Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

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

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

Guiyu Hong
School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, guiyuhong@link.cuhk.edu.cn

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

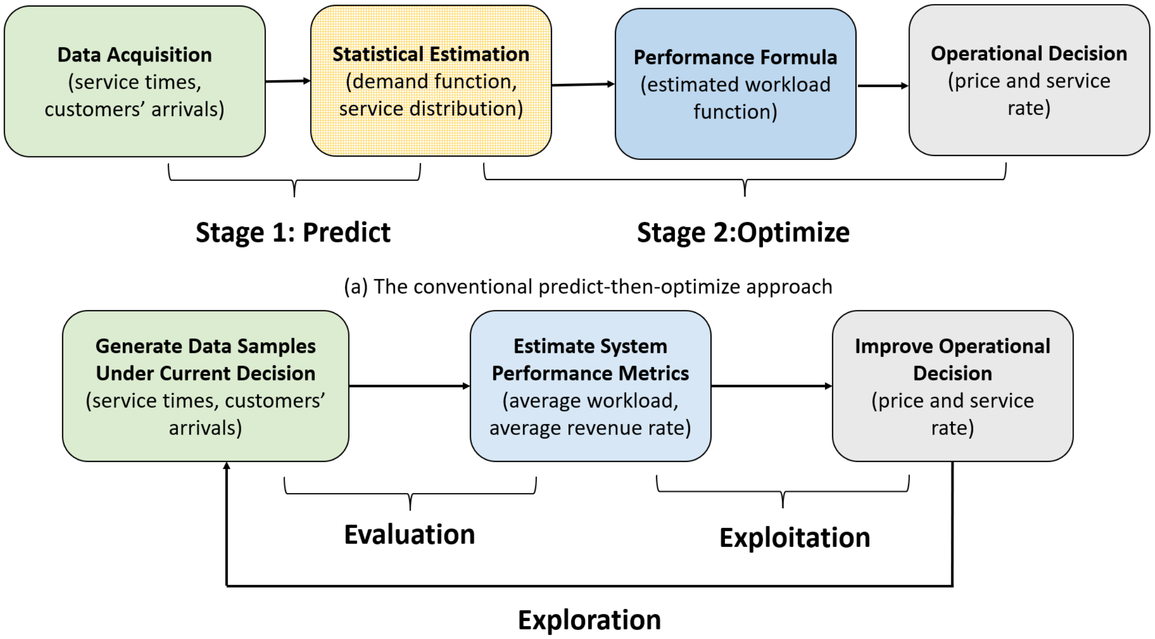
The conventional performance analysis and optimization in queueing systems require the precise knowledge of certain distributional information of the arrival process and service times. For example, consider the $M/GI/1$ queue having Poisson arrivals and general service times, the expected steady-state 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 insightful 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 many 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 so-called *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 main issue of PTO is that the required queueing formulas can be quite sensitive to the estimation errors of the input parameters (e.g., λ and μ), especially when the system's congestion level is high. 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 may place a heavy burden on the "predict" step for obtaining near-perfect estimates of the input parameters. Otherwise, solution methods driven by these seemly convenient formulas may "backfire".

The aforementioned performance deficiency of PTO (especially when the system is in heavy traffic) is attributed to its incapability to correctly account for the parameter estimation error along with the significant impact of this error on the quality of the "optimized" decision variables. To 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 without needing to set up the parameter estimation as a separate step as in PTO.* In this paper, we solve a pricing and capacity sizing problem in an $M/GI/1$ queue, where the service provider seeks the optimal service fee p and service rate μ so as


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.

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. (2020) 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 a good estimate of the function $\lambda(p)$ across the full spectrum of the price p is not straightforward and can be time consuming and costly. In contrast, the online learning approach does not require such information apriori so it is good at "learning from scratch". Second, unlike the two-step PTO procedure, the online learning approach is an *integrated* method that automatically incorporates estimation errors in observed data in the decision prescription process, so it is able to make a better use of the data in order to achieve improved decisions that are more effective and robust.

On the other hand, the online learning methodology in queue systems is by no means a quick extension of online learning in other fields because it needs to account for the unique features in queueing systems. First, as the control policy is updated at the beginning of a cycle, the previously established near steady-state dynamics breaks down and the system undergoes a new transient period of which the dynamics endogenously depends on the control policy. Such an effect gives rise to the so-called *regret of nonstationary.* Next, convergence of the decision iterations heavily relies on statistical efficiency in the evaluation step and properties of queueing data. Towards this, the unique features of queueing dynamics brings new challenges. Unlike the standard online leanring settings (e.g., stochastic bandit), queueing data such as waiting time and queue length are biased, unbounded and sequentially dependent. All the above features that are unique to queueing models bring new challenges to the design and analysis of online learning methodologies in queues.

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. For the 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.
- 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 the traffic intensity ρ^*, and other model and algorithm parameters such as service distributions and updating step sizes. In addition, we make a full-fledged comparison between LiQUAR and PTO under different system loading. We also generalize our algorithm to the $GI/GI/1$ model.

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 6, we conduct numerical experiments to confirm the effectiveness and robustness of LiQUAR; we provide a direct comparison between LiQUAR and PTO. In Section 7 we give the proofs of our main results. We provide concluding remarks in Section 8. Supplementary results are given in the e-Companion.

2. Related Literature

The present paper is related to the following three 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 pricing and capacity sizing problem in a processor sharing queue motivated by internet applications; Kumar and Randhawa (2010) considers a single-server system with nonlinear delay cost; Nair et al. (2016) studies $M/M/1$ and $M/M/k$ systems with network effect among customers; Kim and Randhawa (2018) considers a dynamic pricing problem in a single-server system. The specific problem that we consider here is related to Lee and Ward (2014), which considers joint 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.

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. Krishnasamy et al. (2021) develops bandit methods for scheduling problems in a multi-server queue with unknown service rates. Zhong et al. (2022) 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. (2020) 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. Our paper is most closely related to Jia et al. (2022a) 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. (2022b) extends the results in Jia et al. (2022a) to the problem setting with a continuous price space and considers linear demand functions. Similar to Jia et al. (2022a,b), we also study a queueing control problem with unknown demand. The major distinction is that in addition to maximizing the service profit as by Jia et al. (2022a), 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 gives rise to some new technical challenges to the algorithm design and regret analysis (e.g, treating queueing data with bias and auto-correlation). Besides, the present paper considers more general service distributions and demand functions.

