k}^h\frac{N_{2k}^h}{T_k^h} + h\int_{\alpha T_k^h}^{T_k^h} Q_{2k-1}^h(t) - Q_{2k}^h(t)dt\right]$$

and the corresponding finite difference

$$\frac{f_h(p_{2k-1}^h) - f_h(p_{2k+1}^h)}{2\delta_k^h} \equiv Df_h(\bar{p}_k^h),$$

where

$$p_{2k-1}^h = \bar{p}_k^h + \delta_k^h, \quad p_{2k}^h = \bar{p}_k^h - \delta_k^h.$$

Following the main paper, we define the bias and variance of the gradient estimator as

$$B_k^h \equiv \mathbb{E}[(\mathbb{E}[H_k^h - f'(\bar{p}_k^h)|\mathcal{F}_k])^2], \quad \mathcal{V}_k^h \equiv \mathbb{E}[(H_k^h)^2].$$

For the simplicity of notation, we will denote all positive constants that are independent of h and T_0 by C in the following analysis.

Lemma EC.10 **(Moment Bounds)**. *Under any control sequence p_l^h,*

$$\mathbb{E}\left[(Q_l^h(t))^m\right] \le Ch^{-m/2}, \text{ for all } l \ge 1 \text{ and } t \in [0, T_k].$$

Proof of Lemma EC.10 Let $\tilde{Q}_h(\cdot)$ be the stationary queue length process of an $M/M/1$ queue with service rate 1 and arrival rate $\lambda(p^* + c_1\sqrt{h})$. Then, for arbitrary control sequence p_l^h, we have

$$Q_l^h(t) \le_{st} \tilde{Q}_h(t),$$

for all $t \geq 0$. Therefore, it is sufficient to show that

$$\mathbb{E}[\tilde{Q}_h(t)^m] \leq Ch^{-m/2}$$

for some $C > 0$ and $1 \leq m \leq 4$. By Taylor expansion, $\lambda(p^* + c_1\sqrt{h}) = 1 + \lambda'(p^* + \theta c_1\sqrt{h})c_1\sqrt{h}$ with some $\theta \in (0,1)$, so the corresponding traffic intensity satisfies

$$1 - \rho = -\lambda'(p^* + \theta c_1\sqrt{h})c_1\sqrt{h} \leq c_1 \cdot C_0\sqrt{h},$$

with $C_0 = -\arg\min_{p\in\mathcal{B}_1} \lambda'(p)$. Then, by the stationary distribution of $M/M/1$ queue, the moment bounds are valid. □

Lemma EC.11 **(Transient Bias Bound)**. *Suppose* $\bar{Q}_l^h(\cdot)$ *is a stationary queue length process synchronously coupled with* $Q_l^h(\cdot)$*. Then, conditional on their initial values,*

$$\mathbb{E}[|Q_l^h(t) - \bar{Q}_l^h(t)||Q_l^h(0), \bar{Q}_l^h(0)] \leq |Q_l^h(0)^2 - \bar{Q}_l^h(0)^2| \cdot \frac{2Ct^{-3/2}}{h} \exp(-ht/2C).$$

Proof of Lemma EC.11 Consider an M/M/1 queue with traffic intensity ρ and i customers in the system at time 0. Let τ be the first hitting time when the system gets empty. Following theorem 3.1 in Abate and Whitt (1988b),

$$P((1-\rho)^2\tau > t) = \int_t^\infty f(s; i, 0)ds,$$

with

$$f(t; i, 0) = (i/t)\rho^{1/2}\exp(-2t/(1+\sqrt{\rho})^2)\exp(-4\rho^{1/2}t/(1-\rho)^2)I_i(4\rho^{1/2}t/(1-\rho)^2).$$

Here $I_i(x)$ is the modified Bessel function of the first kind such that $I_i(x) \leq I_0(x)$ for any integer $i \geq 0$. By Olivares et al. (2018), for all $x > 0$,

$$I_0(x) \leq 1.006 \cdot \frac{e^x + e^{-x}}{2(1+x^2/4)^{1/4}} \frac{1+0.24273x^2}{1+0.43023x^2} \leq 1.006 \cdot \frac{e^x}{(1+x^2/4)^{1/4}} \leq 1.006 \cdot e^x \cdot (1 \wedge \sqrt{2/x}).$$

We bound $f(t; i, 0)$ by

$$f(t; i, 0) \leq 1.006 \cdot (i/t)\exp(-t/2) \cdot \left(1 \wedge (1-\rho)\sqrt{1/t}\right),$$

if $\rho > 1/4$. Therefore, for $t \geq 1$,

$$\begin{aligned}
P(\tau > t) &= P\left((1-\rho)^2\tau > (1-\rho)^2 t\right) = \int_{(1-\rho)^2 t}^\infty f(s; i, 0)ds \\
&\leq \int_{(1-\rho)^2 t}^\infty 1.006 \cdot \frac{i}{s}\exp(-s/2)(1-\rho)\sqrt{1/s}\ ds \\
&\leq 2.012(1-\rho)is^{-3/2}\exp(-s/2)|_{s=(1-\rho)^2 t} \\
&= \frac{2.012i}{(1-\rho)^2}t^{-3/2}\exp(-(1-\rho)^2 t/2)
\end{aligned}$$

The last inequality comes from integral by part. Suppose we synchronously couple an M/M/1 queue length process $Q(t)$ with a stationary one $\bar{Q}(t)$ and denote by $\bar{\tau}$ the first hitting time to 0 of $\bar{Q}(t)$. Then, we have

$$\begin{aligned} E[|Q(t)-\bar{Q}(t)||Q(0)=i] &\le E[|i-\bar{Q}(0)|1(\tau\vee\bar{\tau}>t)] \\ &\le E\left[\frac{2.012|i-\bar{Q}(0)|(i+\bar{Q}(0))}{(1-\rho)^2}t^{-3/2}\exp(-(1-\rho)^2t/2)\right] \\ &\le E[|i^2-\bar{Q}(0)^2|]\cdot\frac{2.012}{(1-\rho)^2}t^{-3/2}\exp(-(1-\rho)^2t/2). \end{aligned}$$

Note that for $p\in\mathcal{B}_h$, $1-\rho=O(\sqrt{h})$. Then, setting $Q(t),\bar{Q}(t)$ being the $Q_l^h(t),\bar{Q}_l^h(t)$ closes the proof. □

Lemma EC.12 **(Variance Bound)**. *For all h and l, the stationary queue satisfies*

$$Var\left[\int_0^T \bar{Q}_l^h(t)ds\right]\le\frac{CT}{h^2}.$$

Proof of Lemma EC.12 Let $c_q(t)=corr(\bar{Q}_l^h(0),\ \bar{Q}_l^h(2t/(1-\rho)^2)\ )$, with $\rho=1-\lambda(p_l^h)$ and thus $1-\rho\ge C\sqrt{h}$. According to corollary 5 of Abate and Whitt (1988a),

$$\int_0^\infty c_q(t)dt=\frac{1+\rho}{2}\le 1.$$

Consequently, we have

$$\int_0^\infty Cov(\bar{Q}_l^h(0),\bar{Q}_l^h(2t/(1-\rho)^2))dt\le\mathbb{E}[\bar{Q}_l^h(0)]^2=\frac{\rho(1+\rho)}{(1-\rho)^2}\le\frac{C}{h}.$$

By changing of variables, we would see

$$\int_0^\infty Cov(\bar{Q}_l^h(0),\bar{Q}_l^h(t))dt\le\frac{C}{h^2}.$$

Now, we have

$$\begin{aligned} Var\left[\int_0^T \bar{Q}_l^h(t)ds\right] &= \int_0^T\int_0^T Cov(\bar{Q}_l^h(t),\bar{Q}_l^h(s))dtds \\ &\le 2\int_0^T\int_0^\infty Cov(\bar{Q}_l^h(t),\bar{Q}_l^h(t+s))dsdt\le\frac{CT}{h^2}. \end{aligned}$$

□

Lemma EC.13 **(FD Approximation Error Bound)**.

$$|Df_h(p_k^h)-f_h'(p_k^h)|\le Ck^{-2/3}.$$

Proof of Lemma EC.13 For fixed h, $p \in \mathcal{B}_h$ and $\delta > 0$,

$$f_h(p+\delta) = f_h(p) + \delta f_h'(p) + \frac{\delta^2}{2} f_h''(p) + \frac{\delta^3}{6} f_h'''(p_1)$$
$$f_h(p-\delta) = f_h(p) - \delta f_h'(p) + \frac{\delta^2}{2} f_h''(p) - \frac{\delta^3}{6} f_h'''(p_2)$$

Therefore,

$$\frac{f_h(p+\delta) - f_h(p-\delta)}{2\delta} = f_h'(p) + \frac{\delta^2 f_h'''(p_3)}{6}.$$

Note that

$$f_h'''(p) = 3\lambda''(p) + p\lambda'''(p) - \frac{6h\lambda'(p)^3}{(1-\lambda(p))^4} - \frac{6h\lambda''(p)\lambda(p)}{(1-\lambda(p))^3} - \frac{h\lambda'''(p)}{(1-\lambda(p))^2}.$$

As $1 - \lambda(p) = O(\sqrt{h})$, we can conclude that

$$f_h'''(p) = O(h^{-1}).$$

As $\delta = O(\sqrt{h} k^{-1/3})$, we conclude that the FD approximation error is of order $O(k^{-2/3})$. □

Lemma EC.14 **(Bounds on Gradient Estimator Bias and Variance)**. *For all h and k,*

$$B_k^h \le C \cdot k^{-2/3}, \quad \mathcal{V}_k^h \le C.$$

Proof of Lemma EC.14 We first prove the bias term and then we prove the variance term.

