Online Learning and Pricing with Reusable Resources: Linear Bandits with Sub-Exponential Rewards

Huiwen Jia[1] Cong Shi[1] Siqian Shen[1]

Abstract

We consider a price-based revenue management problem with reusable resources over a finite time horizon T. The problem finds important applications in car/bicycle rental, ridesharing, cloud computing, and hospitality management. Customers arrive following a price-dependent Poisson process and each customer requests one unit of c homogeneous reusable resources. If there is an available unit, the customer gets served within a price-dependent exponentially distributed service time; otherwise, she waits in a queue until the next available unit. The decision maker assumes that the inter-arrival and service intervals have an unknown linear dependence on a d_f-dimensional feature vector associated with the posted price. We propose a rate-optimal online learning and pricing algorithm, termed Batch Linear Confidence Bound (BLinUCB), and prove that the cumulative regret is $\tilde{O}(d_f\sqrt{T})$. In establishing the regret, we bound the transient system performance upon price changes via a coupling argument, and also generalize linear bandits to accommodate sub-exponential rewards.

1. Introduction

Revenue management with reusable resources finds a wide range of applications in modern economy that heavily involves resource sharing. In these applications, a firm is endowed with a finite capacity of reusable products and each customer requests a product, uses it for some random time, and returns it to the firm, at which point the product unit can be used by other customers (see Levi & Radovanović, 2010; Chen et al., 2017; Rusmevichientong et al., 2020). For example, firms such as Amazon and Microsoft offer cloud computing services and users utilize virtual machines for a certain duration of time to finish computing tasks (see Kaewpuang et al., 2013; Püschel et al., 2015). The exact usage duration of a request in a cloud is not specified *a priori* and thus the departure of demand is also stochastic. The firm needs to dynamically decide what price to offer to users based on the current capacity utilization as well as the received requests (see Doan et al., 2020). More examples can be found in car/bicycle rental business (Oliveira et al., 2018), parking facility management (Owen & Simchi-Levi, 2018), and on-demand service platforms, such as ride-hailing services including Uber and Lyft (Banerjee et al., 2015; Liu & Li, 2017; Bimpikis et al., 2019). In all the aforementioned applications, customers' willingness for using the service is affected by the price being offered; however, the decision maker may not know the demand distribution and how demand reacts to price changes (see Xu & Li, 2013). The problem is even more challenging when the candidate price set is continuous (i.e., learning an optimal pricing strategy over a continuous action space).

Brief Problem Statement. We consider the following problem: A service-providing firm is endowed with a pool of c homogeneous reusable resources and can dynamically post prices (for this pool) over a finite horizon T. Customers arrive following a price-dependent Poisson process, and request one unit from c reusable resources. If there is a resource unit available for use, the arriving customer gets the unit and spends a random time that is exponentially distributed. Upon service completion, this customer pays the firm the amount of the posted price multiplied by the actual service time, and the resource unit is freed up to serve other customers. If all units are occupied upon arrival, the customer waits in a queue until the next available unit. The goal is to find the optimal pricing policy that maximizes the total expected revenue. In this paper, we assume that the inter-arrival and service intervals have an unknown linear dependence on a d_f-dimensional feature vector associated with the posted price. Thus, we need to learn the underlying linear functions, while maximizing the total expected revenue on the fly. The performance measure is cumulative regret, which is the difference between the revenue attained by a learning algorithm and by a clairvoyant optimal pricing policy under full distributional information.

[1] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109. Correspondence to: Cong Shi <shicong@umich.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Main Result and Contributions. We propose a Batch Linear Confidence Bound (BLinUCB) algorithm and prove that the cumulative regret is $\tilde{O}(d_f\sqrt{T})$, which matches the lower bound up to a logarithmic factor.

The state-of-the-art literature in revenue management predominantly focuses on perishable resources (i.e., the unit sold is gone and cannot be re-used by other customers) and the problem then belongs to the general class of online knapsack problems (e.g., Ferreira et al., 2018; Chen & Shi, 2019b). In contrast, the reusable resource setting naturally embeds multi-server queueing systems that are more difficult to analyze. One major complication is that whenever we change a price, the system still contains "old" customers from the previously posted price and also takes time to reach steady state when the new stream of customers comes in (so as to extract the performance under the newly posted price).

On a high level, our algorithm separates the learning horizon into successive batches and selects a price using past sales collected in previous batches. For each batch, we need to further divide it into two intervals. The first interval is for (i) completing serving the existing customers (under previously posted prices) and reaching the steady state of the corresponding queue under the newly posted price and (ii) eliminating the heavy-tailed effects brought by sub-exponential observations. The system maintains a steady state in the second interval.

This paper develops a new linear-bandits-based approach for revenue management of reusable resources under a *continuous* price set. We highlight two main techniques in establishing the optimal regret upper bound. First, leveraging the coupling arguments developed by Jia et al. (2020), we bound the *loss of nonstationarity* due to transient system performance. Whenever a new price is posted, the underlying queueing dynamics change, and it takes time for the system to reach steady state. We analyze the length of this transient period for M/M/c queues upon price changes (see Propositions 4 and 5). Second, the typical linear bandits assume sub-Gaussian errors, which allow these algorithms to be directly integrated with linear (ridge) regression techniques. Our regret analysis provides a theoretical study of linear bandits with *sub-exponential* observations and contributes to the broad landscape of linear bandit literature by delineating the impact of heavy-tailed rewards. Our BLinUCB accumulates a carefully designed set of observations for invoking Bernstein's inequality. We use the empirical statistics as data points to update the estimations of regression coefficients with a probabilistic guarantee (see Propositions 2 and 3).

2. Literature Review

Our work is closely related to the following streams of literature.

Revenue Management with Reusable Resources. The majority of the prior revenue management literature studies perishable resources (den Boer, 2015; Deng et al., 2020; 2021). Here we only focus on the literature of revenue management with reusable resources. One key challenge in the reusable resource setting is the need to dynamically match demands with the varying inventory of reusable resources. Several recent studies developed provably near-optimal heuristic admission controls (see, e.g, Levi & Radovanović, 2010; Chen & Shi, 2017; Chen et al., 2017). Besides admission control, there has been a stream of literature considering static and dynamic pricing for both single resource setting (see, e.g., Maglaras, 2006; Araman & Caldentey, 2009; Xu & Li, 2013; Besbes et al., 2022) and multi-resource setting (Doan et al., 2020; Lei & Jasin, 2020; Owen & Simchi-Levi, 2018; Rusmevichientong et al., 2020). The aforementioned literature assumed that the distributional information of the underlying model is known to the decision-maker *a priori* and there has been very little work considering the incomplete information. To the best of our knowledge, the only two learning papers are by Chen et al. (2020b) and Jia et al. (2020). The former developed a stochastic gradient descent algorithm, which is different from our proposed linear bandit algorithm in this paper. Jia et al. (2020) applied multi-armed bandit techniques and involved intertemporal dynamics (i.e., customers waiting, getting served, and leaving) but it was restricted to a finite discrete price set. In contrast, our work considers a continuous price space, which is highly non-trivial to analyze. The concentration bounds for linear bandits under our setting are new and the analysis for sub-optimality is also new (showing how the feature vector norm over the data matrix evolves).

Linear Bandits. There have been several well-developed studies on linear bandits (see, e.g., Li et al., 2010; Abbasi-Yadkori et al., 2011; Tao et al., 2018; Alieva et al., 2021), including various focuses, such as generalization to nonlinear bandits (Filippi et al., 2010), extensions to delayed rewards (Zhou et al., 2019), multi-task linear bandit (Cella et al., 2020; Hu et al., 2021), ans so on. All these general linear bandit algorithms assume sub-Gaussian errors (see, e.g., Assumption 1 in Rusmevichientong & Tsitsiklis, 2010), which allow these algorithms to be directly integrated with linear (ridge) regression techniques. To the best of our knowledge, our regret analysis provides the first theoretical study of linear bandits with sub-exponential observations (which can also be reviewed as sub-exponential rewards) and contributes to the broad landscape of linear bandit literature by delineating the impact of heavy-tailed rewards.

Bandit Problems with Infrequent Action Changes. Bandit problems with infrequent action changes have also been studied in recent years, varying from static batch design (Perchet et al., 2016) to adaptively determined batch size (Gao et al., 2019). Several literature considers switching cost explicitly (Cesa-Bianchi et al., 2013; Simchi-Levi & Xu, 2019; Han et al., 2020). Specifically, Cheung et al. (2017); Chen & Chao (2019); Chen et al. (2020a) discussed practical reasons of infrequent price changes in revenue management and developed pricing algorithms with limited price switches. In our work, we consider a static batch design with $O(\log(T))$ batches, which is in the same order as those in Auer et al. (2002); Auer & Ortner (2010); Gao et al. (2019). The departure from related literature is that we do not enforce a strict budget on action changes nor an objective for minimizing the action changes.

Reinforcement Learning via MDP. The process that a firm decides prices based on the current number of customers in the system can also be abstracted as a Markov Decision Process (MDP). Reinforcement Learning (RL) is widely used for solving MDP (Szepesvári, 2010; Sutton & Barto, 2018). However, popular RL approaches, such as Q-Learning (Even-Dar et al., 2003), UCRL2 (Auer et al., 2008), and Thompson Sampling for RL (Russo et al., 2018), all assume a discounted summation relationship between immediately observable rewards and the state-action value function, which does not hold in our setting. The key point is that adapting any of the above methods to our setting would require estimation of the value function under steady state conditions, which necessitates the coupling analysis. Our framework opens many doors to conducting RL in complex stochastic systems.

Transient Analysis of Queues. The analysis of M/M/c queues under transient state has a long history. Although the distribution is explicitly known in terms of modified Bessel functions of the first kind, further studies were conducted to understand how the queue evolves over time (Ledermann & Reuter, 1954; Abate & Whitt, 1987; 1988; Bailey, 1954; Kelton & Law, 1985; Morisaku, 1976; Parthasarathy & Sharafali, 1989). Kelton & Law (1985) pointed out that a queue needs to run "long-enough" to dissipate the initial starting point and to collect observations during the ensuing "steady-state" portion of the run. Our result gives the first finite sample bound on mixing times of M/M/c queues (upon action changes). Specifically, we develop a coupling argument between this target system and a virtual system starting from a state sampled from the steady-state distribution and maintaining the steady state thereafter.

3. Problem Formulation

We consider a service-providing firm who supplies customers with reusable resources of finite capacity c over a finite horizon T. At the beginning of period $t \leq T$, the firm posts a price p_t between a fixed range $[p_L, p_U]$ to maximize the total cumulative revenue. Under the posted price p, $\forall p \in [p_L, p_U]$, customers arrive at the system according to a Poisson process with rate λ_p and they are served on a first-arrive-first-serve basis by occupying one unit of the resource with the service time following an exponential distribution with rate μ_p. Customers pay p for per unit time of service. If there is not enough resource capacity, the customer will join the queue until being served. Note that the service system under any posted price $p \in [p_L, p_U]$ can be reduced to an M/M/c queue (when the system only contains customers arriving under this price, i.e., after finishing serving other customers arriving under the previously posted prices). The assumption of exponential service time is standard in the literature (see, e.g., Savin et al., 2005; Gans & Savin, 2007; Owen & Simchi-Levi, 2018).

We consider a linear relationship between feature vectors and arrival/service rates. More precisely, for every price $p \in [p_L, p_U]$, the firm can observe a feature vector $\mathbf{x}_p \in \mathbb{R}^{d_f}$. For example, a simple non-trivial case is when the rates are typically linear on price itself (i.e., $d_f = 1$), and this example is well supported by literature (see Mankiw, 2014); and $d_f = 3$ can include price itself, a major competitor's price, and neighborhood income level and willing to pay. The arrival rate λ_p and the service rate μ_p have the following relationship with the feature vector $\mathbf{x}_p$:

$$\frac{1}{\lambda_p} = \theta_\lambda^T \mathbf{x}_p \text{ and } \frac{1}{\mu_p} = \theta_\mu^T \mathbf{x}_p,$$

where $\theta_\lambda \in \mathbb{R}^{d_f}$ and $\theta_\mu \in \mathbb{R}^{d_f}$ are two unknown coefficient vectors. Recall that $1/\lambda_p$ is the average time interval between two consecutively arrived customers and $1/\mu_p$ is the average service time under price p.

We assume that the firm does not know the underlying coefficients θ_λ and θ_μ *a priori*. The firm aims at finding a periodic review pricing policy as $\pi : \{(n,t) : n = 0, 1, \ldots, \infty,\ t = 1, \ldots, T\} \to [p_L, p_U]$, where the firm selects price $\pi(n,t) \in [p_L, p_U]$ for period t when there are n customers in the system at the beginning of period t. The goal is to maximize the expected total revenue. The expected revenue under pricing policy π during period t is denoted by J_t^π and the cumulative expected revenue over periods $\{1, \ldots, T\}$ is denoted by $J^\pi = \sum_{t=1}^T J_t^\pi$.

Assumption 1.

1. The utilization factor $\rho_p = \frac{\lambda_p}{c\mu_p} < 1$ for any candidate price $p \in [p_L, p_U]$.
2. $\log(T) \geq 4$.
3. A constant $r_{\max}$ is known such that $r_{\max} \geq \max_{p \in [p_L, p_U]} \lambda_p / \mu_p$.

Assumption 1 says the following: 1) the system is stable for any candidate price; 2) we require the planning horizon to be sufficiently long, since it takes time for the underlying service system to reach a steady state under any posted price; 3) this assumption can be simply satisfied by letting

$r_{\max} = c$ and it further implies that a valid upper bound of the stationary revenue rate of all candidate prices is known.

Assumption 2. For any two candidate prices $p_h,\ p_l \in [p_L, p_U]$ and $p_h > p_l$, we have:

1. $\lambda_{p_l} \geq \lambda_{p_h}$.
2. $\lambda_{p_l} - \lambda_{p_h} \leq \frac{\rho_l}{-3e \log \rho_l}$.
3. if $\mu_{p_l} > \mu_{p_h}$, then $\rho_l \geq \rho_h^2$.

Assumption 2 says the following: 1) the arrival rate under a lower price is higher than or equal to that under a higher price; 2) the difference in arrival rates between two prices is bounded (by a large constant); and 3) if the service rate under a lower price is higher than that under a higher price, the utilization factor in the former case must be greater than the latter squared. This condition (which is later used to justify inequalities (18g) and (24c)) is rather mild in the following sense: if $\mu_{p_l} \leq \mu_{p_h}$, we do not need any conditions; otherwise, if $\mu_{p_l} > \mu_{p_h}$, a simple sufficient condition to guarantee that $\rho_l \geq \rho_h^2$ is $\rho_l \geq \rho_h$, which reduces to $\lambda_{p_l}/\lambda_{p_h} \geq \mu_{p_l}/\mu_{p_h}$, meaning that the ratio of arrival-rate changes (with respect to price) exceeds the ratio of change in service-rate changes.

3.1. Regret, Relaxed Regret, and LP Benchmark

The notion of cumulative regret is commonly used in online learning (see Shalev-Shwartz et al., 2011) to evaluate the performance of a policy if the decision-maker has limited information of the system against the optimal performance under full information. In our problem, the full information means that the firm knows the underlying mappings between price $p,\ \forall p \in [p_L, p_U]$ and the associated arrival rate λ_p and service rate μ_p, i.e., the firm knows θ_λ and θ_μ. Under full information, we have state-dependent optimal policy $\pi^* = \text{argmax}_\pi J^\pi$ and state-dependent optimal expected revenue $J^* = J^{\pi^*}$. Thus, we can define the regret of any pricing policy π as the difference between the total state-dependent optimal expected revenue J^* and the total expected revenue under the given policy π, i.e., J^π. In short, we aim at finding heuristic pricing policies that lead to a small regret.

