



Operations Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Online Learning and Pricing for Service Systems with Reusable Resources

Huiwen Jia, Cong Shi, Siqian Shen

To cite this article:

Huiwen Jia, Cong Shi, Siqian Shen (2022) Online Learning and Pricing for Service Systems with Reusable Resources. Operations Research

Published online in Articles in Advance 10 Nov 2022

. <https://doi.org/10.1287/opre.2022.2381>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Methods

Online Learning and Pricing for Service Systems with Reusable Resources

Huiwen Jia,^a Cong Shi,^{a,*} Siqian Shen^a

^aDepartment of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109

*Corresponding author

Contact: hwjia@umich.edu,  <https://orcid.org/0000-0002-2633-9278> (HJ); shicong@umich.edu,  <https://orcid.org/0000-0003-3564-3391> (CS); siqian@umich.edu,  <https://orcid.org/0000-0002-2854-163X> (SS)

Received: January 16, 2021

Revised: December 10, 2021; May 4, 2022;
August 23, 2022

Accepted: September 1, 2022

Published Online in Articles in Advance:
November 10, 2022

Area of Review: Stochastic Models

<https://doi.org/10.1287/opre.2022.2381>

Copyright: © 2022 INFORMS

Abstract. We consider a price-based revenue management problem with finite reusable resources over a finite time horizon T . 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, the customer waits in a queue until the next available unit. In this paper, we assume that the firm does not know how the arrival and service rates depend on posted prices, and thus it makes adaptive pricing decisions in each period based only on past observations to maximize the cumulative revenue. Given a discrete price set with cardinality P , we propose two online learning algorithms, termed batch upper confidence bound (BUCB) and batch Thompson sampling (BTS), and prove that the cumulative regret upper bound is $\tilde{O}(\sqrt{PT})$, which matches the regret lower bound. In establishing the regret, we bound the transient system performance upon price changes via a novel coupling argument, and also generalize bandits to accommodate subexponential rewards. We also extend our approach to models with balking and reneging customers and discuss a continuous price setting. Our numerical experiments demonstrate the efficacy of the proposed BUCB and BTS algorithms.

Funding: This research was partially supported by an Amazon research award and the Department of Energy [Award DE-SC0018018].

Keywords: learning • pricing • reusable resources • service systems • multiarmed bandit • coupling analysis

1. Introduction

In recent years, revenue management has achieved considerable success in many application areas including airline management, electric power management, health-care systems, advertising platforms, and online retailing (see, e.g., Talluri and Van Ryzin 2006, Özer and Phillips 2012). Most existing literature focuses on settings with perishable resources in which the product once sold is permanently removed from inventory. There has been limited literature on settings with *reusable resources* that can be repeatedly used to serve customers.

Undoubtedly, revenue management with reusable resources finds many important applications in today's economy, such as cloud computing services, car/bicycle rental services, ride-hailing services, hotel management, project team management, and call center services. In these applications, the firm is endowed with a finite capacity of reusable products. Each customer requests a product, uses it for some random time, and returns it to the firm, at which point the product can be used to serve other customers (see, e.g., Levi and 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 their computational job requests (see, e.g., 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, e.g., Doan et al. 2020). More examples can be found in car/bicycle rental business (Oliveira et al. 2018), parking facility management (Owen and Simchi-Levi 2017), and on-demand service platforms, such as ride-hailing services including Uber and Lyft (Banerjee et al. 2015, Liu and Li 2017, Bimpikis et al. 2019). In all the aforementioned applications, customers' willingness to use the service is affected by the price being offered; however, the decision maker may not know how demand reacts to price changes (see Xu and Li 2013).

1.1. Brief Problem Statement

We consider a general class of revenue management problems with finite reusable resources under incomplete demand-price and service-time information. 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 . The customers arrive following a price-dependent Poisson process and request one unit from c reusable resources on a first-come, first-served basis. 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 demand and service rates have an unknown relation with prices (within a finite and discrete price set). Thus, we need to learn the underlying mapping between prices and rates from past observations, while maximizing the total expected revenue on the fly.

1.2. Main Results and Contributions

We propose new online learning algorithms, termed batch upper confidence bound (BUCB) and batch Thompson sampling (BTS), for service systems with reusable resources. The performance measure is cumulative regret, which is the difference between the revenue attained by our learning algorithms and by a clairvoyant optimal pricing policy under full distributional information. When T is the planning horizon and P is the cardinality of a given discrete price set, we prove in Theorems 1 and 2 that the cumulative regret upper bound of BUCB and BTS is $\tilde{O}(\sqrt{PT})$, which matches the regret lower bound up to a logarithmic factor. We also extend our framework to settings with balking and reneging customers. Our numerical experiments confirm the efficacy of the proposed algorithms. The major contribution of our work is threefold:

a. *Connecting to the multiarmed bandit (MAB) problem via a linear programming benchmark*: The state-of-the-art online learning literature in revenue management predominantly focuses on settings with perishable resources (i.e., the unit once sold is removed from inventory). Their problems belong to the general class of online knapsack problems (e.g., Ferreira et al. 2018, Chen and Shi 2022), which can be naturally linked with multiarmed bandits under knapsack constraints (e.g., Badanidiyuru et al. 2018). In contrast, the reusable resource setting naturally embeds a multiserver queue with intertemporal system dynamics (i.e., customers arriving, waiting, getting served, and leaving). This seems to be a general reinforcement learning (RL) problem, and it is not a priori clear how to

connect the problem to MABs. We derive a *linear programming* upper bound on the clairvoyant optimal revenue, which gives rise to a static price. The linear programming optimality gap is shown to be $o(\sqrt{T})$, which is insignificant compared with $O(\sqrt{T})$ loss due to learning. This allows us to naturally cast the problem as an MAB problem (to learn the optimal static price) through this linear programming benchmark.

b. *Batched learning with mixing times and infrequent actions*: The most critical challenge in the design and analysis of our online learning algorithms lies in the following fact: Whenever the algorithm posts a new price, it is difficult to quantify its impact on the cumulative regret. 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 (to quantify the performance under the newly posted price). We resolve this challenge by proposing a batched MAB approach. More precisely, our algorithm separates the learning phase into successive batches and selects a price using observed information from previous batches. For each batch, we further divide it into two intervals. In the first interval, the system completes serving the previous customers and reaches the steady state under the newly posted price. In the second interval, the system maintains the steady state (exploiting the newly posted price).

There are two major advantages of this batched approach. First, it allows for infrequent $O(\log T)$ price changes, which is more implementable, cost-effective, and customer friendly in practical settings (see Cheung et al. 2017). Second, in establishing the regret, 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 by developing a new coupling analysis for $M/M/c$ queues upon price changes (see Propositions 4 and 5). To the best of our knowledge, this is the first result giving a finite sample bound on mixing times of $M/M/c$ queues, which could be of independent interest.

c. *Heavy-tailed rewards and pruning instability*: The observed information, that is, the interarrival time and service time of an $M/M/c$ queueing system, are exponentially distributed, thereby inducing *unbounded and heavy-tailed random rewards*. This is in stark contrast with the sub-Gaussian rewards generally required in the MAB literature. The key is to achieve similar-order concentration bounds as for light-tailed sub-Gaussian observations. Our algorithms include a warm-up phase, to perform a uniform information collection over all candidate prices, which invokes Bernstein’s inequality and eliminates the heavy-tailed effects of subexponential random variables (see Proposition 2). Another key benefit of uniform exploration is to prune undesirable prices that would result in unstable queues (see Proposition 3).

In addition to the high-level summary of the key contributions above, we also discuss the differences between our approach and the existing MAB/RL approaches in Section 3.2. We believe that our algorithmic framework could open up new directions of learning complex stochastic systems.

1.3. Literature Review

Our work is closely related to following streams of literature.

1.3.1. Revenue Management with Reusable Resources. The literature on revenue management with perishable resources is vast (see Talluri and Van Ryzin 2006, Özer and Phillips 2012, den Boer 2015). We refrain from reviewing a prolonged list, and in this section focus only on the literature of revenue management with reusable resources.

Revenue management with reusable resources considers selling (or renting) reusable resources, which requires the firm to properly balance resource utilization across different service cycles. A key challenge in the reusable resource setting is the need to dynamically match demands with the varying inventory of reusable resources. Several recent studies focused on admission control in a stationary infinite-horizon and single-resource setting with a general service time, and the authors developed provably near-optimal heuristic controls with constant factor performance guarantees (see Levi and Radovanović 2010, Chen and Shi 2017, Chen et al. 2017). Besides admission control, there has been a stream of literature considering static and dynamic pricing. Maglaras (2006) studied static pricing (via a fluid model analysis) in a single reusable resource setting with multiple classes of customers. Araman and Caldentey (2009) formulated the problem as a Poisson intensity control problem and derived structural properties of an optimal solution, and suggested an efficient approximated solution. Xu and Li (2013) studied optimal pricing control under the assumption of exponential service time for cloud computing service providers and analyzed some properties of the optimal pricing control and the maximum cumulative revenue. Also for cloud computing service pricing, Doan et al. (2020) considered a multiresource setting under ambiguous distributions of demand and service time, and constructed efficient heuristic fixed-price pricing policies through robust deterministic approximation models. Lei and Jasin (2020) also considered a multiresource setting but under a deterministic service time assumption, and developed an asymptotically near-optimal real-time pricing control algorithm. Besbes et al. (2022) proved that a static pricing policy simultaneously guarantees 78.9% of the profit, market share, and service level from the optimal policy. Owen and Simchi-Levi (2017) considered a joint assortment and pricing problem under a multiresource setting

and provided constant factor performance guarantees for their proposed controls assuming discrete price points and exponential service time. Rusmevichientong et al. (2020) considered a dynamic assortment and pricing problem with reusable products and proposed a provably near-optimal policy based on approximate dynamic programming, which is guaranteed to obtain at least half of the optimal revenue.