Bias term By definition, the bias is defined by

$$(B_k^h)^2 = \mathbb{E}[(\mathbb{E}[H_k^h - f'(\bar{p}_k^h)|\mathcal{F}_k])^2] \le 2\mathbb{E}[\mathbb{E}[f_h'(p_k^h) - Df_h(p_k^h)|\mathcal{F}_k]^2] + 2\mathbb{E}[\mathbb{E}[H_k^h - Df_h(p_k^h)|\mathcal{F}_k]^2].$$

By Lemma EC.13, we have following bound for the first term.

$$\mathbb{E}[\mathbb{E}[f_h'(p_k^h) - Df_h(p_k^h)|\mathcal{F}_k]^2] \le Ck^{-4/3}.$$

We next bound the second term. By Lemma EC.11, we have

$$\mathbb{E}[|Q_l^h(t) - \bar{Q}_l^h(t)||Q_l^h(0), \bar{Q}_l^h(0)] \le |Q_l^h(0)^2 - \bar{Q}_l^h(0)^2| \cdot \frac{2Ct^{-3/2}}{h} \exp(-ht/2C).$$

Consequently, we have

$$\begin{aligned}\mathbb{E}[\hat{f}_h(p_l^h) - f_h(p_l^h)|\mathcal{G}_l] &= \frac{h}{(1-\alpha)T_k^h} \mathbb{E}\left[\int_{\alpha T_k^h}^{T_k^h} Q_l^h(t) - \bar{Q}_l^h(t) dt \Big| \mathcal{G}_l\right] \\ &\le \frac{|Q_l^h(0)^2 - \bar{Q}_l^h(0)^2|}{(1-\alpha)T_k^h} \cdot \int_{\alpha T_k^h}^{T_k^h} 2Ct^{-3/2} \exp(-ht/2C) dt \\ &\le \frac{C|Q_l^h(0)^2 - \bar{Q}_l^h(0)^2|}{\alpha^{3/2}(T_k^h)^{3/2}} \exp(-\alpha h T_k^h/2C),\end{aligned}$$

where the last inequality holds due to the monotonicity of $t^{-3/2}\exp(-ht/2C)$. Therefore, by our choice of T_k^h, δ_k^h, we have

$$\begin{aligned}\mathbb{E}[H_k^h - Df_h(p_k^h)|\mathcal{F}_k] = &\frac{C\mathbb{E}[Q_l^h(0)^2 - \bar{Q}_l^h(0)^2|\mathcal{F}_k]}{\delta_k (T_k^h)^{3/2}}\exp(-\alpha h T_k^h/2C)\\ \le & C\frac{h^{-1}}{\sqrt{h}k^{-1/3}h^{-3/2}k^{1/2}}\exp(-\alpha k^{1/3}/2C) \le C\cdot k^{-2/3},\end{aligned}$$

for sufficient large k. This closes the proof of Bias.

Variance Term For the variance term, we have

$$\mathbb{E}[H_k^2] \le 3(\delta_k^h)^{-2}\sum_{l=2k-1}^{2k}\mathbb{E}[\hat{f}_h(p_l^h) - f_h(p_l^h)^2] + 3(\delta_k^h)^{-2}\mathbb{E}[f_h(p_{2k}^h) - f_h(p_{2k-1}^h)^2].$$

For the second term, we calculate that for $p \in \mathcal{B}_h$,

$$f_h'(p) = -p\lambda'(p) - \lambda(p) + h\frac{\lambda'(p)}{(1-\rho(p))^2} = O(1).$$

Consequently, we have

$$(\delta_k^h)^{-2}\mathbb{E}[f_h(p_{2k}^h) - f_h(p_{2k-1}^h)^2] \le \max_{p\in\mathcal{B}_h}\|f_h'(p)\| = O(1).$$

For the first term, we have

$$\mathbb{E}[(\hat{f}_h(p_l^h) - f_h(p_l^h))^2] \le 2\mathbb{E}\left[\left(p_l^h\frac{N_l}{T_k} - p_l^h\lambda(p_l^h)\right)^2\right] + 2\frac{h^2}{((1-\alpha)T_k^h)^2}\mathbb{E}\left[\left(\int_{\alpha T_k^h}^{T_k^h} Q_l^h(t) - \mathbb{E}[\bar{Q}_l^h(t)]\right)^2\right].$$

Let's denote $\bar{Q}_l^h(t)$ as a stationary version of queueing process synchronously coupled with $Q_l^h(t)$, and define $\tau, \bar{\tau}$ as the first hitting time of them to the empty states. Note that

$$\begin{aligned}&\mathbb{E}\left[\left(\int_{\alpha T_k^h}^{T_k^h} Q_l^h(t) - \mathbb{E}[\bar{Q}_l^h(t)]dt\right)^2\right]\\ \le&\mathbb{E}\left[\left(\int_{\alpha T_k^h}^{T_k^h} \bar{Q}_l^h(t) - \mathbb{E}[\bar{Q}_l^h(t)]dt\right)^2\right] + \mathbb{E}\left[\left(\int_{\alpha T_k^h}^{T_k^h} Q_l^h(t) - \mathbb{E}[\bar{Q}_l^h(t)]dt\right)^2 \mathbf{1}(\tau\vee\bar{\tau} > \alpha T_k^h)\right]\\ \overset{(a)}{\le}&\frac{C(1-\alpha)T_k^h}{h^2} + (1-\alpha)T_k^h\mathbb{E}\left[\int_{\alpha T_k^h}^{T_k^h}(Q_l^h(t) - \mathbb{E}[\bar{Q}_l^h](t))^2 dt\mathbf{1}(\tau\vee\bar{\tau} > \alpha T_k^h)\right]\\ \le&\frac{C(1-\alpha)T_k^h}{h^2} + CT_k^h\cdot\frac{T_k^h}{h}\cdot\mathbb{P}(\tau\vee\bar{\tau} > \alpha T_k^h)^{1/2}\\ \le&\frac{C(1-\alpha)T_k^h}{h^2} + \frac{C}{h^2}T_k^h\cdot hT_k^h\cdot\frac{\sqrt{h}\mathbb{E}[Q_l^h(0) + \bar{Q}_l^h(0)]}{(hT_k^h)^{3/2}}e^{-hT_k^h/2C}\\ \le&\frac{C}{h^2}T_k^h.\end{aligned}$$

Here, the inequality (a) comes from Lemma EC.12 and the Cauchy-Schwartz inequality, and the last inequality comes from the fact that $hT_k^h \to \infty$ and $\sqrt{h}\mathbb{E}[Q_l^h(0)+\bar{Q}_l^h(0)] = O(1)$. Consequently, we have

$$\mathbb{E}[(\hat{f}_h(p_l^h) - f_h(p_l^h))^2] \le \frac{C}{T_k},$$

for some C large enough. Therefore, we have

$$\mathbb{E}[H_k^2] \le \max\left(\frac{C}{T_k^h \delta_k^2}, C\right) = C.$$

□

Lemma EC.15 **(Convexity)**. *There exists a constant $K_0 > 0$ independent of h such that, for all $p \in \mathcal{B}_h$,*

$$f_h''(p) > h^{-1/2} K_0.$$

Proof of Lemma EC.15 Note that for all $p \in \mathcal{B}_h$, the traffic intensity $1 - \rho(p) = O(1/\sqrt{h})$. Then, by direct calculation and Polleczk-Khinchine formula, we have

$$f_h''(p) = (-p\lambda(p))'' + \frac{h}{(1-\rho(p))^3}\left(2(\lambda'(p)^2 + (1-\rho(p))\lambda''(p))\right) > h^{-1/2} K_0,$$

with $K_0 = 2\min_{p\in\mathcal{B}_1} 2|\lambda'(p)|^2$. □

Given Lemmas EC.14 and EC.15, we are ready to provide an upper bound on the L_2 distance $\mathbb{E}[(\bar{p}_k^h - p_h^*)^2]$ following the analysis of main paper.

Lemma EC.16 **(Suboptimal Regret)**. *The suboptimal regret could be bounded by*

$$R_1^h(L) \le C \cdot \frac{L^{2/3}}{\sqrt{h}}.$$

Proof of Lemma EC.16 For all $h > 0$ and $k \ge 1$, we denote

$$b_k^h \equiv h^{-1}(\bar{p}_k^h - p_h)^2.$$

For a given h small enough, we omit the superscript h for the simplicity of notation and obtain

$$\begin{aligned}
hb_{k+1} &= \mathbb{E}[(\bar{p}_{k+1} - p^*)^2] \le \mathbb{E}[(\bar{p}_k - p^* - \eta_k H_k)] \\
&= \mathbb{E}[(\bar{p}_k - p^*)^2 - 2\eta_k f'(\bar{p}_k)\cdot(\bar{p}_k - p^*)] - 2\eta_k \mathbb{E}[(H_k - f'(\bar{p}_k))\cdot(\bar{p}_k - p^*)] + 2\eta_k^2 \mathbb{E}[H_k^2] \\
&\le (1 - 2\eta_k h^{-1/2} K_0)\mathbb{E}[(\bar{p}_k - p^*)^2] + \sqrt{h}\eta_k B_k(1 + b_k) + 2\eta_k^2 V_k \\
&= (1 - 2c_\eta K_0 k^{-1}) hb_k + hc_\eta k^{-1} B_k + hc_\eta k^{-1} B_k b_k + 2hc_\eta^2 V_k \\
&\le h \cdot \left[(1 - 2c_\eta K_0 k^{-1}) b_k + Ck^{-5/3} + Ck^{-5/3} b_k + Ck^{-2}\right].
\end{aligned}$$