Definition 1. The cumulative regret by the end of period T of policy π is defined as:

$$\text{Regret}(\pi, T) = J^* - J^\pi.$$

However, finding the state-dependent optimal pricing policy π^* requires the firm to solve a dynamic program with an infinite number of potential states, in the form of (n, t), under uncertain parameters, which is computationally intractable. Moreover, the transient system performance is complex and hard to analyze when the offered price changes over periods. Thus, it is impractical to solve this dynamic program, and further impossible to obtain π^* and J^*. To tackle this problem, we develop an upper bound on the expected revenue achievable by any policy, and therefore, also an upper bound of the state-dependent optimal expected revenue.

We uniformly discretize $[p_L, p_U]$ as a set $\mathcal{P}$ where the discretizing interval can be arbitrarily small. Consider continuous decision variables π_{nt}^p for $n = 0, \ldots, \infty,\ t = 1, \ldots, T,\ p \in \mathcal{P},\ a_{nt}$ for $n = 0, \ldots, \infty,\ t = 1, \ldots, T$, and J_{nt} for $n = 0, \ldots, \infty,\ t = 1, \ldots, T$. Let variable a_{nt} represent the probability that the system has n customers at the beginning of period t under an arbitrary policy; let variable π_{nt}^p represent the probability that this policy chooses price p at state (n, t); and constraints (1e) show that variable J_{nt} take the value at most the expected revenue collected before the end of period T from the customers who arrive during period t at state (n, t). Thus, we can formulate a linear program (LP) as follows.

$$J^{\text{LP}} = \tag{1a}$$

$$\max_{\pi, a, J} \sum_{n=0}^{\infty} \sum_{t=1}^{T} a_{nt} J_{nt} \tag{1b}$$

$$\text{s.t.} \sum_{n=0}^{\infty} a_{nt} = 1 \ \forall t = 1, \ldots, T \tag{1c}$$

$$\sum_{p \in \mathcal{P}} \pi_{nt}^p = 1 \ \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T \tag{1d}$$

$$J_{nt} \leq \sum_{p \in \mathcal{P}} \pi_{nt}^p \frac{\lambda_p}{\mu_p} p \ \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T \tag{1e}$$

$$0 \leq a_{nt} \leq 1 \ \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T \tag{1f}$$

$$0 \leq \pi_{nt}^p \leq 1 \ \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T,\ \forall p \in \mathcal{P}. \tag{1g}$$

Proposition 1. *The LP solution provides an upper bound on the optimal revenue, i.e., $J^* \leq J^{LP}$.*

Proof sketch: We show that the decisions variables (π, a, J) associated with any admissible policy satisfy the constraints of the linear program (1), and the expected revenue of this admissible policy is exactly the corresponding objective value in (1). Recall that the discretizing interval of $\mathcal{P}$ can be arbitrarily small and thus $J^{\text{LP}} \geq J^\pi$ for any admissible policy π as well as the optimal state-dependent policy π^*. The detailed proof can be found in Appendix A.

By defining $\tilde{p} = \text{argmax}_{p \in [p_L, p_U]} \frac{\lambda_p}{\mu_p} p$, we find an optimal solution $(\tilde{\pi}, \tilde{a}, \tilde{J})$ to the LP in (1) as follows (we discretize $[p_L, p_U]$ as $\mathcal{P}$ such that $\tilde{p} \in \mathcal{P}$).

$$\begin{aligned}
&\tilde{\pi}_{nt}^{\tilde{p}} = 1, \quad \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T \\
&\tilde{\pi}_{nt}^{p} = 0, \quad \forall n = 0, \ldots, \infty,\ \forall t = 1, \ldots, T,\ \forall p \in \mathcal{P} \backslash \{\tilde{p}\} \\
&\tilde{J}_{0t} = \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}} \tilde{p} \quad \forall t = 1, \ldots, T \\
&\tilde{J}_{nt} = 0 \quad \forall n = 1, \ldots, \infty,\ \forall t = 1, \ldots, T \\
&\tilde{a}_{0t} = 1 \quad \forall t = 1, \ldots, T \\
&\tilde{a}_{nt} = 0 \quad \forall n = 1, \ldots, \infty,\ \forall t = 1, \ldots, T.
\end{aligned}$$

We can compute $J^{\text{LP}} = T\frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}}\tilde{p}$. Note that $\frac{\lambda_p}{\mu_p}p$ is the stationary revenue rate of price p, denoted as $r(\lambda_p, \mu_p, p)$. Thus, we call the price $\tilde{p}$ the (static) optimal state-independent price, i.e., the price with the highest stationary revenue rate. This result can be seen as the "reusable analog" of the classical static pricing upper bound for revenue management with perishable resources (see Gallego & Van Ryzin, 1994).

Corollary 1. *The single static price policy $\pi^{\tilde{p}} : \pi(n, t) = \tilde{p},\ \forall n = 0, 1, \ldots, \infty,\ \forall t = 1, \ldots, T$ is asymptotically optimal as $T \to \infty$, and we also have $J^{LP} - J^{\pi^{\tilde{p}}} \leq o(\sqrt{T})$.*

This corollary follows directly from Propositions 1 and 4 and we provide the detailed proof in Appendix A. From Corollary 1, for any finite T, the loss between the static benchmark and any arbitrarily complex state-dependent policy is roughly on the order of $\log(T)$, which is insignificant compared to the tight $\sqrt{T}$ loss due to learning a static policy. As a result, learning a static pricing policy is already provably near-optimal and also arguably more implementable and fairer in practical settings.

Definition 2. The relaxed regret is defined as

$$\overline{\text{Regret}}(\pi, T) = J^{\text{LP}} - J^{\pi}.$$

Based on Proposition 1, we have that for any policy π, the corresponding relaxed regret is a valid upper bound of its regret, i.e., $\overline{\text{Regret}}(\pi, T) \geq \text{Regret}(\pi, T)$.

4. Online Learning and Pricing: BLinUCB

Without loss of generality, we consider a known feature generation function $F(\cdot) : [p_L, p_U] \to \mathbb{R}^{d_f}$, which gives the firm the feature vector $\mathbf{x}_p$ of a price $p \in [p_L, p_U]$.

4.1. M/M/c Queue Parameters under Linear Relationship

Consider a price $p \in [p_L, p_U]$ with a feature vector $\mathbf{x}_p$. Without loss of generality, we first analyze the arrival process. The arrival time interval is a random variable following an exponential distribution with mean $1/\lambda_p = \theta_\lambda^T \mathbf{x}_p$. Consider $n_m(p)$ observations of arrival times $\hat{d}_i(p)$, $i = 1, \ldots, n_m(p)$ and denote $\bar{d}_p$ as the empirical mean of arrival time intervals $\bar{d}_p = \sum_{i=1}^{n_m(p)} \hat{d}_i(p)/n_m(p)$. Then the random variable $\bar{d}_p$ follows an Erlang distribution, $\text{Erlang}(n_m(p), n_m(p)\lambda_p)$. For each implemented price p, we have a set of correspondent data with it, which can be represented by a tuple $(p, \mathbf{x}_p, \bar{d}_p, n_m(p))$.

Let SE denote sub-exponentials. By Lemma 2, $\hat{d}_i(p) \sim \text{SE}(4/\lambda_p^2, 2/\lambda_p)$. Then derived from Lemmas 3 and 4 (in Appendix B), we have

$$\bar{d}_p \sim \text{SE}(4/(n_m(p)\lambda_p^2), 2/(n_m(p)\lambda_p)).$$

Therefore, we can equivalently write it as

$$\bar{d}_p = \theta_\lambda^T \mathbf{x}_p + \epsilon_p, \tag{2}$$

where the random error term

$$\epsilon_p \sim \text{SE}\left(4/(n_m(p)\lambda_p^2), 2/(n_m(p)\lambda_p)\right).$$

Furthermore, by analyzing the Erlang distribution, we can derive that the mean of ϵ_p is 0 and the variance of ϵ_p is $1/(n_m(p)\lambda_p^2)$, where $1/\lambda_p = \theta_\lambda^T \mathbf{x}_p$. Based on the weighted least squares approach (Seber & Lee, 2012) for linear regression models with heteroscedasticity, we estimate the coefficient θ_λ as follows. Consider we have observations of N different prices and $N \geq d_f$. Define a diagonal matrix Ω where the i^{th} element is the variance of the error term of the i^{th} implemented price, i.e., $1/(n_m(p_i)\lambda_{p_i}^2)$. Note that λ_p is unknown and we use an approximate matrix $\hat{\Omega}$ to substitute Ω later. We use a matrix $\mathbf{X} \in \mathbb{R}^{N \times d_f}$ to denote the features of implemented prices and use a vector $\mathbf{d} \in \mathbb{R}^N$ to denote the empirical arrival time means of implemented prices. Then we can estimate the unknown coefficients by

$$\hat{\theta}_\lambda = (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Omega^{-1} \mathbf{d}. \tag{3}$$

By letting $\vec{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$, we can write $\hat{\theta}_\lambda = \theta_\lambda + (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Omega^{-1} \vec{\epsilon}$. In the following analysis, we assume that the dependency between X and $\vec{\epsilon}$ is negligible, due to batching. If such dependency is not assumed away, one can develop a sub-exponential self-normalized bound of vector-valued martingales for $||\mathbf{X}^T \Omega^{-1} \vec{\epsilon}||^2_{(\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1}}$ following the scheme in Abbasi-Yadkori et al. (2011).

With $N \geq d_f$, $\mathbf{X}^T \Omega^{-1} \mathbf{X}$ is non-singular and thus $\hat{\theta}_\lambda$ is well-defined. In addition, we can easily prove that $\mathbb{E}[\hat{\theta}_\lambda] = \theta_\lambda$ and $\text{Var}(\hat{\theta}_\lambda) = (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1}$. For any price p' with a feature vector $\mathbf{x}'$, we can estimate $\theta_\lambda^T \mathbf{x}'$ by $\hat{\theta}_\lambda^T \mathbf{x}'$ and further derive an upper confidence bound for $\theta_\lambda^T \mathbf{x}'$.

Proposition 2. ***(Sub-exponential Tail Bound for Poisson Process under Linear Relationship.)*** *Consider N implemented prices with $N \geq d_f$ and $n_m(p) \geq 8\log(T)$ for any implemented price p. Then, for any new valid feature vector $\mathbf{x}'$ and $\hat{\theta}_\lambda$ computed in (3), one has*

$$\mathbb{P}\left(\frac{|\hat{\theta}_\lambda^T \mathbf{x}' - \theta_\lambda^T \mathbf{x}'|}{\sqrt{\mathbf{x}'^T (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{x}'}} \geq \sqrt{32\log(T)}\right) \leq \frac{2}{T^4}. \tag{4}$$

Proof sketch: Let matrix $A = (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Omega^{-1}$. Use column vector $a_{\cdot i}$ to denote the i^{th} column of A and a_{di} to denote elements of A. We firstly show that the d^{th} element of $\hat{\theta}_\lambda$ follows $\text{SE}(\sum_{i=1}^N a_{di}^2(4/n_i\lambda_i^2), \max_{i=1,\ldots,N} \quad 2|a_{di}|/(n_i\lambda_i))$ with mean $\theta_{\lambda,d}$. We further show that $\hat{\theta}_\lambda^T \mathbf{x}'$ follows $\text{SE}(4\mathbf{x}'^T(\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1}\mathbf{x}', \max_{i=1,\ldots,N} 2|\mathbf{x}'^T a_{\cdot i}|/(n_i\lambda_i))$

with mean $\theta_\lambda^T \mathbf{x}'$. With $n_m(p) \geq 8\log(T)$, we can further prove that it meets the requirement for applying the sub-Gaussian behavior type of concentration bound for SE variables (this inequality can be found in literature, e.g., Boucheron et al., 2013; Rigollet & Hütter, 2015). We provide the detailed proof in Appendix B. The essential idea here is to eliminate the heavy-tailed effects by accumulating observations (see Jia et al., 2021).

In the above analysis, we consider a known covariance matrix Ω. However, in the reality, this matrix is also unknown. Therefore, we estimate it with $\hat{\Omega}$ (also denoted by $\hat{\Omega}_\lambda$ for arrival process). We estimate the i^{th} element, i.e., the variance of the error term of the i^{th} implemented price, by $\bar{d}_{p_i}^2/(n_m(p_i))$. By the linear regression analysis with heteroscedasticity (see, e.g., Seber & Lee, 2012), the estimated coefficient

$$\hat{\theta}_\lambda = (\mathbf{X}^T \hat{\Omega}_\lambda^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Omega}_\lambda^{-1} \mathbf{d} \tag{5}$$

shares the same properties as the estimated coefficient in (3). Therefore, Proposition 2 also holds for $\hat{\theta}$ defined in (5). In the following analysis, we use $\hat{\Omega}$ to substitute Ω for both arrival and service processes.

We apply the same technique to the service process. Use $n_m^s(p)$ to denote the number of customers who have been successfully serviced and $\hat{g}_i(p)$ to denote the observed service time of customer i. Define $\hat{\Omega}_\mu$ and $\mathbf{y}_\mu$ similarly and therefore we can compute an estimation as

$$\hat{\theta}_\mu = (\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{y}_\mu. \tag{6}$$

Therefore, we can reach the same concentration result:

$$\begin{aligned} &\mathbb{P}\left(\frac{|\hat{\theta}_\mu^T \mathbf{x}' - \theta_\mu^T \mathbf{x}'|}{\sqrt{\mathbf{x}'^T (\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X})^{-1} \mathbf{x}'}} \geq \sqrt{32\log(T)} \right) \\ &\leq 2\exp\left(-\frac{32\log(T) \cdot \mathbf{x}'^T (\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X})^{-1} \mathbf{x}'}{8\mathbf{x}'^T (\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X})^{-1} \mathbf{x}'} \right) \leq \frac{2}{T^4}. \end{aligned} \tag{7}$$

Proposition 3. *For price p with a feature vector $\mathbf{x}$, we have:*

$$\mathbb{P}\left(\left| \frac{\lambda_p}{\mu_p} - \frac{\hat{\theta}_\mu^T \mathbf{x}}{\hat{\theta}_\lambda^T \mathbf{x}} \right| \leq \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T \mathbf{x}} \mathcal{G} \right) \geq 1 - \frac{4}{T^4},$$

where

$$\mathcal{G} = \left(r_{\max} \sqrt{\mathbf{x}^T (\mathbf{X}^T \hat{\Omega}_\lambda^{-1} \mathbf{X})^{-1} \mathbf{x}} + \sqrt{\mathbf{x}^T (\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X})^{-1} \mathbf{x}} \right).$$

The proof of Proposition 3 is based on Proposition 2. The LHS can be lowered bounded by the product of two terms and both the two terms are in the same format as the LHS of Proposition 2. We provide detailed proof in Appendix B.

4.2. BLinUCB

We present Batch LinUCB (BLinUCB) in Algorithm 1. We divide the total horizon into two phases, *Warm-up* Phase and *Learning* Phase. In the Warm-up Phase, the algorithm gives a start of valid estimations of parameters θ_λ and θ_μ as computed in (5) and (6). To initiate invertible matrices $\mathbf{X}^T \hat{\Omega}_\mu^{-1} \mathbf{X}$ and $\mathbf{X}^T \hat{\Omega}_\lambda^{-1} \mathbf{X}$, we select d_f number of prices, whose feature vectors form a basis for span$(\mathbf{x}_p, p \in [p_L, p_U])$, and collect $8\log(T)$ number of arrival and service time observations. Denote this set of basis prices as $\mathcal{P}_b$. In the Learning Phase, the algorithm separates the time as consecutive batches where the length of batch $m = 1, \ldots, M$ is $I_m \tau$ with $I_m = 2^m$ and $\tau = (\log(T))^2$. At the beginning of each batch, BLinUCB selects a price over the range $[p_L, p_U]$ with the highest upper confidence bound of the revenue rate.