The above literature assumed that the distributional information of the underlying model was known to the decision maker a priori. There has been little work considering online learning for service systems with reusable resources. The closest work to ours is by Chen et al. (2020b), who proposed an online learning framework for a dynamic pricing and capacity sizing problem in a $GI/GI/1$ queue to maximize the cumulative profit. Their model focused on a single-server queue whereas our model focuses on a multiserver queue. Methodologically, their work developed a stochastic gradient descent algorithm to iterate the price and capacity decisions to approach the optimal decision, whereas we develop a batched multiarmed bandit approach. There have been other recent studies in queueing network controls adopting more general reinforcement learning approaches, such as those by Liu et al. (2019), Shah et al. (2020), and Dai and Gluzman (2022).

1.3.2. MAB Problem with Heavy Tails and with Infrequent Action Changes. The multiarmed bandit literature predominantly assumes that rewards follow sub-Gaussian (including bounded) distributions and proves tight regret bounds based on concentration inequalities of the light-tailed sub-Gaussians (see, e.g., Ledoux 2001, Boucheron et al. 2013). Bubeck et al. (2013) was the first to study the MAB problem with heavy-tailed reward distributions. They considered rewards with moments of order $1 + \epsilon$ for some $\epsilon \in (0, 1]$. Their general approach was to replace the empirical mean in the upper confidence index with other robust estimators, such as the truncated empirical mean, Catoni's M-estimator, and the median-of-means estimator. Their results also imply that moments of order two (i.e., finite variance) are sufficient to obtain regret bounds of the same order as under sub-Gaussian reward distributions. Besides Bubeck et al. (2013), there have been other studies devoted to the MAB problem with heavy-tailed reward distributions. For example, Vakili et al. (2013) proposed sequential phases of exploration and exploitation; Korda et al. (2013) and Dubey and Pentland (2019) modified Thompson sampling (TS) algorithms for heavy-tailed rewards; and Xia et al. (2016) studied heavy-tailed rewards under the multiplayer setting, and Dubey and Pentland (2020) considered the multiagent setting. A recent work by Jia et al. (2021) developed three UCB algorithms to solve the MAB problem with sub-exponential rewards. However, their plain MAB setting

does not take into account intertemporal system dynamics (customers arriving, waiting, getting served, and leaving).

MAB problems with infrequent action changes have also been studied in recent years, motivated by the fact that many real-world constraints can prevent the decision maker from conducting extensive experimentation and making frequent action changes. Specifically, Cheung et al. (2017), Chen and Chao (2019), and Chen et al. (2020a) discussed practical reasons for infrequent price changes in revenue management and developed pricing algorithms with limited price switches. Before a systematic analysis of the batched algorithms developed, UCB2 (Auer et al. 2002) and Improved-UCB (Auer and Ortner 2010) were implicitly M -batch policies with $M = \Theta(\log(T))$. Later, Perchet et al. (2016) studied batched MAB problems in the two-armed case under a static batch design and Gao et al. (2019) extended the algorithms and results to the general multiple-armed case and adaptively determined batch size. There has also been a rich literature focusing on batched MAB problems with switching costs. For example, Cesa-Bianchi et al. (2013) considered a uniform switching cost in the objective, and Simchi-Levi and Xu (2019) treated the finite switching cost budget as a hard constraint. In our work, we consider a static batch design that is similar to the batch length structure of UCB2 (Auer et al. 2002) and geometric grid design in Gao et al. (2019), and our algorithms consider $O(\log(T))$ number of batches. The difference is that we do not consider a strict budget on action changes nor an objective for minimizing the action changes. We have this batched action framework to control the loss of the nonstationarity (resulting from the transient time after each action change).

1.3.3. Reinforcement Learning via Markov Decision Process. The process where 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 is widely used for solving MDPs (Szepesvári 2010, Sutton and Barto 2018). However, the state-of-the-art RL approaches, such as UCRL2 (Jaksch et al. 2010), optimistic Q -learning (Jin et al. 2018), and Thompson sampling for RL (Russo et al. 2018), all assume tabular settings with finite state space (or functional approximations). In contrast, the states in our problem include the total number of customers in the system (in service and in the queue), which could be infinite. Also, the existing RL approaches mostly assume immediately observable rewards on the state–action value function. In contrast, our reward observations (of prescribing any new action) are delayed because of the need to stabilize the queueing system (which necessitates a new coupling analysis). We shall discuss these key differences in greater details in Section 3.2.

1.3.4. Transient Analysis of Queues. In the queueing literature, the analysis of $M/M/c$ queues under transient states has a long history. The queue length of the transient $M/M/1$ queue was studied first (see, e.g., Ledermann and Reuter 1954). Although the distribution was explicitly known in terms of modified Bessel functions of the first kind, further studies were conducted to understand how the queue evolves over time. For example, researchers approximated the moments of the queue length (see, e.g., Abate and Whitt (1987) using exponential and hyperexponential approximations and Abate and Whitt (1988) using Laplace transforms) and investigated generating functions for the partial differential equations (Bailey 1954). The transient analysis techniques were further extended to $M/M/c$ queues. Kelton and Law (1985) evaluated the transient distribution function of the delay in queue of the n th arriving customer based on the analysis of the $M/M/1$ queue in Morisaku (1976). Parthasarathy and Sharafali (1989) derived a time-dependent solution for the queue length based on partial differential equations. Kelton and Law (1985) pointed out that one needs to allow the queue to run “long enough” for the effect of the initial starting point to have dissipated and to collect observations during the ensuing “steady-state” portion of the run, and they investigated different starting points to promote rapid convergence. Our work develops a new coupling argument between our system with an arbitrary starting state and a virtual system with a starting state sampled from the steady-state distribution. To the best of our knowledge, our result gives the first finite-time high probability bound on mixing times of $M/M/c$ queues (upon action changes).

1.4. Structure of This Paper

The remainder of this paper is organized as follows. In Section 2, we first describe our problem and define the regret via the linear programming benchmark. In Section 3, we describe the algorithms BUCB and BTS. In Section 4, we analyze the regret bounds of BUCB and BTS, where we first provide in Section 4.1 a coupling analysis as an intermediate key result of algorithmic performance analysis. In Section 5, we extend our results (with a discrete price set) to settings where customers may balk or renege when the system does not have enough capacity. In Section 6, we test the numerical performance of our algorithms and benchmark algorithms using diverse instances. In Section 7, we discuss several future research directions. For better readability, we provide a summary of major notation in Tables A.1, A.2, and A.3 in Appendix A.

2. Problem Formulation

We consider a service-providing firm who serves customers with reusable resources of finite capacity c over a finite horizon T . At the beginning period $t \leq T$, the

firm posts a price p_t from a given candidate set $\mathcal{P}$ to maximize the total cumulative revenue. We assume that $\mathcal{P}$ is finite and discrete and use $P = |\mathcal{P}|$ to denote its cardinality. Under the posted price p , $\forall p \in \mathcal{P}$, customers arrive at the system according to a Poisson process with rate λ_p and request one unit from c reusable resources on a first-come, first-served basis. If there is a resource unit available for use, the arriving customer gets the unit and spends a random time that is exponentially distributed with rate μ_p . Upon service completion, this customer pays the firm the amount of the posted price p 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. Note that the service system under any posted price $p \in \mathcal{P}$ can be reduced to an $M/M/c$ queue model (when the system contains only customers arriving under this price, that is, after finishing serving other customers arriving under the previously posted prices). The assumption of exponential service times is standard in the literature (see, e.g., Savin et al. 2005, Gans and Savin 2007, Owen and Simchi-Levi 2017). We assume that the firm does not know the underlying arrival rate λ_p and service rate μ_p under price p , $\forall p \in \mathcal{P}$, a priori.

Assumption 1. We make the following assumptions on system parameters and time:

1. A constant $r_{\max}$ is known such that $r_{\max} \geq \max_{p \in \mathcal{P}} \lambda_p / \mu_p$;
2. $\log T \geq 4$.

Assumption 1 states the following: (1) A valid upper bound of the ratio between arrival and service rates of all candidate prices is known. Or, equivalently, a valid upper bound on the maximum revenue rate is known, which is a mild and standard assumption in the learning literature. (2) We require the total planning horizon T to be sufficiently long, because it takes time for the underlying service system to reach steady state under any given posted price.

Assumption 2. For any two candidate prices $p_h, p_l \in \mathcal{P}$, 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 states 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). (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 used to justify inequalities (E.5g) and

(E.11c)) 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 $\rho_l \geq \rho_h^2$ 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 change in arrival rates (with respect to price) exceeds the ratio of change in service rates.

The firm aims at finding a periodic-review pricing policy as $\pi : \{(n, t) : n = 0, 1, \dots, \infty, t = 1, \dots, T\} \rightarrow \mathcal{P}$, where the firm selects price $\pi(n, t) \in \mathcal{P}$ for period t when there are n customers in the system at the beginning of period t , with the goal of maximizing the expected total revenue. We discuss the sufficiency of using (n, t) as the state variable in Appendix C. The expected revenue under pricing policy π during period t , $\forall t = 1, \dots, T$, is denoted by J_t^π , and the cumulative expected revenue over periods $\{1, \dots, T\}$ is denoted by $J^\pi = \sum_{t=1}^T J_t^\pi$.

2.1. Relaxed Regret via a Linear Programming Benchmark

The notion of cumulative regret is commonly used in online learning (see, e.g., Shalev-Shwartz 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.

Under full information, the firm knows the underlying arrival rate λ_p and service rate μ_p for all the candidate prices $p \in \mathcal{P}$. If the decision maker posts some price p but the utilization factor $\rho_p = \lambda_p / \mu_p \geq 1$, then the system will be unstable (i.e., the queue length will grow indefinitely). Even if this particular p could generate more revenue (than other stable prices) in the long run, the outcome would be highly undesirable from the system perspective. To avoid such an issue, we define what we call the $(1 - \varepsilon_T)$ -stable price set as follows:

$$\mathcal{P}_{\varepsilon_T} = \{p \in \mathcal{P} : \rho_p \leq 1 - \varepsilon_T\} \subseteq \mathcal{P}, \quad \text{where } \varepsilon_T = \bar{\varepsilon} / \log(T),$$

where $\bar{\varepsilon}$ is a hyperparameter chosen by the firm. In other words, all the prices from $\mathcal{P}_{\varepsilon_T}$ will ensure that the system is stable. We refer to the pricing policies that are restricted to select prices from $\mathcal{P}_{\varepsilon_T}$ as $(1 - \varepsilon_T)$ -stable pricing policies. Thus, we have $(1 - \varepsilon_T)$ -stable state-dependent optimal policy $\pi_{\varepsilon_T}^* = \arg \max_{\pi : (n,t) \rightarrow \mathcal{P}_{\varepsilon_T}} J^\pi$ and $(1 - \varepsilon_T)$ -stable state-dependent optimal expected revenue $J_{\varepsilon_T}^* = J^{\pi_{\varepsilon_T}^*}$. Therefore, we can define the ε_T -regret of any pricing policy π as the difference between $J_{\varepsilon_T}^*$ and the total expected revenue under this policy π , that is, J^π .