Following the proof of theorem 2 in the main paper, we can prove by induction that, there exists a constant $C > 0$ independent of h such that $b_k \le Ck^{-2/3}$, and therefore, we can conclude

$$\mathbb{E}[(\bar{p}_k^h - p_h^*)^2] \le C \cdot hk^{-2/3}.$$

As a result, we have

$$\begin{aligned} R_1^h(L) =& \sum_{k=1}^{L} \mathbb{E}\left[(f(\bar{p}_k^h) - f(p_h^*))T_k^h\right] \\ \le& \sum_{k=1}^{L} \mathbb{E}\left[\nabla^2 f(p^* + c_1\sqrt{h})(\bar{p}_k^h - p_h^*)^2 T_k^h\right] \\ \le& \sum_{k=1}^{L} C\sqrt{h}k^{-2/3}T_k^h \le C \cdot \frac{L^{2/3}}{\sqrt{h}} \end{aligned}$$

□

Lemma EC.17 **(Non-stationary Regret)**. *The non-stationary regret could be bounded by*

$$R_2^h(L) \le C \cdot \frac{L^{2/3}\log(L)}{\sqrt{h}}.$$

Proof of Lemma EC.17 Following the decomposition of non-stationary regret in the main paper, we have

$$\begin{aligned} R_{2k}^h =& \sum_{l=2k-1}^{2k} h\mathbb{E}\left[\int_0^{T_k^h} Q_l^h(t) - \bar{Q}_l^h(t)dt\right] \\ =& \sum_{l=2k-1}^{2k} h\mathbb{E}\left[\int_0^{t_k^h} Q_l^h(t) - \bar{Q}_l^h(t)dt\right] + h\mathbb{E}\left[\int_{t_k^h}^{T_k^h} Q_l^h(t) - \bar{Q}_l^h(t)dt\right], \end{aligned}$$

with $t_k^h = \frac{2\log k}{h}$. In this way, we following the similar analysis in our main paper. For the second term, by Lemma EC.11, we have

$$h\int_{t_k^h}^{T_k^h} \mathbb{E}[|Q_l^h(t) - \bar{Q}_l^h(t)|]dt \le \int_{t_k^h}^{\infty} \frac{C\mathbb{E}[|Q_l^h(0)^2 - \bar{Q}_l^h(0)|^2]}{ht_k^{3/2}} \exp(-ht/2C)dht \overset{(b)}{\le} \frac{C}{h^2 t_k^{3/2}} \cdot k^{-1} \le \frac{C}{\sqrt{h}k}. \tag{EC.11}$$

The inequality (b) comes from the fact that $\mathbb{E}[Q_l^h(0)^2], \mathbb{E}[\bar{Q}_l^h(0)^2] = O(h^{-1})$. For the first term, we decompose $\mathbb{E}[Q_l^h(t) - Q_l^h(t)]$ into $\mathbb{E}[\bar{Q}_{l-1}^h(t) - \bar{Q}_l^h(t)]$ and $\mathbb{E}[Q_l^h(t) - \bar{Q}_{l-1}(t)]$ as we did in the main paper. By Polleczk-Khinchine formula, we have

$$\mathbb{E}[\bar{Q}_{l-1}^h(t) - \bar{Q}_l^h(t)] = \mathbb{E}\left[\frac{\lambda(p_l^h) - \lambda(p_{l-1}^h)}{(1 - \lambda(p_l^h))(1 - \lambda'(p_{l-1}^h))}\right] \le \frac{C}{h}\mathbb{E}[|p_l^h - p_{l-1}^h| \le \frac{C}{h}\mathbb{E}[|p_l^h - p_{l-1}^h|^2]^{1/2}.$$

Next, following the same argument in Lemma 7 in the main paper, we define $\tilde{Q}_l^h(\cdot)$ and $\tilde{X}_l^h(\cdot)$ as the queue length and busy period process with arrival rate 1 and service rate $1/\lambda(p_l^h)$. Then, by the same analysis in Lemma 7 in the main paper,

$$
\begin{aligned}
&\int_0^{t_k^h} \mathbb{E}[Q_l^h(t) - \bar{Q}_{l-1}^h(t)]dt \\
\leq & \frac{1}{\lambda_l}\int_0^{\lambda_{l-1}t_k^h} \mathbb{E}[|\tilde{Q}_l^h(t) - \tilde{Q}_{l-1}(T_{k(l-1)}^h + t)dt|] + \mathbb{E}\left[\left|\frac{1}{\lambda_l} - \frac{1}{\lambda_{l-1}}\right| \int_0^{\lambda_{l-1}t_k^h} \tilde{Q}_{l-1}(t)dt\right] + \mathbb{E}\left[\frac{1}{\lambda_l}\int_{\lambda_{l-1}t_k^h}^{\lambda_l t_k^h} \tilde{Q}_l(t)dt\right] \\
\leq & \mathbb{E}\left[\left|\frac{1}{\lambda(p_l^h)} - \frac{1}{\lambda(p_{l-1}^h)}\right|^2\right]^{1/2} \mathbb{E}[\tilde{X}_l(t)^2]^{1/2} t_k^h + C\mathbb{E}\left[\left|\frac{1}{\lambda(p_l^h)} - \frac{1}{\lambda(p_{l-1}^h)}\right|^2\right]^{1/2} \mathbb{E}[\tilde{Q}_{l-1}^h(t)^2]^{1/2} t_k^h \\
\leq & C\mathbb{E}[|p_l^h - p_{l-1}^h|^2]^{1/2} \frac{t_k^h}{h}
\end{aligned}
$$

Therefore, by Lemma EC.12, we have

$$
h\mathbb{E}\left[\int_0^{t_k^h} Q_l^h(t) - \bar{Q}_l^h(t)dt\right] \leq C \cdot \mathbb{E}[|p_l^h - p_{l-1}^h|^2]^{1/2} t_k^h \leq C t_k^h \cdot \max(\eta_k\sqrt{\mathcal{V}_k}, \delta_k) = C\frac{\log k}{\sqrt{h}k^{1/3}}. \quad \text{(EC.12)}
$$

Combining equations (EC.12) and (EC.11), we have $R_{2k}^h \leq C\frac{\log k}{\sqrt{h}k^{1/3}}$. Therefore, we have

$$
R_2^h(L) \leq C\frac{L^{2/3}\log L}{\sqrt{h}}.
$$

□

Lemma EC.18 **(Finite-Difference Regret)**. *The finite-difference regret could be bounded by*

$$
R_3^h(L) \leq C \cdot \frac{L^{2/3}}{\sqrt{h}}.
$$

Proof of Lemma EC.18 By calculation, we have

$$
R_3^h(L) = \sum_{k=1}^{L} \mathbb{E}\left[(f(\bar{p}_{2k-1}) + f(\bar{p}_{2k}) - 2f(p_h^*))T_k^h\right] \leq C\sum_{k=1}^{L} \frac{1}{\sqrt{h}}(\delta_k^h)^2 T_k^h \leq C \cdot \frac{L^{2/3}}{\sqrt{h}}.
$$

□

Summing up all three regrets, the total regret in the first L cycle is

$$
R^h(L) = R_1^h(L) + R_2^h(L) + R_3^h(L) \leq C\frac{L^{2/3}\log L}{\sqrt{h}}.
$$

Note that the total time that used is $T^h = \frac{T_0}{h} = \frac{L^{4/3}}{h}$, and therefore,

$$
R^h(T_0/h) \leq C\sqrt{h}^{-1}\sqrt{T_0 \log T_0}.
$$

□

EC.2.4. Proof of Proposition 6

We neglect the superscribe h in the following analysis to ease the burden of notation. The proof of Proposition 6 basically follows the proof of proposition 2 in Besbes and Zeevi (2009). Let $\Delta_0 = \max_{p\in\mathcal{B}_h} f^h(p) - f^h(p^*) = O(\sqrt{h})$, and denote p^*_G as the optimal points in the testing pricing grid. The regret can be decomposed according to three sources of cost: exploration cost, stochastic error, and discrete grid cost.

$$\begin{aligned}R^h(T_0) &\le \Delta_0 t_0 + (T-t_0)\mathbb{E}[f(\hat{p}^*) - f(p^*)]\\ &\le \underbrace{\Delta_0 t_0}_{\text{Exploration Cost}} + \mathbb{E}[T\cdot[\underbrace{f(\hat{p}^*) - f(p^*_G)}_{\text{Stochastic Error}} + \underbrace{f(p^*_G) - f(p^*)}_{\text{Discrete Grid Cost}}]]\end{aligned}$$

We treat the first two terms following in the same way as in Besbes and Zeevi (2009). For the third term, we apply second order Taylor expansion (rather than the first order in Besbes and Zeevi (2009)) as $\nabla f(p^*) = 0$ in our problem. Therefore, using the fact that the grid length is at most $|p^*_G - p^*| \le |\mathcal{B}_h|/\kappa = O(\sqrt{h}/\kappa)$, we have

$$\begin{aligned}R^h(T_0) &\le \Delta_0 t_0 + CT\cdot\sqrt{\frac{\kappa\log T}{t_0}} + CT\cdot\frac{\nabla^2 f(\xi)}{2}|p^*_G - p^*|^2\\ &\le C\sqrt{h}t_0 + CT\cdot\sqrt{\frac{\kappa\log T}{t_0}} + C\frac{T}{\sqrt{h}}\cdot\left(\frac{\sqrt{h}}{\kappa}\right)^2.\end{aligned}$$

By optimizing the regret order, we choose

$$t_0 = O\left(\frac{T_0^{5/7}\log(T)^{2/7}}{h}\right), \quad \kappa = O\left(\frac{T_0^{1/7}}{\log T}\right),$$

such that

$$R^h(T_0) = O\left(\frac{T_0^{5/7}\log(T_0/h)^{2/7}}{\sqrt{h}}\right).$$

□

EC.3. Regret Lower Bound

In this part, we prove that, when the demand function is unknown, the worst-case suboptimal regret of any pricing and capacity sizing policy is at least of order $\Omega(\sqrt{T})$ where T is the total time elapsed.