Definition 3. The upper confidence bound of the revenue rate associated with price p by the end of batch m is:

$$U_m(p) = \left(\frac{\hat{\theta}_\mu^T \mathbf{x}}{\hat{\theta}_\lambda^T \mathbf{x}} + \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T \mathbf{x}} \mathcal{G} \right) p, \tag{8}$$

and the lower confidence bound of the revenue rate of price p by the end of batch m is:

$$L_m(p) = \left(\frac{\hat{\theta}_\mu^T \mathbf{x}}{\hat{\theta}_\lambda^T \mathbf{x}} - \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T \mathbf{x}} \mathcal{G} \right) p. \tag{9}$$

Define $\mathbf{Rad}_m(p) = \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T \mathbf{x}} \mathcal{G} p$ as the confidence radius of price p by end of batch m.

5. Performance Analysis of BLinUCB

We analyze the performance by the relaxed regret defined in Definition 2. First, we derive the probability that a system reaches the steady state after a certain amount of time starting from (i) an empty state (i.e., zero customers in the system) under a fixed price (Proposition 4) and (ii) the steady state under another price (Proposition 5). Then, based on the probability analysis, we provide the regret bound for BLinUCB in Theorem 1.

5.1. High Probability Bound for Mixing Time

When the firm starts service from the empty state or switches to another price from the currently offered price, the service (queueing) system enters a transient state and needs a certain amount of time to reach the steady state again. Therefore, the time spent on reaching the steady state, which is also referred to as the mixing time, is crucial for computing the practical revenue under a specific policy and thus crucial for the regret analysis. To the best of our knowledge, this

Algorithm 1 Online Batch LinUCB Algorithm (BLinUCB).

```
Input: $T, p_L, p_U, d_f$.
Initialize: $\tau$, $I_m$, $M$, $\mathcal{P}_b$ as in Section 4.2.
Warm-up Phase:
for $p \in \mathcal{P}_b$ do
    Offer price $p$, record $\hat{d}_i(p)$ for arriving customers and
    $\hat{g}_i(p'),\ \forall p' \in [p_L, p_U]$ for leaving customers.
    if $n^s_m(p) \ge 8\log(T)$ then
        Update $\mathbf{X}, \hat{\Omega}_\lambda, \hat{\Omega}_\mu, \mathbf{d}, \mathbf{y}_\mu$
        Continue.
    end if
end for
Compute $\hat{\theta}_\lambda$ and $\hat{\theta}_\mu$ by (5) and (6)
Learning Phase:
for $m = 1, \ldots, M$ do
    Choose $p_m = \text{argmax}_{p \in [p_L, p_U]} U_{m-1}(p)$.
    Offer $p_m$ in batch $m$, i.e., for $I_m\tau$ periods.
    Record $\hat{d}_i(p_m)$ for arriving customers and
    $\hat{g}_i(p),\ \forall p \in \mathcal{P}$ for leaving customers.
    Update $\mathbf{X}, \hat{\Omega}_\lambda, \hat{\Omega}_\mu, \mathbf{d}, \mathbf{y}_\mu$; Compute $\hat{\theta}_\lambda$ and $\hat{\theta}_\mu$.
end for
```

is the first result giving a finite-time high probability bound on mixing times of M/M/c queues (upon action changes). Specifically, we develop a coupling argument between this target system and a virtual system starting from a state sampled from the steady-state distribution and maintaining the steady state thereafter.

Without loss of generality, we first focus on a system under fixed arrival rate λ and service rate μ. Let S_t denote the target system that starts from the empty state and $\hat{S}_t$ denote the virtual system that starts with a state sampled from steady-state distribution and remains thereafter. Let S_∞ denote the steady-state overshoot. The random variable $S_t\ (\hat{S}_t),\ t \ge 0$ denotes the number of customers in the system at time t of system $S_t\ (\hat{S}_t)$. The analysis in this section relies on busy period $A_n = \min\{t\colon n$ customers in the system at time 0^+ and $n-1$ customers in the system at time $t\},\ \forall n = \{1, \ldots, \infty\}$, (see Omahen & Marathe, 1978; Daley & Servi, 1998) and the first-order stochastic dominance between two random variables (see Hadar & Russell, 1969; Seth & Yalonetzky, 2014).

Proposition 4. ***(Adapted from Proposition 4 in Jia et al. (2020), Coupling Probability of M/M/c Queue starting from Empty State)*** *For $t \ge \tau$, where $\tau = (\log(T))^2$, then*

$$\mathbb{P}(S_t = \hat{S}_t \textit{ after time } \tau) \ge 1 - \frac{2}{T^2}.$$

The proof of Proposition 4 is provided in Appendix C, which follows the law of total probability, the length of the time horizon T in Assumption 1, the first-order stochastic dominance in Lemma 5, and the concentration inequality for independent samples in Lemma 6.

When implementing our learning algorithms, the M/M/c queue does not really start from empty state. This is because whenever a new price is posted, the customers who arrived under the previously posted price remain in the system. We need to ensure that the coupling occurs with high probability even with price changes (with the aid of Assumption 2), which is encapsulated in Proposition 5.

Proposition 5. ***(Adapted from Proposition 5 in Jia et al. (2020), Coupling Probability of M/M/c Queue when Price Changes)*** *The probability that the system reaches the steady state within 2τ after the price changes is bounded by $1 - \frac{4}{T^2}$, where $\tau = (\log(T))^2$.*

The proof of Proposition 5 is based on Proposition 4 and Assumption 2. Specifically, we decompose possible price changes into two cases, i.e., when the price changes from a lower price p_l to a higher price p_h and vice versa. For each case, we further consider two sub-cases, (i) the unit service rate $\mu_{p_l} > \mu_{p_h}$ and (ii) $\mu_{p_l} \le \mu_{p_h}$. For each case, we construct a virtual system and derive the coupling probability of the corresponding virtual system. Then we show that the actual coupling probability is bounded from below by the probability with which the virtual system reaches the steady state. The full proof is given in Appendix C.

5.2. Regret Bound

Theorem 1. *The T-period cumulative regret of BLinUCB is bounded by $\tilde{O}\left(d_f\sqrt{T}\right)$.*

For comparison, the state-of-the-art regret lower bound for linear bandits is of order $\Omega(d_f\sqrt{T})$ (see, e.g., Rusmevichientong & Tsitsiklis, 2010). Our result matches this lower bound up to a logarithmic factor.

Proof sketch: The proof consists of two parts, where we bound the regret of Warm-up and Learning Phases separately. The regret of the Warm-up phase is at most linear on the length of the Warm-up Phase, i.e., $O(d_f\log(T))$. The regret of the Learning Phase can be further decomposed into the loss of nonstationarity and suboptimality.

Loss of nonstationarity. With the key coupling results from Propositions 4 and 5, we can use the steady-state revenue rate to compute the expected revenue of each batch with at most a loss linear on 2τ. As a result, we have $J^{\text{LP}}_{\text{Learning, m}} - J^{\pi_{\text{BLinUCB}}}_{\text{Learning, m}} \le \Delta(p_m)I_m\tau + O(\tau)$, where $\Delta(p_m) = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$.

Loss of suboptimality. By Defintion 3, we can derive that $\Delta(p_m)$ can be upper bounded by the sum of three terms: $\Delta(p_m) \le \min\left\{\left(U_m(p_m) - L_m(p_m)\right), r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\right\} + \left(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - U_m(\tilde{p})\right) + \left(L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)\right)$. The first term is analyzed with the help of Lemmas 7 and

8, which are provided in Appendix D. The second and third terms can be easily bounded with the help of Proposition 3. We derive that the regret during the Learning Phase is $O\left(\log(T)\sqrt{d_f T\log(T)}\right)$. Combine the results together, we have the regret of BLinUCB algorithm is $\tilde{O}\left(d_f\sqrt{T}\right)$.

6. Numerical Experiments

Experimental Setup. The total operation time horizon is 8000 periods and the capacity of the reusable resource is $c = 100$. We choose the price from a fixed range of $[10, 18]$, of which the corresponding service rates are equal. We consider a three-dimensional feature vector $(p, \phi(p), 1)$ for price p, where the second feature vector is defined as (we plot the value below in Figure 1):

$$\phi(p) = \frac{1}{1 + \exp\left(-\frac{5(2p-(p_L+p_U))}{p_U-p_L}\right)}.$$

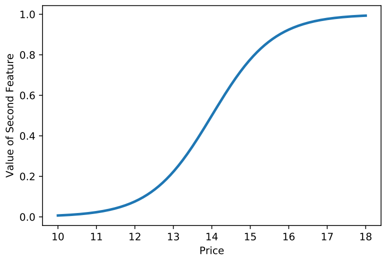
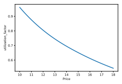
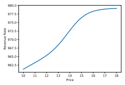
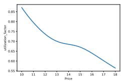
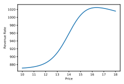
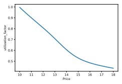
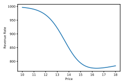


Figure 1: Function value of $\phi(p)$.

We consider three scenarios of the arrival rates associated with candidate prices and thus the corresponding system dynamics (three instances correspondingly). In Instance #1, the state-independent optimal price has a relatively low utilization factor. In Instance #2, the state-independent optimal price has a relatively moderate utilization factor. In Instance #3, the state-independent optimal price has the highest utilization factor. Instance details are in Table 1.

Benchmark. We adopt the ϵ-greedy algorithm (a commonly used benchmark that balances exploration and exploitation; see, e.g., Filippi et al. (2010)) to our settings as the benchmark. The ϵ-greedy algorithm will estimate the coefficients through the same regression step as that of BLinUCB. It will choose the price with the best estimated revenue rate with probability $1-\epsilon$ and randomly select a price with probability ϵ. Further, we use the same batch framework as that of BLinUCB for the benchmark and the algorithm makes pricing decisions for each batch. Algorithm 2 presents the algorithmic details of ϵ-greedy benchmark we used in this section. To conclude, for each instance, we implement four pricing algorithms: BLinUCB and three benchmark policies with $\epsilon = 0.3$, 0.2, and 0.1, i.e., the probability for conducting exploration.

Table 1: Parameter Settings and Experimental Design.

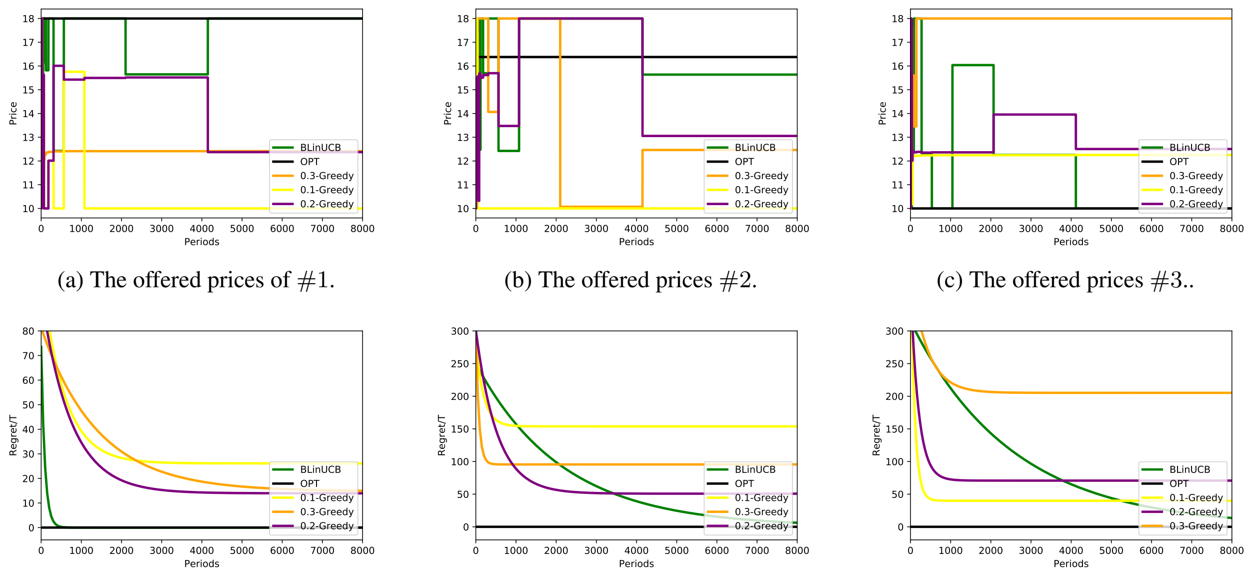
	Arrival Coef. θ_λ	Service Coef. θ_μ	Utilization Factor	Revenue Rate
Inst. #1	$(0.0102, -0.0018, 0.0020)$	$(0, 0, 10)$		
Inst. #2	$(0.0115, -0.03, 0)$	$(0, 0, 10)$		
Inst. #3	$(0.01, 0.05, 0)$	$(0, 0, 10)$		

Algorithm 2 ϵ-greedy Benchmark.

```
Input: $T$, $p_L$, $p_U$, $d_f$, $\epsilon$.
Initialize: $\tau$, $I_m$, $M$, $\mathcal{P}_b$.
for $p \in \mathcal{P}_b$ do
    Offer price $p$, record $\hat{d}_i(p)$ for arriving customers and $\hat{g}_i(p')$, $\forall p' \in [p_L, p_U]$ for leaving customers.
    if $n_m^s(p) \geq 8\log(T)$ then
        Update $\mathbf{X}, \hat{\Omega}_\lambda, \hat{\Omega}_\mu, \mathbf{d}, \mathbf{y}_\mu$
        Continue.
    end if
end for
Compute $\hat{\theta}_\lambda$ and $\hat{\theta}_\mu$ by (5) and (6)
for $m = 1, \ldots, M$ do
    With probability $1-\epsilon$: Choose $p_m = \text{argmax}_{p\in[p_L,p_U]}\, p \cdot \frac{\hat{\theta}_\mu^T \mathbf{x}_p}{\hat{\theta}_\lambda^T \mathbf{x}_p}$.
    Otherwise, Choose $p_m \in [p_L, p_U]$ uniformly.
    Offer $p_m$ in batch $m$, i.e., for $I_m\tau$ periods.
    Record $\hat{d}_i(p_m)$ for arriving customers and $\hat{g}_i(p)$, $\forall p \in \mathcal{P}$ for leaving customers.
    Update $\mathbf{X}, \hat{\Omega}_\lambda, \hat{\Omega}_\mu, \mathbf{d}, \mathbf{y}_\mu$; Compute $\hat{\theta}_\lambda$ and $\hat{\theta}_\mu$.
end for
```

We compare the results of the above four pricing policies with state-independent optimal price (OPT). We present two figures for the results of each instance (see Figure 2): the first row shows the offered price over periods of each algorithm and the second row depicts the cumulative time-average relaxed regret, i.e., $(\sum_{t'=1}^{t} J_{t'}^{\text{LP}} - \sum_{t'=1}^{t} J_{t'}^{\pi})/t$.

Compared with ϵ-greedy Benchmarks. From the numerical results, BLinUCB performs the best for all instances. BLinUCB ends up with the state-independent optimal price in Instances #1 and #3, and with a near-optimal price in Instance #2. The difference of the revenue rate between the state-independent optimal price and the last price chosen by BLinUCB of Instance #2 is small, where the former is 1023



(a) The offered prices of #1.

(b) The offered prices #2.