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). 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

$$C_1 \equiv 2\max\left(g(\bar{\mu})\frac{\lambda''(p)}{\lambda'(p)}, g(\underline{\mu})\frac{\lambda''(p)}{\lambda'(p)}\right)\lambda(p) - \frac{4\lambda(p)(\underline{\mu}-\lambda(p))}{h_0 C}, \quad C_2 \equiv -\max\left(\sqrt{\frac{0\vee(-\lambda''(p)(\bar{\mu}-\lambda(p)))}{2}}, \frac{p\lambda''(p)}{2}\right),$$

$g(\mu) = \frac{\mu}{\mu-\lambda(p)} - \frac{p(\mu-\lambda(p))}{h_0 C}$ *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 $0 < b < 4\underline{\lambda}(\underline{\mu}-\bar{\lambda})/h_0 C$; (2) quadratic demand $\lambda(p) = c - ap^2$ with $a, c > 0$, and $\frac{\bar{\mu}-c}{3\underline{p}^2} < a < \left(\frac{3(\underline{\mu}-\bar{\lambda})p}{h_0 C} - \frac{\mu}{\underline{\mu}-\bar{\lambda}}\right)\frac{\lambda}{\bar{p}^2}$; (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.2 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).

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 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

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

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:

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

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.

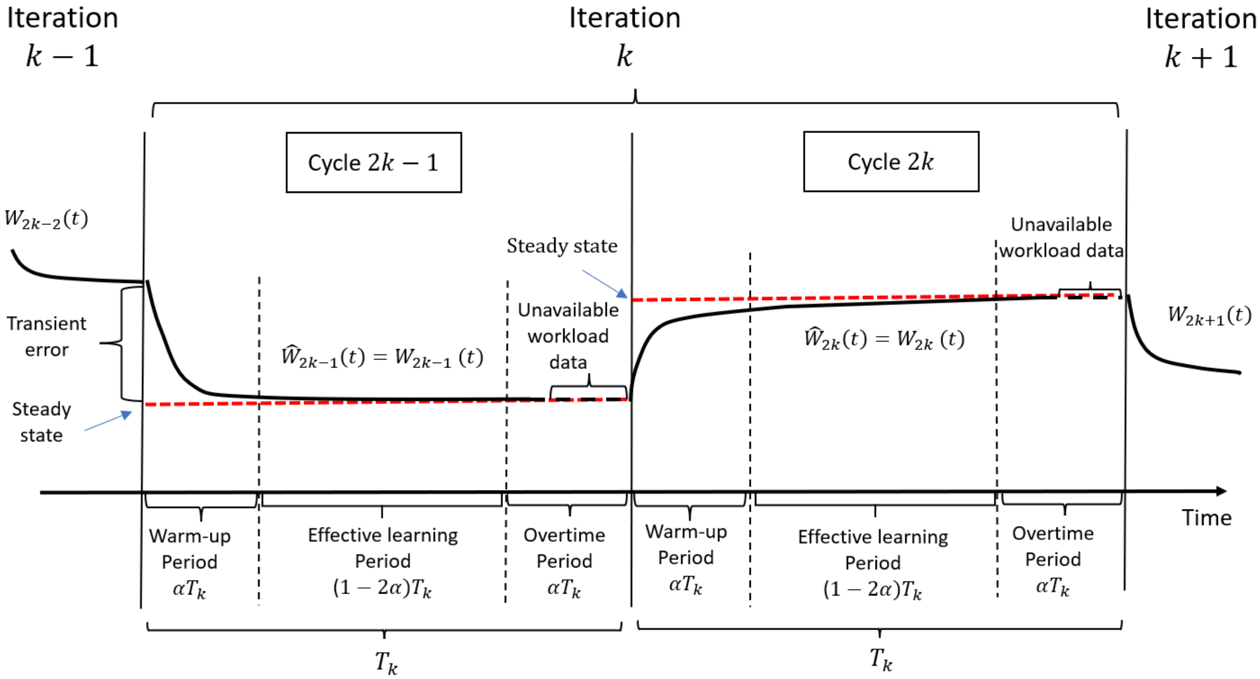

Iteration $k-1$
Iteration k
Iteration $k+1$
Cycle $2k-1$
Cycle $2k$
$W_{2k-2}(t)$
Transient error
Steady state
$\widehat{W}_{2k-1}(t) = W_{2k-1}(t)$
Steady state
Unavailable workload data
$\widehat{W}_{2k}(t) = W_{2k}(t)$
Unavailable workload data
$W_{2k+1}(t)$
Time
Warm-up Period αT_k
Effective learning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

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 universal 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 a uniform moment bound for the workload process under arbitrary control policies, and θ_0 is a small number. Their values are specified in Appendix EC.1.2.

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

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

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 will choose these hyperparameters in Section 5 to minimize the regret bound.