In particular, we construct a specific demand function with a unknown parameter. The proof is then based on the analysis of KL divergence that measures the uncertainty on this unknown parameter. Intuitively, the proof basically says that, on the one hand, if the uncertainty on the parameter is high, the regret is also high because of the uncertainty (Lemma EC.21). On the other hand, it is shown that to reduce uncertainty, a learning cost must be paid (Lemma EC.20). As a consequence, there is a lower bound for the regret caused by the uncertainty of the parameter.

To make the analysis more intuitive, we consider T as an integer and decompose the total time T into T periods with unit period length. We restrict the policy class so that any admissible policy can only change the price and service capacity at the beginning of each period. This simplification is reasonable because changing policy is usually costly for service providers in reality, and this restriction does not lose generality for our intuition in practice. Note that LiQUAR also belongs to this class. We can formally describe the admissible policies as follows. Denote ω_0 as the initial decision (μ_0, p_0) and $\omega_t,\ t \geq 1$, as the arrivals and corresponding job sizes in t-th period and let $\boldsymbol{\omega}_t = (\omega_0, \omega_1, \cdots, \omega_t)$. We denote the corresponding filters as $\{(\Omega_t, \mathcal{F}_t)\}_{t=0}^T$. An admissible policy is defined by a sequence of decision functions $\pi = \{\pi_1, \cdots, \pi_T\},\ \pi_t : \Omega_{t-1} \to \mathbb{R}_+^2$. We denote these non-anticipating policy class as Ψ.

Theorem 4 ***(Theoretic Lower Bound of Regret)*** *There exists a demand function $\lambda(p)$ satisfying Assumption 1 in our main paper and a positive constant C_2 such that for any admissible policy $\pi \in \Psi$ and $T \geq 2$,*

$$R_2(T) \geq C_2\sqrt{T}.$$

We remark that the regret of the nonstationarity could be negative in general and it remains open to provide a full regret lower bound containing regret of nonstationarity and the regret of suboptimality.

Next, we first introduce the demand class and some key properties of problem class $\mathcal{C}$ in Section EC.3.1. Based on these properties, we prove two critical lemmas craving the trade-off between learning cost and uncertainty cost in Section EC.3.2. The lower bound is the direct consequence of these two lemmas.

EC.3.1. Demand Class and Its Properties

We consider a parametric problem class $\mathcal{C}$ where the demands are linear functions with slope z as parameter

$$\lambda(p; z) = 4 - z(p - 5.5), \tag{EC.13}$$

We set $z \in \mathcal{Z} = [0.95, 1.05]$ and $\mathcal{B} = [5, 6.5] \times [5.4, 6]$ and the queueing system is $M/M/1$. Moreover, we set $h_0 = 1$ and $c(\mu) = \mu$. In this case, the objective function is

$$f(\mu, p; z) = -p\lambda(p; z) + \frac{\lambda(p; z)}{\mu - \lambda(p; z)} + \mu.$$

Denote optimal decision under demand $\lambda(p;z)$ as

$$(\mu^*(z), p^*(z)) = \arg\min_{(\mu,p)\in\mathcal{B}} f(\mu,p;z).$$

The corresponding suboptimal regret is

$$R_2(z,\pi,t) = \mathbb{E}^{z,\pi}\left[\sum_{k=1}^{t}\left(f(\mu_k,p_k;z) - f(\mu^*(z),p^*(z);z)\right)T_k\right].$$

In the next lemma, we summarize the key properties of this demand class, which we will use in lower bound analysis.

Lemma EC.19. *The problem instance class $\mathcal{C}$ has the following properties:*

1. ***Uninformative point*** *All demand curves cross an uninformative point, i.e., $\lambda(5.5;z)=4$ for all $z\in\mathcal{Z}$. Moreover, $p^*(1)=5.5$.*
2. ***Strongly convex*** *For any $z\in\mathcal{Z}$, the objective function $f(\mu,p;z)$ is strongly convex. As a result, there exists a constant $K_5>0$, such that*

$$|f(\mu^*(z),p^*(z)) - f(p,\mu;z)| \geq K_5\left((p-p^*(z)^2 + (\mu-\mu^*(z))^2\right)$$

3. ***Uniform stability*** *The system is uniformly stable for all problem instances, i.e.,*

$$\sup_{p,z}\lambda(p;z) = \lambda(5.4;0.95) < \underline{\mu}.$$

4. ***Continuity of demand function*** *The difference between two demand curves can be represented by difference of z and z_0*

$$|\lambda(p;z) - \lambda(p;z_0)| = |(p-5.5)(z-1)|.$$

5. ***Separability between optimal solutions*** *There exists a constant K_1 such that $|p^*(z)'|\geq K_1$ for all $z\in\mathcal{Z}$. Therefore,*

$$|p^*(z) - 5.5| \geq K_6|z-1|.$$

Proof of Lemma EC.19 Properties 1, 3 and 4 are obvious by direct calculation. For property 2, notice that the demands $\lambda(p;z)$ are linear functions and by direct calculation, we have the strongly convexity result. For property 5, by the first-order condition $\nabla f(\mu^*(z),p^*(z)) = 0$, the optimal solution is given by

$$\begin{cases} \mu^*(z) & = \lambda(p^*(z);z) + \sqrt{\lambda(p^*(z);z)} \\ 1 & = (2p^*(z) - 6.5 - 4z^{-1})^2(4 + 5.5z - p^*(z)z). \end{cases}$$

To show property 5, we define an auxiliary function $g(p,z) = (2p - 6.5 + 4z^{-1})^2(4 + 5.5z - pz)$. By direct calculation, there is an $p^*(z) \in [5.4, 6]$ satisfying $g(p^*(z), z) = 1$. In addition, by direct calculation, in our problem instance,

$$\frac{\partial}{\partial p} g(p,z) = [16 + 22z - 6p + 6.5 + 4z^{-1}](2p - 6.5 - 4z^{-1}) > 0,$$
$$\frac{\partial}{\partial z} g(p,z) = 5.5(2p - 6.5 - 4z^{-1})^2 + \frac{8}{z^2}(4 + 5.5z - p)(2p - 6.5 - 4z^{-1}) > 0.$$

Note that

$$\frac{d}{dz} g(p^*(z), z) = \frac{\partial}{\partial p} g(p^*(z), z) p^*(z)' + \frac{\partial}{\partial z} g(p^*(z), z) = 0,$$

which implies that $p^*(z)' < 0$ for all $z \in \mathcal{Z}$. Since $\mathcal{Z}$ is compact, there is a constant $K_6 > 0$ satisfying the statement in this property. This closes the proof. □

According to Lemma EC.19, this problem class has an uninformative point at $p = 5.5$, where all demands cross. It's also the optimal price for $z = 1$. As a consequence, when $z = 1$, the algorithm needs to step away from the uninformative point to learn the demand, which will incur suboptimal cost. On the other hand, if one algorithm performs very well when $z = 1$, it seldom learns any information and thus cannot perform well under other z. The above observations lead to our proof of the lower bound.

EC.3.2. Proof for Lower Bound

We denote $p_0^* = 5.5$ and $z_0 = 1$. We shall introduce two lemmas to describe the trade-off between learning cost and the cost of uncertainty. We use Kullback-Leibler divergence to measure the information gain. Let $\mathbb{P}_t^{\pi,z}$ denote the probability measure of $\boldsymbol{\omega}_t$ under demand $\lambda(p;z)$ with policy π. We measure the knowledge of demand by

$$\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z}).$$

The following lemma craves the learning cost. Denote $\underline{\lambda} \equiv \inf_{z,p} \lambda(p;z) = \lambda(6.5; 1.05) = 2.95$.

Lemma EC.20. *For any $z \in \mathcal{Z}$, $T > 0$ and any piecewise constant policy $\pi \in \Psi$,*

$$\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z}) \le \frac{(z - z_0)^2}{2\underline{\lambda} K_5} R_2(z_0, \pi, T)$$

Proof of Lemma EC.20 We decompose the KL-divergence in T into conditional KL-divergence in each periods. By chain rule of KL divergence,

$$\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z}) = \sum_{t=1}^{T} \mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z} | \boldsymbol{\omega}_{t-1})$$
$$\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z} | \boldsymbol{\omega}_{t-1}) = \int_{\boldsymbol{\omega}_t} \log \left(\frac{d\mathbb{P}_t^{\pi,z_0}(\omega_t | \boldsymbol{\omega}_{t-1})}{d\mathbb{P}_t^{\pi,z}(\omega_t | \boldsymbol{\omega}_{t-1})} \right) d\mathbb{P}_t^{\pi,z_0}(\boldsymbol{\omega}_t)$$

