

E-Companion

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

EC.1. Main Proofs

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

EC.1.1. Proof of Proposition 1

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

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

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

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

Then, following (7),

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

where the last inequality follows from Lemma EC.1. \square

EC.1.2. Proof of Proposition 2

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

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

where $w_l = \mathbb{E}[W_\infty(\mu_l, p_l)]$ is the steady-state mean workload. To bound the moments of this difference, which correspond to the bias and MSE of $\hat{f}^G(\mu_l, p_l)$, we construct a stationary workload process $\bar{W}_l(t)$

for $0 \leq t \leq T_k$. 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

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

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

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

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

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

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

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

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

Using this lemma, we can compute

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

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

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

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

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

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, $\mathbb{E}_l[E_1^2] = p_l^2 \lambda(p_l) T_k^{-1} \leq \bar{p}^2 \bar{\lambda} T_k^{-1}$.

For E_2 , we have

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

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

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

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

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

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

Following (EC.4), we write

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

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

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

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

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

then, we have

$$\mathbb{E}_l[(\hat{f}^G(\mu_l, p_l) - f(\mu_l, p_l))^2] \leq 2\mathbb{E}_l[E_1^2] + 2\mathbb{E}_l[E_2^2] \leq K_M T_k^{-1} (W_l(0)^2 + 1) \exp(\theta_0 W_l(0)).$$

□

EC.1.3. Proof of Proposition 3

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

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

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

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,

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

Therefore,

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

By Proposition 2, the bias of estimated system performance

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

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

$$\begin{aligned} & \mathbb{E}[\mathbb{E}_l[\hat{f}^G(\mu_l, p_l) - f(\mu_l, p_l)]^2] \\ & \leq \frac{4 \exp(-2\theta_1 \alpha T_k)}{(1-2\alpha)^2 \theta_1^2 T_k^2} (4M^4 \mathbb{E}[\exp(2\theta_0 W_l(0)) + W_l(0)^2] + 4M^2 \mathbb{E}[W_l(0)^2 \exp(2\theta_0 W_l(0))] + 4M^6) \\ & \leq \frac{4 \exp(-2\theta_1 \alpha T_k)}{(1-2\alpha)^2 \theta_1^2 T_k^2} \cdot (8M^5 + 4M^3 + 4M^6) = O(\exp(-2\theta_1 \alpha T_k)). \end{aligned}$$

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

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

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

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

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

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

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

EC.1.4. Proof of Theorem 1

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

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

- (a) $(\mathbf{x} - \mathbf{x}^*)^T \nabla f(x) \geq K_0 \|\mathbf{x} - \mathbf{x}^*\|^2$,
- (b) $|\partial_\mu^3 f(\mathbf{x})|, |\partial_p^3 f(\mathbf{x})| \leq K_1$.

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

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

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

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

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

EC.1.5. Proof of Proposition 4

The regret of nonstationarity

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

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

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

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

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

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

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

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

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

As a consequence,

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

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

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

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. \square

EC.1.6. Proof of Theorem 2

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

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

$$R_{3k} \leq K_4 T_k \delta_k^2. \quad (\text{EC.6})$$

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

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

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 $R(L) = O(\sqrt{T(L)} \log(T(L)))$. \square

EC.2. Proofs

EC.2.1. Full Proof of Theorem 1

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

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

Note that

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

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

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

Then, we obtain the recursion

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

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

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

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

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

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

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

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

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

Let

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

Then we have $\|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 \leq K_2 k^{-\beta}$ for all $1 \leq k \leq k_0$, and we can conclude by induction that, for all $k \geq k_0$,

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

□

EC.2.2. Proofs of Technical Lemmas

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

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

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

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

Proof of Lemma EC.9 We consider a $M/GI/1$ system under a stationary policy such that $\mu_l \equiv \underline{\mu}$ and $p_l \equiv \underline{p}$ for all $l \geq 1$. We call this system the dominating system and denote its workload process by $W_l^D(t)$. In addition, we set $W_1^D(0) \stackrel{d}{=} W_\infty(\underline{\mu}, \underline{p})$ so that $W_l^D(t) \stackrel{d}{=} W_\infty(\underline{\mu}, \underline{p})$ 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

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

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

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

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

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

Under Assumption 2, we have

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

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

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

for $m = 1, 2$.