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
```

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

where $\theta_1\equiv\min(\gamma,\theta_0\underline{\mu}/2)$ and γ and K_M are two positive and universal 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 7, where the details of the two constants γ and K_M are also specified. The key step in the proof is to bound the transient bias (from the steady-state distribution) and auto-correlation of the workload process $\{W_l(t):0\le t\le T_k\}$, 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 5 in Section 7 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 1 (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

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

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

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

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 8 in Section 7).

The remaining regret analysis will focus on the nonstationary regret 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 7.6.

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 universal constant specified in Lemma 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 2 (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.

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. First, we test the performance of LiQUAR using an $M/M/1$ base example with logit demand functions in Section 6.1. Next, we conduct sensitivity analysis in the algorithm's hyperparameters including T_k and η_k (Section 6.2). Then, we compare the performance of LiQUAR and PTO and investigate the impact of traffic intensity ρ^* on both methods in Section 6.3. Finally, we extend LiQUAR to queues with non-Poisson arrivals in Section 6.4.

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)

$$\lambda(p) = M_0 \cdot \frac{\exp(a - bp)}{1 + \exp(a - bp)}, \tag{17}$$

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

$$c(\mu) = c_0\mu. \tag{18}$$

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

$$\max_{\mu,p} \left\{ p\lambda(p) - h_0 \frac{\lambda(p)/\mu}{1-\lambda(p)/\mu} - c_0\mu \right\}. \tag{19}$$

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 $(\mathcal{P}(\mu^*, p^*) - \mathcal{P}(\hat{\mu}_\epsilon, \hat{p}_\epsilon))/\mathcal{P}(\mu^*, p^*)$, 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 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 6.3 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 (19), 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 $\eta_k =$

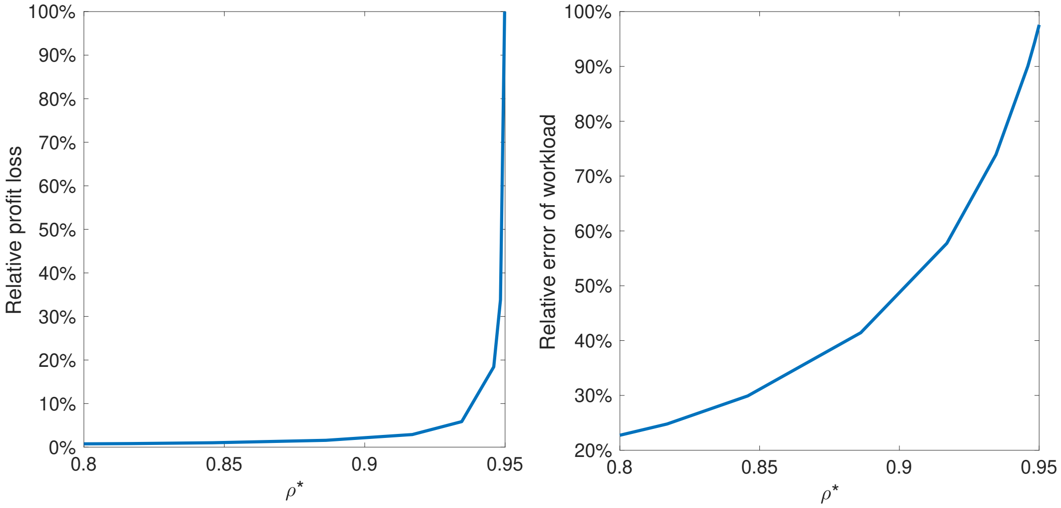


Figure 3 Relative profit loss (left) and workload error (right) for the $M/M/1$ example with $M_0 = 10, a = 4.1$, and $b = 1$ and linear staffing cost $c(\mu) = \mu$.

$4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$ 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 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.

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:

$$\eta_k = c \cdot 4k^{-1}, \quad \text{and} \quad \delta_k = \min(0.1, c \cdot 0.5k^{-1/3}). \tag{20}$$

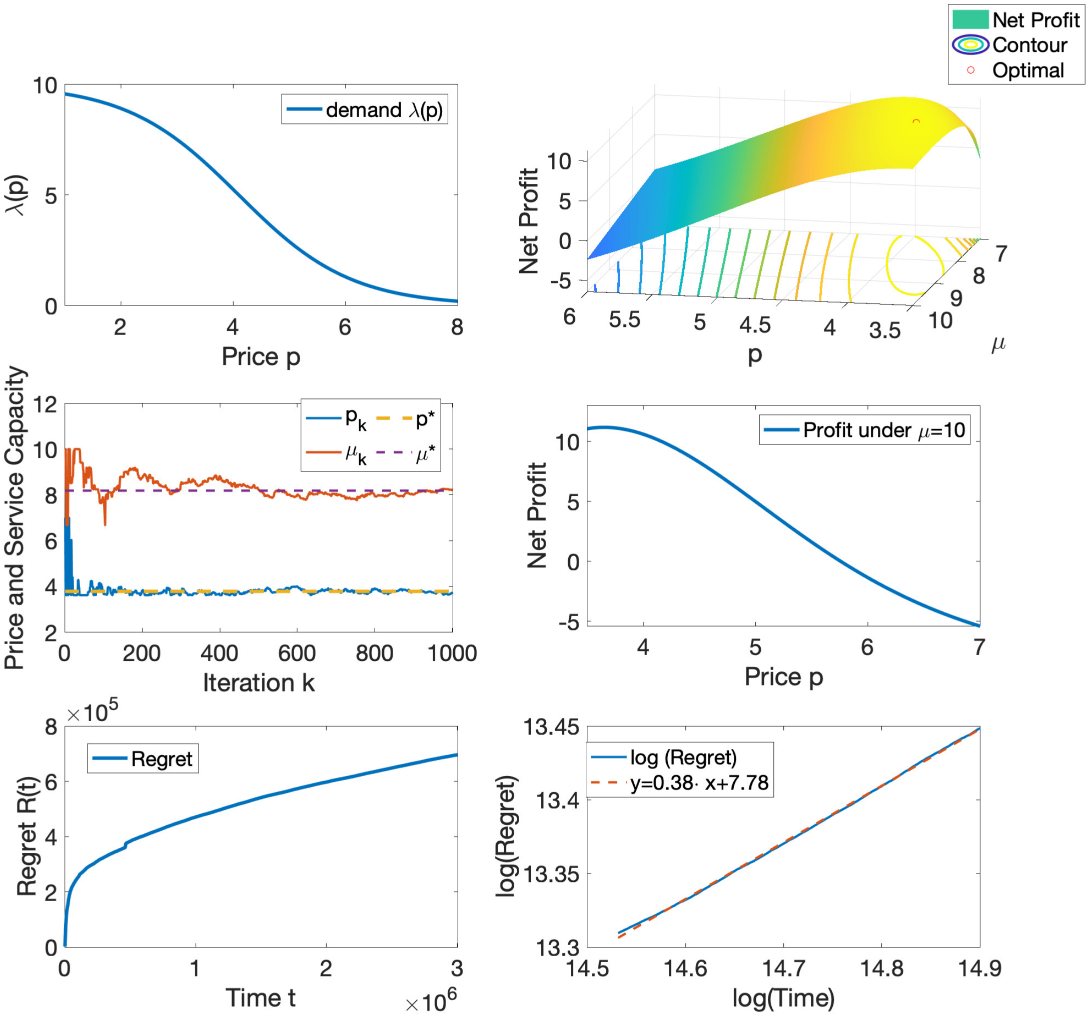


Figure 4 Joint pricing and staffing in the $M/M/1$ logistic demand base example with $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$, $p_0 = 5$, $\mu_0 = 10$ and $\alpha = 0.1$: (i) Demand function $\lambda(p)$ (top left panel); (ii) net profit function (top right panel); (iii) 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.

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

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

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 $c = 0.6$

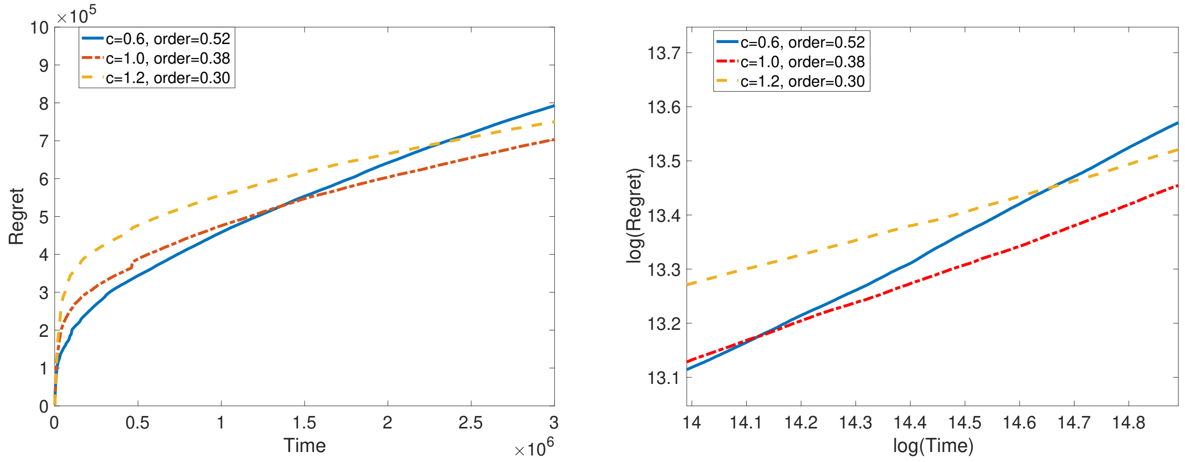


Figure 5 Regret under different $c \in \{0.6, 1.0, 1.2\}$: (i) average regret from 100 independent runs (left panel); (ii) regret curve in logarithm scale, with $T_k = 200k^{1/3}$, $\eta_k = c \cdot 4k^{-1}, \delta_k = \min(0.1, c \cdot 0.5k^{-1/3})$ and $\alpha = 0.1$.

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 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).

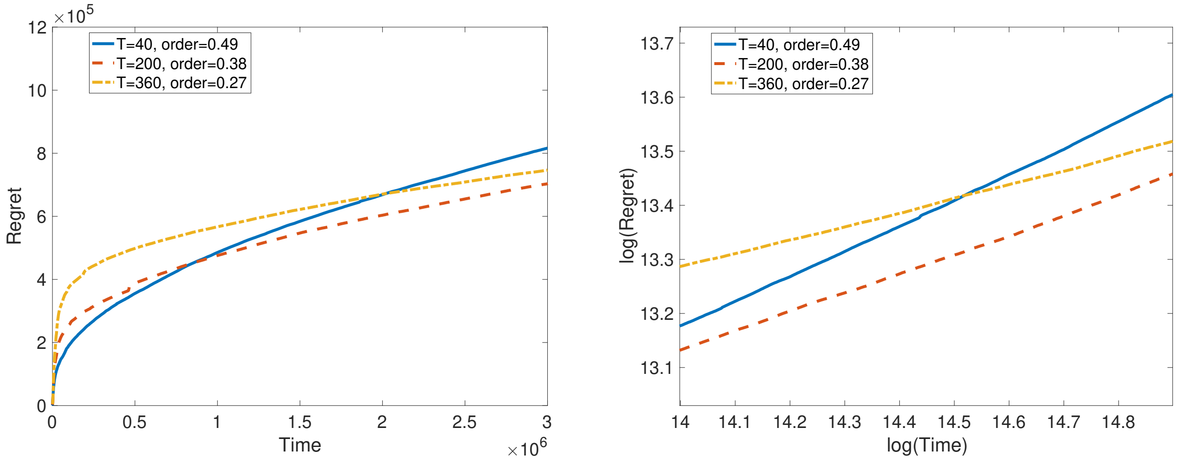


Figure 6 Regret under different $T \in \{40, 200, 360\}$: (i) average regret from 100 independent runs (left panel); (ii) regret curve in logarithm scale, with $T_k = T \cdot k^{1/3}$, $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$ and $\alpha = 0.1$.

6.3. LiQUAR vs. PTO

In this subsection, we conduct numerical analysis to contrast the performance of LiQUAR to that of the conventional predict-then-optimize (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 $\widehat{\lambda}(p)$.