Conditional on $\boldsymbol{\omega}_{t-1}$, the arrivals in cycle t follows Poisson process with rate $\lambda_t^z \equiv \lambda(p_t; z)$ and we denote the density function of individual work load V by $g(\cdot)$. Then, using the conditional density of Poisson arrivals, we have

$$
\begin{aligned}
&\mathcal{K}(\mathbb{P}_t^{\pi,z_0} \| \mathbb{P}_t^{\pi,z} | \boldsymbol{\omega}_{t-1}) \\
=& \int_{\boldsymbol{\omega}_{t-1}} \int_{\omega_t} \log\left(\frac{d\mathbb{P}_t^{\pi,z_0}(\omega_t | \boldsymbol{\omega}_{t-1})}{d\mathbb{P}_t^{\pi,z}(\omega_t | \boldsymbol{\omega}_{t-1})} \right) d\mathbb{P}_t^{\pi,z_0}(\omega_t | \boldsymbol{\omega}_{t-1}) d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) \\
=& \int_{\boldsymbol{\omega}_{t-1}} \sum_{k=0}^{\infty} \int_{v_1,\cdots,v_k} \frac{(\lambda_t^{z_0})^k e^{-\lambda_t^{z_0}}}{k!} \log\left(\frac{(\lambda_t^{z_0})^k \exp(-\lambda_t^{z_0})(k!)^{-1} 1^{-k} \prod_{i=1}^k g(v_i)}{(\lambda_t^{z})^k \exp(-\lambda_t^{z})(k!)^{-1} 1^{-k} \prod_{i=1}^k g(v_i)} \right) dv_1 \cdots dv_k d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) \\
=& \int_{\boldsymbol{\omega}_{t-1}} (\lambda_t^z - \lambda_t^{z_0}) + \lambda_t^{z_0} \log\left(\frac{\lambda_t^{z_0}}{\lambda_t^z} \right) d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) \\
=& \int_{\boldsymbol{\omega}_{t-1}} (\lambda_t^z - \lambda_t^{z_0}) - \lambda_t^{z_0} \log\left(1 + \frac{\lambda_t^z - \lambda_t^{z_0}}{\lambda_t^{z_0}} \right) d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) \\
\overset{(a)}{\leq}& \int_{\boldsymbol{\omega}_{t-1}} \frac{(\lambda_t^z - \lambda_t^{z_0})^2}{2\underline{\lambda}} d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) = \frac{(z - z_0)^2}{2\underline{\lambda}} \int_{\boldsymbol{\omega}_{t-1}} (p_t - p_0^*)^2 d\mathbb{P}_{t-1}^{\pi,z_0} \\
\overset{(b)}{\leq}& \frac{(z - z_0)^2}{2K_5\underline{\lambda}} \mathbb{E}^{\pi,z_0} \left[f(\mu_t, p_t; z_0) - f(u^*(z_0), p^*(z_0); z_0) \right]
\end{aligned}
$$

Here (a) uses the fact that $-\log(1+x) \leq -x + \frac{x^2}{2}$, and (b) uses the strongly convex property (Lemma EC.19) of our problem case. Therefore, summing up all t, we have the result. □

The next lemma describes the cost of uncertainty.

Lemma EC.21. *For any integer $T \geq 1$, set $z_1 = z_0 + K_7 T^{-1/4}$ with some $K_7 \neq 0$. Then, for any policy $\pi \in \Psi$, we have*

$$
R_2(z_0, \pi, T) + R_2(z_1, \pi, T) \geq \frac{K_5 K_6^2 K_7^2}{18} T^{1/2} e^{-\mathcal{K}\left(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1}\right)}.
$$

Lemma EC.21 directly follows lemma 3.4 in Broder and Rusmevichientong (2012), so we omit the proof here. With these two lemmas, we now complete the proof of Theorem 4.

Proof of Theorem 4 Let $z_1 = z_0 + K_7 T^{-1/4}$ and by Lemma EC.20, we have

$$
R_2(z_0, T, \pi) + R_2(z_1, T, \pi) \geq \frac{2\underline{\lambda} K_5}{K_7^2} \sqrt{T} \mathcal{K}(\mathbb{P}_T^{\pi,z_0}, \| \mathbb{P}_T^{\pi,z_1}).
$$

By Lemma EC.21, we also have

$$
R_2(z_0, T, \pi) + R_2(z_1, T, \pi) \geq \frac{K_5 K_6^2 K_7^2}{18} \sqrt{T} e^{-\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1})}.
$$

Therefore, set $C_2 = \frac{1}{4} \min\left\{ \frac{2\underline{\lambda} K_5}{K_7^2}, \frac{K_5 K_6^2 K_7^2}{18} \right\}$ and we have

$$
\begin{aligned}
\max_{z \in \{z_0, z_1\}} R_2(z, \pi, T) \geq& \frac{R_2(z_0, \pi, T) + R_2(z_1, \pi, T)}{2} \\
\geq& \frac{\sqrt{T}}{4} \left(\frac{2\underline{\lambda} K_5}{K_7^2} \mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1}) + \frac{K_5 K_6^2 K_7^2}{18} e^{-\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1})} \right) \\
\geq& C_2 \sqrt{T} \left(\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1}) + e^{-\mathcal{K}(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1})} \right) \\
\geq& C_2 \sqrt{T}
\end{aligned}
$$

The last inequality is because $x + e^{-x} \geq 1$ for all x. This finishes the proof of the lower bound. □

EC.3.3. Lower Bound in Heavy-Traffic

In this section, we further extend the regret lower bound to the heavy-traffic case in the regime of Section 5.3. Specifically, we show that for sufficiently large T_0 and sufficiently small h_0, for any algorithm π, there is a problem instance such that the suboptimal regret of the algorithm is at least in the order of $\Omega(\sqrt{T_0/h})$, which is consistent with our regret upper bound in heavy-traffic.

Theorem 5 ***(Theoretic Lower Bound of Regret in Heavy-traffic)*** *For any algorithm and any sufficiently small h, there exists a demand function $\lambda_h(p)$ satisfying Assumptions 1 to 3 and a positive constant C_3 such that*

$$R_1^h(T_0) \geq C_3\sqrt{T_0/h}.$$

The proof technique almost follows the regret lower bound with $h = 1$ (base traffic case) with two lemmas measuring the cost of learning cost and the cost of uncertainty. Following the structure of the lower bound with $h = 1$, we first describe the problem instances $\mathcal{C}_h$ for the h-th system as follows.

Let $p_0 \equiv 2$ and $z_0 \equiv 1$. For the h-th system, denote $p_h^*(z_0)$ as the unique solution of the equation

$$(p-2)^2(2p-3) = h \tag{EC.14}$$

such that $p > p_0 = 2$. Then, the demand function for the h-th system is defined as

$$\lambda_h(p;z) \equiv 3 - p_h^*(z_0) - z(p - p_h^*(z_0)), \quad z \in \mathcal{Z} \equiv [0.95, 1.00].$$

The objective function at the h-th system with $\lambda_h(p;z)$ is defined as

$$f_h(p;z) = -p\lambda_h(p;z) + h \cdot \frac{\lambda_h(p;z)}{1-\lambda_h(p;z)},$$

and with slight abuse of notation, the optimal solution is determined denoted by

$$p_h^*(z) \equiv \arg\min_{p > p_0} f_h(p;z).$$

It could be verified that $\lambda(p;z_0) = 3 - p$ and $p_h^*(z_0)$ is indeed the unique solution of equation (EC.14) for $h \leq 1$. Similar to the proof of lower bound in base traffic case, the regret of suboptimality in our heavy-traffic regime for the h-th system for any policy π with demand $\lambda_h(p;z)$ is given by

$$R_1^h(z,\pi,T_0) \equiv R_1(z,\pi,T_0^h) = \mathbb{E}^{z,\pi}\left[\sum_{t=1}^{T_0^h} f_h(p_k;z) - f_h(p_h^*(z);z)\right].$$

Following the structure of the regret lower bound in base traffic case, we give the key properties of this demand family class.

LEMMA EC.22. *The problem instance class $\mathcal{C}_h$ has the following properties:*

1. ***Uninformative point.*** *All demand curves cross an uninformative point, i.e., $\lambda(p_h^*(z_0);z) = 3 - p_h^*(z_0) = \lambda(p_h^*(z_0);z_0)$ for all $z \in \mathcal{Z}$. Moreover, $p_h^*(z_0) = 2 + \sqrt{h} + o(\sqrt{h})$.*
2. ***Strongly convex.*** *For any $z \in \mathcal{Z}$, the objective function $f_h(p;z)$ is strongly convex. In addition, there exists a constant $K_8 > 0$ independent of h, such that*
$$|f_h(p_h^*(z);z) - f_h(p;z)| \geq \frac{K_8}{\sqrt{h}}\left(p - p_h^*(z)^2\right).$$
3. ***Uniform stability.*** *The system is uniformly stable for all problem instances, i.e.,*
$$\sup_{p \in \mathcal{B}_h} \lambda_h(p;z) < 1.$$
4. ***Continuity of demand function.*** *The difference between two demand curves can be represented by difference of z and z_0*
$$|\lambda_h(p;z) - \lambda_h(p;z_0)| = |(p - p_h^*(z_0))(z - z_0)|.$$
5. ***Separability between optimal solutions.*** *There exists a constant K_9 independent of h such that $|p_h^*(z)'| \geq K_9\sqrt{h}$ for all $z \in \mathcal{Z}$. Therefore,*
$$|p_h^*(z) - p_h^*(1)| \geq \sqrt{h}K_9|z - 1|.$$

The proof of Lemma EC.22 exactly follows that of Lemma EC.19 by taking the h into account, and thus, we omit the proof to avoid redundancy.

Similarly, we could provide the following two lemmas describing the cost of learning and the cost of information uncertainty.

LEMMA EC.23. *For any $z \in \mathcal{Z}$, $T_0 > 0$ and any piecewise constant policy $\pi \in \Psi$,*
$$\mathcal{K}(\mathbb{P}_{T_0^h}^{\pi,z_0} \| \mathbb{P}_{T_0^h}^{\pi,z}) \leq \sqrt{h} \cdot \frac{(z - z_0)^2}{2K_8} R_1^h(z_0, \pi, T_0)$$

Proof of Lemma EC.23 Following the same analysis in the proof of Lemma EC.20, we have
$$\mathcal{K}(\mathbb{P}_t^{\pi,z_0} \| \mathbb{P}_t^{\pi,z} | \boldsymbol{\omega}_{t-1}) \leq \int_{\boldsymbol{\omega}_{t-1}} \frac{(\lambda_t^z - \lambda_t^{z_0})^2}{2\underline{\lambda}} d\mathbb{P}_{t-1}^{\pi,z_0}(\boldsymbol{\omega}_{t-1}) = \frac{(z - z_0)^2}{2\underline{\lambda}} \int_{\boldsymbol{\omega}_{t-1}} (p_t - p_0^*)^2 d\mathbb{P}_{t-1}^{\pi,z_0}$$
$$\leq \sqrt{h}\frac{(z - z_0)^2}{2K_8\underline{\lambda}} \mathbb{E}^{\pi,z_0}\left[f_h(p_t;z_0) - f_h(p^*(z_0);z_0)\right].$$

Summing up all t and by the fact that $\lambda(p;z) > \mu = 1$, we have the result. □

LEMMA EC.24. *For any $T_0 \geq 1$, set $z_1 = z_0 - K_{10}T_0^{-1/4}$ with $K_{10} = 0.05$. Then, for any policy $\pi \in \Psi$, we have*
$$R_1^h(z_0, \pi, T_0) + R_1^h(z_1, \pi, T_0) \geq \frac{1}{\sqrt{h}} \frac{K_8 K_9^2 K_{10}^2}{18} T_0^{1/2} e^{-\mathcal{K}\left(\mathbb{P}_T^{\pi,z_0} \| \mathbb{P}_T^{\pi,z_1}\right)}.$$

The proof of Lemma EC.24 is exactly the same as Lemma EC.21 by replacing K_5, K_6, T with $K_8/\sqrt{h}, \sqrt{h}K_9, T_0/h$.