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

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

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 $M = C \vee (\mathbb{E}[X_\infty(1, \underline{\mu}/\lambda(\underline{p}))]/\lambda(\bar{p}))$ and we can conclude that $\mathbb{E}[X_l(t)] \leq M$. \square

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

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

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

$$|W(t) - \bar{W}(t)| \leq |W(0) - \bar{W}(0)| \mathbf{1}(t < \tau \vee \bar{\tau}). \quad (\text{EC.8})$$

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

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

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

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

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

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

Then, it satisfies that

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

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

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

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

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

□

Proof of Lemma EC.3 We first analyze the conditional expectation $\mathbb{E}_l[(W_l(t) - w_l)(W_l(s) - w_l)]$ for each given pair of (s, t) such that $0 \leq s \leq t \leq T_k$. 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

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

Then, we have

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

By Lemma EC.2,

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

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

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

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,

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

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 EC.2,

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

As a consequence,

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

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

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

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

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

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

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 EC.5 By Pollaczek-Khinchin formula and PASTA,

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

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

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

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

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

The second-order derivatives are

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

By Condition (a) of Assumption 1, we have

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

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

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

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

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

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

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

Therefore,

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

□

Proof of Lemma EC.7 Statement (1) is a direct corollary of Pollaczek–Khinchine formula. The proof of Statement (2) involves coupling workload processes with different parameters. Let us first explain the coupling in detail. Suppose $W^1(t)$ and $W^2(t)$ are two workload processes on $[0, T]$ with parameters (μ_1, λ_1)

and (μ_2, λ_2) respectively. Let $W^1(0)$ and $W^2(0)$ be the given initial states. We construct two workload processes $\tilde{W}^1(t)$ and $\tilde{W}^2(t)$ on $[0, \infty)$ with parameters $(\mu_1/\lambda_1, 1)$ and $(\mu_2/\lambda_2, 1)$ such that $\tilde{W}^i(0) = W^i(0)$ for $i = 1, 2$. The two processes $\tilde{W}^1(t)$ and $\tilde{W}^2(t)$ are coupled such that they share the same Poisson arrival process $N(t)$ with rate 1 and the same sequence of individual workload V_n .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

For the second term,

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

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

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

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

$$R_{3k} = \mathbb{E}[T_k (f(\mathbf{x}_{2k-1}) + f(\mathbf{x}_{2k}) - 2f(\bar{\mathbf{x}}_k))] = \mathbb{E}[T_k (f''(\mathbf{x}') + f''(\mathbf{x}'')) \delta_k^2] \leq K_4 T_k \delta_k^2,$$

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

EC.2.3. Proof of Theorem 3

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

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

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

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

and the corresponding finite difference

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

with

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Consequently, we have

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

By changing of variables, we would see

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

Now, we have

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

\square

LEMMA EC.13 (FD Approximation Error Bound).

$$|Df_h(p_k^h) - f'_h(p_k^h)| \leq Ck^{-2/3}.$$

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

$$\begin{aligned} f_h(p + \delta) &= f_h(p) + \delta f'_h(p) + \frac{\delta^2}{2} f''_h(p) + \frac{\delta^3}{6} f'''_h(p_1) \\ f_h(p - \delta) &= f_h(p) - \delta f'_h(p) + \frac{\delta^2}{2} f''_h(p) - \frac{\delta^3}{6} f'''_h(p_2) \end{aligned}$$