In later sections, under incomplete information, we can start with any price set $\mathcal{P}$ (with a potential subset of “unstable” prices), and then prune all the unstable prices $\mathcal{P} \setminus \mathcal{P}_{\varepsilon_T}$ in the uniform exploration phase with high probability. Hence, our online learning algorithms do not need to enforce any uniform stability assumption on $\mathcal{P}$. Instead, the algorithms are “smart” enough to automatically remove potentially unstable prices and converge to the $(1 - \varepsilon_T)$ -stable optimal price.

Definition 1. The cumulative ε_T -regret by the end of period T of an online learning policy π (which does not know the arrival and service rates) is defined as

$$\text{Regret}(\pi, \varepsilon_T, T) = J_{\varepsilon_T}^* - J^\pi. \quad (1)$$

Under full information, finding the state-dependent optimal pricing policy π_ε^* requires solving a dynamic program with an infinite number of potential states, in the form of (n, t) , which is computationally intractable even if all the system parameters are known. Moreover, the transient system performance is complex and hard to analyze when the offered price changes over periods. Thus, it is impractical or intractable to obtain $\pi_{\varepsilon_T}^*$ and $J_{\varepsilon_T}^*$. To tackle this problem, we develop an upper bound on the expected revenue achievable by any $(1 - \varepsilon_T)$ -stable policy, and, therefore, also an upper bound of the $(1 - \varepsilon_T)$ -stable optimal expected revenue.

Given a specific ε_T , we can formulate a linear program (LP) as follows. Consider continuous decision variables π_{nt}^p for $n = 0, \dots, \infty, t = 1, \dots, T, p \in \mathcal{P}_{\varepsilon_T}$; a_{nt} for $n = 0, \dots, \infty, t = 1, \dots, T$; and J_{nt} for $n = 0, \dots, \infty, t = 1, \dots, T$. Then we can formulate the following LP: We also discuss the sufficiency of using (n, t) as the state variable in Appendix C.

$$J_{\varepsilon_T}^{\text{LP}} = \max_{\pi, a, J} \sum_{n=0}^{\infty} \sum_{t=1}^T a_{nt} J_{nt} \quad (2a)$$

$$\text{s.t. } \sum_{n=0}^{\infty} a_{nt} = 1 \quad \forall t = 1, \dots, T; \quad (2b)$$

$$\sum_{p \in \mathcal{P}_{\varepsilon_T}} \pi_{nt}^p = 1 \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T; \quad (2c)$$

$$J_{nt} \leq \sum_{p \in \mathcal{P}_{\varepsilon_T}} \pi_{nt}^p \frac{\lambda_p}{\mu_p} \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T; \quad (2d)$$

$$0 \leq a_{nt} \leq 1 \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T; \quad (2e)$$

$$0 \leq \pi_{nt}^p \leq 1 \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T, \forall p \in \mathcal{P}_{\varepsilon_T}. \quad (2f)$$

Proposition 1. The optimal solution of the LP provides an upper bound on the $(1 - \varepsilon_T)$ -stable optimal revenue, that is, $J_{\varepsilon_T}^* \leq J_{\varepsilon_T}^{\text{LP}}$.

Proof Sketch. Consider an arbitrary $(1 - \varepsilon_T)$ -stable policy $\hat{\pi}$. Let variable $\hat{a}_{nt}$ take the value of probability that the system has n customers at the beginning of period t under this policy. Let variable $\hat{\pi}_{nt}^p = 1$ if the policy $\hat{\pi}$ chooses price p at state (n, t) , and otherwise, $\hat{\pi}_{nt}^p = 0$. Let variable $\hat{J}_{nt}$ take the value of the expected revenue collected before the end of period T from the customers who arrive during period t at state (n, t) . Then we can show that the decision variables $(\hat{\pi}, \hat{a}, \hat{J})$ associated with any admissible $(1 - \varepsilon_T)$ -stable policy satisfy the constraints of the LP in (2), and the expected

revenue of this admissible policy is exactly the corresponding objective value in (2a). Thus, the optimal objective value $J_{\varepsilon_T}^{\text{LP}}$ is greater than or equal to the objective value $J_{1-\varepsilon_T}^*$ of a feasible solution associated any admissible $(1 - \varepsilon_T)$ -stable policy π_{ε_T} as well as the optimal $(1 - \varepsilon_T)$ -stable state-dependent policy $\pi_{\varepsilon_T}^*$. The detailed proof can be found in Appendix C. $\square$

We define

$$\tilde{p} := \tilde{p}_{\varepsilon_T} = \arg \max_{p \in \mathcal{P}_{\varepsilon_T}} \frac{\lambda_p}{\mu_p} p.$$

Given fixed values of ε and T , ε_T is also fixed. Thus, to minimize notation, unless there is ambiguity, we drop the subscript and write $\tilde{p}_{\varepsilon_T} = \tilde{p}$ throughout this paper. Then, an optimal solution $(\tilde{\pi}, \tilde{a}, \tilde{J})$ to the LP defined in (2) is given as follows:

$$\tilde{\pi}_{nt}^{\tilde{p}} = 1, \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T; \quad (3a)$$

$$\tilde{\pi}_{nt}^p = 0, \quad \forall n = 0, \dots, \infty, \forall t = 1, \dots, T, \forall p \in \mathcal{P}_{\varepsilon_T} \setminus \{\tilde{p}\}; \quad (3b)$$

$$\tilde{J}_{0t} = \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}} \tilde{p}, \quad \forall t = 1, \dots, T; \quad (3c)$$

$$\tilde{J}_{nt} = 0, \quad \forall n = 1, \dots, \infty, \forall t = 1, \dots, T; \quad (3d)$$

$$\tilde{a}_{0t} = 1, \quad \forall t = 1, \dots, T; \quad (3e)$$

$$\tilde{a}_{nt} = 0, \quad \forall n = 1, \dots, \infty, \forall t = 1, \dots, T. \quad (3f)$$

Then, we can compute

$$J_{\varepsilon_T}^{\text{LP}} = T \frac{\lambda_{\tilde{p}}}{\mu_{\tilde{p}}} \tilde{p} = \frac{T \lambda_{\tilde{p}}}{\mu_{\tilde{p}}} \tilde{p}. \quad (4)$$

Note that if $\rho_p = \lambda_p / \mu_p < 1$, we use

$$r(\lambda_p, \mu_p, p) := p \lambda_p / \mu_p$$

to denote the stationary revenue rate of price p . We refer to the price $\tilde{p}$ as the (static) optimal $(1 - \varepsilon_T)$ -stable price, that is, the price in $\mathcal{P}_{\varepsilon_T}$ with the highest stationary revenue rate. This result can be seen as the “reusable analog” of the classical static pricing upper bound for the canonical revenue management problem with perishable resources (see Gallego and Van Ryzin 1994).

A standard Markov chain analysis reveals that for the $M/M/c$ queue in steady state (under a stable price p), the average number of customers in service is precisely $\frac{\lambda_p}{\mu_p}$, which is equal to $c \rho_p$. Thus, the long-run average revenue rate for price p is $\frac{\lambda_p}{\mu_p} p$. It is natural to choose $\tilde{p}$ that maximizes $\frac{\lambda_p}{\mu_p} p$, and the corresponding long-run revenue $J^{\text{LP}} = \frac{T \lambda_{\tilde{p}}}{\mu_{\tilde{p}}} \tilde{p}$ (in which c does not show up explicitly).

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

This corollary follows directly from Propositions 1 and 4. This result spells out the fact that the static price is a near-optimal benchmark in the sense that the optimality gap is only $o(\sqrt{T})$, which is insignificant compared with the tight square-root learning loss (developed later).

Definition 2. The relaxed ε_T -regret is defined as

$$\overline{\text{Regret}}(\pi, \varepsilon_T, T) = J_{\varepsilon_T}^{\text{LP}} - J^\pi. \quad (5)$$

By Proposition 1, we have that for any policy π , the relaxed regret provides a valid upper bound of its regret, that is, $\overline{\text{Regret}}(\pi, \varepsilon_T, T) \geq \text{Regret}(\pi, \varepsilon_T, T)$. Note that the relaxed regret is as tight as the original regret because of the $o(\sqrt{T})$ suboptimality gap in Corollary 1 (which is insignificant compared with the tight square-root learning loss developed later). In the following sections, we demonstrate the performance of proposed algorithms by analyzing the relaxed regret.

3. Online Learning and Pricing Algorithms: BUCB and BTS

We propose two new online learning algorithms, namely, the online batch upper confidence bound algorithm, in Section 3.1, and online batch Thompson sampling algorithm, in Section 3.3. These two algorithms can be seen as batched MAB problems with mixing times and infrequent action changes. We provide interested readers with some preliminary knowledge of MAB problems in Appendix B. We compare our algorithms and existing learning approaches in Section 3.2.

3.1. Online Batch Upper Confidence Bound Algorithm

We describe how BUCB works. We separate T periods into two phases, a warm-up phase and a learning phase. During the algorithm implementation, we have the following observations associated with each $p \in \mathcal{P}$ on hand. We use $n(p)$ to denote the number of arrived customers under price p . For each customer i arrived under the price p , we use $\hat{d}_i(p)$ to denote the observed interarrival time between customers $i-1$ and i . Then, we compute $\check{d}(p) = \frac{n(p)}{\sum_{i=1}^{n(p)} \hat{d}_i(p)}$ as an empirical estimate of the arrival rate. Let $n^s(p)$ represent the number of customers who have been successfully served. For these customers, we use $\hat{g}_i(p)$ to denote the observed service time of customer i and compute the empirical mean of the service time as $\bar{g}(p) = \frac{\sum_{i=1}^{n^s(p)} \hat{g}_i(p)}{n^s(p)}$.