Therefore, combining the Lemma EC.23 and Lemma EC.24, using the same proof of Theorem 4, we have the Theorem 5.

EC.4. Examples of the Demand Function

In this part, we verify that the following two inequalities in Condition (a) of Assumption 1 hold for a variety of commonly-used demand functions.

$$-\lambda'(p) > \max\left(\sqrt{\frac{0\vee(-\lambda''(p)(\bar{\mu}-\lambda(p)))}{2}}\,,\ \frac{p\lambda''(p)}{2}\right), \tag{EC.15}$$

$$\lambda'(p) > \max_{\mu\in[\underline{\mu},\bar{\mu}]}\left(2g(\mu)\frac{\lambda''(p)\lambda(p)}{\lambda'(p)} - \frac{4\lambda(p)(\mu-\lambda(p))}{h_0 C}\right). \tag{EC.16}$$

Example EC.1 (Linear Demand). Consider a linear demand function

$$\lambda(p) = a - bp, \qquad \text{with } 0 < b < \frac{4\underline{\lambda}(\underline{\mu}-\bar{\lambda})}{h_0 C}.$$

Then, inequality (EC.15) holds immediately as $\lambda''(p)\equiv 0$. Inequality (EC.16) is equivalent to

$$-b > -\frac{4\lambda(p)(\underline{\mu}-\lambda(p))}{h_0 C},$$

which also holds as $\lambda(p)(\underline{\mu}-\lambda(p)) \geq \underline{\lambda}(\underline{\mu}-\bar{\lambda})$.

Example EC.2 (Quadratic Demand). Consider a quadratic demand function

$$\lambda(p) = c - ap^2, \qquad \text{with } a, c > 0 \text{ and } 0 < \frac{\bar{\mu}-c}{3\underline{p}^2} < a < \left(\frac{3(\underline{\mu}-\bar{\lambda})\underline{p}}{h_0 C} - \frac{\underline{\mu}}{\underline{\mu}-\bar{\lambda}}\right)\frac{\underline{\lambda}}{\bar{p}^2}.$$

Inequality (EC.15) is equivalent to $3a^2p^2 > a(\bar{\mu}-c)$, which holds as $a > \frac{\bar{\mu}-c}{3\underline{p}^2}$. For inequality (EC.16), note that $\lambda'' = -2a$ and $\lambda' = -2ap$. So, for any $\mu\in[\underline{\mu},\bar{\mu}]$, we have

$$\begin{aligned}
&\lambda'(p) - 2g(\mu)\frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu-\lambda(p))}{h_0 C}\\
=&-2ap - 2\left(\frac{\mu}{\mu-\lambda} - \frac{(\mu-\lambda)p}{h_0 C}\right)\frac{\lambda}{p} + \frac{4\lambda(\mu-\lambda)}{h_0 C}\\
=&2p\left(\frac{\lambda}{p^2}\left(\frac{3(\mu-\lambda)p}{h_0 C} - \frac{\mu}{\mu-\lambda}\right) - a\right).
\end{aligned}$$

Note that $\frac{3(\mu-\lambda)p}{h_0C} - \frac{\mu}{\mu-\lambda} > \frac{3(\underline{\mu}-\bar{\lambda})\underline{p}}{h_0C} - \frac{\underline{\mu}}{\underline{\mu}-\bar{\lambda}} > 0$ by our assumption, and consequently,

$$\frac{\lambda}{p^2}\left(\frac{3(\mu-\lambda)p}{h_0 C} - \frac{\mu}{\mu-\lambda}\right) - a > \left(\frac{3(\underline{\mu}-\bar{\lambda})\underline{p}}{h_0 C} - \frac{\underline{\mu}}{\underline{\mu}-\bar{\lambda}}\right)\frac{\underline{\lambda}}{\bar{p}^2} - a > 0,$$

which shows that (EC.16) holds.

Example EC.3 (Exponential Demand). Consider an exponential demand function

$$\lambda(p) = \exp(a - bp), \qquad \text{with } b > 0 \text{ and } b\bar{p} < 2.$$

Then $\lambda'(p) = -b\lambda(p)$ and $\lambda''(p) = b^2\lambda(p) > 0$. Therefore, inequality (EC.15) is automatically satisfied as $b < 2/\bar{p}$. For inequality (EC.16), given that $p \le \bar{p} < 2/b$, we have, for any $\mu \in [\underline{\mu}, \bar{\mu}]$,

$$\begin{aligned}
& \lambda'(p) - 2g(\mu)\frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu - \lambda(p))}{h_0 C} \\
= & -b\lambda(p) - 2\frac{\mu}{\mu - \lambda} \cdot \frac{b^2\lambda^2(p)}{-b\lambda(p)} + \frac{4\lambda(\mu - \lambda) - 2bp\lambda(\mu - \lambda)}{h_0 C} \\
> & -b\lambda(p) + 2\frac{\mu}{\mu - \lambda} b\lambda(p) > b\lambda(p) > 0.
\end{aligned}$$

Therefore, (EC.16) holds as well.

Example EC.4 (Logit Demand). Consider a logit demand function

$$\lambda(p) = c \cdot \exp(a - bp)/(1 + \exp(a - bp)), \qquad \text{with } a - b\bar{p} < \log(1/2) \text{ and } 0 < b < 2/\bar{p}.$$

We have

$$\lambda'(p) = -\frac{b}{1+e}\lambda(p), \ \lambda''(p) = \frac{b^2(1-e)}{(1+e)^2}\lambda(p), \ \text{with } e \equiv \exp(a - bp).$$

As a result, inequality (EC.15) becomes $2 > bp(1-e)/(1+e)$ if $e < 1$. Since $a - bp < \log(1/2)$, $e < 1/2$ and (EC.15) holds accordingly. We next show that (EC.16) holds as well. For any $\mu \in [\underline{\mu}, \bar{\mu}]$,

$$\begin{aligned}
& \lambda'(p) - 2g(\mu)\frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu - \lambda(p))}{h_0 C} \\
= & \left(-\frac{b}{1+e} + \frac{2\mu(1-e)b}{(\mu - \lambda)(1+e)} - \frac{2p(\mu - \lambda)}{h_0 C} \cdot \frac{b(1-e)}{1+e} + \frac{4(\mu - \lambda)}{h_0 C}\right) \cdot \lambda \\
> & \left(-\frac{b}{1+e} + \frac{\mu b}{(\mu - \lambda)(1+e)} - \frac{2bp(1-e)}{1+e}\frac{(\mu - \lambda)}{h_0 C} + \frac{4(\mu - \lambda)}{h_0 C}\right) \cdot \lambda > 0,
\end{aligned}$$

where the first inequality holds as $0 < e < 1/2$ and the second inequality holds as long as $b < 2/p$. So (EC.16) holds as well.

EC.5. Additional Numerical Experiments

EC.5.1. Robustness of LiQUAR

In this section, we give more discussion on the robustness of LiQUAR via numerical examples. Specifically, we test the performance of LiQUAR in a set of model settings with different values of optimal traffic intensity ρ^* and service time distributions.

We consider an $M/GI/1$ model with phase-type service-time distribution and the logistic demand function in (20) with $M_0 = 10, a = 4.1$ and $b = 1$. We fix staffing cost coefficient $c_0 = 1$ in (21) in this experiment. By PK formula and PASTA, the service provider's problem reduces to

$$\min_{\mu,p}\left\{f(\mu,p) = -p\lambda(p) + \frac{h_0(1+c_s^2)}{2}\cdot\frac{\lambda(p)/\mu}{1-\lambda(p)/\mu} + \mu\right\},$$

where c_s^2 is SCV of the service time. We investigate the impact on performance of LiQUAR of the following two factors: (i) the optimal traffic intensity ρ^* (which measures the level of heavy traffic), and the service-time SCV c_s^2 (which quantifies the stochastic variability in service and in the overall system).