Therefore,

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

Note that

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

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

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

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

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

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

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

Bias term By definition, the bias is defined by

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

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

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

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

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

Consequently, we have

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

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

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

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

Variance Term For the variance term, we have

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

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

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

Consequently, we have

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

For the first term, we have

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

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

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

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

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

for some C large enough. Therefore, we have

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

□

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As a result, we have

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

\square

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

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

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

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

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

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

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

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

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

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

Therefore, by Lemma EC.12, we have

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

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

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

□

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

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

Proof of Lemma EC.18 By calculation, we have

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

□

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

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

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

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

□

EC.2.4. Proof of Proposition 6

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

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

We treat the first two terms following in the same way as in Besbes and Zeevi (2009). For the third term, we apply second order Taylor expansion (rather than the first order in Besbes and Zeevi (2009)) as $\nabla f(p^*) = 0$

in our problem. Therefore, using the fact that the grid length is at most $|p_G^* - p^*| \leq |\mathcal{B}_h|/\kappa = O(\sqrt{h}/\kappa)$, we have

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

By optimizing the regret order, we choose

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

such that

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

□

EC.3. Regret Lower Bound

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

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

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

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

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

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

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

EC.3.1. Demand Class and Its Properties

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

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

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

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

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

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

The corresponding suboptimal regret is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that

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

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

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

EC.3.2. Proof for Lower Bound

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

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

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

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

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

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

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

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

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

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

The next lemma describes the cost of uncertainty.

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

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

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

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

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

By Lemma EC.21, we also have

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

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

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

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

EC.3.3. Lower Bound in Heavy-Traffic

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. **Uninformative point.** All demand curves cross an uninformative point, i.e., $\lambda(p_h^*(z_0); z) = 3 - p_h^*(z_0) = \lambda(p_h^*(z_0); z_0)$ for all $z \in \mathcal{Z}$. Moreover, $p_h^*(z_0) = 2 + \sqrt{h} + o(\sqrt{h})$.
2. **Strongly convex.** For any $z \in \mathcal{Z}$, the objective function $f_h(p; z)$ is strongly convex. In addition, there exists a constant $K_8 > 0$ independent of h , such that

$$|f_h(p_h^*(z); z) - f_h(p; z)| \geq \frac{K_8}{\sqrt{h}} (p - p_h^*(z))^2.$$

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

$$\sup_{p \in \mathcal{B}_h} \lambda_h(p; z) < 1.$$

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

$$|\lambda_h(p; z) - \lambda_h(p; z_0)| = |(p - p_h^*(z_0))(z - z_0)|.$$

5. **Separability between optimal solutions.** There exists a constant K_9 independent of h such that $|p_h^*(z)'| \geq K_9 \sqrt{h}$ for all $z \in \mathcal{Z}$. Therefore,

$$|p_h^*(z) - p_h^*(1)| \geq \sqrt{h} K_9 |z - 1|.$$

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

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

LEMMA EC.23. *For any $z \in \mathcal{Z}$, $T_0 > 0$ and any piecewise constant policy $\pi \in \Psi$,*

$$\mathcal{K}(\mathbb{P}_{T_0}^{\pi, z_0} \parallel \mathbb{P}_{T_0}^{\pi, z}) \leq \sqrt{h} \cdot \frac{(z - z_0)^2}{2K_8} R_1^h(z_0, \pi, T_0)$$

Proof of Lemma EC.23 Following the same analysis in the proof of Lemma EC.20, we have

$$\begin{aligned} \mathcal{K}(\mathbb{P}_t^{\pi, z_0} \parallel \mathbb{P}_t^{\pi, z} | \omega_{t-1}) &\leq \int_{\omega_{t-1}} \frac{(\lambda_t^z - \lambda_t^{z_0})^2}{2\lambda} d\mathbb{P}_{t-1}^{\pi, z_0}(\omega_{t-1}) = \frac{(z - z_0)^2}{2\lambda} \int_{\omega_{t-1}} (p_t - p_0^*)^2 d\mathbb{P}_{t-1}^{\pi, z_0} \\ &\leq \sqrt{h} \frac{(z - z_0)^2}{2K_8\lambda} \mathbb{E}^{\pi, z_0} [f_h(p_t; z_0) - f_h(p^*(z_0); z_0)]. \end{aligned}$$