Algorithm 1 (Online Batch Upper Confidence Bound Algorithm (BUCB))

1: **Input:** $T, \mathcal{P}, \bar{\varepsilon}, r_{\max}, c$

2: **Initialize:** $\{I_m = 2^m\}, m = 1$, and

$$\varepsilon_T = \bar{\varepsilon} / \log T, \quad \delta = \frac{\bar{\varepsilon}}{(2(1 + r_{\max}/c) \log(T))},$$

$$N_W = \max\left\{8 \log T, \frac{\log(T)}{2\delta^2}\right\},$$

$$\tau = (\log T)^2, \quad \rho_W = 1 - \frac{\delta(1 + r_{\max}/c)}{1 + \delta}.$$

3: **Warm-up phase:**

4: **for** $p \in \mathcal{P}$ **do**

5: Offer p , record $\hat{d}_i(p)$ for arriving customers and $\hat{g}_i(p')$, $\forall p' \in \mathcal{P}$ for leaving customers.

6: **if** $n_W^s(p) \geq N_W$ **then**

7: Continue.

8: **end if**

9: **end for**

10: Compute $\check{d}_W(p)$, $\bar{g}_W(p)$, and $\hat{\rho}_W(p)$, $\forall p \in \mathcal{P}$.

11: Prune “unstable” prices $\{p \in \mathcal{P} : \hat{\rho}_W(p) > \rho_W\}$ and retain $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$.

12: Compute $U_0(p) = (\bar{g}_W(p)\check{d}_W(p) + 4(r_{\max} + \bar{g}_W(p)\check{d}_W(p))\sqrt{\frac{2\log(T)}{n_W^s(p)}}p$, $\forall p \in \mathcal{P}_L$.

13: **Learning phase:**

14: Compute the length of batch m by $\min\{I_m\tau$, the number of remaining periods}.

15: **while** the number of remaining periods > 0 **do**

16: Choose $p_m = \arg \max_{p \in \mathcal{P}_L} U_{m-1}(p)$ and offer p_m in batch m .

17: Record $\hat{d}_i(p_m)$ for arriving customers and $\hat{g}_i(p)$, $\forall p \in \mathcal{P}_L$ for leaving customers.

18: Compute and update $\check{d}_m(p)$, $\bar{g}_m(p)$, $n_m(p)$, and $n_m^s(p)$, $\forall p \in \mathcal{P}_L$.

19: Compute $U_m(p) = (\bar{g}_m(p)\check{d}_m(p) + 4(r_{\max} + \bar{g}_m(p)\check{d}_m(p))\sqrt{\frac{2\log(T)}{n_m^s(p)}}p$, $\forall p \in \mathcal{P}_L$.

20: Update $m = m + 1$ and the number of remaining periods.

21: Compute the length of batch m by $\min\{I_m\tau$, the number of remaining periods}.

22: **end while**

3.1.1. Warm-Up Phase. The first phase is the warm-up phase, where we offer each price one by one and collect the information. For notational simplicity, let $\delta = \bar{\varepsilon} / (2(1 + r_{\max}/c) \log(T))$. We collect N_W number of interarrival and service time observations, where $N_W = \max\{8 \log T, (\log T)/2\delta^2\}$. We denote the observations by the end of the warm-up phase by $(n_W(p), \check{d}_W(p), n_W^s(p), \bar{g}_W(p))$. Therefore, the total length of this warm-up phase is $O(P(\log T)^3)$. Then, we compare the empirical utilization factor $\hat{\rho}_W(p) = \bar{g}_W(p)\check{d}_W(p)$ and a threshold value $\rho_W = 1 - \delta(1 + r_{\max}/c) / (1 + \delta)$. If the empirical utilization factor of a candidate price $p \in \mathcal{P}$ is larger than this

threshold, that is, $\hat{\rho}_W(p) > \rho_W$, then this price is unstable with high probability and is therefore *pruned* from the candidate price set. We refer to the resulting price set after pruning as the learning price set and denote it by $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$. Intuitively, $\mathcal{P}_L$ retains all the stable prices with high probability.

This warm-up phase (uniform exploration) is judiciously designed for the following two reasons:

a. First, it can eliminate the heavy-tail effects caused by the exponential interarrival and service times of $M/M/c$ queues and therefore achieve a similar-order tight concentration bound as for sub-Gaussian observations (see Proposition 2).

b. Second, it can prune all the “unstable” candidate prices whose utilization factors are larger than one and retain all the “stable” candidate prices whose utilization factors are less than $1 - \varepsilon_T$ with high probability (see Proposition 3).

Proposition 2. For any price $p \in \mathcal{P}$, if the number of service time observations $n^s \geq 8\log(T)$, then we have

$$\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}(p)\check{d}(p)\right| \leq 4\left(r_{\max} + \bar{g}(p)\check{d}(p)\right)\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \geq 1 - \frac{4}{T^4}.$$

Proposition 3. We have following properties of $\mathcal{P}$ and $\mathcal{P}_L$ after the warm-up phase:

1. For price $p \in \mathcal{P}_L$, $\mathbb{P}(\rho_p \geq 1) \leq 4/\sqrt{T}$.
2. For price $p \in \mathcal{P}$ with $\rho_p \leq 1 - \varepsilon_T$, $\mathbb{P}(p \in \mathcal{P}_L) \geq 1 - 4/\sqrt{T}$.

The proofs of Propositions 2 and 3 are based on the concentration inequality for subexponential random variables (see, e.g., Boucheron et al. 2013) and are provided in Appendix D. This result shows that after $8\log T$ observations, the negative effects of the heavy-tailed distribution can be eliminated and the concentration inequality is as tight as that for sub-Gaussians. Based on the value of N_W , (i) the tight concentration inequality in Propositions 2 can be used after the warm-up phase, and (ii) any price in the learning price set has a steady state with high probability.

3.1.2. Learning Phase. The second phase is the learning phase. We separate the learning phase into M number of batches. Define $I_m = 2^m$, $\forall m = 1, \dots, M$. Each batch m contains $I_m \tau$ periods, where $\tau = (\log(T))^2$. Note that the batch size is exponentially increasing, resulting in $O(\log T)$ infrequent action changes. We denote the observations by the end of batch m by $(n_m(p), \check{d}_m(p), n_m^s(p), \bar{g}_m(p))$. Therefore, we have that $n_m(p) \propto n_m^s(p)$, $\forall m = 1, \dots, M, p \in \mathcal{P}_L$.

Based on Proposition 2, we define the corresponding upper and lower confidence bounds of the stationary revenue rate associated with price $p \in \mathcal{P}_L$ by the end of batch m as follows.

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

$$U_m(p) = \left(\bar{g}_m(p)\check{d}_m(p) + 4\left(r_{\max} + \bar{g}_m(p)\check{d}_m(p)\right)\sqrt{\frac{2\log(T)}{n_m^s(p)}}\right)p, \quad (6)$$

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

$$L_m(p) = \left(\bar{g}_m(p)\check{d}_m(p) - 4\left(r_{\max} + \bar{g}_m(p)\check{d}_m(p)\right)\sqrt{\frac{2\log(T)}{n_m^s(p)}}\right)p. \quad (7)$$

Define $\text{Rad}_m(p) = 4\left(r_{\max} + \bar{g}_m(p)\check{d}_m(p)\right)\sqrt{\frac{2\log(T)}{n_m^s(p)}}p$ as the confidence radius of price p by end of batch m . Then, for some $\gamma > 0$, two properties hold (see, e.g., Slivkins 2019):

$$\mathbb{E}[(r(\lambda_p, \mu_p, p) - U_m(p))^+] \leq \frac{\gamma}{PT}, \quad (8)$$

$$\mathbb{E}[(L_m(p) - r(\lambda_p, \mu_p, p))^+] \leq \frac{\gamma}{PT}. \quad (9)$$

At the beginning of each batch $m + 1$, BUCB chooses price p with the highest $U_m(p)$ and offers this price during batch $m + 1$. We present the algorithmic details in Algorithm 1. When offering price p , it is possible that one current customer in service was accepted under another price p' , which was provided in previous periods. Therefore, we record arrival information of price p and also service information of any other price $p' \in \mathcal{P}_L$.

3.2. Necessity of BUCB with Mixing Time Analysis and Comparison Against Existing MAB/RL Approaches

It is worth pausing to explain why any existing MAB/RL approach will not work for our setting. Our approach, however, fully leverages the underlying structure of the stochastic system and provides the most parsimonious way of tackling this problem (with the additional benefit of $O(\log T)$ infrequent action changes, which is more implementable in practical settings).

3.2.1. BUCB Compared with Naïve UCB. One cannot run UCB (or simple variants) on the revenue rate once a new interarrival or service time observation is obtained. If the goal of the algorithm were to purely identify the optimal static price after T time periods (i.e., solving the so-called “best-arm identification” problem in MAB problems), then, because of the simple revenue rate structure, a simple UCB (with minor adaptations) without the batch design would suffice. However, estimating the revenue rate is *not* equivalent to collecting running reward in the MAB process. The goal of our

algorithm is to minimize the cumulative regret compared with the clairvoyant optimal policy. Therefore, we need to explicitly account for the revenue collected during the algorithm implementation. Once the price changes, the system immediately enters a transient state, and the revenue collected under this state is intractable. Thus, to evaluate the performance of the algorithm, we require the system to reach steady state in the learning phase to quantify the expected revenue that is collected when implementing the algorithm. This evaluation necessitates an exact finite time analysis on mixing times of the underlying queueing system upon price changes (see the detailed analysis in Section 4.1).

3.2.2. BUCB Compared with Explore-Then-Commit. We also discuss the popular explore-then-commit algorithm that is widely implemented in practical settings. For our system, one could first learn the arrival rates and service times (and thereby the revenue rates) for each candidate price $p \in \mathcal{P}$ and then optimize. Whenever a new price is posted, there is complication of waiting for the “old” customers who arrived under the previous posted price to exit the system. However, consider a much simpler fictitious system that could “boot” customers in the queue at will without any penalty costs. Consider the following simple algorithm based on uniform exploration:

1. Exploration phase: Try each price N times and collect the rates; whenever posting a new price, boot any old customers in the queue (who arrived under the previous price).
2. Select the price $\hat{p}$ with the highest average empirical reward (breaking ties arbitrarily).
3. Exploitation phase: Stick to price $\hat{p}$ in all remaining times $T - PN$.