(c) The offered prices #3..

(d) Time-average regret of #1.

(e) Time-average regret #2..

(f) Time-average regret #3..

Figure 2: Offered Price and Time-average Cumulative Regret of Policies in Instances #1 to #3.

and the latter is 1020. One interesting observation is that all three ϵ-greedy benchmarks fail to identify the optimal price for these three instances. One possible reason could be the lack of exploration in the region where feature vectors are more vertical to those of implemented prices, resulting in inaccurate estimation of the coefficients in the linear relationship. The performance of three ϵ-greedy benchmarks, parametrized by ϵ, varies among different instances.

7. Conclusion

We consider a price-based revenue management problem with a single reusable resource over a finite time horizon under incomplete information and give the first rate-optimal online learning and pricing algorithm that admits a regret bound of $\tilde{O}(d_f\sqrt{T})$. Numerical results demonstrate that BLinUCB converges to optimality very fast and outperforms other benchmark algorithms.

There are several future research avenues. First, one may consider multi-product settings (see, e.g., Owen & Simchi-Levi, 2018; Doan et al., 2020). Second, one may consider generalizing the current model to accommodate general arrival and service distributions. However, this would require developing new coupling arguments to bound the loss of nonstationarity. Lastly, one may consider settings with nonstationary demand and/or inhomogeneous personal activities (see, Borgs et al., 2014; Besbes et al., 2015; Lei & Jasin, 2020). Extensions to any of the above settings would require new methods and techniques.

Acknowledgments

The authors would like to thank anonymous reviewers for their detailed and helpful comments. This research is partially supported by an Amazon research award and Department of Energy award #DE-SC0018018.

References

Abate, J. and Whitt, W. Transient behavior of the M/M/l queue: Starting at the origin. *Queueing Systems*, 2(1): 41–65, 1987.

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

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

Alieva, A., Cutkosky, A., and Das, A. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, pp. 187–195. PMLR, 2021.

Araman, V. F. and Caldentey, R. Dynamic pricing for nonperishable products with demand learning. *Operations Research*, 57(5):1169–1188, 2009.

Auer, P. and Ortner, R. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Bailey, N. T. A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):288–291, 1954.

Banerjee, S., Johari, R., and Riquelme, C. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, pp. 639, New York, NY, USA, 2015. Association for Computing Machinery.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Besbes, O., Elmachtoub, A. N., and Sun, Y. Static pricing: Universal guarantees for reusable resources. *Operations Research*, 70(2):1143–1152, 2022.

Bimpikis, K., Candogan, O., and Saban, D. Spatial pricing in ride-sharing networks. *Operations Research*, 67(3): 744–769, 2019.

Borgs, C., Candogan, O., Chayes, J., Lobel, I., and Nazerzadeh, H. Optimal multiperiod pricing with service guarantees. *Management Science*, 60(7):1792–1811, 2014.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, UK, 2013.

Cella, L., Lazaric, A., and Pontil, M. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pp. 1360–1370. PMLR, 2020.

Cesa-Bianchi, N., Dekel, O., and Shamir, O. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pp. 1160–1168, 2013.

Chen, B. and Chao, X. Parametric demand learning with limited price explorations in a backlog stochastic inventory system. *IISE Transactions*, 51(6):605–613, 2019.

Chen, B. and Shi, C. Tailored base-surge policies in dual-sourcing inventory systems with demand learning. Technical report, University of Michigan, Ann Arbor, MI (Available at SSRN 3456834), 2019a.

Chen, B., Chao, X., and Wang, Y. Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research*, 68(5):1445–1456, 2020a.

Chen, X., Liu, Y., and Hong, G. An online learning approach to dynamic pricing and capacity sizing in service systems. Technical report, Chinese University of Hong Kong, Shenzhen, China, 2020b.

Chen, Y. and Shi, C. Optimal pricing policy for service systems with reusable resources and forward-looking customers. Technical report, University of Michigan, Ann Arbor, MI, 2017.

Chen, Y. and Shi, C. Network revenue management with online inverse batch gradient descent method. Technical report, University of Michigan, Ann Arbor, MI (Available at SSRN 3331939), 2019b.

Chen, Y., Levi, R., and Shi, C. Revenue management of reusable resources with advanced reservations. *Production and Operations Management*, 26(5):836–859, 2017.

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

Daley, D. J. and Servi, L. Idle and busy periods in stable M/M/k queues. *Journal of Applied Probability*, 35(4): 950–962, 1998.

den Boer, A. V. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1): 1–18, 2015.

Deng, Y., Lahaie, S., and Mirrokni, V. Robust pricing in dynamic mechanism design. In *International Conference on Machine Learning*, pp. 2494–2503. PMLR, 2020.

Deng, Y., Lahaie, S., Mirrokni, V., and Zuo, S. Revenue-incentive tradeoffs in dynamic reserve pricing. In *International Conference on Machine Learning*, pp. 2601–2610. PMLR, 2021.

Doan, X. V., Lei, X., and Shen, S. Pricing of reusable resources under ambiguous distributions of demand and service time with emerging applications. *European Journal of Operational Research*, 282(1):235–251, 2020.

Even-Dar, E., Mansour, Y., and Bartlett, P. Learning rates for q-learning. *Journal of Machine Learning Research*, 5 (1), 2003.

Ferreira, K. J., Simchi-Levi, D., and Wang, H. Online network revenue management using Thompson sampling. *Operations Research*, 66(6):1586–1602, 2018.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, pp. 586–594, 2010.

Foss, S., Korshunov, D., Zachary, S., et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, Berlin/Heidelberg, Germany, 2011.

Gallego, G. and Van Ryzin, G. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

Gans, N. and Savin, S. Pricing and capacity rationing for rentals with uncertain durations. *Management Science*, 53(3):390–407, 2007.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Hadar, J. and Russell, W. R. Rules for ordering uncertain prospects. *The American Economic Review*, 59(1):25–34, 1969.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. Sequential batch learning in finite-action linear contextual bandits. Technical report, Stanford University, Stanford, CA (Available at arXiv:2004.06321), 2020.

Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pp. 4349–4358. PMLR, 2021.

Jia, H., Shi, C., and Shen, S. Online learning and pricing for service systems with reusable resources. Technical report, University of Michigan, Ann Arbor, MI (Available at SSRN 3755902), 2020.

Jia, H., Shi, C., and Shen, S. Multi-armed bandit with sub-exponential rewards. *Operations Research Letters*, 49(5): 728–733, 2021.

Kaewpuang, R., Niyato, D., Wang, P., and Hossain, E. A framework for cooperative resource management in mobile cloud computing. *IEEE Journal on Selected Areas in Communications*, 31(12):2685–2700, 2013.

Kelton, W. D. and Law, A. M. The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

Ledermann, W. and Reuter, G. E. H. Spectral theory for the differential equations of simple birth and death processes. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 246 (914):321–369, 1954.

Lei, Y. and Jasin, S. Real-time dynamic pricing for revenue management with reusable resources, advance reservation, and deterministic service time requirements. *Operations Research*, 68(3):676–685, 2020.

Levi, R. and Radovanović, A. Provably near-optimal LP-based policies for revenue management in systems with reusable resources. *Operations Research*, 58(2):503–507, 2010.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Liu, Y. and Li, Y. Pricing scheme design of ridesharing program in morning commute problem. *Transportation Research Part C: Emerging Technologies*, 79:156–177, 2017.

Maglaras, C. Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research*, 54(5):914–932, 2006.

Mankiw, N. G. *Principles of economics*. Cengage Learning, Boston, MA, 2014.

Morisaku, T. *Techniques for data-truncation in digital computer simulation*. PhD thesis, University of Southern California, Los Angeles, CA, USA, 1976.

Oliveira, B. B., Carravilla, M. A., and Oliveira, J. F. Integrating pricing and capacity decisions in car rental: A matheuristic approach. *Operations Research Perspectives*, 5:334–356, 2018.

Omahen, K. and Marathe, V. Analysis and applications of the delay cycle for the M/M/c queueing system. *Journal of the ACM*, 25(2):283–303, 1978.

Owen, Z. and Simchi-Levi, D. Price and assortment optimization for reusable resources. Technical report, MIT, Cambridge, MA (Available at SSRN 3070625), 2018.

Parthasarathy, P. and Sharafali, M. Transient solution to the many-server poisson queue: A simple approach. *Journal of Applied Probability*, 26(3):584–594, 1989.

Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. Batched bandit problems. *The Annals of Statistics*, 44(2): 660–681, 2016.

Püschel, T., Schryen, G., Hristova, D., and Neumann, D. Revenue management for cloud computing providers: Decision models for service admission control under non-probabilistic uncertainty. *European Journal of Operational Research*, 244(2):637–647, 2015.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. *Lecture Notes for Course 18S997*, 813:814, 2015.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Rusmevichientong, P., Sumida, M., and Topaloglu, H. Dynamic assortment optimization for reusable products with random usage durations. *Management Science*, 66(7): 2820–2844, 2020.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Savin, S. V., Cohen, M. A., Gans, N., and Katalan, Z. Capacity management in rental businesses with two customer bases. *Operations Research*, 53(4):617–631, 2005.

Seber, G. A. and Lee, A. J. *Linear regression analysis*, volume 329. John Wiley & Sons, Hoboken, NJ, 2012.

Seth, S. and Yalonetzky, G. Stochastic dominance with parametric distributions. Technical report, Oxford University, Oxfordshire, UK, 2014.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Simchi-Levi, D. and Xu, Y. Phase transitions and cyclic phenomena in bandits with switching constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, Cambridge, MA, 2018.

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

Tao, C., Blanco, S., and Zhou, Y. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pp. 4877–4886. PMLR, 2018.

Xu, H. and Li, B. Dynamic cloud pricing for revenue maximization. *IEEE Transactions on Cloud Computing*, 1(2): 158–171, 2013.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32: 5197–5208, 2019.

A. Appendix for Section 3

Proof of Proposition 1. We show that the decisions variables (π, a, J) associated with any admissible policy must satisfy the constraints of the linear program (1), and the expected revenue of this admissible policy is exactly the corresponding objective value in (1).

Consider an arbitrary policy $\hat{\pi}$. We obtain the values of decision variables by the following steps. Variable $\hat{a}_{nt}$ takes the value of probability that the system has n customers at the beginning of period t under the policy $\hat{\pi}$ for $n = 0, \dots, \infty$ and $t = 1, \dots, T$. Therefore, constraints (1c) are immediately satisfied. Variable $\hat{\pi}^{p}_{nt} = 1$ if the policy $\hat{\pi}$ chooses price p at state (n, t), otherwise, $\hat{\pi}^{p}_{nt} = 0$, for $n = 0, \dots, \infty$, $t = 1, \dots, T$, $p \in \mathcal{P}$. As a result, constraints (1d) hold because in the left-hand side we have one variable equal to 1 and others all zero. Variable $\hat{J}_{nt}$ takes the value of expected revenue collected before the end of period T from the customers who arrive during period t, when the system has n customers at the beginning of period t, for $n = 0, \dots, \infty$, $t = 1, \dots, T$. Consider arbitrary n and t, we denote the price chosen at state (n, t) by the policy $\hat{\pi}$ as $\hat{p}$. Then, by plugging in $\hat{\pi}^{\hat{p}}_{nt} = 1$, in the right-hand side of constraints (1e), we have $\frac{\lambda_{\hat{p}}}{\mu_{\hat{p}}}\hat{p}$, where $\lambda_{\hat{p}}$ is the average number of arrived customers during period t, $\frac{1}{\mu_{\hat{p}}}$ is the average service time for each customer, and $\hat{p}$ is the revenue when one customer occupies one unit of resource for one unit of time. Therefore, $\frac{\lambda_{\hat{p}}}{\mu_{\hat{p}}}\hat{p}$ denotes the expected revenue collected from customers who arrive during period t (where assuming all arrived customers being served successfully). Consequently, we can conclude that the right-hand side of (1e) is as large as the left-hand side, because the right-hand side may also compute the revenue collected after the end of period T. Hence, constraints (1e) are also satisfied. By the value assignment steps of $\hat{a}_{nt}$ and $\hat{J}_{nt}$, for $n = 0, \dots, \infty$, $t = 1, \dots, T$, one has that the corresponding LP objective value at solution $(\hat{\pi}, \hat{a}, \hat{J})$ is $J^{\hat{\pi}}$. Recall that the discretizing interval of $\mathcal{P}$ can be arbitrarily small and thus $J^{\text{LP}} \geq J^{\pi}$ for any admissible policy π as well as the optimal state-dependent policy π^{*}. □

Proof of Corollary 1. This corollary follows directly from Propositions 1 and 4 in Section 5.1. Based on the above constructed optimal solution $(\tilde{\pi}, \tilde{a}, \tilde{J})$, one can compute that the difference between the revenue upper bound J^{LP} and the collected revenue under the single-price policy $J^{\pi^{\tilde{p}}}$ is from the transient performance of the system when it starts serving customers from an empty state at $t = 1$. The length of the transient state is less than $(\log(T))^2$ with probability higher than $1 - 2/T^2$ by Proposition 4. Thus, one can derive that when $T \to \infty$, $J^{\text{LP}} - J^{\pi^{\tilde{p}}} \leq O((\log(T))^2)$, which suggests that $\pi^{\tilde{p}}$ is asymptotically optimal. □

B. Appendix for Section 4

Definition 4. A random variable X with mean $\mathbb{E}[X]$ is (τ^2, b)-sub-exponential (SE) if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2\tau^2}{2}\right) \quad \text{for } |\lambda| \leq \frac{1}{b}.$$

Examples of sub-exponential (SE) variables include (i) exponential random variables and (ii) χ^2 random variables. We refer readers to Foss et al. (2011) for more properties of sub-exponential distributions.

We invoke a standard concentration inequality for sub-exponentials from Bernstein's inequality (see, e.g., Rigollet & Hütter (2015); Boucheron et al. (2013)).

Lemma 1 (Concentration of sub-exponentials from Bernstein's Inequality)**.** *Let X be (τ^2, b)-sub-exponential. Then for a non-negative number $t \geq 0$:*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \left\{2\exp\left(-\frac{t^2}{2\tau^2}\right) \ \textit{if}\ 0 \leq t \leq \frac{\tau^2}{b};\ 2\exp\left(-\frac{t}{2b}\right) \ \textit{if}\ t \geq \frac{\tau^2}{b}\right\}.$$

Lemma 2. ***(Sub-exponential Property.)*** *If a random variable X follows an exponential distribution with mean $1/\lambda$, then the random variable $X - \frac{1}{\lambda}$ is $(\frac{4}{\lambda^2}, \frac{2}{\lambda})$-sub-exponential.*

Proof of Lemma 2. According to the definition of sub-exponential random variables, we need to prove that

$$\mathbb{E}\left[\exp\left(s\left(X - \frac{1}{\lambda}\right)\right)\right] \leq \exp\left(\frac{s^2\frac{4}{\lambda^2}}{2}\right), \quad \forall s \leq \frac{\lambda}{2}.$$

By the moment generating function of exponential distribution, we have for $s \leq \frac{\lambda}{2} < \lambda$:

$$\mathbb{E}\left[\exp\left(s\left(X - \frac{1}{\lambda}\right)\right)\right] = \mathbb{E}\left[\exp\left(sX\right)\right] \cdot \exp\left(-\frac{s}{\lambda}\right) = \frac{\lambda}{\lambda - s} \cdot \exp\left(-\frac{s}{\lambda}\right).$$

Let $t = \frac{s}{\lambda}$, we have:

$$\log\left(\frac{\lambda}{\lambda - s} \cdot \exp\left(-\frac{s}{\lambda}\right)\right) = -\log(1-t) - t \leq 2t^2 = \log\left(\exp\left(\frac{s^2 \frac{4}{\lambda^2}}{2}\right)\right).$$

□

Lemma 3. *Consider a random variable $X_i \sim SE(\nu^2, \alpha)$ and β is a non-zero scalar, then $\beta X_i \sim SE(\beta^2\nu^2, |\beta|\alpha)$.*

Lemma 4. *Consider independent random variables $X_i \sim SE(\nu_i^2, \alpha_i)$ for $i = 1, \ldots, n$, then $X = \sum_{i=1}^n X_i$ follows $SE(\sum_{i=1}^n \nu_i^2, \max_i \ \alpha_i)$.*

The proofs of Lemma 3 and Lemma 4 follow directly from the definitions of SE variables and we omit the details here.