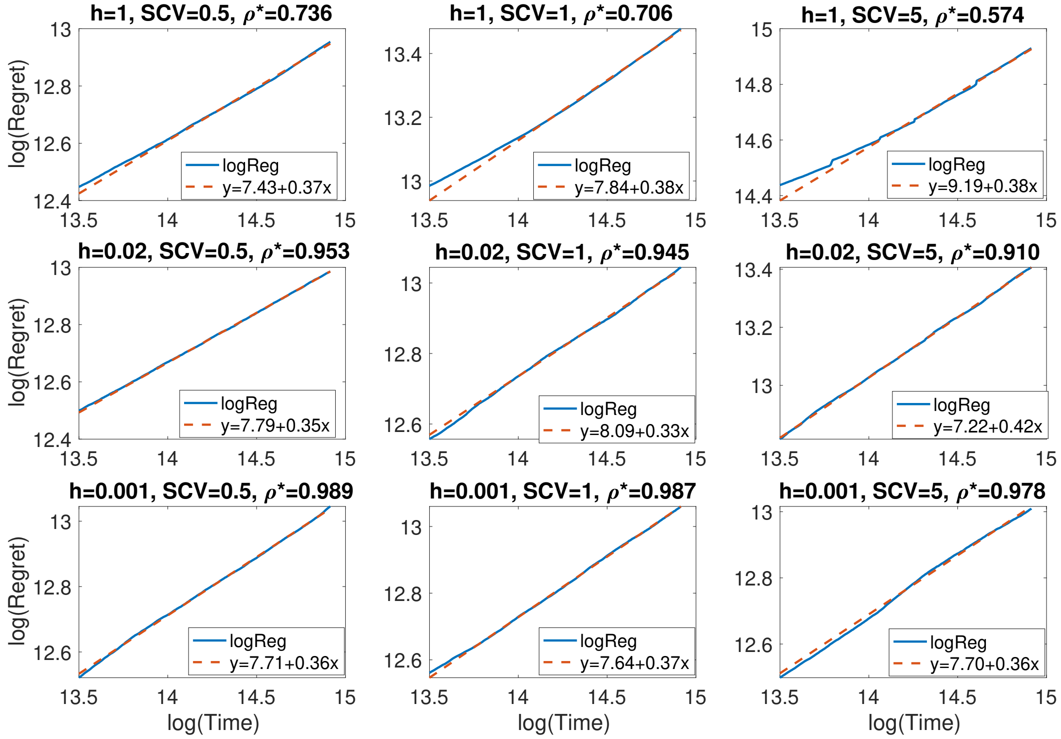


Figure EC.1 The regret curve in logarithm scale and a linear fit for the $M/GI/1$ model, under different traffic intensity $\rho^* \in [0.547, 0.989]$ and service-time SCV $c_s^2 = 0.5$ (E_2 service), 1 (M service) and 5 (H_2 service). All curves are estimated by averaging 100 independent runs.

To obtain different values of ρ^*, we vary the holding cost $h_0 \in \{0.001, 0.02, 1\}$. For the SCV, we consider $c_s^2 = 0.5, 1, 5$ using Erlang-2, exponential and hyperexponential service time distributions. In Figure EC.1 we plot the regret curves in logarithm scale along with their linear fits in all above-mentioned settings. We set $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$ and $\alpha = 0.1$. For all 9 cases, we run LiQUAR for $L = 1000$ iterations and estimate the regret curve by averaging 100 independent runs.

Note that the optimal traffic intensity ρ^* ranges from 0.547 to 0.987. In all the cases, the linear fitted regret curve has a slope below the theoretic bound 0.5, ranging in $[0.35, 0.42]$. Besides, the intercept (which measures the constant term of the regret) does not increase significantly in ρ^* and ranges in $[7.64, 7.79]$ for $\rho^* > 0.95$. The results imply that the performance of LiQUAR is not too sensitive to the traffic intensity ρ^* and service-time SCV.

EC.5.2. Relaxing the Uniform Stable Condition

In this section, we conduct numerical experiments with relaxed uniformly stable condition. We consider a modified version of LiQUAR and test the performance of LiQUAR in our heavy-traffic examples in Section 7. The modified version of LiQUAR is nearly the same as LiQUAR but with an additional early-stop and backtracking step if it finds that the system is too busy. Specifically, the systems have two additional hyperparameters, threshold τ and an anchoring price p_a, under which the system is known to be stable. In the each cycle, the system continues track the observed workload and if the workload of the system is larger than τ, then the system early-stop this cycle and set the next price as the midpoint between the current price and anchor price (backtracking); otherwise the systems works identically with LiQUAR. For more details, please see Algorithm 2.

Then, we test the performance of LiQUAR with backtracking under the setting in Section 7 but without uniformly stable assumption. Specifically, we consider a pricing problem for $M/M/1$ queue having exponential demand function

$$\lambda(p) = \exp(a - bp)$$

with $a = 1 + \log(2)$ and $b = 1$. In addition, we set $h = 0.005$. We set the initial $\bar{p}_0 = 1.55$ so that the initial traffic intensity $\rho_0 = \lambda(p_0)/\mu = 1.15 > 1$ succeeds the critical level one, leading to an unstable system. Following the analysis in Section 5.3, we set the hyperparameters $T_k = 2000k^{1/3}, \delta_k = 0.07k^{-1/3}, \eta_k = 0.21k^{-1}$ and $\tau = 141$ with the anchoring point $p_a = 1.84$. As is shown in the first two panels of EC.2, the pricing policy remain convergent to p^*. Consistently, the resulting traffic intensity ρ_k is quickly controlled to decrease below 1 although the system is unstable in the initial cycles.

Next, we further numerically investigate the impact of the uniformly stable conditions. For this purpose, we consider two scenarios: (i) the service provider does not know a uniformly stable region and use LiQUAR with backtracking; (ii) the service provider knows a uniformly stable region and directly use LiQUAR. Setting the hyperparameters for LiQAUR as $T_k = 2000k^{1/3}, \delta_k = 0.07k^{-1/3}, \eta_k = 0.21k^{-1}$, we draw the regret curves of LiQUAR under two scenarios in the bottom penal of Figure EC.2. From Figure EC.2, we observe that the uniform stable condition does help the convergence of LiQUAR, leading to a smaller regret in the initial cycles. On the other hand,

Algorithm 2: LiQUAR with backtracking

```
Input: number of iterations $L$, threshold $\tau$, anchor price $p_a$;
parameters $0 < \alpha < 1$, and $T_k$, $\eta_k$, $\delta_k$ for $k = 1, 2, .., L$;
initial value $\bar{p}_1$, $Q_1(0) = 0$;
for $k = 1, 2, ..., L$ do
    Set control parameter $p_{2k-1} = \bar{p}_k - \delta_k/2$ and Stable Sign= 0;
    while $\frac{1}{t}\int_0^t Q(t)dt < \tau$ for $t < T_k$ do
        Run Cycle $2k-1$: Run the system under $p_{2k-1}$ ;
    end
    Set control parameter $p_{2k} = \bar{p}_k + \delta_k/2$ ;
    while $\frac{1}{t}\int_0^t Q(t)dt < \tau$ for $t < T_k$ do
        Run Cycle $2k$: Run the system under $p_{2k}$ ;
        Stable Sign=1;
    end
    if Stable Sign=1 then
        Compute FD gradient estimator:
            $H_k = \frac{1}{\delta_k}\left[\frac{h_0}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k} (Q_{2k}(t) - Q_{2k-1}(t))\, dt - \frac{p_{2k}N_{2k} - p_{2k-1}N_{2k-1}}{T_k})\right]$
         Update $\bar{p}_{k+1} = \Pi_{[0,\infty)}(\bar{p}_k - \eta_k H_k)$.
    end
    else
        Backtracking:$\bar{p}_{k+1} = \frac{\bar{p}_k + p_a}{2}$
    end
end
```

after passing through the unstable region, LiQUAR with backtracking has a flat regret growth rate, indicating a fast convergence when it is in stable region, which leads to a comparable regret compared to LiQUAR.

Overall, this experiment is an initial study of how to conduct LiQUAR without using uniformly stable conditions. From the result, we find that it is doable by considering a slightly modified version of LiQUAR. However, to theoretically analyze LiQUAR with backtracking, we need to additionally bound the growth of the regret in the unstable region and deal with the a huge initial workload in each cycle accumulated in the unstable regions, which remains challenging in the current analysis framework. In addition, we also need to take good trade-off in the threshold τ, so that it will not backtracking too frequently and could still detect the unstable region agilely. We leave the further exploration of this direction to future research.

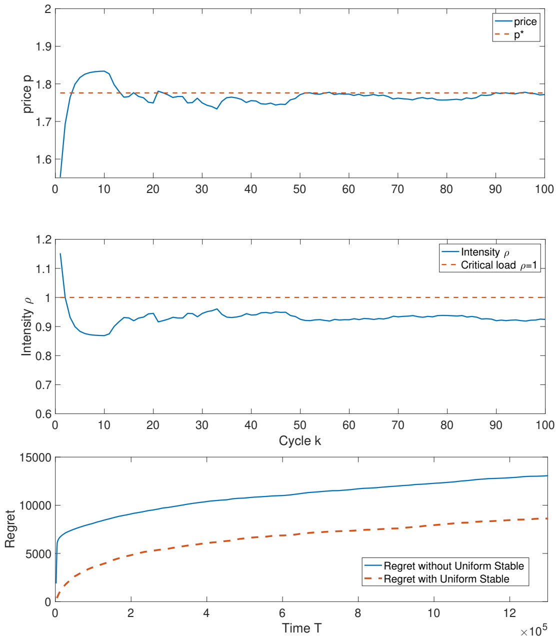


Figure EC.2 Pricing for the $M/M/1$ model without uniform stable condition for LiQUAR with backtracking: (i) sample path of price for LiQUAR with backtracking (top panel); (ii) sample path of traffic intensity ρ for LiQUAR with backtracking; (iii) regret comparison between LiQUAR with backtracking but no uniformly stable condition and LiQUAR with uniformly stable condition. The hyperparameter choices are $T_k = 2000k^{1/3}, \delta_k = 0.07k^{-1/3}, \eta_k = 0.21k^{-1}$ and $\tau = 141$ with the anchoring point $p_a = 1.84$ and $h = 0.005$.