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

LEMMA EC.24. *For any $T_0 \geq 1$, set $z_1 = z_0 - K_{10}T_0^{-1/4}$ with $K_{10} = 0.05$. Then, for any policy $\pi \in \Psi$, we have*

$$R_1^h(z_0, \pi, T_0) + R_1^h(z_1, \pi, T_0) \geq \frac{1}{\sqrt{h}} \frac{K_8 K_9^2 K_{10}^2}{18} T_0^{1/2} e^{-\mathcal{K}(\mathbb{P}_T^{\pi, z_0} \parallel \mathbb{P}_T^{\pi, z_1})}.$$

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

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

EC.4. Examples of the Demand Function

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

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

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

EXAMPLE EC.1 (LINEAR DEMAND). Consider a linear demand function

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

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

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

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

EXAMPLE EC.2 (QUADRATIC DEMAND). Consider a quadratic demand function

$$\lambda(p) = c - ap^2, \quad \text{with } a, c > 0 \text{ and } 0 < \frac{\bar{\mu} - c}{3p^2} < a < \left(\frac{3(\mu - \bar{\lambda})p}{h_0C} - \frac{\mu}{\mu - \bar{\lambda}} \right) \frac{\lambda}{p^2}.$$

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

$$\begin{aligned} & \lambda'(p) - 2g(\mu) \frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu - \lambda(p))}{h_0C} \\ &= -2ap - 2 \left(\frac{\mu}{\mu - \lambda} - \frac{(\mu - \lambda)p}{h_0C} \right) \frac{\lambda}{p} + \frac{4\lambda(\mu - \lambda)}{h_0C} \\ &= 2p \left(\frac{\lambda}{p^2} \left(\frac{3(\mu - \lambda)p}{h_0C} - \frac{\mu}{\mu - \lambda} \right) - a \right). \end{aligned}$$

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

$$\frac{\lambda}{p^2} \left(\frac{3(\mu - \lambda)p}{h_0C} - \frac{\mu}{\mu - \lambda} \right) - a > \left(\frac{3(\underline{\mu} - \bar{\lambda})p}{h_0C} - \frac{\underline{\mu}}{\underline{\mu} - \bar{\lambda}} \right) \frac{\lambda}{\bar{p}^2} - a > 0,$$

which shows that (EC.16) holds.

EXAMPLE EC.3 (EXPONENTIAL DEMAND). Consider an exponential demand function

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

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

$$\begin{aligned} & \lambda'(p) - 2g(\mu) \frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu - \lambda(p))}{h_0C} \\ &= -b\lambda(p) - 2 \frac{\mu}{\mu - \lambda} \cdot \frac{b^2\lambda^2(p)}{-b\lambda(p)} + \frac{4\lambda(\mu - \lambda) - 2bp\lambda(\mu - \lambda)}{h_0C} \\ &> -b\lambda(p) + 2 \frac{\mu}{\mu - \lambda} b\lambda(p) > b\lambda(p) > 0. \end{aligned}$$

Therefore, (EC.16) holds as well.

EXAMPLE EC.4 (LOGIT DEMAND). Consider a logit demand function

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

We have

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

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

$$\begin{aligned} & \lambda'(p) - 2g(\mu) \frac{\lambda''(p)\lambda(p)}{\lambda'(p)} + \frac{4\lambda(p)(\mu - \lambda(p))}{h_0C} \\ &= \left(-\frac{b}{1+e} + \frac{2\mu(1-e)b}{(\mu-\lambda)(1+e)} - \frac{2p(\mu-\lambda)}{h_0C} \cdot \frac{b(1-e)}{1+e} + \frac{4(\mu-\lambda)}{h_0C} \right) \cdot \lambda \\ &> \left(-\frac{b}{1+e} + \frac{\mu b}{(\mu-\lambda)(1+e)} - \frac{2bp(1-e)}{1+e} \frac{(\mu-\lambda)}{h_0C} + \frac{4(\mu-\lambda)}{h_0C} \right) \cdot \lambda > 0, \end{aligned}$$

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

EC.5. Additional Numerical Experiments

EC.5.1. Robustness of LiQUAR

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

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

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

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

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.