Proof of Proposition 2. We first show the SE property of $\hat{\theta}_\lambda^T \mathbf{x}'$ and then derive the tail bound.

SE property of $\hat{\theta}_\lambda$. By the computing equation of $\hat{\theta}_\lambda$ in (3) and $\mathbf{d} = \mathbf{X}\theta_\lambda + \vec{\epsilon}$, we can derive

$$\hat{\theta}_\lambda = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\mathbf{X}\theta_\lambda + (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\vec{\epsilon} = \theta_\lambda + (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\vec{\epsilon},$$

where $\vec{\epsilon} \in \mathbb{R}^N$ is a vector of all error terms ϵ_i of implemented prices in (2). Let matrix $A = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}$ and thus $\hat{\theta}_\lambda = \theta_\lambda + A\vec{\epsilon}$. Use row vector $a_{d\cdot}$ to denote the d^{th} row of A and a_{di} to denote elements of A. The d^{th} element of $A\vec{\epsilon}$ is $a_{d\cdot}\vec{\epsilon}$, which follows SE$(\sum_{i=1}^N a_{di}^2(4/n_i\lambda_i^2), \max_{i=1,\ldots,N} \ 2|a_{di}|/(n_i\lambda_i))$ by Lemmas 3 and 4. Therefore, the d^{th} element of $\hat{\theta}_\lambda$ follows SE$(\sum_{i=1}^N a_{di}^2(4/n_i\lambda_i^2), \max_{i=1,\ldots,N} \ 2|a_{di}|/(n_i\lambda_i))$ with mean $\theta_{\lambda,d}$.

SE property of $\hat{\theta}_\lambda^T\mathbf{x}'$. One has

$$\hat{\theta}_\lambda^T\mathbf{x}' = \mathbf{x}'^T(\theta_\lambda + A\vec{\epsilon}) = \mathbf{x}'^T\theta_\lambda + \mathbf{x}'^T A\vec{\epsilon}.$$

Then we focus on the second term. Similarly, use column vector $a_{\cdot i}$ to denote the i^{th} column of A, and thus we have $\mathbf{x}'^T A = [\mathbf{x}'^T a_{\cdot 1}, \mathbf{x}'^T a_{\cdot 2}, \ldots, \mathbf{x}'^T a_{\cdot n}]$. By Lemmas 3 and 4, we can easily derive the second term follows SE$(\sum_{i=1}^N(\mathbf{x}'^T a_{\cdot i})^2(4/n_i\lambda_i^2), \max_{i=1,\ldots,N} 2|\mathbf{x}'^T a_{\cdot i}|/(n_i\lambda_i))$. By the definitions of matrices Ω and A, the first SE parameter can be rewritten as

$$\sum_{i=1}^N (\mathbf{x}'^T a_{\cdot i})^2(4/n_i\lambda_i^2) = 4\mathbf{x}'^T A\Omega A^T\mathbf{x}' = 4\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'.$$

Hence we can rewrite the SE distribution as SE$(4\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}', \max_{i=1,\ldots,N} 2|\mathbf{x}'^T a_{\cdot i}|/(n_i\lambda_i))$. Therefore, $\hat{\theta}_\lambda^T\mathbf{x}'$ follows SE$(4\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}', \max_{i=1,\ldots,N} 2|\mathbf{x}'^T a_{\cdot i}|/(n_i\lambda_i))$ with mean $\theta_\lambda^T\mathbf{x}'$.

Concentration bound for SE random variable $\hat{\theta}_\lambda^T\mathbf{x}'$. We apply the concentration inequality for SE variables. For SE variable X with parameters (ν^2, α), then $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2\exp(-t^2/(2\nu^2))$ if $t \leq \nu^2/\alpha$ (this concentration inequality can be found in, e.g., Boucheron et al., 2013). We can show

$$\left(\max_{i=1,\ldots,N} \frac{|\mathbf{x}'^T a_{\cdot i}|}{n_i\lambda_i}\right)^2 = \max_{i=1,\ldots,N} \frac{(\mathbf{x}'^T a_{\cdot i})^2}{n_i^2\lambda_i^2} \leq \sum_{i=1,\ldots,N} \frac{(\mathbf{x}'^T a_{\cdot i})^2}{n_i^2\lambda_i^2} \tag{10a}$$

$$\leq \frac{1}{8\log(T)} \cdot \sum_{i=1,\ldots,N} \frac{(\mathbf{x}'^T a_{\cdot i})^2}{n_i\lambda_i^2} = \frac{\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}{8\log(T)}, \tag{10b}$$

where the first inequality in (10b) is derived from $n_i \geq 8\log(T)$. Further, we have

$$\sqrt{32\log(T) \cdot \mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'} = \frac{4\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}{\frac{2\sqrt{\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}}{\sqrt{8\log(T)}}} \leq \frac{4\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}{\max_{i=1,\ldots,N} 2|\mathbf{x}'^T a_{\cdot i}|/(n_i\lambda_i)}. \tag{11a}$$

Therefore, we meet the requirement for applying the sub-Gaussian behavior type of concentration bound and we have

$$\mathbb{P}\left(\frac{|\hat{\theta}_\lambda^T \mathbf{x}' - \theta_\lambda^T \mathbf{x}'|}{\sqrt{\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}} \geq \sqrt{32\log(T)}\right) \leq 2\text{exp}(-\frac{32\log(T)\cdot \mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}{8\mathbf{x}'^T(\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{x}'}) \leq \frac{2}{T^4}. \tag{12}$$

□

Proof of Proposition 3. The proof is based on Proposition 2.

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{\hat{\theta}_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T\mathbf{x}}\mathcal{G}\right)\\
=&\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}} + \frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}} - \frac{\hat{\theta}_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T\mathbf{x}}\left(r_{\max}\sqrt{\mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{x}} + \sqrt{\mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\mu^{-1}\mathbf{X})^{-1}\mathbf{x}}\right)\right)\\
\geq&\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{r_{\max}\sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{x}}}{\hat{\theta}_\lambda^T\mathbf{x}}\right)\cdot\mathbb{P}\left(\left|\frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}} - \frac{\hat{\theta}_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{\sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\mu^{-1}\mathbf{X})^{-1}\mathbf{x}}}{\hat{\theta}_\lambda^T\mathbf{x}}\right).
\end{aligned}$$

We analyze the two terms separately. For the first term, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{r_{\max}\sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{x}}}{\hat{\theta}_\lambda^T\mathbf{x}}\right)\\
=&\mathbb{P}\left(\frac{\lambda_p\left|\theta_\lambda^T\mathbf{x} - \hat{\theta}_\lambda^T\mathbf{x}\right|}{\mu_p\cdot\hat{\theta}_\lambda^T\mathbf{x}} \leq \frac{r_{\max}\sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{x}}}{\hat{\theta}_\lambda^T\mathbf{x}}\right)\\
\geq&\mathbb{P}\left(\left|\theta_\lambda^T\mathbf{x} - \hat{\theta}_\lambda^T\mathbf{x}\right| \leq \sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{x}}\right) \geq 1 - \frac{2}{T^4}.
\end{aligned}$$

For the second term, we have the following inequality by (7):

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\theta_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}} - \frac{\hat{\theta}_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{\sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\mu^{-1}\mathbf{X})^{-1}\mathbf{x}}}{\hat{\theta}_\lambda^T\mathbf{x}}\right)\\
=&\mathbb{P}\left(\left|\theta_\mu^T\mathbf{x} - \hat{\theta}_\mu^T\mathbf{x}\right| \leq \sqrt{32\log(T)\ \mathbf{x}^T(\mathbf{X}^T\hat{\Omega}_\mu^{-1}\mathbf{X})^{-1}\mathbf{x}}\right) \geq 1 - \frac{2}{T^4}.
\end{aligned}$$

Therefore, we have

$$\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{\hat{\theta}_\mu^T\mathbf{x}}{\hat{\theta}_\lambda^T\mathbf{x}}\right| \leq \frac{\sqrt{32\log(T)}}{\hat{\theta}_\lambda^T\mathbf{x}}\mathcal{G}\right) \geq \left(1 - \frac{2}{T^4}\right)^2 \geq 1 - \frac{4}{T^4}.$$

□

C. Appendix for Section 5.1

The analysis in this section relies on busy period A_n, $\forall n = \{1, \ldots, \infty\}$, defined as the minimum time that the system takes to have $n-1$ customers if starting with n customers in the system, i.e., $A_n = \min\{t\colon n$ customers in the system at time 0^+ and $n-1$ customers in the system at time $t\}$. Note that $A_n,\ n = 1, \ldots, c$, are random variables following different distributions due to difference in the number of servers in use, while $A_n,\ n = c, \ldots, \infty$, are i.i.d. random variables (see Omahen & Marathe, 1978; Daley & Servi, 1998). The analysis also considers the first-order stochastic dominance between two random variables defined in Definition 1 in Hadar & Russell (1969) and in Seth & Yalonetzky (2014).

Lemma 5. ***(Lemma 4 in Jia et al. (2020), First-order Stochastic Dominance between A_i and A_j in M/M/c Queue)*** *For random variables A_i and A_j where $i < j \leq c$, A_i first-order dominates A_j, i.e., $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$ for any $x > 0$.*

Proof of Lemma 5 (Adapted from Jia et al. (2020)). To compare random variables A_i and A_j where $i < j$ of an M/M/c queue process Q, we construct an auxiliary process $\hat{Q}$, which is the same to Q except that when the system has n customers with $j \leq n \leq c + j - i$, the service rate of $\hat{Q}$ is $n - j + i$. Similarly, we define $\hat{A}_n$ for $n = 1, \dots, \infty$ of $\hat{Q}$. Lemma 1 in Omahen & Marathe (1978); Daley & Servi (1998) shows the recursion between A_n, for $n = 1, \dots, \infty$. A_n can be decomposed into intervals shown in Figure 3. For $n = 1, \dots, \infty$, define events $\mathcal{B}_n^a$ as where a new customer arrives before any current customers leave starting from n customers in the system, and t_n^a as the time when the new customer arrives. Similarly, define events $\mathcal{B}_n^l$ as where a current customer leaves before any new customer arrives, starting from n customers in the system and define t_n^l as the time when the first current customer leaves. Therefore, for $n = 1, \dots, \infty$, the following recursion holds:

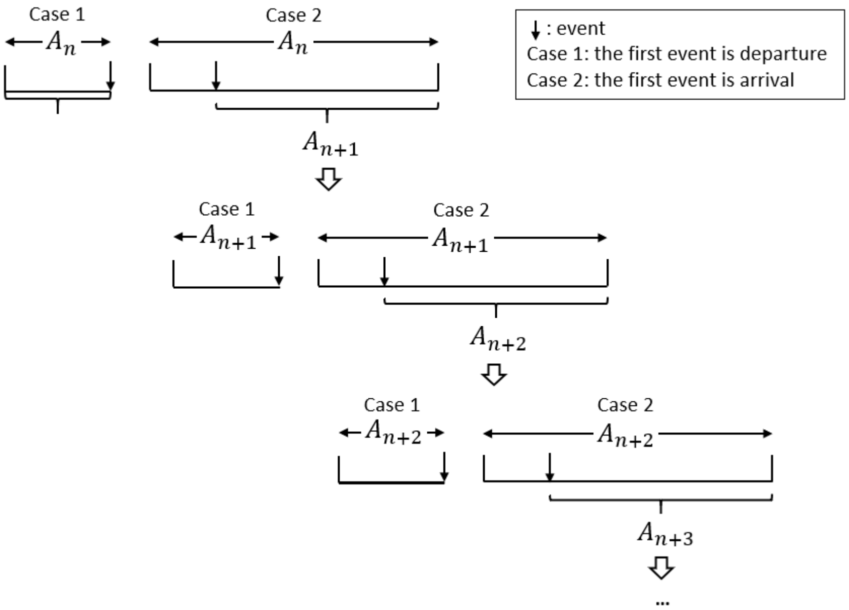


Figure 3: Intervals within Busy Period A_n.

$$\bar{A}_n = \mathbb{P}\left(\mathcal{B}_n^a\right) \cdot \left(\mathbb{E}\left[t_n^a\right] + \bar{A}_{n+1} + \bar{A}_n\right) + \mathbb{P}\left(\mathcal{B}_n^l\right) \cdot \mathbb{E}\left[t_n^l\right]. \tag{13}$$

Comparing Q and $\hat{Q}$, we conclude that:

1. $\hat{A}_j$ and A_i have the same distribution, because they have the same arrival and departure rates in every interval in Figure 3.
2. $\mathbb{P}(A_j \leq x) > \mathbb{P}(\hat{A}_j \leq x)$, because $\hat{A}_j$ has the same arrival rate, but a smaller (when $n < c + j - i$) or same (when $n \geq c + j - i$) departure rate in the intervals in Figure 3.

Therefore, $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$. By Definition 1 in Seth & Yalonetzky (2014), one further has $\mathbb{E}_{A_j}[u(x)] < \mathbb{E}_{A_i}[u(x)]$ for all strictly increasing, continuous utility functions $u(x) : [0, \infty] \to \mathbb{R}$. This completes the proof. □

Lemma 6. ***(Lemma 5 in Jia et al. (2020), Concentration Inequality for Independent Samples)*** *This lemma is similar to Lemma 11 in Chen & Shi (2019a). Let ξ_i be i.i.d. random variables with mean 0 and standard deviation σ. If the moment generating function of ξ_i around 0 is finite, i.e., there exists a constant $\delta > 0$ such that for any $s \in (-\delta, \delta)$ it holds that*

$$\mathbb{E}[e^{s\xi_i}] < \infty,$$

then

$$\mathbb{P}\left(\frac{1}{\bar{S}} \sum_{i=1}^{\bar{S}} \xi_i > 2\sigma\sqrt{-\log \rho}\right) \leq e^{-\log(T)^{3/2}}.$$

where $\bar{S} = \left\lfloor \frac{\log(T)^{3/2}}{-\log\rho} \right\rfloor$.

Proof of Lemma 6 (Adapted from Jia et al. (2020)). For s in $[-\delta, \delta]$ define

$$\Phi(s) = \log \mathbb{E}[e^{s\xi_i}].$$

For $x > 0$ and $s \in [0, \rho)$, by Markov's inequality one has

$$\mathbb{P}\left(\frac{1}{\bar{S}}\sum_{i=1}^{\bar{S}} \xi_i > x\right) \leq e^{\bar{S}(\Phi(s) - sx)}.$$

Let $s^* = \frac{x}{\sigma^2}$ and apply Taylor's expansion to the third order around 0, one has

$$\Phi(s^*) = \frac{1}{2}\sigma^2(s^*)^2 + \frac{1}{6}\Phi'''(0)(s^*)^3.$$

Therefore, one has

$$\Phi(s^*) - s^*x \leq -\frac{x^2}{2\sigma^2} + C_3\frac{x^3}{\sigma^6} \leq -\frac{x^2}{4\sigma^2}.$$

Plug in $x = 2\sigma\sqrt{-\log\rho}$ then we have

$$\mathbb{P}\left(\frac{1}{\bar{S}}\sum_{i=1}^{\bar{S}} \xi_i > 2\sigma\sqrt{-\log\rho}\right) \leq e^{\bar{S}\left(-\frac{x^2}{4\sigma^2}\right)} = e^{-\log(T)^{3/2}}.$$

□

For $t = 0, \ldots, T$, define events

$$\mathcal{A}_t = \{S_t = \hat{S}_t \text{ after time } t\}.$$

Therefore, systems S_t and $\hat{S}_t$ couple after t is equivalent to that event $\mathcal{A}_t$ happens, $\forall t \geq 0$.