EC.6. Details of PG type Algorithms in Section 8

EC.6.1. Base Policy Gradient

In this section, we provide the detailed description for Policy Gradient algorithms in Algorithm 3, the outline of which is described in Section 8. Specifically, in Algorithm 3, Policy Gradient algorithm organizes time by cycles, with each cycle containing L episodes. In each episode, the system operates the system following π_θ for episode length T time units. At the end of each episode, a gradient gradient estimator in this episode $\hat{\nabla}_{i,t}$ is calculated using the policy gradient formula (Sutton and Barto 2018, p.339) and the closed form of Gaussian parameterization (line 9 in Algorithm 3). Then, at the end of each cycle, an overall policy gradient estimator is obtained by averaging over all the episodic policy gradient estimators in the cycle (line 11 in Algorithm 3). The full algorithm is given in Algorithm 3.

EC.6.2. Policy Gradient with SAGE

In this section, we introduce the policy gradient algorithm with Score-Aware Gradient Estimates (PG-SAGE) in Comte et al. (2023). Following Comte et al. (2023), similar to the policy gradient

Algorithm 3: PG-base

```
Input: normal parameterization $\pi(a|\theta)$, step size $\eta > 0$, initial policy parameter
       $\theta : (\bar{p}_1, \bar{\mu}_1, \sigma^2_{p,1}, \sigma^2_{\mu,1})$, cycle length $L$ (how many episodes in each episode), episode
       length $T$ (how many time slots in each episode);
for each cycle do
    for episode $i = 1 : L$ do
        Generate an episode $Q_1, (p_1, \mu_1), R_1, \cdots, Q_{T-1}, (p_T, \mu_T), R_T$ following $\pi_\theta$;
        $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_T$;
        for $t = 1, \cdots, T$ do
            $G = \sum_{k=t}^{T} R_k - \bar{R}$;
            $\hat{\nabla}_{i,t} \leftarrow G \cdot \begin{pmatrix} (p_t - \bar{p})/\sigma_p^2 \\ (\mu_t - \bar{\mu})/\sigma_\mu^2 \\ \left[(p_t - \bar{p})^2 - \sigma_p^2\right]/\sigma_p^3 \\ \left[(\mu_t - \bar{p})^2 - \sigma_\mu^2\right]/\sigma_\mu^3 \end{pmatrix}$;
        end
        $\hat{\nabla}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{\nabla}_{i,t}$;
    end
    $\theta \leftarrow \theta + \eta \cdot \frac{1}{L}\sum_{i=1}^{L} \hat{\nabla}_i$;
end
```

theorem, $\nabla_\theta h_{\pi_\theta}$ could be represented as

$$\nabla_\theta h_{\pi_\theta} = \mathbb{E}[R_t \nabla \ln \pi(A_t|S_t, \theta)] + \mathbb{E}[R_t \nabla \log P_\infty^\theta(S_t)],$$

where $P_\infty^\theta(S_t)$ is the on-policy steady-state distribution of states S_t under policy π_θ. The key idea of SAGE is that, if the distribution $P_\infty^\theta(s)$ is of exponential family, i.e.,

$$P_\infty^\theta(s) = Z(\theta)^{-1}\Psi(s)\rho(\theta)^{S(s)},$$

for some uniformization function $Z(\theta)$, scalar function $\Psi(s)$, load function $\rho(\theta)$ and sufficient statistics $S(s)$, then the baseline could be represented as follows

$$\mathbb{E}[\nabla \log P_\infty^\theta(s)] = \mathrm{Cov}(R_t, S(S_t)) \cdot \nabla \log \rho(\theta).$$

In our specific case, the on-policy steady-state distribution could be approximated by the steady-state distribution of $M/M/1$ queue with arrival rate $\lambda(\bar{p})$ and service rate $\bar{\mu}$. In this case, the sufficient statistics is queue length $S(s) = s$ and the load function is $\rho(\theta) = \frac{\lambda(\bar{p})}{\bar{\mu}}$. Consequently, we have

$$\nabla \log \rho(\theta) = \left(\frac{\lambda'(\bar{p})}{\lambda(\bar{p})}, -1/\bar{\mu}, 0, 0\right).$$

To apply the PG-SAGE, we need to have information on $\lambda'(\bar{p})$ and $\lambda(\bar{p})$, which is not attainable in our case. Nevertheless, we consider a situation favoring the PG-SAGE where at the end of each cycle, a score oracle would tell the true value of score $\lambda'(\bar{p})/\lambda(\bar{p})$ to PG-SAGE and PG-SAGE is achieved by plugging-in this score. The full algorithm is given in Algorithm 4.

Algorithm 4: PG-SAGE

```
Input: normal parameterization $\pi(a|\theta)$, step size $\eta > 0$, initial policy parameter
       $\theta : (\bar{p}_1, \bar{\mu}_1, \sigma^2_{p,1}, \sigma^2_{\mu,1})$, cycle length $L$ (how many episodes in each episode), episode
       length $T$ (how many time slots in each episode), score oracle: $\mathcal{S}(p) = \lambda'(p)/\lambda(p)$;
for each cycle do
    for episode $i = 1 : L$ do
        Generate an episode $Q_1, (p_1, \mu_1), R_1, \cdots, Q_{T-1}, (p_T, \mu_T), R_T$ following $\pi_\theta$;
        $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_T$;
        for $t = 1, \cdots, T$ do
            $\hat{\nabla}_{i,t} \leftarrow R_t \cdot \begin{pmatrix} (p_t - \bar{p})/\sigma_p^2 \\ (\mu_t - \bar{\mu})/\sigma_\mu^2 \\ \left[(p_t - \bar{p})^2 - \sigma_p^2\right]/\sigma_p^3 \\ \left[(\mu_t - \bar{p})^2 - \sigma_\mu^2\right]/\sigma_\mu^3 \end{pmatrix}$;
        end
        $\hat{\nabla}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{\nabla}_{i,t}$;
        $\hat{\nabla}_i^{SAGE} = \mathrm{Cov}(R_t, S_t, t = 1 : T) \cdot \big(\mathcal{S}(\bar{p}); -1/\bar{\mu}; 0; 0\big)^T$;
    end
    $\theta \leftarrow \theta + \eta \cdot \frac{1}{L}\sum_{i=1}^{L}(\hat{\nabla}_i + \hat{\nabla}_i^{SAGE})$;
end
```

In Figure 10, we report the best regret curves of PG-base and PG-SAGE methods for better exposition. From the Figure 10, we find out that compared with PG-base method, the PG-SAGE with constant step sizes does improve the performance of policy gradient with lower regret. In addition, the PG-SAGE method with time-dependent step sizes seems to have a better convergence in the end (flat regret growth) but has a larger regret in the beginning due to the larger step sizes in the beginning. Overall, LiQUAR outperforms the Policy Gradient methods in this experiments, as LiQUAR's design has carefully taken the structure of queueing systems into consideration.

	Notation	Description
Model parameters and functions	$\mathcal{B} = [\underline{\mu}, \bar{\mu}] \times [\underline{p}, \bar{p}]$	Feasible region
	$c(\mu)$	Staffing cost
	$c_s^2 = Var(S)/\mathbb{E}[S]^2$	Squared coefficient of variation (SCV) of the service times
	$C = \frac{1+c_s^2}{2}$	Variational constant in PK formula
	$f(\mu, p)$	Objective (loss) function
	h_0	Holding cost of workload
	$\lambda(p)$	Underlying demand function
	μ	Service rate
	p	Service fee
	θ, γ_0, η	Parameters of light-tail assumptions (Assumption 2)
	V_n	Individual workload
	$W_\infty(\mu, p)$	Stationary workload under decision (μ, p)
	$\boldsymbol{x}^* = (\mu^*, p^*)$	Optimal decision service rate and fee
Algorithmic parameters and variables	α	Warm-up and overtime rate
	$\delta_k, (\delta_k^h)$	Exploration length in iteration k (of h^{th} system)
	$\eta_k, (\eta_k^h)$	Step length for gradient update in iteration k (of h^{th} system)
	$\boldsymbol{H}_k$	Gradient estimator in iteration k
	$\hat{f}^G(\mu_l, p_l)$	Estimation of objective function in cycle l
	$Q_k^h(t)$	Queue length at time t in cycle k of the h^{th} system
	$T_k, T_{k(l)}, (T_k^h)$	Cycle length of iteration k and cycle l (of h^{th} system)
	$W_l(t)(\hat{W}_l(t))$	(Estimated) workload at time t in cycle l
	$X_l(t)$	Observed busy time at time t in cycle l
	$\bar{\boldsymbol{x}}_k$	Control parameter in iteration k
	$\boldsymbol{Z}_k$	Updating direction in iteration k
Constants and bounds in regret analysis	$B_k, \mathcal{V}_k$	Bias and Variance upper bound for H_k
	c	Constant for noise-free FD error in Lemma EC.4
	c_η, c_T, c_δ	Coefficient of hyperparameters in Theorem 2
	C	Constant in Theorem 3 irrelevant to h
	C_0	Constant in Lemma EC.7
	M	Upper bound for queueing functions in Lemma EC.9
	γ	Ergodicity rate constant in Lemma EC.2
	K_0, K_1	Convex and smoothness constant of objective function in Lemma EC.5
	K_2, K_3	Constants in the proof of Theorem 1 in Appendix EC.2.1
	K_4	Constant in Lemma EC.8
	K_5, K_6, K_7	Constants in Theorem 4 in Section EC.3
	K_V	Constant of auto-correlation in Lemma EC.3
	K_M	Constant of MSE of $\hat{f}^G$ in Proposition 2
	$R(L), R_1(L), R_2(L), R_3(L)$	Total regret, regret of sub-optimality, non-stationarity, finite difference
	θ_0	Constant in Lemma EC.9
	$\theta_1 = \min(\gamma, \theta_0\underline{\mu}/2)$	Constant in Proposition 3
	$\bar{W}_l(t)$	Stationary workload process coupled from the beginning of cycle l
	$\bar{W}_l^s(t)$	Stationary workload process coupled from time s of cycle l (in Appendix)
	$W_l^D(t), X_l^D(t)$	Workload and observed busy time for the dominating queue (in Appendix)

Table EC.1 Glossary of key notations