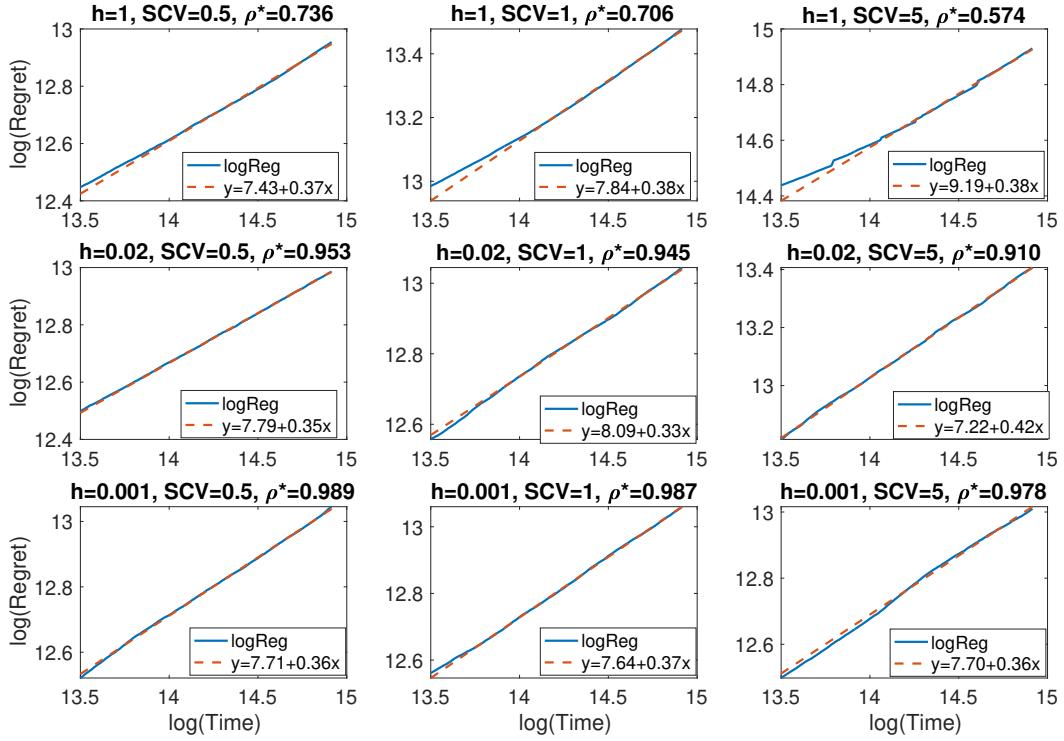


Figure EC.1 The regret curve in logarithm scale and a linear fit for the $M/GI/1$ model, under different traffic intensity $\rho^* \in [0.547, 0.989]$ and service-time SCV $c_s^2 = 0.5$ (E_2 service), 1 (M service) and 5 (H_2 service). All curves are estimated by averaging 100 independent runs.

EC.5.2. Relaxing the Uniform Stable Condition

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

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

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

with $a = 1 + \log(2)$ and $b = 1$. In addition, we set $h = 0.005$. We set the initial $\bar{p}_0 = 1.55$ so that the initial traffic intensity $\rho_0 = \lambda(p_0)/\mu = 1.15 > 1$ succeeds the critical level one, leading to an unstable system.

Algorithm 2: LiQUAR with backtracking

Input: number of iterations L , threshold τ , anchor price p_a ;
parameters $0 < \alpha < 1$, and T_k, η_k, δ_k for $k = 1, 2, \dots, L$;
initial value $\bar{p}_1, Q_1(0) = 0$;

```

1 for  $k = 1, 2, \dots, L$  do
2   Set control parameter  $p_{2k-1} = \bar{p}_k - \delta_k/2$  and Stable Sign= 0;
3   while  $\frac{1}{t} \int_0^t Q(t)dt < \tau$  for  $t < T_k$  do
4     | Run Cycle  $2k - 1$ : Run the system under  $p_{2k-1}$  ;
5   end
6   Set control parameter  $p_{2k} = \bar{p}_k + \delta_k/2$  ;
7   while  $\frac{1}{t} \int_0^t Q(t)dt < \tau$  for  $t < T_k$  do
8     | Run Cycle  $2k$ : Run the system under  $p_{2k}$  ;
9     | Stable Sign=1;
10  end
11 if Stable Sign=1 then
12   | Compute FD gradient estimator:
13   | 
$$H_k = \frac{1}{\delta_k} \left[ \frac{h_0}{(1-2\alpha)T_k} \int_{\alpha T_k}^{(1-\alpha)T_k} (Q_{2k}(t) - Q_{2k-1}(t)) dt - \frac{p_{2k}N_{2k} - p_{2k-1}N_{2k-1}}{T_k} \right]$$

14   | Update  $\bar{p}_{k+1} = \Pi_{[0, \infty)}(\bar{p}_k - \eta_k H_k)$ .
15 end
16 else
17   | Backtracking:  $\bar{p}_{k+1} = \frac{\bar{p}_k + p_a}{2}$ 
18 end
19 end