Proof of Proposition 4 (Adapted from Jia et al. (2020)). For any $t \geq \tau$, one has

$$\mathbb{P}(\mathcal{A}_\tau) = \mathbb{P}(\hat{S}_\tau = S_\tau),$$

where the equality holds because once $\hat{S}_\tau = S_\tau$, then the two processes have an equal number of customers at any time thereafter and thus the two processes couple. Define $\bar{S} = \left\lfloor \frac{(\log(T))^{3/2}}{-\log\rho} \right\rfloor$, we have

$$\mathbb{P}(\hat{S}_\tau = S_\tau) \geq \mathbb{P}(\hat{S}_\tau = S_\tau | \hat{S}_0 \leq \bar{S})\mathbb{P}(\hat{S}_0 \leq \bar{S}) \geq \mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_n \leq \tau\right)\mathbb{P}(\hat{S}_0 \leq \bar{S}). \tag{14a}$$

Define the stationary distribution for n customers in the system as $\ell_n,\ n \geq 0$, one has (by Assumption 1)

$$\mathbb{P}(S_0 \leq \bar{S}) = 1 - \frac{\ell_0 c^c}{c!}\rho^{\bar{S}+1} \geq 1 - \frac{1}{T^2}.$$

By Lemma 5, A_1 first-order dominates A_n where $n > 1$, and thus one can derive that $\mathbb{P}(\sum_{n=1}^{\bar{S}} A_1 \leq \tau)$ is a valid lower bound of $\mathbb{P}(\sum_{n=1}^{\bar{S}} A_n \leq \tau)$. Define $\bar{A}_1$ as the mean and σ_1 as the standard deviation of random variable A_1. Based on the recursive analysis of M/M/c queue (see Omahen & Marathe, 1978; Daley & Servi, 1998), $\bar{A}_1$ can be derived by $\bar{A}_c$, where $\bar{A}_c = \frac{1}{c\mu - \lambda}$. By Lemma 6, we have

$$\mathbb{P}(\hat{S}_\tau = S_\tau) \geq \mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_1 \leq \tau\right)\left(1 - \frac{1}{T^2}\right) \tag{15a}$$

$$= \left(1 - \mathbb{P}\left(\sum_{n=1}^{\bar{S}} \left(A_1 - \bar{A}_1\right) > \tau - \bar{S}\bar{A}_1\right)\right)\left(1 - \frac{1}{T^2}\right) \tag{15b}$$

$$= \left(1 - \mathbb{P}\left(\sum_{n=1}^{\bar{S}} \left(A_1 - \bar{A}_1\right) > (\log(T))^2 - \bar{A}_1\left\lfloor\frac{(\log(T))^{3/2}}{-\log\rho}\right\rfloor\right)\right)\left(1 - \frac{1}{T^2}\right) \tag{15c}$$

$$\geq \left(1 - \mathbb{P}\left(\sum_{n=1}^{\bar{S}} \left(A_1 - \bar{A}_1\right) > C_3(\log(T))^{\frac{7}{4}}\right)\right)\left(1 - \frac{1}{T^2}\right) \tag{15d}$$

$$\geq \left(1 - \mathbb{P}\left(\frac{1}{\bar{S}}\sum_{n=1}^{\bar{S}} \left(A_1 - \bar{A}_1\right) > C_4(\log(T))^{\frac{1}{4}}\right)\right)\left(1 - \frac{1}{T^2}\right) \tag{15e}$$

$$\geq \left(1 - \mathbb{P}\left(\frac{1}{\bar{S}}\sum_{n=1}^{\bar{S}} \left(A_1 - \bar{A}_1\right) > 2\sigma_1\sqrt{-\log\rho}\right)\right)\left(1 - \frac{1}{T^2}\right) \tag{15f}$$

$$\geq \left(1 - e^{-(\log(T))^{3/2}}\right)\left(1 - \frac{1}{T^2}\right) \tag{15g}$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \geq 1 - \frac{2}{T^2}, \tag{15h}$$

where C_3 and C_4 are positive constants. □

Proof of Proposition 5 (Adapted from Jia et al. (2020)). Without loss of generality, we consider two cases, (i) the price changes from a lower price p_l to a higher price p_h and (ii) the price changes from a higher price p_h to a lower price p_l. For each case, we further consider two sub-cases, (i) the unit service rate $\mu_{p_l} > \mu_{p_h}$ and (ii) $\mu_{p_l} \leq \mu_{p_h}$. For each case, we construct a virtual system and show that the actual probability is bounded by the probability with which the virtual system reaches the steady state, respectively. Denote the actual system as S and the virtual system as $\breve{S}$. For $t \in \{\tau, 2\tau\}$, define events:

$$\mathcal{C}_t = \{S \text{ reaches the steady state under the new price before } t\}, \tag{16}$$

$$\breve{\mathcal{C}}_t = \{\breve{S} \text{ reaches the steady state under the new price before } t\}. \tag{17}$$

Case (1.1): from p_l to p_h and $\mu_{p_l} > \mu_{p_h}$. Without loss of generality, the system starts from $t = 0$ with S_0 customers in the system under service rate μ_{p_l}. New customers arrive under rate λ_{p_h} with an exponential service time with mean $1/\mu_{p_h}$. Consider a virtual system $\breve{S}$ starts with S_0 customers. The future customers arrive following the same process as the actual system, i.e., with rate λ_{p_h}. All the existing customers and future customers have an exponential service time with mean $1/\mu_{p_h}$. Therefore, the probability of the virtual system reaching the steady state before 2τ is the probability that an M/M/c queue under price p_h reaching the steady state with the number of initial customers as S_0. Therefore, we can compute the probability as:

$$\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\right) \tag{18a}$$

$$\geq \mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\middle|S_0 \leq 2\bar{S}_h\right) \cdot \mathbb{P}\left(S_0 \leq 2\bar{S}_h\right) \tag{18b}$$

$$\geq \mathbb{P}\left(\sum_{n=1}^{2\bar{S}_h} \bar{A}_{1h} \leq 2\tau\right) \cdot \mathbb{P}\left(S_0 \leq 2\bar{S}_h\right) \tag{18c}$$

$$\geq \left(\mathbb{P}\left(\sum_{n=1}^{\bar{S}_h} \bar{A}_{1h} \leq \tau\right)\right)^2 \cdot \mathbb{P}\left(S_0 \leq 2\bar{S}_h\right) \tag{18d}$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \left(1 - \frac{\ell_{0l}c^c}{c!}\rho_l^{2\bar{S}_h+1}\right) \tag{18e}$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \left(1 - \rho_l^{\frac{2}{\log\rho_h}(-(\log(T))^{3/2})}\right) \tag{18f}$$

$$\geq \left(1-\frac{1}{T^2}\right)^3 \tag{18g}$$

$$\geq 1-\frac{3}{T^2}, \tag{18h}$$

where $\bar{S}_h = \frac{(\log(T))^{3/2}}{-\log \rho_{p_h}}$ and ℓ_{0l} is the steady-state probability of the system with zero customers under price p_l. The inequality (18g) holds because of Assumption 2.3.

The virtual system considers a longer service time for the existing customers arrived under a lower price in the previous batch. Therefore, in probabilistic sense, the virtual system takes longer time to reach steady state:

$$\mathbb{P}(\mathcal{C}_{2\tau}) \geq \mathbb{P}(\breve{\mathcal{C}}_{2\tau}) \geq 1-\frac{3}{T^2}. \tag{19}$$

Case (1.2): from p_l to p_h and $\mu_{p_l} \leq \mu_{p_h}$. Consider a virtual system $\breve{S}$ starts with S_0 customers. The future customers arrive as the process under p_l, i.e., with rate λ_{p_l}. All the existing customers and future customers have an exponential service time with mean $\frac{1}{\mu_{p_l}}$. Therefore, the virtual system is still an M/M/c queue under p_l. Denote the first time that the virtual system has zero customers as $\hat{t}$. After $\hat{t}$, the further customers arrive with rate λ_{p_h} have an exponential service time with mean $1/\mu_{p_h}$. Denote the time starting from $\hat{t}$ that the virtual system uses to reach the steady state as $\breve{t}$. Therefore, the probability of the virtual system reaching the steady state under p_l before 2τ is the probability that $\hat{t}+\breve{t} \leq 2\tau$. Therefore, we can compute the probability as:

$$\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\right) \tag{20a}$$

$$=\mathbb{P}\left(\breve{t}+\hat{t} \leq 2\tau\right) \tag{20b}$$

$$\geq \mathbb{P}\left(\breve{t}+\hat{t} \leq 2\tau | S_0 \leq 2\bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq 2\bar{S}_l\right) \tag{20c}$$

$$\geq \mathbb{P}\left(\breve{t} \leq \tau\right) \cdot \mathbb{P}\left(\hat{t} \leq \tau | S_0 \leq 2\bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq 2\bar{S}_l\right) \tag{20d}$$

$$\geq \left(1-\frac{2}{T^2}\right)^2 \tag{20e}$$

$$\geq 1-\frac{4}{T^2}, \tag{20f}$$

where (20e) is by Proposition 4. The virtual system considers a longer (or equal) service time for the existing customers arrived under a lower price in the previous batch. Therefore, in probabilistic sense, the virtual system takes longer (or equal) time to reach steady state:

$$\mathbb{P}(\mathcal{C}_{2\tau}) \geq \mathbb{P}(\breve{\mathcal{C}}_{2\tau}) \geq 1-\frac{4}{T^2}. \tag{21}$$

Case (2.1): from p_h to p_l and $\mu_{p_l} > \mu_{p_h}$. Without loss of generality, the system starts from $t=0$ with S_0 customers in the system under service rate μ_{p_h}. New customers arrive under rate $\lambda_{p_l} = \lambda_{p_h} + \Delta_\lambda$, where $\Delta_\lambda \geq 0$ by Assumption 2.1 and $\Delta_\lambda \leq \frac{\rho_{p_l}}{-3e\log p_l}$ by Assumption 2.2. We can manually split the arriving process of new customers as two independent Poisson arrival processes with rates λ_{p_h} and Δ_λ. Consider a virtual system that starts with S_0 customers. The arrival processes of the new customers are the same as the actual system, i.e., λ_{p_h} and Δ_λ. The customers arrive according to the process with rate λ_{p_h} enter the system with an exponential service time with mean $1/\mu_{p_h}$. The customers arriving according to the process with rate Δ_λ are waiting in another buffer queue. Denote the first time that the virtual system (except for the buffer queue) has zero customers as $\hat{t}$ and the length of the buffer queue at this time as $\breve{S}_{\hat{t}}^{\text{buffer}}$. Starting from $\hat{t}$, the buffer queue merges with the virtual service system. All the customers in the buffer queue and further customers have an exponential service time with mean $\frac{1}{\mu_{p_l}}$. Therefore, after $\hat{t}$, the system is an M/M/c queue with balking and reneging under price p_l with an initial number of customers as $\breve{S}_{\hat{t}}^{\text{buffer}}$. We analyze the virtual system before and after $\hat{t}$ separately. Before $\hat{t}$, the virtual system starts with a steady state under price p_h and keeps in the steady state until $\hat{t}$. As in Proposition 4, the virtual system reaches an empty state within time τ with a high probability:

$$\mathbb{P}\left(\hat{t} \leq \tau\right) \geq 1-\frac{2}{T^2}. \tag{22}$$

Similarly, by Proposition 4, if the starting length of the system $\breve{S}_{\hat{t}}^{\text{buffer}}$ is less than $\frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}$, then the virtual system reaches the steady state under price p_l after τ with probability $1 - \frac{1}{T^2}$, starting at time $\hat{t}$. Therefore, we can compute the probability that the virtual queue reaches the steady state within 2τ as:

$$
\begin{aligned}
&\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\right) && \text{(23a)}\\
\geq &\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau} | \hat{t} \leq \tau\right) \cdot \mathbb{P}\left(\hat{t} \leq \tau\right) && \text{(23b)}\\
\geq &\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau} \middle| \hat{t} \leq \tau, \breve{S}_{\hat{t}}^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right) \cdot \mathbb{P}\left(\breve{S}_{\hat{t}}^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}} \middle| \hat{t} \leq \tau\right) \cdot \mathbb{P}\left(\hat{t} \leq \tau\right) && \text{(23c)}\\
= &\left(1 - \frac{2}{T^2}\right)\left(1 - \frac{1}{T^2}\right) \cdot \mathbb{P}\left(\breve{S}_{\hat{t}}^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}} \middle| \hat{t} \leq \tau\right) && \text{(23d)}\\
\geq &\left(1 - \frac{3}{T^2}\right) \cdot \mathbb{P}\left(\breve{S}_{\tau}^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right). && \text{(23e)}
\end{aligned}
$$

The random variable $\breve{S}_{\tau}^{\text{buffer}}$ follows a Poisson distribution with mean $\tau\Delta_{p_l}$ and thus we have:

$$
\begin{aligned}
&\mathbb{P}\left(\breve{S}_{\tau}^{\text{buffer}} \geq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right) && \text{(24a)}\\
\leq &\min_a \; e^{\tau\Delta_{p_l} e^a - \tau\Delta_{p_l} - a\frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}} && \text{(24b)}\\
\leq &e^{-(\log(T))^{3/2}\left(\frac{1+\log(-\log \rho_{p_l}\Delta_\lambda \sqrt{\log(T)})}{\log \rho_{p_l}}\right)} && \text{(24c)}\\
\leq &e^{-(\log(T))^{3/2}} && \text{(24d)}\\
\leq &\frac{1}{T^2}, && \text{(24e)}
\end{aligned}
$$

where (24b) is by Chernoff bound. The inequality (24c) holds because $\frac{1+\log(-\log \rho_{p_l}\Delta_\lambda \sqrt{\log(T)})}{\log \rho_{p_l}} \geq 1$ by Assumption 2.3. Therefore, we can future derive that

$$
\mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\right) \geq \left(1 - \frac{3}{T^2}\right)\left(1 - \frac{1}{T^2}\right) \geq 1 - \frac{4}{T^2}. \tag{25}
$$

The virtual system considers a longer service time for part of the new customers and holds the rest part of the new customers in the buffer queue even there are available resources before $\hat{t}$. Therefore, in probabilistic sense, the virtual system takes longer time to reach steady state:

$$
\mathbb{P}\left(\mathcal{C}_{2\tau}\right) \geq \mathbb{P}\left(\breve{\mathcal{C}}_{2\tau}\right) \geq 1 - \frac{4}{T^2}. \tag{26}
$$