For simplicity, in the subsequent PTO experiments we consider a setting where **the decision maker already has the knowledge of the form of the demand function so the PTO decision maker focuses only on the estimation of the values of the parameters of the demand function**. We dub this setting the *parametric PTO* (pPTO). Of course, such a comparison between pPTO and LiQUAR appears to be unfair because knowing the full form of $\lambda(p)$ gives pPTO an obvious advantage over LiQUAR which does not have this information. Nevertheless, as we will show next, the already information-favored pPTO is still outperformed by LiQUAR, not to mention the nonparametric version of PTO of which the performance can only be worse.

Define $\theta \in (0,1)$ as the exploration ratio and T as the total time (or budget). We set up the pPTO approach by splitting the total time horizon T into two intervals: the interval $[0,\theta T]$ in which we focus on learning the parameters of the demand function, and the interval $[\theta T, T]$ where we operates the system using decisions optimized based on the estimated parameters. We give the details of the two steps below:

- **Prediction.** Suppose m parameters of the demand function are to be estimated, we uniformly select $p_1,\cdots,p_m \in [\underline{p},\bar{p}]$ as experimentation variables. In order to generate samples of the arrival data, we sequentially operate the system under each of the experimentation variable for $\theta T/m$ units of time. Using the arrival data, we give an estimation for the demand parameters using a standard least square method. Specifically, we have

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{m} (\lambda(p_k;\boldsymbol{\beta}) - N_k m/(\theta T))^2,$$

  where $\lambda(\cdot;\boldsymbol{\beta})$ is the parametric form of the demand function, and p_k, N_k are the price and the observed number of arrivals at experimentation point p_k, respectively.
- **Optimization.** Next, we obtain the PTO-optimal policy $\hat{x}^*$ by solving a deterministic optimization problem using the P-K formula with $\widehat{\lambda}(p;\boldsymbol{\beta}^*)$ in place of $\lambda(p)$, and then implement such a policy for the rest of the time interval $[\theta T, T]$.

Experiment settings. We consider our base logit example in Section 6.1 having demand function (17) 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 robustness of pPTO and LiQUAR in the system's congestion level, we consider two scenarios specified by the optimal traffic intensity ρ^*: (i) the light-traffic case with $\rho^* = 0.709$ ($h_0 = 1$) and (ii) the 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 pPTO, 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 represent different level of exploration. For the clarity of the figures, we only present the three pPTO curves having the lowest regrets.

Experiment results. In Figure 7, we report results of regret under LiQUAR and pPTO. According to the left-hand panels in Figure 7, we conclude that the exploration ratio has a significant impact on pPTO performance. Specifically, the pPTO regret exhibits a piecewise linear structure: in the prediction phase, it grows rapidly due to the periodic exploration across all experimentation variables; in the optimization phase, it continues to grow linearly with a less steep slope, because the system is now operated using the pPTO-optimal solution which is not optimal. A bigger (smaller) θ incurs a bigger (small) regret in the prediction phase and produces a more (less) accurate model, leading to decisions that are more (less) optimal and a flatter (steeper) regret curve in the optimization phase. Apparently, LiQUAR performs more effectively than pPTO by yielding a smaller regret.

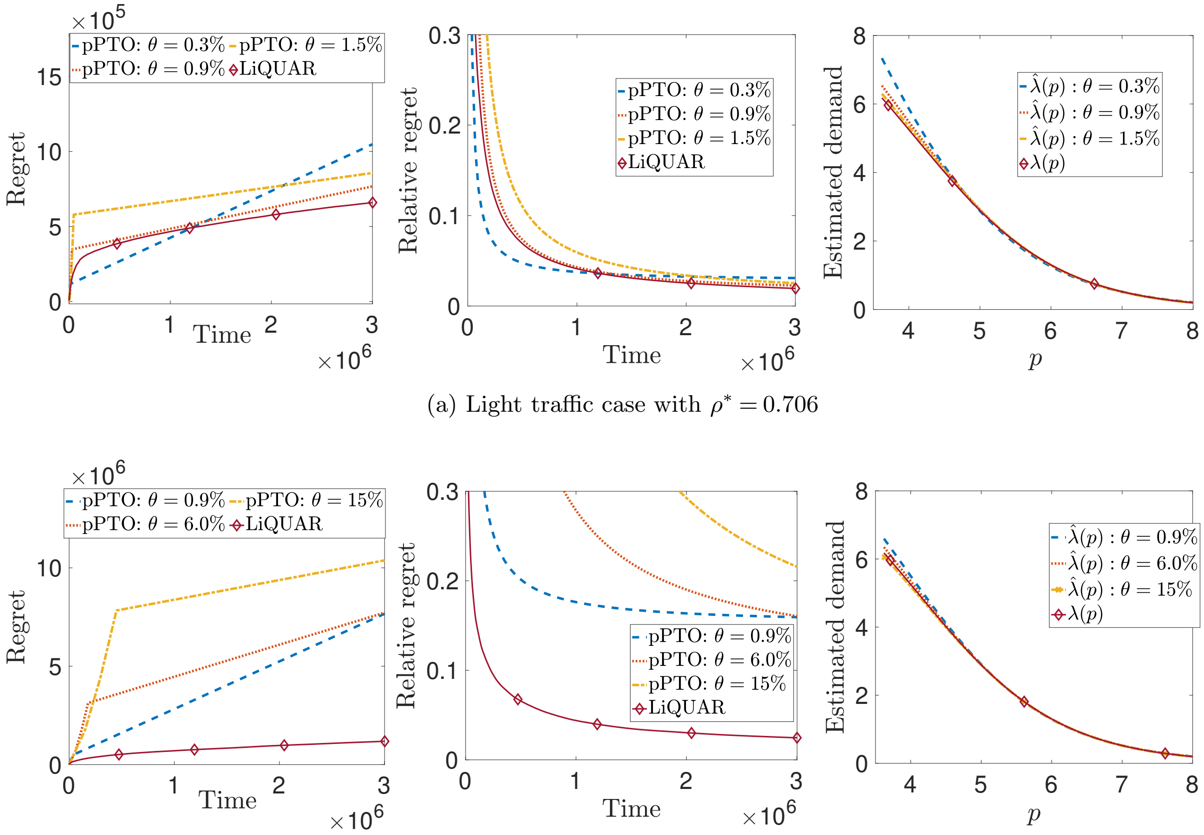
Next, we investigate how the system's congestion level (i.e., ρ^*) impacts the performance of the two algorithms. To facilitate a fair comparison, we consider a relative regret instead of the absolute regret in order to exclude the influence of ρ^* on the optimal profit values. The relative regret is defined as the ratio of the cumulative regret to the cumulative optimal profit by t, that is,

$$\tilde{R}_{\rho^*}(t) \equiv \frac{R_{\rho^*}(t)}{\mathcal{P}_{\rho^*}(\mu^*, p^*)t} \tag{21}$$

where where the subscript ρ^* is appended in order to emphasize the dependence on the traffic intensity. By contrasting the middle panels in subfigures (a) and (b), we find that pPTO becomes less effective in the heavy-traffic case than in the light-traffic case (all pPTO settings yield a much bigger relative regret when ρ^* is large) although fitted demand curves (right-hand panels) appear to be accurate in both cases, which echos the motivating example in Section 6.1.1. Such a performance degradation in heavy traffic is due to the fact that a smaller error in the estimated demand curve can be disproportionally amplified by the formula-based PTO method. On the other hand, LiQUAR is much more robust to ρ^* of which the regret exhibits nearly no distinction in the two cases. This is because LiQUAR is an integrated method of which its design has automatically incorporated the parameter estimation errors in the decision prescription process.