It is not difficult to show that we should set $N = \tilde{O}((T/P)^{2/3})$, inducing a regret of $\tilde{O}(T^{2/3})$, which is not optimal (see, e.g., Slivkins 2019). A caveat is that for the exploitation phase in the third step, we still need to invoke a similar coupling analysis for $\hat{p}$ (to reach steady state and then quantify regret), even for this very simple algorithm design. On a tangential note, Besbes and Zeevi (2009) gave a regret bound $\tilde{O}(T^{3/4})$ for a similar explore-then-commit strategy for pricing perishable resources. Our batched MAB approach not only achieves the optimal theoretical regret, but also significantly outperforms the explore-then-commit strategies numerically (see the comparison in Section 6).

3.2.3. BUCB Compared with UCRL2. We briefly discussed Section 1.3 why any existing RL approach cannot be readily adapted for our setting. Here we pick the celebrated UCRL2 algorithm for a more detailed comparison. UCRL2 (Jaksch et al. 2010) and its improved variants (Bartlett and Tewari 2009, Bourel et al. 2020,

Fruit et al. 2020) leverage the key concept of “optimism in the face of uncertainty” to construct a confidence set of MDPs based on collected observations and use the optimistic MDP policy (which gives the highest reward among the constructed set) in a phased manner. One may attempt to write our problem in an MDP formation and solve it using UCRL2. Unfortunately, this will not work for the following reasons.

First, UCRL2 (Jaksch et al. 2010) and its improved variants (Bartlett and Tewari 2009, Bourel et al. 2020, Fruit et al. 2020) require a finite state space in constructing the algorithm (more specifically, constructing the radii of the set of plausible MDPs) and also in proving its regret bound. Unfortunately, for our problem, the states include the total number of customers in the system (in service and in queue), which could be *infinite*. Because every state in the underlying Markov chain is positive recurrent, any simple adaptations of UCRL2 will fail to work (in that it is unclear how to construct the radii of the set of plausible MDPs).

Second, UCRL2 imposes a very strong finite-diameter assumption on its regret analysis. It essentially requires a finite “hitting time” from one state to any other state. In contrast, in our analysis, we connect to a simpler batched bandit setting via a linear programming upper bound. Then, when switching prices at the start of each batch, we can bound the “coalescence time” of two stochastic processes (one with an arbitrary starting state and the other fictitious one already in steady state), which does not need any strong finite-diameter assumption. Our analysis is made possible by connecting the problem to the MAB problem via an upper-bounding LP as well as a novel finite-sample analysis on the mixing time of $M/M/c$ queues. The standard RL approaches do not exploit these structures fully, and their regret bounds depend linearly on the diameter and the size of state space, which would become vacuous in our setting.

Third, the definitions of “transient state” and “mixing time” in UCRL2 are very different from ours. In the UCRL2 setting, a transient state refers to a state that is not reachable under any policy, whereas in our setting, a transient (system) state refers to the state of the underlying queueing system not yet reaching steady state. (Here, a steady state means that the distribution of the number of customers within the system stays the same over time.) In the UCRL2 setting, the mixing time (also called the diameter) refers to the average steps between two arbitrary states s and s' , whereas in our setting, the mixing time (in the conventional queueing sense) refers to the time duration for the underlying queueing system to reach steady state starting from an arbitrary initial state. As a result, UCRL2’s so-called “good” and “bad” states are different from ours, resulting in totally disparate regret analyses.

Fourth, the reward structures are different. UCRL2 relies on the immediate (and bounded) reward observations on

the state-action value function. However, our reward observations (of prescribing any new action) are delayed because of the need to stabilize the queueing system (which would necessitate the coupling analysis). Also, the batch size in UCRL2 is adaptive, whereas ours is predetermined and nonadaptive.

Finally, from the viewpoint of practical implementation, UCRL2 requires solving extended value iterations of a potentially large set of plausible MDPs to extract the optimistic MDP policy (that gives the highest reward) in each batch. It is well known in the literature that this step is theoretically convenient but computationally very expensive, especially when the radii of the constructed confidence sets (for the plausible MDPs) are large and continuous. We instead take an entirely different approach of leveraging the problem structure to cast the problem as an MAB problem (via an upper-bounding LP). In each batch, the decision maker needs only to compute the UCB of each price and post the price with the highest index. Our approach is simple to implement, computationally very light, and, therefore, can be well appreciated by corporate partners and high-level managers.

3.3. Online Batch Thompson Sampling Algorithm

Consider that the firm has some prior information, which equivalently means that the problem instance is drawn from a known prior distribution. To solve this revenue management (RM) problem under prior information, we develop an online batch Thompson sampling algorithm. By the conjugate prior theory (see, e.g., Minton et al. 1962, Diaconis and Ylvisaker 1979), we assume that the firm holds Gamma prior distributions for the rates. Specifically, for price $p \in \mathcal{P}$, we denote the prior distribution of λ_p by $\text{Gamma}(\alpha^{\lambda_p}, \beta^{\lambda_p})$ and the prior distribution of μ_p , $\forall p \in \mathcal{P}$, as $\text{Gamma}(\alpha^{\mu_p}, \beta^{\mu_p})$. The posterior distribution of λ_p is updated by $\text{Gamma}(\alpha^{\lambda_p} + 1, \beta^{\lambda_p} + \hat{d}_i(p))$, after observing one interarrival time $\hat{d}_i(p)$. Similarly, the posterior distribution of μ_p is updated by $\text{Gamma}(\alpha^{\mu_p} + 1, \beta^{\mu_p} + \hat{g}_i(p))$ after observing one service time $\hat{g}_i(p)$.

BTS has two phases: (i) a warm-up phase, which is identical to the warm-up phase of BUCB, and (ii) a batched learning phase, which is specifically designed with regard to Gamma prior distributions. Regarding the warm-up phase, because we use the same algorithm design as for BUCB, Propositions 2 and 3 are still valid for the warm-up phase of BTS. Regarding the learning phase, as in BUCB, we separate the learning phase into M number of batches, and the length of batch m is $I_m \tau$. At the beginning of batch m , $\forall m = 1, \dots, M$, we sample parameters $\hat{\lambda}_p$ and $\hat{\mu}_p$ independently following the updated posterior distributions and then offer the price with the maximum sampled stationary revenue rate $r(\hat{\lambda}_p, \hat{\mu}_p, p) = \frac{\hat{\lambda}_p}{\hat{\mu}_p} p$. We present the algorithmic details in Algorithm 2.

Algorithm 2 (Online Batch Thompson Sampling Algorithm (BTS))

- 1: **Input:** $T, \mathcal{P}, \bar{\varepsilon}, r_{\max}, c, \alpha^{\lambda_p}, \beta^{\lambda_p}, \alpha^{\mu_p}, \beta^{\mu_p}, \forall p \in \mathcal{P}$.
- 2: **Initialize:** $\{I_m = 2^m\}, m = 1$, and

$$\varepsilon_T = \bar{\varepsilon} / \log T, \quad \delta = \frac{\bar{\varepsilon}}{(2(1 + r_{\max}/c) \log(T))'}$$

$$N_W = \max \left\{ 8 \log T, \frac{\log(T)}{2\delta^2} \right\},$$

$$\tau = (\log T)^2, \quad \rho_W = 1 - \frac{\delta(1 + r_{\max}/c)}{1 + \delta}.$$

3: Warm-up phase:

4: for $p \in \mathcal{P}$ do

5: Offer p , record $\hat{d}_i(p)$ for arriving customers and $\hat{g}_i(p')$, $\forall p' \in \mathcal{P}$ for leaving customers.

6: if $n_W^s(p) \geq N_W$ then

7: Continue.

8: end if

9: end for

10: Compute $\check{d}_W(p)$, $\bar{g}_W(p)$, and $\hat{\rho}_W(p)$, $\forall p \in \mathcal{P}$.

11: Prune “unstable” prices $\{p \in \mathcal{P} : \hat{\rho}_W(p) > \rho_W\}$ and retain $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$.

12: Update $\alpha^{\lambda_p} = \alpha^{\lambda_p} + n_W(p)$, $\beta^{\lambda_p} = \beta^{\lambda_p} + \sum_{i=1}^{n_W(p)} \hat{d}_i(p)$,
 $\alpha^{\mu_p} = \alpha^{\mu_p} + n_W^s(p)$, $\beta^{\mu_p} = \beta^{\mu_p} + \sum_{i=1}^{n_W^s(p)} \hat{g}_i(p)$, $\forall p \in \mathcal{P}_L$.

13: Learning phase:

14: Compute the length of batch m by $\min\{I_m \tau, \text{the number of remaining periods}\}$.

15: while the number of remaining periods > 0 do

16: Sample $\hat{\lambda}_p \sim \text{Gamma}(\alpha^{\lambda_p}, \beta^{\lambda_p})$ and $\hat{\mu}_p \sim \text{Gamma}(\alpha^{\mu_p}, \beta^{\mu_p})$, $\forall p \in \mathcal{P}$.

17: Choose $p_m = \arg \max_{p \in \mathcal{P}_L} r(\hat{\lambda}_p, \hat{\mu}_p, p)$ and offer p_m in batch m .

18: Record $\hat{d}_i(p_m)$ for arriving customers and $\hat{g}_i(p)$, $\forall p \in \mathcal{P}_L$, for leaving customers.

19: Update $\alpha^{\lambda_p} = \alpha^{\lambda_p} + (n_m(p) - n_{m-1}(p))$, $\beta^{\lambda_p} = \beta^{\lambda_p} + \sum_{i=n_{m-1}(p)+1}^{n_m(p)} \hat{d}_i(p)$, $\alpha^{\mu_p} = \alpha^{\mu_p} + (n_m^s(p) - n_{m-1}^s(p))$,
 $\beta^{\mu_p} = \beta^{\mu_p} + \sum_{i=n_{m-1}^s(p)+1}^{n_m^s(p)} \hat{g}_i(p)$, $\forall p \in \mathcal{P}_L$.