Case (2.2): from p_h to p_l and $\mu_{p_l} \leq \mu_{p_h}$. Without loss of generality, the system starts from $t = 0$ with S_0 customers in the system under service rate μ_{p_h}. Consider a virtual system $\breve{S}$ starts with S_0 customers in the system under service rate μ_{p_l}. The future customers arrive in the process under rate λ_{p_l} with an exponential service time with mean $\frac{1}{\mu_{p_l}}$. To compare the coupling speed of the virtual queue, we consider the speed conditional on $S_0 \leq \bar{S}_l$. Consider another virtual system $\hat{S}$ starts with $\hat{S}_0$, sampled from the steady-state distribution under p_l and keeps in steady-state thereafter. Therefore, similar to Proposition 4, the probability that the actual system S and the virtual system $\breve{S}$ reach the steady-state equals to the probability that these two systems couple with the virtual system $\hat{S}$. We have shown that once they reach empty state together, then they couple with each other. For this case, we show an even stronger result that the actual system reaches the steady-state under p_l within τ. Define events:

$$
\breve{\mathcal{C}}_{\tau}^{0} = \{\breve{S} \text{ reaches the empty state before } \tau\}, \tag{27}
$$

$$
\hat{\mathcal{C}}_{\tau}^{0} = \{\hat{S} \text{ reaches the empty state before } \tau\}, \tag{28}
$$

$$
\mathcal{C}_{\tau}^{0} = \{S \text{ reaches the empty state before } \tau\}. \tag{29}
$$

Based on the above analysis, we can derive:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{C}_\tau\right) && \text{(30a)}\\
\geq&\mathbb{P}\left(\mathcal{C}_\tau|S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq \bar{S}_l\right) && \text{(30b)}\\
\geq&\mathbb{P}\left(\mathcal{C}_\tau^0|S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(\hat{\mathcal{C}}_\tau^0\right) && \text{(30c)}\\
\geq&\mathbb{P}\left(\breve{\mathcal{C}}_\tau^0|S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(\hat{\mathcal{C}}_\tau^0\right) && \text{(30d)}\\
\geq&\mathbb{P}\left(\breve{\mathcal{C}}_\tau^0|S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(S_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(\hat{\mathcal{C}}_\tau^0|\hat{S}_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(\hat{S}_0 \leq \bar{S}_l\right) && \text{(30e)}\\
\geq&\mathbb{P}\left(\breve{\mathcal{C}}_\tau^0|S_0 \leq \bar{S}_h\right) \cdot \mathbb{P}\left(S_0 \leq \bar{S}_h\right) \cdot \mathbb{P}\left(\hat{\mathcal{C}}_\tau^0|\hat{S}_0 \leq \bar{S}_l\right) \cdot \mathbb{P}\left(\hat{S}_0 \leq \bar{S}_l\right) && \text{(30f)}\\
\geq&\left(1-\frac{1}{T^2}\right)^4 && \text{(30g)}\\
\geq&1-\frac{4}{T^2}, && \text{(30h)}
\end{aligned}$$

where $\bar{S}_l = \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}$ and $\bar{S}_h = \frac{(\log(T))^{3/2}}{-\log \rho_{p_h}}$. The inequality (30f) is by Assumption 2.1 and $\mu_{p_l} < \mu_{p_h}$. The inequality (30g) is by the intermediate results we derive in Proposition 4. □

D. Appendix for Section 5.2

For notation simplicity, we use $\|\mathbf{x}\|_M = \sqrt{x^T M x}$ to denote the matrix norm of vector $\mathbf{x}$ induce by a positive semidefinite matrix M. Let $\mathcal{M}_{\lambda,m} = \mathbf{X}_m^T \hat{\Omega}_{\lambda,m}^{-1} \mathbf{X}_m$ and $\mathcal{M}_{\mu,m} = \mathbf{X}_m^T \hat{\Omega}_{\mu,m}^{-1} \mathbf{X}_m$, where the subscript m represents the results and information at the beginning of batch m. Therefore, at the beginning of batch m with a set of implemented prices $p = 1, \ldots, d_f + m - 1$, we can rewrite $U(p)$ as

$$\left(\frac{\hat{\theta}_{\mu,m}^T \mathbf{x}}{\hat{\theta}_{\lambda,m}^T \mathbf{x}} + \frac{\sqrt{32\log(T)}}{\hat{\theta}_{\lambda,m}^T \mathbf{x}} \left(r_{\max} \|\mathbf{x}\|_{\mathcal{M}_{\lambda,m}^{-1}} + \|\mathbf{x}\|_{\mathcal{M}_{\mu,m}^{-1}}\right)\right) p \tag{31}$$

for any new price p with a feature vector $\mathbf{x}$. We also rewrite the confidence radius of price p at the beginning of batch m as

$$\mathbf{Rad}_m(p) = \frac{\sqrt{32\log(T)}}{\hat{\theta}_{\lambda,m}^T \mathbf{x}} \left(r_{\max} \|\mathbf{x}\|_{\mathcal{M}_{\lambda,m}^{-1}} + \|\mathbf{x}\|_{\mathcal{M}_{\mu,m}^{-1}}\right) p.$$

For the sake of regret analysis, we shall further assume that the following assumptions hold.

Assumption 3. The norm of covariates in $\{\mathbf{x}_p,\ p \in [p_L, p_U]\}$ is bounded: there exists $\mathcal{C}_\lambda, \mathcal{C}_\mu < \infty$ such that for all price $p \in [p_L, p_U]$, the corresponding feature vector $\mathbf{x}_p$ and the underlying arrival and service rates satisfy $\lambda_p \|\mathbf{x}_p\|_2 \leq \mathcal{C}_\lambda$ and $\mu_p \|\mathbf{x}_p\|_2 \leq \mathcal{C}_\mu$.

Assumption 4. The smallest eigenvalue of $\sum_{p \in \mathcal{P}_b} \lambda_m^2 \mathbf{x}_p \mathbf{x}_p^T$ is lower-bounded by l_λ and the smallest eigenvalue of $\sum_{p \in \mathcal{P}_b} \mu_m^2 \mathbf{x}_p \mathbf{x}_p^T$ is lower-bounded by l_μ.

These two assumptions are mild, only requiring information about boundary cases of unknown system parameters. Assumption 3 can be satisfied if the firm knows upper bounds of the covariates of any context vectors $\|\mathbf{x}_p\|_2 \leq \mathcal{C}_f$, any arrival rates $\lambda_p \leq \lambda_{\max}$, and any service rates $\mu_p \leq \mu_{\max}$. Then $\mathcal{C}_\lambda$ can take the value of $\lambda_{\max}\mathcal{C}_f$ and $\mathcal{C}_\mu$ can take the value of $\mu_{\max}\mathcal{C}_f$. Assumption 4 can be satisfied if the firm knows an lower bound l_f for the smallest eigenvalue of $\sum_{p \in \mathcal{P}_b} \mathbf{x}_p \mathbf{x}_p^T$, and lower bounds for any arrival rates $\lambda_p \geq \lambda_{\min}$ and any service rates $\mu_p \geq \mu_{\min}$. Then l_λ can take the value of $\lambda_{\min}^2 l_f$ and l_μ can take the value of $\mu_{\min}^2 l_f$. Both assumptions are used in the proof of Lemma 8.

Proof of Theorem 1. We bound the regret of Warm-up and Learning Phases separately.

Regret of the Warm-up Phase. The difference between $J^*_{\text{Warm-up}}$ and $J^{\text{BLinUCB}}_{\text{Warm-up}}$ is at most linear on the length of the Warm-up Phase, which is $O(d_f \log(T))$. Therefore we can directly have

$$\overline{\text{Regret}}_{\text{Warm-up}}(\pi_{\text{BLinUCB}}, T) = O(d_f \log(T)). \tag{32}$$

Regret of nonstationarity during the Learning Phase. The expected revenue collected in batch m can be separated into two parts: the revenue collected during the transient state (i.e., the first 2τ periods) and during the steady state (i.e., the rest periods in this batch). We denote the offered price of batch m in BLinUCB as p_m. Therefore, the expected revenue collected in batch m is

$$J_m(p_m) = J_{m,0\sim 2\tau}(p_m) + J_{m,2\tau\sim I_m\tau}(p_m),$$

where $J_{m,0\sim 2\tau}(p_m)$ denotes the expected revenue collected in the first 2τ periods and $J_{m,2\tau\sim I_m\tau}(p_m)$ denotes the expected revenue collected from $(2\tau+1)$th period to the end of batch m. We define event $\mathcal{C}_{2\tau}$ as that the system reaches the steady state under the new price within the first 2τ periods and define $\bar{\mathcal{C}}_{2\tau}$ as the complementary event of $\mathcal{C}_{2\tau}$. By Proposition 5, we have $\mathbb{P}[\mathcal{C}_{2\tau}] \geq 1 - \frac{4}{T^2}$. Consequently, we define $J_{m,\mathcal{C}_{2\tau}}(p_m)$ to be the expected value of $J_{m,2\tau\sim I_m\tau}(p_m)$ conditional on $\mathcal{C}_{2\tau}$ and $J_{m,\bar{\mathcal{C}}_{2\tau}}(p_m)$ similarly. As a result, we can express the expected revenue during batch m as:

$$\begin{aligned}
J_m(p_m) &= J_{m,0\sim 2\tau}(p_m) + J_{m,\mathcal{C}_{2\tau}}(p_m)\mathbb{P}(\mathcal{C}_{2\tau}) + J_{m,\bar{\mathcal{C}}_{2\tau}}(p_m)\big(1-\mathbb{P}(\mathcal{C}_{2\tau})\big) \\
&\geq O(\tau) + \left(\frac{p_m\lambda_{p_m}}{\mu_{p_m}}\right)(I_m\tau - 2\tau)\left(1-\frac{4}{T^2}\right) + O\left(\frac{I_m\tau - 2\tau}{T^2}\right) \\
&= \left(\frac{p_m\lambda_{p_m}}{\mu_{p_m}}\right) I_m\tau - O(\tau) \\
&= r(\lambda_{p_m}, \mu_{p_m}, p_m) I_m\tau - O(\tau).
\end{aligned} \tag{33}$$

Define $\Delta(p_m) = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$. Recall $J^{\text{LP}}_{\text{Learning, m}} = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) I_m\tau$. Therefore, the relaxed regret of batch m can be bounded by

$$J^{\text{LP}}_{\text{Learning, m}} - J^{\pi_{\text{BLinUCB}}}_{\text{Learning, m}} \leq \Delta(p_m) I_m\tau + O(\tau).$$

Regret of suboptimality during the Learning Phase. We start by analyzing the regret of one batch. If the algorithm selects p_m for batch m, then we can still compute the time average regret of batch m (i.e., the loss of the revenue rate due to suboptimality). Furthermore, we can have $\Delta(p_m)$ bounded by

$$\min\Big\{\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - U_m(\tilde{p})\big) + \big(U_m(p_m) - L_m(p_m)\big) + \big(L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)\big), r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\Big\} \tag{34a}$$

$$\leq \min\Big\{\big(U_m(p_m) - L_m(p_m)\big), r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\Big\} + \big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - U_m(\tilde{p})\big) + \big(L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)\big), \tag{34b}$$

where (34a) is because the revenue rates are all non-negative. In the following proof we will consider the above three terms separately.

We further analyze the first term of (34b).

$$\begin{aligned}
&\min\Big\{\big(U_m(p_m) - L_m(p_m)\big), r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\Big\} \\
=&\min\left\{2\mathbf{Rad}_m(p_m), \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}}\tilde{p}\right\} \\
=&\min\left\{2r_{\max}p_m\sqrt{32\log(T)}\left(\frac{1}{\hat{\theta}^T_{\lambda,m}\mathbf{x}}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\lambda,m}} + \frac{1}{r_{\max}\hat{\theta}^T_{\mu,m}\mathbf{x}_m}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\mu,m}}\right), \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}}\tilde{p}\right\} \\
\leq&\min\left\{2r_{\max}\tilde{p}\sqrt{32\log(T)}\left(\frac{1}{\hat{\theta}^T_{\lambda,m}\mathbf{x}_m}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\lambda,m}} + \frac{1}{r_{\max}\hat{\theta}^T_{\mu,m}\mathbf{x}_m}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\mu,m}}\right), r_{\max}\tilde{p}\right\} \\
\leq& r_{\max}\tilde{p}\cdot\min\left\{2\sqrt{32\log(T)}\left(\frac{1}{\hat{\theta}^T_{\lambda,m}\mathbf{x}_m}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\lambda,m}} + \frac{1}{r_{\max}\hat{\theta}^T_{\mu,m}\mathbf{x}_m}\|\mathbf{x}_m\|_{\mathcal{M}^{-1}_{\mu,m}}\right), 1\right\},
\end{aligned}$$

with $r_{\max} = \max_p \lambda_p/\mu_p$ and $\mathbf{Rad}_m(p_m)$ defined in (31). Take the expectation of both sides, we can derive that the expected value of this time average suboptimality loss is upper bounded by

$$\mathbb{E}\left[\min\left\{2\mathbf{Rad}_m(p_m), \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}}\tilde{p}\right\}\right] \tag{35a}$$

$$\le \mathbb{E}\left[r_{\max}\tilde{p}\cdot\min\left\{2\sqrt{32\log(T)}\left(\mathbb{E}\left[\frac{1}{\hat{\theta}_{\lambda,m}^{T}\mathbf{x}_m}\right]\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}+\mathbb{E}\left[\frac{1}{r_{\max}\hat{\theta}_{\lambda,m}^{T}\mathbf{x}_m}\right]\|\mathbf{x}_m\|_{\mathcal{M}_{\mu,m}^{-1}}\right),1\right\}\right] \tag{35b}$$

$$\le \mathbb{E}\left[r_{\max}\tilde{p}\cdot\min\left\{2\sqrt{32\log(T)}\left(\lambda_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}+\frac{\lambda_{p_m}}{r_{\max}}\|\mathbf{x}_m\|_{\mathcal{M}_{\mu,m}^{-1}}\right),1\right\}\right] \tag{35c}$$

$$\le \mathbb{E}\left[r_{\max}\tilde{p}\cdot\min\left\{2\sqrt{32\log(T)}\left(\lambda_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}+\mu_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\mu,m}^{-1}}\right),1\right\}\right]. \tag{35d}$$

Now we can analyze the cumulative regret by the end of the learning phase.

$$\overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BLinUCB}},T)=\sum_{m=1}^{M}\left(J_{\text{Learning, m}}^{\text{LP}}-J_{\text{Learning, m}}^{\pi_{\text{BLinUCB}}}\right) \tag{36a}$$

$$\le\sum_{m=1}^{M}I_m\tau\min\left\{\big(U_m(p_m)-L_m(p_m)\big),r(\lambda_{\tilde{p}},\mu_{\tilde{p}},\tilde{p})\right\}+\sum_{m=1}^{M}I_m\tau\big(r(\lambda_{\tilde{p}},\mu_{\tilde{p}},\tilde{p})-U_m(\tilde{p})\big) \tag{36b}$$

$$+\sum_{m=1}^{M}I_m\tau\big(L_m(p_m)-r(\lambda_{p_m},\mu_{p_m},p_m)\big)+\sum_{m=1}^{M}O(\tau).$$

We first analyze the first term of (36b). The total number of batches is $\log_2\left(1+\frac{T-d_f\log T}{2(\log(T))^2}\right)=O\big(\log(T)\big)$. Plug in the result in (35) we have:

$$\sum_{m=1}^{M}I_m\tau\min\left\{\big(U_m(p_m)-L_m(p_m)\big),r(\lambda_{\tilde{p}},\mu_{\tilde{p}},\tilde{p})\right\}$$

$$\le\sum_{m=1}^{M}I_m\tau\cdot r_{\max}\tilde{p}\cdot\min\left\{2\sqrt{32\log(T)}\left(\lambda_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}+\mu_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\mu,m}^{-1}}\right),1\right\}+O(\tau\log(T))$$

$$=r_{\max}\tilde{p}\cdot\sum_{m=1}^{M}I_m\tau\cdot\left(\min\left\{2\sqrt{32\log(T)}\lambda_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}},1\right\}+\min\left\{2\sqrt{32\log(T)}\mu_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\mu,m}^{-1}},1\right\}\right)+O(\tau\log(T)). \tag{37}$$

By Cauchy-Schwarz inequality, we can derive that

$$\sum_{m=1}^{M}I_m\tau\cdot\min\left\{2\sqrt{32\log(T)}\lambda_{p_m}\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}},1\right\} \tag{38a}$$

$$\le\sqrt{T}\cdot\sqrt{\sum_{m=1}^{M}I_m\tau\cdot\min\left\{128\log(T)\lambda_{p_m}^2\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}^2,1\right\}} \tag{38b}$$

$$\le\sqrt{T}\cdot\sqrt{\sum_{m=1}^{M}I_m\tau\cdot128\log(T)\cdot\min\left\{\lambda_{p_m}^2\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}^2,\frac{1}{128\log(T)}\right\}} \tag{38c}$$

$$\le\sqrt{T}\cdot\sqrt{\sum_{m=1}^{M}I_m\tau\cdot128\log(T)\left(1+\frac{1}{128\log(T)}\right)\cdot\log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}^2\right)} \tag{38d}$$

$$=\sqrt{(128\log(T)+1)T}\cdot\sqrt{\sum_{m=1}^{M}I_m\tau\cdot\log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}^2\right)} \tag{38e}$$

$$=\sqrt{(128\log(T)+1)T}\cdot O\left(\sqrt{\sum_{m=1}^{M}n_m\cdot\log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|_{\mathcal{M}_{\lambda,m}^{-1}}^2\right)}\right) \tag{38f}$$

$$\le\sqrt{(128\log(T)+1)T}\cdot O\left(\sqrt{O(\tau)+2\mathcal{K}_{\lambda}d_f\log\left(\frac{\mathcal{C}_{\lambda}^2O(T)}{8l_{\lambda}\log(T)}\right)}\right), \tag{38g}$$

where (38d) is by the fact that $x \leq (\alpha+1)\log(1+x)$ with $0 \leq x \leq \alpha$, (38f) is by $O(I_m\tau) = O(n_m)$, and (38g) is from Lemma 7 (see Appendix D).