6.4. 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 $\mathbb{E}[U_n] = 1$.



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

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

Figure 7 Comparison between PTO and LiQUAR under logit demand, with varying ρ^*: (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.

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 (17) with $M = 10, a = 4.1, b = 1$ and linear staffing cost function (18). 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 $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$,

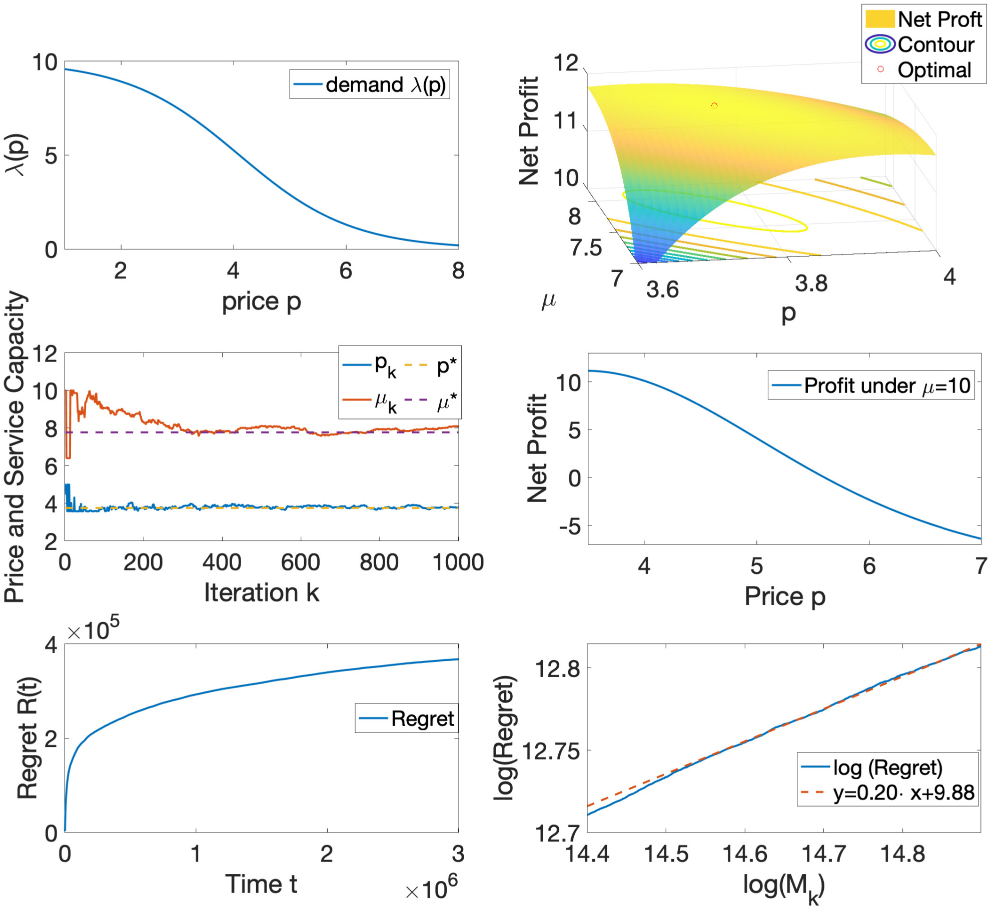


Figure 8 Joint pricing and staffing in the $E_2/M/1$ queue with $\eta_k = 4k^{-1}, \delta_k = \min(0.1, 0.5k^{-1/3})$, $T_k = 200k^{1/3}$, $\alpha = 0.1$, $p_0 = 5$ and $\mu_0 = 10$.

and $\alpha = 0.1$. Figure 8, as an analog to Figure 4, shows that the refined LiQUAR continues to be effective, exhibiting a rapid converge to the optimal decision and a slowly growing regret curve (bottom left panel of Figure 8). 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. Main Proofs

In this section, we provide the proofs of the main theorems and propositions. Proofs of technical lemmas are given in the E-Companion.

7.1. Proof of Proposition 1

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

Lemma 1 **(Uniform Moment Bounds)**. *Under Assumptions 1 and 2, there exist some universal 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 1. □

7.2. 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 $0 \leq t \leq T_k$. 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

$$\begin{aligned}
&\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]\\
=&\ \frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\left(\mathbb{E}_l\left[\hat{W}_l(t)\right]-\mathbb{E}_l[\bar{W}_l(t)]\right)dt\le\frac{1}{(1-2\alpha)T_k}\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l\left[|\hat{W}_l(t)-\bar{W}_l(t)|\right]dt.\\
\le&\ \frac{1}{(1-2\alpha)T_k}\left(\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l[|\hat{W}_l(t)-W_l(t)|]dt+\int_{\alpha T_k}^{(1-\alpha)T_k}\mathbb{E}_l[|W_l(t)-\bar{W}_l(t)|]dt\right).
\end{aligned}\tag{22}$$

The first term in (22) is the error caused by delayed observation. Following the same analysis as in Section 7.1,

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

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 1, $\bar{W}_l(t)$ is stochastic bounded by the stationary workload with parameter $(\underline{\mu},\underline{p})$. Therefore,

$$\begin{aligned}
&\mathbb{E}_l\left[|\hat{W}_l(t)-W_l(t)|\right]\le\mathbb{E}_l[W_l(t)^2]^{1/2}\cdot\exp(-\theta_0\mu_l(T_k-t)/2)\mathbb{E}_l[\exp(\theta_0W_l(t))]^{1/2}\\
\le&\ \exp(-\theta_0\mu_l(T_k-t)/2)(W_l(0)^2+2W_l(0)\mathbb{E}_l[\bar{W}_l(t)]+\mathbb{E}_l[\bar{W}_l(t)^2])^{1/2}\exp(\theta_0W_l(0))\mathbb{E}_l[\exp(\theta_0\bar{W}_l(t))]^{1/2}\\
\le&\ \exp(-\theta_0\mu_l(T_k-t)/2)(W_l(0)^2+2MW_l(0)+M^2)^{1/2}\exp(\theta_0W_l(0))M^{1/2}\\
\le&\ \exp(-\theta_0\mu_l(T_k-t)/2)M(M+W_l(0))\exp(\theta_0W_l(0)).
\end{aligned}\tag{23}$$

The last inequality holds as $M\ge1$. The second term in (22) will be bounded using the following lemma on convergence rate of two synchronously coupled workload processes.

Lemma 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*

$$\mathbb{E}\left[|W(t)-\bar{W}(t)|^m\mid W(0),\bar{W}(0)\right]\le e^{-\gamma t}(e^{\theta_0W(0)}+e^{\theta_0\bar{W}(0)})|W(0)-\bar{W}(0)|^m.$$

Using this lemma, we can compute

$$\begin{aligned}
\mathbb{E}_l[|W_l(t)-\bar{W}_l(t)|]&\le\exp(-\gamma t)\mathbb{E}_l\left[|W_l(0)-\bar{W}_l(0)|(\exp(\theta_0W_l(0))+\exp(\theta_0\bar{W}_l(0)))\right]\\
&\le\exp(-\gamma t)\left(W_l(0)\exp(\theta_0W_l(0))+MW_l(0)+M\exp(\theta_0W_l(0))+M\right)\\
&\le\exp(-\gamma t)(M+W_l(0))(\exp(\theta_0W_l(0))+M).
\end{aligned}\tag{24}$$

Let $\theta_1=\min(\gamma,\theta_0\underline{\mu}/2)$. Plugging inequalities (23) and (24) into (22), we obtain the following bound for the bias

$$\left|\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]\right|\le\frac{1}{(1-2\alpha)T_k}\cdot\frac{2\exp(-\theta_1\alpha T_k)}{\theta_1}\cdot M(M+W_l(0))(\exp(\theta_0W_l(0))+M).$$

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 $0 \le s \le t \le T_k$,

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

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

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

$$\mathbb{E}_l[(W_l(t) - w_l)(W_l(s) - w_l)] \le K_V\left(\exp(-\gamma(t-s)) + \exp(-\gamma s)\right)(W_l(0)^2 + 1)\exp(\theta_0 W_l(0)). \quad (25)$$

Following (25), 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

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