20: Update $m = m + 1$ and the number of remaining periods.

21: Compute the length of batch m by $\min\{I_m \tau, \text{the number of remaining periods}\}$.

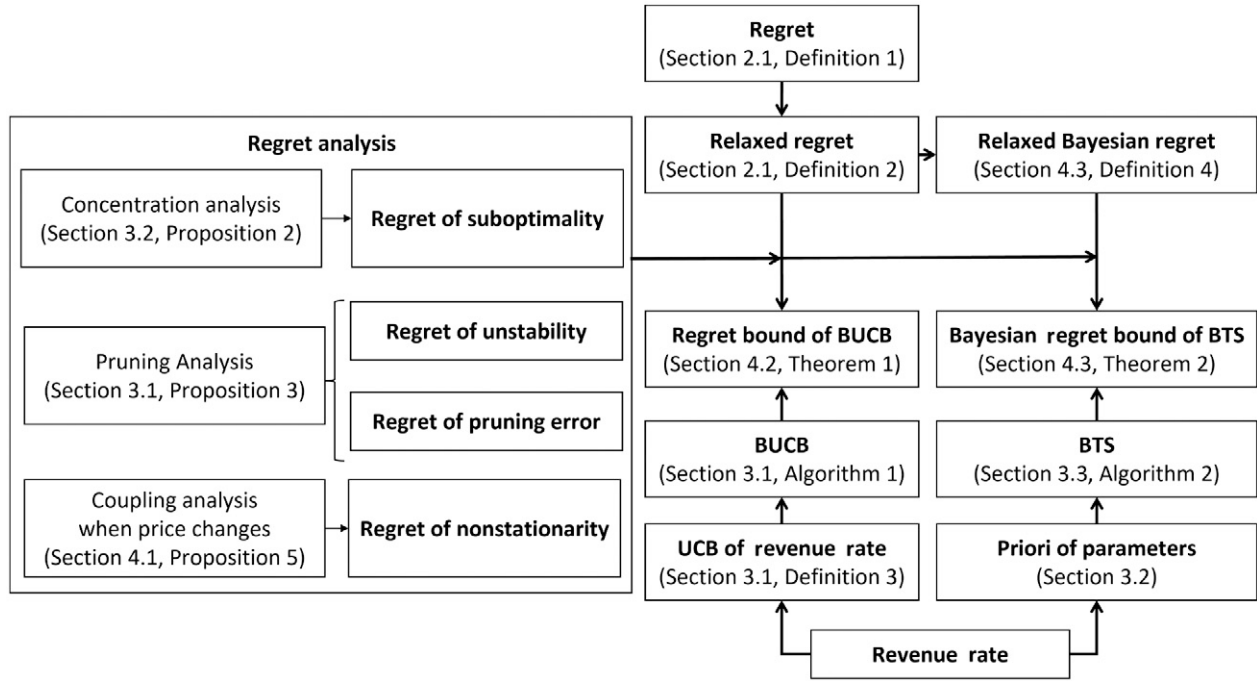
22: end while

4. Performance Analysis of BUCB and BTS

We analyze the performance by the relaxed ε -regret defined in Definition 2, which is defined as the difference between the revenue collected of our algorithms and $T\tilde{p}\lambda_{\tilde{p}}/\mu_{\tilde{p}}$ with $\tilde{p} = \arg \max_{p \in \mathcal{P}_{\varepsilon_T}} \frac{\lambda_p}{\mu_p} p$.

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

Figure 1. Road Map of Regret Analysis for BUCB and BTS



steady state under another price (Proposition 5) in Section 4.1. Then, based on the probability analysis, we provide the regret bound for BUCB in Section 4.2 and for BTS in Section 4.3. A high-level road map of our regret analysis is shown in Figure 1.

4.1. High-Probability Bound for Mixing Time When the Price Changes

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 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 new 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.

We first study the coupling analysis for the $M/M/c$ queue starting from the empty state. By Proposition 3, we focus on a system under fixed arrival rate λ and service rate μ with $\rho < 1$. Let S_t denote the target system that starts from the empty state, and let $\hat{S}_t$ 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 $S_t(\hat{S}_t)$, $t \geq 0$ 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 , $\forall n = \{1, \dots, \infty\}$, defined as the minimum time that the system takes to have $n - 1$ customers if starting with n customers in the system, that is, $A_n = \min\{t: n \text{ customers in the system at time } 0^+ \text{ and } n - 1 \text{ customers in the system at time } t\}$. Note that A_n , $n = 1, \dots, c$, are random variables following different distributions due to difference in the number of servers in use, whereas A_n , $n = c, \dots, \infty$, are independent and identically distributed (i.i.d.) random variables (Omahen and Marathe 1978, Daley and Servi 1998).

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

$$\mathcal{A}_t = \{S_t = \hat{S}_t \text{ after time } t\}. \quad (10)$$

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

Proposition 4 (Coupling Probability of the $M/M/c$ Queue Starting from the Empty State). *For $t \geq \tau$, where $\tau = (\log(T))^2$, the probability that $\hat{S}_t$ and S_t couple is bounded as follows:*

$$\mathbb{P}(\mathcal{A}_t) \geq 1 - \frac{2}{T^2}.$$

The proof of Proposition 4 is provided in Appendix E, which follows the law of total probability, the length of the time horizon T in Assumption 1, the first-order stochastic dominance in Lemma E.1, and the concentration inequality for independent samples in Lemma E.2. Note

that this result gives a uniform lower bound for all system parameters; one could obtain a parametric lower bound from the proof (that is decreasing in ρ), which suggests that it is slower to reach stationarity when the utilization factor ρ increases.

Then, we extend the results to the probability bound when the system is starting from the steady state of the previous price. We prove that the new steady state can be reached uniformly fast with high probability, which is encapsulated in Proposition 5.

Proposition 5 (Coupling Probability of the $M/M/c$ Queue when the 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 $\tau = \log(T)^2$.*

Proof Sketch. When implementing our learning algorithms, whenever a new price is posted, the customers that arrived under the previous posted price still remain in the system. Therefore, we decompose possible price changes into the following two cases: the first is when the price changes from a lower price p_l to a higher price p_h , and the second is the reverse. For each case, we further consider two subcases: (i) the unit service rate $\mu_{p_l} > \mu_{p_h}$ and (ii) $\mu_{p_l} \leq \mu_{p_h}$. For each case and subcase, we construct an alternative (bridging) system and derive the probability bound of the mixing time of the corresponding alternative system. Specifically, we develop a new coupling argument between this alternative system and a virtual system starting from a state sampled from the steady-state distribution and maintaining the steady state thereafter. Last, we show that the actual mixing probability is bounded from below by the probability of the alternative system. The full proof is given in Appendix E. To the best of our knowledge, this is the first result giving a finite-time high-probability bound on mixing times of $M/M/c$ queues (upon action changes). $\square$

4.2. Performance of BUCB

With the high-probability bound for mixing time developed in Section 4.1 and the key result in Proposition 5, we are ready to carry out the regret analysis of BUCB.

Theorem 1. *The T -period cumulative ε_T -regret of BUCB is bounded by $\tilde{O}(\sqrt{PT})$.*

Note that the regret lower bound for the nominal MAB is $\Omega(\sqrt{PT})$ (see, e.g., theorem 2.1 in Slivkins 2019), which suggests that our regret bound is tight, up to a logarithmic factor. Regarding the problem-dependent bound, one could follow a similar scheme to establish an $O(\log T \cdot \frac{P}{\kappa})$ regret upper bound, where κ is revenue rate gap between the first and second best prices (capturing the hardness of the problem). We omit the details here in the interest of space.

Proof Sketch. The proof consist 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, that is, $O(P(\log(T))^3)$. The regret of the learning phase can be further decomposed into four parts, the loss of instability, nonstationarity, pruning error, and suboptimality:

a. The loss of instability is induced by the risk that a (low) price in $\mathcal{P}$ has a utilization factor larger than one, which will potentially make the queue length grow indefinitely. By Proposition 3, we can bound the probability that all the selected prices in the learning phase have utilization factors less than one to be at least $1 - 4P/\sqrt{T}$. Thus, this algorithm suffers $O(P\sqrt{T})$ regret loss of instability.

b. The loss of nonstationarity is induced by the intractable revenue collected during the transient state when the price changes. The analysis is conditional on that all stayed prices have utilization factors less than one. With the key mixing time result from Proposition 5, we can use the steady-state revenue rate to compute the expected revenue of each batch with at most a nonstationarity loss linear on 2τ . Define $\Delta_{\tilde{p}}(p_m) = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$; then the loss of batch m can be bounded from above: $J_{\text{Learning},m}^{\text{LP}} - J_{\text{Learning},m}^{\text{BUCB}} \leq \Delta_{1-\varepsilon}(p_m) I_m \tau + O(\tau)$.

c. For the loss of pruning error, define $p_L^* = \arg \max_{p \in \mathcal{P}_L} p\lambda_p/\mu_p$ as learning the upper-bound price, and $\Delta_L(p_m) = r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$. This loss is induced by the mismatch between p_L^* and $\tilde{p}$, resulting from pruning $\tilde{p}$ from $\mathcal{P}_L$. By Proposition 3, this loss from pruning error $\Delta_{\tilde{p}}(p_m)$ is less than the suboptimality gap $\Delta_L(p_m)$ with probability at least $1 - 4/\sqrt{T}$.

d. For the loss of suboptimality, by Definition 3, we can derive that $\Delta_L(p_m)$ can be upper bounded by the sum of three terms: $\Delta_L(p_m) \leq (U_m(p_m) - L_m(p_m)) + (r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)) + (L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m))$. The first term is analyzed with the help of Lemma F.1 and Jensen's inequality. The second and third terms can be easily bounded with the help of Proposition 2.

To conclude, we can show that the regret during the learning phase is $\tilde{O}(\sqrt{PT})$. Combining with the warm-up phase, the result follows. The detailed proof is provided in Appendix F. $\square$

4.3. Performance of Online Batch Thompson Sampling

BTS holds prior distributions for the arrival and service rates, and thus we use the Bayesian regret as our new performance measure, which is defined as the expected regret of the instance in expectation over all problem instances (drawn from prior distributions). Similarly, we directly define a relaxed $(1 - \varepsilon_T)$ -stable Bayesian regret as a valid upper bound of the Bayesian regret. We prove a relaxed $(1 - \varepsilon_T)$ -stable Bayesian regret bound of BTS in the same

order as that we obtain for the BUCB approach in Section 4.2.