With (38), we can bound the regret in (37) as:

$$\begin{aligned}
&\sum_{m=1}^{M} I_m\tau \min\Big\{\big(U_m(p_m) - L_m(p_m)\big), r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\Big\} \\
\leq & r_{\max}\tilde{p} \cdot \Bigg[\sqrt{(128\log(T)+1)T} \cdot O\Bigg(\sqrt{O(\tau) + 2d_f\mathcal{K}_\lambda \log\left(\frac{\mathcal{C}_\lambda^2 O(T)}{8l_\lambda \log(T)}\right)} + \\
&\sqrt{O(\tau) + 2d_f\mathcal{K}_\mu \log\left(\frac{\mathcal{C}_\mu^2 O(T)}{8l_\mu \log(T)}\right)}\Bigg)\Bigg] + O(\tau\log(T)) \\
\leq & O\left(\log(T)\sqrt{d_f T\log(T)}\right).
\end{aligned}$$

Below we analyze the second term of (36b), and the same logic applies to the third term, of which we omit the details here.

$$\sum_{m=1}^{M} \mathbb{E}\Big[\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - U_m(\tilde{p})\big)\Big] I_m\tau \tag{39a}$$

$$\leq \sum_{m=1}^{M} \mathbb{E}\Big[\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - U_m(\tilde{p})\big)\mathbb{1}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big)\Big] I_m\tau \tag{39b}$$

$$\leq \sum_{m=1}^{M} \mathbb{E}\Big[r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\mathbb{1}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big) - U_m(\tilde{p})\mathbb{1}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big)\Big] I_m\tau \tag{39c}$$

$$= \sum_{m=1}^{M} r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\mathbb{P}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big) I_m\tau - \sum_{m=1}^{M} \mathbb{E}\Big[U_m(\tilde{p})\mathbb{1}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big)\Big] I_m\tau \tag{39d}$$

$$\leq \sum_{m=1}^{M} r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})\mathbb{P}\big(r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) > U_m(\tilde{p})\big) I_m\tau \tag{39e}$$

$$\leq O\left(\frac{1}{T^3}\right), \tag{39f}$$

where (39e) is because $U_m(\tilde{p}) > 0$ by Definition 3, and (39f) follows Proposition 3.

Summing over the four terms of (36b) we bound the relaxed regret during the learning phase as:

$$\begin{aligned}
\overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BLinUCB}}, T) = & \sum_{m=1}^{M} \mathbb{E}\Big[\Delta(p_m) I_m\tau\Big] + \sum_{m=1}^{M} O(\tau) \\
= & O\left(\log(T)\sqrt{d_f T\log(T)}\right) + O\left(\frac{1}{T^3}\right) + O\big(((\log(T))^3\big) \\
= & O\left(\log(T)\sqrt{d_f T\log(T)}\right).
\end{aligned}$$

Total Regret of BLinUCB algorithm. By adding up the relaxed regret during the Warm-up and Learning Phases, we have the regret of BLinUCB algorithm as:

$$\overline{\text{Regret}}(\pi_{\text{BLinUCB}}, T) = \overline{\text{Regret}}_{\text{Warm-up}}(\pi_{\text{BLinUCB}}, T) + \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BLinUCB}}, T) \tag{40a}$$

$$= O(d_f \log(T)) + O\left(\log(T)\sqrt{d_f T\log(T)}\right) \tag{40b}$$

$$= O\left(d_f \log(T)\sqrt{T\log(T)}\right) \tag{40c}$$

$$= \tilde{O}\left(d_f\sqrt{T}\right). \tag{40d}$$

□

Lemma 7. *Consider $\mathcal{K}_\lambda$ is a positive constant such that for any two prices p_1 and p_2, the collected arrival time observations $n_m(p_1)$ and $n_m(p_2)$ during a certain length of periods ($\geq 2\tau$) has $n_m(p_1) \leq \mathcal{K}_\lambda n_m(p_2)$ with a high probability, i.e., $1 - 1/T^4$. Constant $\mathcal{K}_\lambda$ is defined similarly for the number of service time observations. For BLinUCB algorithm, one has*

$$\sum_{m=1}^{M} n_m \cdot \log\left(1 + \lambda_{p_m}^2 \|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m}}\right) \leq O(\tau) + 2\mathcal{K}_\lambda d_f \log\left(\frac{\mathcal{C}_\lambda^2 O(T)}{8l_\lambda \log(T)}\right)$$

and

$$\sum_{m=1}^{M} n_m \cdot \log\left(1 + \mu_{p_m}^2 \|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\mu,m}}\right) \leq O(\tau) + 2\mathcal{K}_\mu d_f \log\left(\frac{\mathcal{C}_\mu^2 O(T)}{8l_\mu \log(T)}\right).$$

Proof of Lemma 7. Without loss of generality, it is enough to prove the equation for the arrival process, i.e., the equation related to λ. By the definition of $\mathcal{M}_{\lambda,m}$, we have

$$\mathcal{M}_{\lambda,m+1} = \mathcal{M}_{\lambda,m} + n_{\lambda,m}\lambda_m^2 \mathbf{x}_m \mathbf{x}_m^T.$$

We further consider notation $\mathcal{M}_{\lambda,m,i} = \mathcal{M}_{\lambda,m} + i \cdot \lambda_m^2 \mathbf{x}_m \mathbf{x}_m^T$ and thus $\mathcal{M}_{\lambda,m,0} = \mathcal{M}_{\lambda,m}$ and $\mathcal{M}_{\lambda,m,n_m} = \mathcal{M}_{\lambda,m+1}$. Therefore, we have

$$\log\left(1 + \lambda_{p_{m+1}}^2 \|\mathbf{x}_{m+1}\|^2_{\mathcal{M}^{-1}_{\lambda,m+1}}\right) \leq \log\left(1 + \lambda_{p_{m+1}}^2 \|\mathbf{x}_{m+1}\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right), \ \forall i = 1, \ldots, n_{\lambda,m}.$$

By the theory of projection and by the definition of $\mathcal{M}_{\lambda,m,i}$, we further have

$$\log\left(1 + \lambda_{p_{m+1}}^2 \|\mathbf{x}_{m+1}\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right) \leq \log\left(1 + \lambda_{p_m}^2 \|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right), \ \forall i = 1, \ldots, n_{\lambda,m}.$$

The number of observations in one batch is in the same order as the length of the batch, i.e, $n_m = O(I_m\tau)$. Recall the batch size is $I_m\tau = 2^m\tau$, thus with a high probability that $n_m \leq 2\mathcal{K}_\lambda n_{m-1}$. Further we can derive the following bound for $m = 2, \ldots, M$:

$$\begin{aligned} & n_m \cdot \log\left(1 + \lambda_{p_m}^2 \|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m}}\right) \\ \leq & 2\sum_{i=1}^{n_m/2} \log\left(1 + \lambda_{p_{m-1}}^2 \|\mathbf{x}_{m-1}\|^2_{\mathcal{M}^{-1}_{\lambda,m-1,i}}\right) \\ \leq & 2\mathcal{K}_\lambda \sum_{i=1}^{n_{m-1}} \log\left(1 + \lambda_{p_{m-1}}^2 \|\mathbf{x}_{m-1}\|^2_{\mathcal{M}^{-1}_{\lambda,m-1,i}}\right). \end{aligned}$$

Sum over batches $m = 1, \ldots, M$, we can reach the following result:

$$\sum_{m=1}^{M} n_m \cdot \log\left(1 + \lambda_{p_m}^2 \|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m}}\right) \tag{41a}$$

$$\leq O(\tau) + 2\mathcal{K}_\lambda \sum_{m=2}^{M} \sum_{i=1}^{n_{m-1}} \log\left(1 + \lambda_{p_{m-1}}^2 \|\mathbf{x}_{m-1}\|^2_{\mathcal{M}^{-1}_{\lambda,m-1,i}}\right) \tag{41b}$$

$$\leq O(\tau) + 2\mathcal{K}_\lambda d_f \log\left(\frac{\mathcal{C}_\lambda^2 O(T)}{8l_\lambda \log(T)}\right). \tag{41c}$$

where (41c) is by Lemma 8. □

Lemma 8. *The feature vector of the price selected in batch m is denoted by $\mathbf{x}_m$, the number of arrival time observations of this price is denoted by n_m, and the number of service time observations of this price is denoted by n^s_m. Then*

$$\sum_{m=1}^{M-1}\sum_{i=1}^{n_{m-1}} \log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right) \le d_f \log\left(\frac{\mathcal{C}_\lambda^2 O(T)}{8l_\lambda \log(T)}\right).$$

and

$$\sum_{m=1}^{M-1}\sum_{i=1}^{n_{m-1}} \log\left(1+\mu_{p_m}^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\mu,m,i}}\right) \le d_f \log\left(\frac{\mathcal{C}_\mu^2 O(T)}{8l_\mu \log(T)}\right).$$

Proof of Lemma 8. The proof follows similar ideas of the proof of Proposition 3 in Filippi et al. (2010). Without loss of generality, we firstly prove for arrival processes, i.e., the equation related to λ. By the definition of $\mathcal{M}_{\lambda,m,i}$ for $i=0,1,\ldots,n_m$, we have

$$\mathcal{M}_{\lambda,m,i+1}=\mathcal{M}_{\lambda,m,i}+\lambda_m^2\mathbf{x}_m\mathbf{x}_m^T,\ \forall i=1,\ldots,n_m$$

and two boundary equations, $\mathcal{M}_{\lambda,m,0}=\mathcal{M}_{\lambda,m}$ and $\mathcal{M}_{\lambda,m,n_m}=\mathcal{M}_{\lambda,m+1}$. Therefore, we can compute:

$$\det(\mathcal{M}_{\lambda,m+1}) = \det\left(\mathcal{M}_{\lambda,m,n_m-1}(I+\lambda_m\mathcal{M}^{-1/2}_{\lambda,m,n_m-1}\mathbf{x}_m(\lambda_m\mathcal{M}^{-1/2}_{\lambda,m,n_m-1}\mathbf{x}_m)^T)\right) \tag{42a}$$

$$= \det\left(\mathcal{M}_{\lambda,m,n_m-1}\right)\cdot\det\left(I+\lambda_m\mathcal{M}^{-1/2}_{\lambda,m,n_m-1}\mathbf{x}_m(\lambda_m\mathcal{M}^{-1/2}_{\lambda,m,n_m-1}\mathbf{x}_m)^T\right) \tag{42b}$$

$$= \det\left(\mathcal{M}_{\lambda,m,n_m-1}\right)\cdot\left(1+\lambda_m^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,n_m-1}}\right) \tag{42c}$$

$$= \det\left(\mathcal{M}_{\lambda,m}\right)\prod_{i=0}^{n_m-1}\left(1+\lambda_m^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right) \tag{42d}$$

$$= \det\left(\mathcal{M}_{\lambda,1}\right)\prod_{k=1}^{m}\prod_{i=0}^{n_k-1}\left(1+\lambda_k^2\|\mathbf{x}_k\|^2_{\mathcal{M}^{-1}_{\lambda,k,i}}\right), \tag{42e}$$

where (42c) follows the fact that $1+\lambda_m^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m}}$ is an eigenvalue of the matrix $I+\lambda_m\mathcal{M}^{-1/2}_{\lambda,m}\mathbf{x}_m(\lambda_m\mathcal{M}^{-1/2}_{\lambda,m}\mathbf{x}_m)^T$ and that all the other eigenvalues are equal to 1.

Considering the left-hand side of the equation in the lemma, we have

$$\begin{aligned}
&\sum_{m=1}^{M-1}\sum_{i=1}^{n_{m-1}} \log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right)\\
<&\sum_{m=1}^{M}\sum_{i=1}^{n_{m-1}} \log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right)\\
\le&\log\prod_{m=1}^{M}\prod_{i=0}^{n_m-1}\left(1+\lambda_m^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right)\\
=&\log\left(\frac{\det(\mathcal{M}_{\lambda,M})}{\det(\mathcal{M}_{\lambda,1})}\right).
\end{aligned}$$

Denote the collected arrival time observations during the Warm-up Phase as n_w. Note that the trace of $\mathcal{M}_{m+1}$ is upper bounded by $\mathcal{C}_\lambda^2(n_w+\sum_{k=1}^m n_k)$ according to Assumption 3. Then, since the trace of the positive definite matrix $\mathcal{M}_{m+1}$ is equal to the sum of its eigenvalues and det$(\mathcal{M}_{m+1})$ is the product of its eigenvalues, we have det$(\mathcal{M}_{m+1})\le(\mathcal{C}_\lambda^2(n_w+\sum_{k=1}^m n_k))^{d_f}$. In addition, det$(\mathcal{M}_{\lambda,1})\ge(8\log(T)\cdot l_\lambda)^{d_f}$ by Assumption 4 and the algorithmic design that each price $p\in\mathcal{P}_b$ has more than $8\log(T)$ number of observations during the Warm-up phase. Thus

$$\sum_{m=1}^{M-1}\sum_{i=1}^{n_{m-1}} \log\left(1+\lambda_{p_m}^2\|\mathbf{x}_m\|^2_{\mathcal{M}^{-1}_{\lambda,m,i}}\right) \le d_f\log\left(\frac{\mathcal{C}_\lambda^2\sum_{m=1}^M n_m}{8l_\lambda\log(T)}\right) = d_f\log\left(\frac{\mathcal{C}_\lambda^2 O(T)}{8l_\lambda\log(T)}\right).$$

Similarly, we can prove the part for the service process, i.e., the equation related to μ. □