As $W_l(0)\le(W_l(0)^2+1)/2$ and $M\ge1$, we have $M+W_l(0)\le\frac{2M+1}{2}(1+W_l(0)^2)$. There must exist a finite constant K_M that is large enough such that,

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

□

7.3. Proof of Proposition 3

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

LEMMA 4. *Under Assumption 1, there exists a universal 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

$$\left|\mathbb{E}_l\left[\hat{f}^G(\mu_l,p_l)-f(\mu_l,p_l)\right]\right|\le\frac{2\exp(-\theta_1\alpha T_k)}{(1-2\alpha)\theta_1T_k}\cdot M(M+W_l(0))(\exp(\theta_0W_l(0))+M).$$

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

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

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

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

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

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

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

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

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

7.4. 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 5 **(Convexity and Smoothness of $f(\mu,p)$).** *Suppose Assumption 1 holds. Then, there exist finite positive constants $0<K_0\le 1$ and $K_1>K_0$ such that for all $x=(\mu,p)\in\mathcal{B}$,*
(a) $(\boldsymbol{x}-\boldsymbol{x}^*)^T\nabla f(x)\ge K_0\|\boldsymbol{x}-\boldsymbol{x}^*\|^2$,
(b) $|\partial_\mu^3 f(\boldsymbol{x})|,|\partial_p^3 f(\boldsymbol{x})|\le K_1$.

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.1.1. 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:

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

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.1.1, 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).

7.5. 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 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 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. (2020) 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 7 **(Nonstationary Error in Warm-up Period)**. *Suppose $T_k>t_k\equiv\log(k)/\gamma$ for all $k\geq 1$. Then, there exists a universal 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 6 and Lemma 7, we have

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

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. □

7.6. 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 8 **(Exploration Cost)**. *Under Assumption 1, there exists a constant $K_4 > 0$ such that*

$$R_{3k} \leq K_4 T_k \delta_k^2. \tag{26}$$

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,

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

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 $R(L) = O(\sqrt{T(L)}\log(T(L))$. □

8. Conclusions

In this paper we develop an online learning framework, dubbed LiQUAR, designed for dynamic pricing and staffing in an $M/GI/1$ queue. LiQUAR's main appeal is its "model-free" attribute. Unlike the conventional "predict-then-optimize" approach where precise estimation 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 on 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 explore and then stick to control policies that guarantee a stable system performance, otherwise the decision maker will face a low profit (high cost).

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.

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

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 (2020) An online learning approach to dynamic pricing and capacity sizing in service systems. *Working Paper*.

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.

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.

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.

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

Jia H, Shi C, Shen S (2022a) Online learning and pricing for service systems with reusable resources. *Operations Research* Forthcoming.

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

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.

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.

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

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.

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.

Zhong Y, Birge JR, Ward A (2022) Learning the scheduling policy in time-varying multiclass many server queues with abandonment. *Working Paper*.

E-Companion

This e-companion provides supplementary materials to the main paper. In Section EC.1, we give the technical proofs omitted in the article. In Section EC.2, we verify that the Condition (a) of Assumption 1 holds for some commonly used demand functions. In Section EC.3, 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. Proofs

EC.1.1. 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 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