Definition 4. Denoting all the events before period t by $\mathcal{F}_t$, $\forall t \geq 0$, we define the relaxed $(1 - \varepsilon_T)$ -stable Bayesian Regret of policy π by period T as

$$\overline{\text{BR}}(\pi, \varepsilon_T, T) = \mathbb{E}_{(\lambda, \mu) \sim \text{prior}}[\mathbb{E}[J_{\varepsilon_T}^{\text{LP}} - J^\pi | \mathcal{F}_T]].$$

Theorem 2. The T -period cumulative Bayesian ε_T -regret of BTS is bounded by $\tilde{O}(\sqrt{PT})$.

Proof Sketch. The Bayesian regret of the warm-up phase of BTS is the same as that of BUCB, that is, $O(P(\log(T))^3)$. To analyze the regret of the learning phase of BTS, we follow the same decomposition structure and analyze the following four types of losses: instability, nonstationarity, pruning error, and suboptimality. The analyses of the first three losses are very similar to those for BUCB. The analysis of suboptimality also requires us to bound the difference of the revenue rate between the chosen price with maximum sampled rate and $\tilde{p}$. We bound this difference with the “distribution matching” property of the Thompson sampling-type algorithms. This property indicates that the chosen price p_m and the (static) optimal price $\tilde{p}$ are identically distributed conditional on $\mathcal{F}_m$, and thus we can use $\mathbb{E}[U_m(p_m) - U_m(\tilde{p}) | \mathcal{F}_m] = 0$ in the Bayesian regret analysis. To conclude, we can show that the regret during the learning phase is $\tilde{O}(\sqrt{PT})$, and hence the regret of BTS algorithm is also $\tilde{O}(\sqrt{PT})$. The detailed proof is provided in Appendix G. $\square$

5. Extension to Systems with Balking and Reneging

We now extend our model and algorithms to systems where customers may balk or renege when there are not enough resources. We consider a discrete price set in this section and develop an extended BUCB for solving this extended problem. Under any price p selected from the candidate price set $\mathcal{P}$, the service system can be reduced to an $M/M/c$ queue model with balking and reneging (after finishing serving the customers that arrived under previously offered prices). In addition to unknown arrival and service rates λ_p and μ_p , we further consider a *known* balking and reneging threshold $k_p \geq c$, $\forall p \in \mathcal{P}$. Upon arrival, a customer joins the queue with probability one if the number of customers in the system is less than k_p . If there are k_p customers or more ahead of him or her, then he or she joins the queue with probability η_p and balks with probability $1 - \eta_p$. When the number of customers in the system is larger than k_p , the customers waiting in the queue may get impatient, and then, after a certain length of time, one of the customers in the queue will leave the system. We assume this time follows an exponential distribution with mean $\frac{1}{v_p}$, or, equivalently, the queue loses customers with rate

v_p when there are more than k_p customers in the system. We assume that λ_p , μ_p , η_p , and v_p are unknown to the firm. For the ease of the analysis, we consider Assumptions 1 and 2 and the following assumption.

Assumption 3. A constant $y_{\max}$ is known such that $y_{\max} \geq \max_{p \in \mathcal{P}} v_p / \mu_p$.

For the remainder of this section, all notations are for the extended $M/M/c$ queue with balking and reneging. We describe the model and regret for this extended system in Section 5.1. Then, the extended BUCB and its performance analysis are presented in Sections 5.2 and 5.3, respectively.

5.1. Regret for Extended Systems with Balking and Reneging

We keep the same definition of the utilization factor of the system under price p , that is, $\rho_p = \frac{\lambda_p}{c\mu_p}$. Therefore, the queue with balking and reneging with $\rho < 1$ has an actual utilization rate of the resource capacity less than ρ , and thus also has a stationary revenue rate. Similarly, we are interested in nonexploded prices, and thus we consider $\mathcal{P}_{\varepsilon_T} = \{p \in \mathcal{P} : \rho_p \leq 1 - \varepsilon_T\}$. By Proposition 1, we can compute an upper bound of the collected revenue if a policy is restricted to choose prices only from $\mathcal{P}_{\varepsilon_T}$ and this upper bound is computed as a product between T and the maximum steady-state revenue rate of a single price in $\mathcal{P}_{\varepsilon_T}$.

Consider a price in $\mathcal{P}_{\varepsilon_T}$. Let ℓ_{np} , $n = 0, \dots, \infty$, denote the probability that the system has n customers in the steady state when offering price $p \in \mathcal{P}$, and thus $\sum_{n=0}^{\infty} \ell_{np} = 1$, $\forall p \in \mathcal{P}$. Let $\ell_p^{\text{br}} = \sum_{n=k_p+1}^{\infty} \ell_{np}$, and thus ℓ_p^{br} is the probability that balking and reneging may happen in the steady state. Because the arrival and the departure of the impatient customers without being served are independent, the average customer arrival rate can be expressed as $\lambda_p - (v_p + (1 - \eta_p)\lambda_p)\ell_p^{\text{br}}$. Therefore, one can derive the stationary revenue rate for the extended system as

$$r^E(\lambda_p, \mu_p, v_p, \eta_p, p) = \frac{p}{\mu_p} (\lambda_p - (v_p + (1 - \eta_p)\lambda_p)\ell_p^{\text{br}}). \quad (11)$$

Then, we refer to the price $\tilde{p}_{\varepsilon_T} = \arg \max_{p \in \mathcal{P}_{\varepsilon_T}} r^E(\lambda_p, \mu_p, v_p, \eta_p, p)$ as the (static) optimal $(1 - \varepsilon_T)$ -stable state-independent price. We derive the following upper bound $J_{\varepsilon_T}^{\text{EXT}}$ for the cumulative revenue collected by the optimal state-independent price $\tilde{p}_{\varepsilon_T}$. We omit the subscript of ε_T when referring to $\tilde{p}_{\varepsilon_T}$ in the rest of this section. Thus, we have

$$J_{\varepsilon_T}^{\text{EXT}} = T r^E(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, v_{\tilde{p}}, \eta_{\tilde{p}}, \tilde{p}). \quad (12)$$

Now we can similarly define the performance measure for this extended system.

Definition 5. The relaxed ε_T -regret for the extended system is defined as

$$\overline{\text{Regret}}^{\text{EXT}}(\pi, \varepsilon_T, T) = J_{\varepsilon_T}^{\text{EXT}} - J^\pi. \quad (13)$$

5.2. Extended BUCB for Systems with Balking and Reneging

In addition to the previously defined observations in Section 3, we also record the following values. We let $n^c(p)$ denote the number of arrived customers when the system has more than $k_p + 1$ customers, let $n^b(p)$ denote the number of arrived customers who choose to balk, and let $n^r(p)$ denote the number of customers who join the queue but leave before getting served when offering p . For those customers who renege, we let $\hat{h}_i(p)$ denote the observed waiting time in which a customer gets impatient and leaves the queue, and then compute

$$\bar{h}(p) = \frac{n^r(p)}{\sum_{i=1}^{n^r(p)} \hat{h}_i(p)}$$

as an empirical estimate of the renege-ing rate. Let $\hat{\ell}_p^{br}$ denote the portion of the time when the number of customers in the system is greater than or equal to $k_p + 1$ after the first 2τ periods in each batch. Therefore, $n(p) \propto n^c(p) \propto n^b(p) \propto n^r(p)$, $\forall p \in \mathcal{P}$.

5.2.1. Warm-Up Phase of Extended BUCB. The first phase is the warm-up phase, which is identical to that of BUCB. To conclude, this phase posts each price for collecting N_W number of observations and reaches a learning price set $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$. This warm-up phase (i) eliminates the subexponential tail effects caused by the exponential observations of $M/M/c$ queues, (ii) prunes candidate prices whose utilization factors are larger than one, and (iii) reserves candidate prices whose utilization factors are less than $1 - \varepsilon_T$ with a high probability. We can easily check that Propositions 2 and 3 are still valid here. Moreover, we have the following concentration bound, which directly provides an upper confidence bound of the nominal revenue rate.

Proposition 6. For any price $p \in \mathcal{P}$, if $\min\{n^s(p), n^c(p), n^r(p)\} \geq 8 \log(T)$, then we have

$$\mathbb{P}\left(\left|\left(\frac{1}{\mu_p}(\lambda_p - (v_p + (1 - \eta_p)\lambda_p)\ell_p^{br}) - \left(\bar{g}(p)\left(\bar{d}(p) - \left(\bar{h}(p) + \frac{n_m^b(p)}{n_m^c(p)}\bar{d}(p)\right)\hat{\ell}_p^{br}\right)\right)\right)\right| \leq \Gamma\right) \geq 1 - \frac{8}{T^3},$$

where

$$\Gamma = \frac{4(y_{\max} + r_{\max})}{T^3} + \left(8r_{\max} + 4y_{\max} + 9\bar{h}(p)\bar{g}(p) + 4\bar{h}(p)\bar{g}(p)\right) \sqrt{\frac{2 \log(T)}{\min\{n^s(p), n^c(p), n^r(p)\}'}}$$

and $r_{\max} \geq \max_{p \in \mathcal{P}} \lambda_p / \mu_p$ and $y_{\max} \geq \max_{p \in \mathcal{P}} v_p / \mu_p$.

The proof is similar to that of Proposition 2, which is derived based on concentration inequalities for subexponential random variables and Hoeffding's inequality for bounded random variables, and is provided in Appendix H.

5.2.2. Learning Phase of Extended BUCB. The second phase is the learning phase, which is almost identical to that of BUCB but using a different upper confidence bound. We denote the observations by the end of batch m by $(n_m(p), \bar{d}_m(p), n_m^s(p), \bar{g}_m(p), n_m^c(p), \hat{\ell}_p^{br}, n_m^r(p), \bar{d}_m(p))$. Then, we can define the UCB of the revenue rate of each price below based on Proposition 6.