```

Following the analysis in Section 5.3, we set the hyperparameters $T_k = 2000k^{1/3}, \delta_k = 0.07k^{-1/3}, \eta_k = 0.21k^{-1}$ and $\tau = 141$ with the anchoring point $p_a = 1.84$. As is shown in the first two panels of EC.2, the pricing policy remain convergent to p^* . Consistently, the resulting traffic intensity ρ_k is quickly controlled to decrease below 1 although the system is unstable in the initial cycles.

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

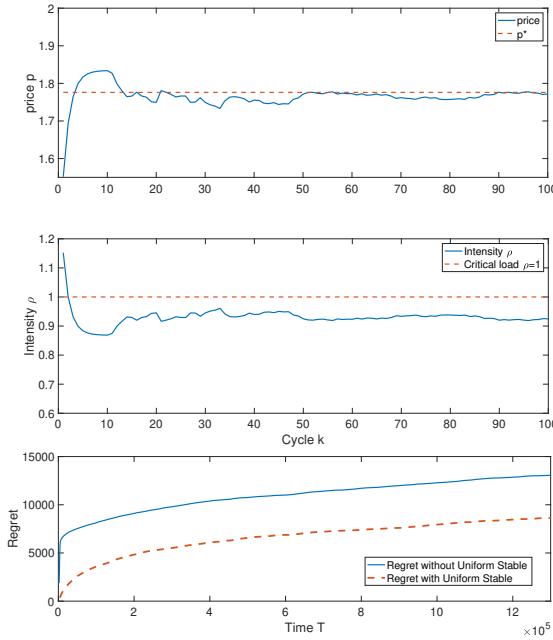


Figure EC.2 Pricing for the $M/M/1$ model without uniform stable condition for LiQUAR with backtracking: (i) sample path of price for LiQUAR with backtracking (top panel); (ii) sample path of traffic intensity ρ for LiQUAR with backtracking; (iii) regret comparison between LiQUAR with backtracking but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

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

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

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

EC.6.1. Base Policy Gradient

In this section, we provide the detailed description for Policy Gradient algorithms in Algorithm 3, the outline of which is described in Section 8. Specifically, in Algorithm 3, Policy Gradient algorithm organizes time by cycles, with each cycle containing L episodes. In each episode, the system operates the system following π_θ for episode length T time units. At the end of each episode, a gradient estimator in this episode

$\hat{\nabla}_{i,t}$ is calculated using the policy gradient formula (Sutton and Barto 2018, p.339) and the closed form of Gaussian parameterization (line 9 in Algorithm 3). Then, at the end of each cycle, an overall policy gradient estimator is obtained by averaging over all the episodic policy gradient estimators in the cycle (line 11 in Algorithm 3). The full algorithm is given in Algorithm 3.