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

The second last inequality follows from Hölder's Inequality, and the last inequality follows from $2a \le 1 + a^2$. 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

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

Next, we prove by mathematical induction that there exists a large constant $K_2 > 0$ such that $b_k \le K_2 k^{-\beta}$ for all $k \ge 1$ using recursion (EC.1). Given that $\eta_k \mathcal{V}_k = O(k^{-\beta})$, we can find a constant

$K_3 > 0$ large enough such that $\eta_k \mathcal{V}_k \le K_3 k^{-\beta}$ for all $k \ge 1$. Then, by the induction assumption that $b_k \le K_2 k^{-\beta}$, we have

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

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

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

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

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

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

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

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 $\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 \le K_2 k^{-\beta}$ for all $1 \le k \le k_0$, and we can conclude by induction that, for all $k \ge k_0$,

$$\mathbb{E}[\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2] \le K_2 k^{-\beta}.$$

□

EC.1.2. Proofs of Technical Lemmas

In addition to the uniform moment bounds for $W_l(t)$ as stated in Lemma 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 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.1 **(Complete Version of Lemma 1)**. *Under Assumptions 1 and 2, there exist some universal constants $\theta_0 > 0$ and $M > 1$ such that, for any sequence of control parameters $\{(\mu_l, p_l) : l \ge 1\}$,*

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

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

Proof of Lemma EC.1 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 universal 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 $t \in [0, T_k]$,

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

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

Proof of Lemma 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

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

Following Lemma 2 of Chen et al. (2020), we have

$$|W(t) - \bar{W}(t)| \leq |W(0) - \bar{W}(0)|\mathbf{1}\left(t < \tau \vee \bar{\tau}\right). \tag{EC.2}$$

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.1 and the value of γ will be specified in (EC.3). Let $\{\mathcal{F}_t\}_{t\geq 0}$ 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

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

We choose

$$\gamma = \underline{\lambda}\left(1 - e^{-\frac{\theta_0\gamma_0}{\theta}}\right)\left(1 + \underline{\mu}\theta_0/\bar{\lambda}\right). \tag{EC.3}$$

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.3). By Fatou's lemma,

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

Similarly, $\mathbb{P}(\bar{\tau} > t|\bar{W}(0)) \leq e^{-\gamma t}\exp(\theta_0 \bar{W}(0))$. Combining these bounds with (EC.2), we can conclude that

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

□

Proof of Lemma 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 $0\leq s\leq t\leq T_k$. To do this, we synchronously couple with $\{W_l(r): s\leq r\leq T_k\}$ a stationary workload process $\{\bar{W}_l^s(r): s\leq r\leq T_k\}$. 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 $s\leq r\leq T_k$, 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 2,

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

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

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

One can check that $W_l(s)\leq W_l(0)+\bar{W}_l(s)$, 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,

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

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

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

On the other hand, by Lemma 2,

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

As a consequence,

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

and we can conclude (25) with $K_V = 2M + 5M^2 + 2M^3$. □

Proof of Lemma 4 By the mean value theorem,

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

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

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

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

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

□

Proof of Lemma 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

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

The second-order derivatives are

$$\partial_{pp} f = \frac{h_0 C \mu}{(\mu-\lambda)^3}\left(2(\lambda')^2 + (\mu-\lambda)\lambda''\right) - p\lambda'' - 2\lambda'$$
$$\partial_{p\mu} f = -\frac{h_0 C(\mu+\lambda)}{(\mu-\lambda)^3}, \quad \partial_{\mu\mu} f = \frac{2h_0 C\lambda}{(\mu-\lambda)^3} + c''(\mu).$$

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,

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

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 6 By Lemma 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 $t \geq t_k$,

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

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 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

$$\begin{aligned}
&\left|\int_0^T W^1(t)dt - \int_0^T W^2(t)dt\ \right| \\
\leq &\frac{1}{\lambda_1}\left|\int_0^{\lambda_2 T} (\tilde{W}^1(t) - \tilde{W}^2(t))dt\right| + \left|\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right|\int_0^{\lambda_2 T} \tilde{W}^2(t)dt + \frac{1}{\lambda_1}\int_{\lambda_2 T}^{\lambda_1 T} \tilde{W}^1(t)dt. \qquad \text{(EC.4)}
\end{aligned}$$

Following a similar argument as in the proof of Lemma 3 in Chen et al. (2020), 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.4) 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.4), 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

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

where $\tilde{X}^D_l(\cdot)$ is the dominant observed busy period defined in the proof of Lemma EC.1. Here inequality (a) follows from the definition of $\tilde{X}^D_l(\cdot)$, inequality (b) from Lemma 6 and equality (c) from Lemma EC.1 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,

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

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

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

As a consequence,

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

□

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

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

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

EC.2. 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.

$$-\lambda'(p)>\max\left(\sqrt{\frac{0\vee(-\lambda''(p)(\bar{\mu}-\lambda(p)))}{2}}\ ,\ \frac{p\lambda''(p)}{2}\right),\tag{EC.5}$$

$$\lambda'(p)>\max_{\mu\in[\underline{\mu},\bar{\mu}]}\left(2g(\mu)\frac{\lambda''(p)\lambda(p)}{\lambda'(p)}-\frac{4\lambda(p)(\mu-\lambda(p))}{h_0C}\right).\tag{EC.6}$$

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

$$\lambda(p)=a-bp,\qquad\text{with }0<b<\frac{4\underline{\lambda}(\underline{\mu}-\bar{\lambda})}{h_0C}.$$

Then, inequality (EC.5) holds immediately as $\lambda''(p)\equiv0$. Inequality (EC.6) is equivalent to

$$-b>-\frac{4\lambda(p)(\underline{\mu}-\lambda(p))}{h_0C},$$

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.5) 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.6), 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.6) 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.5) is automatically satisfied as $b < 2/\bar{p}$. For inequality (EC.6), 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.6) 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.5) becomes $2 > bp(1-e)/(1+e)$ if $e < 1$. Since $a - bp < \log(1/2)$, $e < 1/2$ and (EC.5) holds accordingly. We next show that (EC.6) 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.6) holds as well.

EC.3. Additional Numerical Experiments

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 (17) with $M_0 = 10, a = 4.1$ and $b = 1$. We fix staffing cost coefficient $c_0 = 1$ in (18) 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).

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.

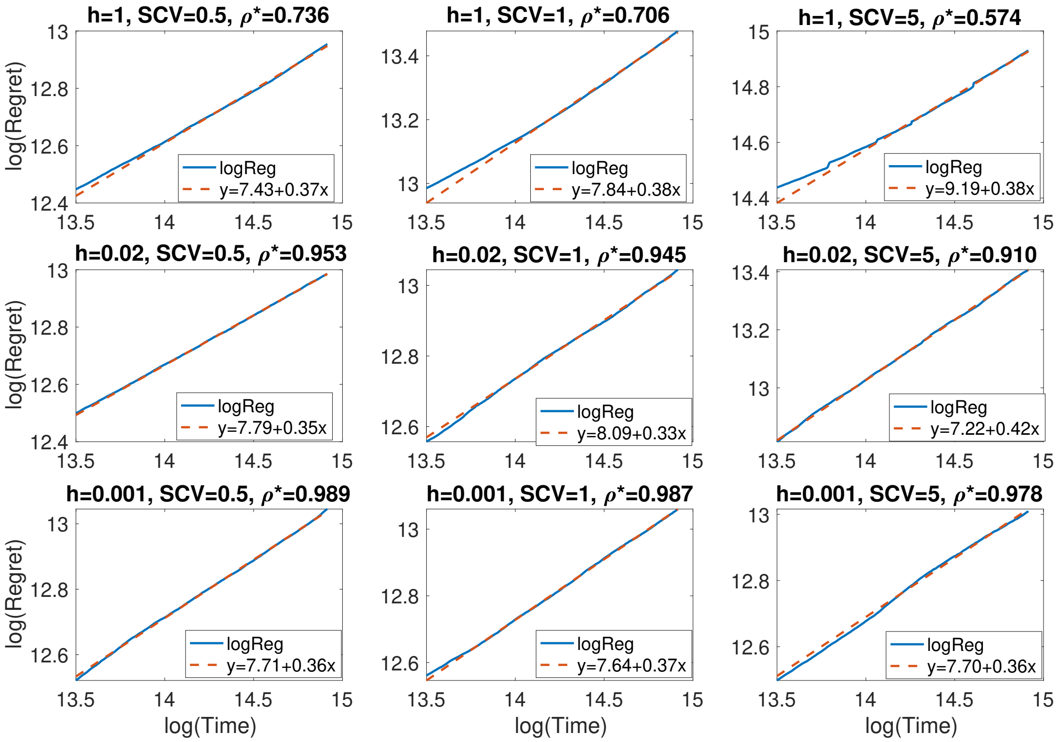


Figure EC.1 The regret curve in logarithm scale and a linear fit for the $M/GI/1$ model, under different traffic intensity $\rho^* \in [0.547, 0.989]$ and service-time SCV $c_s^2 = 0.5$ (E_2 service), 1 (M service) and 5 (H_2 service). All curves are estimated by averaging 100 independent runs.

	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
	δ_k	Exploration length in iteration k
	η_k	Step length for gradient update in iteration k
	$\boldsymbol{H}_k$	Gradient estimator in iteration k
	$\hat{f}^G(\mu_l, p_l)$	Estimation of objective function in cycle l
	$T_k, T_{k(l)}$	Cycle length of iteration k and cycle l
	$W_l(t)(\hat{W}_l(t))$	(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 4
	c_η, c_T, c_δ	Coefficient of hyperparameters in Theorem 2
	C_0	Constant in Lemma 7
	M	Upper bound for queueing functions in Lemma EC.1
	γ	Ergodicity rate constant in Lemma 2
	K_0, K_1	Convex and smoothness constant of objective function in Lemma 5
	K_2, K_3	Constants in the proof of Theorem 1 in Appendix EC.1.1
	K_4	Constant in Lemma 8
	K_V	Constant of auto-correlation in Lemma 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.1
	$\theta_1 = \min(\gamma, \theta_0 \underline{\mu}/2)$	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 notation