Algorithm 3: PG-base

Input: normal parameterization $\pi(a|\theta)$, step size $\eta > 0$, initial policy parameter $\theta : (\bar{p}_1, \bar{\mu}_1, \sigma_{p,1}^2, \sigma_{\mu,1}^2)$, cycle length L (how many episodes in each episode), episode length T (how many time slots in each episode);

```

1 for each cycle do
2   for episode  $i = 1 : L$  do
3     Generate an episode  $Q_1, (p_1, \mu_1), R_1, \dots, Q_{T-1}, (p_T, \mu_T), R_T$  following  $\pi_\theta$ ;
4      $\bar{R} = \frac{1}{T} \sum_{t=1}^T R_t$ ;
5     for  $t = 1, \dots, T$  do
6        $G = \sum_{k=t}^T R_k - \bar{R}$ ;
7        $\hat{\nabla}_{i,t} \leftarrow G \cdot \begin{pmatrix} (p_t - \bar{p})/\sigma_p^2 \\ (\mu_t - \bar{\mu})/\sigma_\mu^2 \\ [(p_t - \bar{p})^2 - \sigma_p^2]/\sigma_p^3 \\ [(\mu_t - \bar{\mu})^2 - \sigma_\mu^2]/\sigma_\mu^3 \end{pmatrix}$ ;
8     end
9      $\hat{\nabla}_i = \frac{1}{T} \sum_{t=1}^T \hat{\nabla}_{i,t}$ ;
10   end
11    $\theta \leftarrow \theta + \eta \cdot \frac{1}{L} \sum_{i=1}^L \hat{\nabla}_i$ ;
12 end

```

EC.6.2. Policy Gradient with SAGE

In this section, we introduce the policy gradient algorithm with Score-Aware Gradient Estimates (PG-SAGE) in Comte et al. (2023). Following Comte et al. (2023), similar to the policy gradient theorem, $\nabla_\theta h_{\pi_\theta}$ could be represented as

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

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

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

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

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

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

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

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

Algorithm 4: PG-SAGE

Input: normal parameterization $\pi(a|\theta)$, step size $\eta > 0$, initial policy parameter $\theta : (\bar{p}_1, \bar{\mu}_1, \sigma_{p,1}^2, \sigma_{\mu,1}^2)$, cycle length L (how many episodes in each episode), episode length T (how many time slots in each episode), score oracle: $\mathcal{S}(p) = \lambda'(p)/\lambda(p)$;

```

1 for each cycle do
2   for episode  $i = 1 : L$  do
3     Generate an episode  $Q_1, (p_1, \mu_1), R_1, \dots, Q_{T-1}, (p_T, \mu_T), R_T$  following  $\pi_\theta$ ;
4      $\bar{R} = \frac{1}{T} \sum_{t=1}^T R_t$ ;
5     for  $t = 1, \dots, T$  do
6        $\hat{\nabla}_{i,t} \leftarrow R_t \cdot \begin{pmatrix} (p_t - \bar{p})/\sigma_p^2 \\ (\mu_t - \bar{\mu})/\sigma_\mu^2 \\ [(p_t - \bar{p})^2 - \sigma_p^2]/\sigma_p^3 \\ [(\mu_t - \bar{p})^2 - \sigma_\mu^2]/\sigma_\mu^3 \end{pmatrix}$ ;
7     end
8      $\hat{\nabla}_i = \frac{1}{T} \sum_{t=1}^T \hat{\nabla}_{i,t}$ ;
9      $\hat{\nabla}_i^{SAGE} = \text{Cov}(R_t, S_t, t = 1 : T) \cdot (\mathcal{S}(\bar{p}); -1/\bar{\mu}; 0; 0)^T$ ;
10    end
11     $\theta \leftarrow \theta + \eta \cdot \frac{1}{L} \sum_{i=1}^L (\hat{\nabla}_i + \hat{\nabla}_i^{SAGE})$ ;
12  end

```

In Figure 10, we report the best regret curves of PG-base and PG-SAGE methods for better exposition. From the Figure 10, we find out that compared with PG-base method, the PG-SAGE with constant step sizes does improve the performance of policy gradient with lower regret. In addition, the PG-SAGE method

with time-dependent step sizes seems to have a better convergence in the end (flat regret growth) but has a larger regret in the beginning due to the larger step sizes in the beginning. Overall, LiQUAR outperforms the Policy Gradient methods in this experiments, as LiQUAR's design has carefully taken the structure of queueing systems into consideration.

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

Table EC.1 Glossary of key notations