Definition 6. The upper confidence bound of the stationary revenue rate associated with price p by the end of batch m is

$$U_m^E(p) = \left(\bar{g}_m(p)\left(\bar{d}_m(p) - \left(\bar{h}_m(p) + \frac{n_m^b(p)}{n_m^c(p)}\bar{d}_m(p)\right)\hat{\ell}_p^{br}\right) + \Gamma' \sqrt{\frac{2 \log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}}\right)p, \quad (14)$$

and the lower confidence bound of the stationary revenue rate associated with price p by the end of batch m is

$$L_m^E(p) = \left(\bar{g}_m(p)\left(\bar{d}_m(p) - \left(\bar{h}_m(p) + \frac{n_m^b(p)}{n_m^c(p)}\bar{d}_m(p)\right)\hat{\ell}_p^{br}\right) - \Gamma' \sqrt{\frac{2 \log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}}\right)p, \quad (15)$$

where $\Gamma' \sqrt{\frac{2 \log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}} \geq \Gamma$.

Similarly, let $\text{Rad}_m^E(p) = \Gamma' \sqrt{\frac{2 \log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}}$ represent the confidence radius by end of batch m . At the beginning of each batch m , the extended BUCB chooses price $p \in \mathcal{P}_L$ with the highest $U_m^E(p)$ and offers this price during batch m , until reaching period T . We present the algorithmic details in Algorithm 3.

Algorithm 3 (Extended Online Batch Upper Confidence Bound Algorithm (Extended BUCB))

- 1: **Input:** $T, \mathcal{P}, k_p, \bar{\varepsilon}, r_{\max}, c$
- 2: **Initialize:** $\{I_m = 2^m\}, m = 1$, and

$$\varepsilon_T = \bar{\varepsilon} / \log T, \quad \delta = \frac{\bar{\varepsilon}}{(2(1 + r_{\max}/c) \log(T))'}$$

$$N_W = \max\left\{8 \log T, \frac{\log(T)}{2\delta^2}\right\},$$

$$\tau = (\log T)^2, \quad \rho_W = 1 - \frac{\delta(1 + r_{\max}/c)}{1 + \delta}.$$

3: **Warm-up phase:**

4: **for** $p \in \mathcal{P}$ **do**

5: Offer p , record $\hat{d}_i(p)$ for arriving customers, $n_m^b(p')$ for balking customers, $\hat{h}_i(p')$ for reneging

- customers, $\hat{g}_i(p')$ for leaving customers, $\forall p' \in \mathcal{P}$, and $\hat{\ell}_p^{br}$ for the system when the total customers is more than k_p .
- 6: **if** $\min\{n_m^s(p), n_m^c(p), n_m^r(p)\} \geq N_W$ **then**
 - 7: Continue.
 - 8: **end if**
 - 9: **end for**
 - 10: Compute and update $\check{d}_W(p), \bar{g}_W(p), \check{h}_W(p), \hat{\ell}_{np}, \hat{\rho}_W(p)$, $\forall p \in \mathcal{P}$.
 - 11: Prune “unstable” prices $\{p \in \mathcal{P} : \hat{\rho}_W(p) > \rho_W\}$ and retain $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$.
 - 12: Compute confidence radius $\mathbf{Rad}_W^E(p) = \Gamma' \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}}$.
 - 13: Compute $U_0^E(p) = (\bar{g}_W(p)(\check{d}_W(p) - (\check{h}_W(p) + \frac{n_m^b(p)}{n_m^c(p)}\check{d}_W(p))\hat{\ell}_p^{br}) + \mathbf{Rad}_W^E(p))p$, $\forall p \in \mathcal{P}_L$.
 - 14: **Learning phase:**
 - 15: Compute the length of batch m by $\min\{I_m\tau$, the number of remaining periods}.
 - 16: **while** the number of remaining periods > 0 **do**
 - 17: Choose $p_m = \arg \max_{p \in \mathcal{P}_L} U_{m-1}^E(p)$ and offer p_m in batch m .
 - 18: Record $\hat{d}_i(p_m)$ for arriving customers, $n_m^b(p)$ for balking customers, $\hat{h}_i(p)$ for reneging customers, $\hat{g}_i(p)$ for leaving customers, $\forall p \in \mathcal{P}$, and $\hat{\ell}_{p_m}^{br}$ for the system when the total customers is more than k_p .
 - 19: Update $\check{d}_m(p), \bar{g}_m(p), \check{h}_m(p), n_m(p), n_m^c(p), n_m^s(p), n_m^r(p)$, $\forall p \in \mathcal{P}$.
 - 20: Compute confidence radius $\mathbf{Rad}_m^E(p) = \Gamma' \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}}}$.
 - 21: Update $U_m^E(p) = (\bar{g}_m(p)(\check{d}_m(p) - (\check{h}_m(p) + \frac{n_m^b(p)}{n_m^c(p)}\check{d}_m(p))\hat{\ell}_p^{br}) + \mathbf{Rad}_m^E(p))p$, $\forall p \in \mathcal{P}_L$.
 - 22: Update $m = m + 1$ and the number of remaining periods.
 - 23: Compute the length of batch m by $\min\{I_m\tau$, the number of remaining periods}.
 - 24: **end while**

5.3. Performance Analysis for Extended BUCB

We analyze the relaxed ε_T -regret defined in Definition 5 for the extended BUCB. The analysis here is following a similar road map as shown in Figure 1. We first establish the coupling analysis for the system with balking and reneging when price changes in the learning phase. Then we prove the regret is in the same order as that of BUCB, that is, $\tilde{O}(\sqrt{PT})$.

For the coupling analysis, we also need the following assumption on the nature of balking and reneging. This assumption is held in several works in the literature, such as Ammar et al. (2013), Al-Seedy et al. (2009), Ward and Glynn (2003, 2005), and Zhang et al. (2005). Specifically, we refer interested readers to a survey

paper by Wang et al. (2010) about queueing systems with balking and reneging and another comprehensive summary by Ward (2012) about systems with reneging.

Assumption 4. An arriving customer joins the queue with probability η_n and balks with probability $1 - \eta_n$ when there are n customers in the system, and the queueing system loses customers with rate v_n when there are n customers in the system:

1. η_n is nonincreasing in n ;
2. v_n is nondecreasing in n .

Proposition 7 (Coupling Probability of the $M/M/c$ Queue with Balking and Reneging when the Price Changes). Under Assumptions 2 and 4, the probability that the system with balking and reneging reaches the steady state within 2τ after the price changes is bounded from below by $1 - \frac{4}{T^2}$, where $\tau = (\log(T))^2$.

The proof of Proposition 7 is similar to those of Propositions 4 and 5. We present the corresponding details in Appendix I.

Theorem 3. The T -period cumulative ε_T -regret of the extended BUCB algorithm (for systems with balking and reneging) can be bounded by $\tilde{O}(\sqrt{PT})$.

Proof Sketch. The proof of Theorem 3 is similar to the proof of Theorem 1. The regret during the warm-up phase is the same. The regret during the learning phase can also be decomposed into four parts: loss due to instability, pruning error, nonstationarity, and suboptimality. The analyses for the loss of the instability and pruning error are the same as those for BUCB, as they all rely on Proposition 3. The analysis for the loss of nonstationarity here is similar to that for BUCB, and the main technical basis is Proposition 7. The results on the loss of suboptimality here are the same as those for BUCB by substituting the new UCB defined in Definition 6. We relegate the full detailed proof to Appendix I. $\square$

5.4. Extended BTS for Systems with Balking and Reneging

Similarly, we develop the extended BTS when the firm holds prior information about the system parameters. We consider a Gamma prior, $\mathbf{Gamma}(\alpha^{v_p}, \beta^{v_p})$, on v_p for each price p , and the posterior distribution is $\mathbf{Gamma}(\alpha^{v_p} + 1, \beta^{v_p} + \hat{h}_i(p))$ after observing one time interval $\hat{h}_i(p)$ when reneging happens. The balking events can be viewed as Bernoulli trials, and thus we set the prior distribution of η_p as $\mathbf{Beta}(\alpha^{\eta_p}, \beta^{\eta_p})$, and the posterior distribution of η_p is $\mathbf{Beta}(\alpha^{\eta_p} + n_m^c(p) - n_m^b(p), \beta^{\eta_p} + n_m^b(p))$ after observing $n_m^b(p)$ out of $n_m^c(p)$ customers balk. The extended BTS has a warm-up phase identical to that of the extended BUCB and reaches a set $\mathcal{P}_L$ by the end of this warm-up phase. At the beginning of batch m in the learning phase, the algorithm samples parameters

following the posterior distributions and offers price $p \in \mathcal{P}_L$ with the maximum sampled revenue rate $r^E(\hat{\lambda}_p, \hat{\mu}_p, \hat{\nu}_p, \hat{\eta}_p, p)$. We present the algorithmic details in Algorithm 4.

Algorithm 4 (Extended Online Batch Thompson Sampling Algorithm (Extended BTS))

- 1: **Input:** $T, \mathcal{P}, \bar{\varepsilon}, r_{\max}, c, k_p, \alpha^{\lambda_p}, \beta^{\lambda_p}, \alpha^{\mu_p}, \beta^{\mu_p}, \alpha^{\nu_p}, \beta^{\nu_p}, \alpha^{\eta_p}, \beta^{\eta_p}, \forall p \in \mathcal{P}$.
- 2: **Initialize:** $\{I_m = 2^m\}, m = 1$, and

$$\varepsilon_T = \bar{\varepsilon} / \log T, \quad \delta = \frac{\bar{\varepsilon}}{(2(1 + r_{\max}/c) \log(T))},$$

$$N_W = \max\left\{8 \log T, \frac{\log(T)}{2\delta^2}\right\},$$

$$\tau = (\log T)^2, \quad \rho_W = 1 - \frac{\delta(1 + r_{\max}/c)}{1 + \delta}.$$

3: **Warm-up phase:**

4: **for** $p \in \mathcal{P}$ **do**

- 5: Offer p , record $\hat{d}_i(p)$ for arriving customers, $n_m^b(p')$ for balking customers, $\hat{h}_i(p')$ for reneging customers, $\hat{g}_i(p')$ for leaving customers, $\forall p' \in \mathcal{P}$.

- 6: **if** $\min\{n_m^s(p), n_m^c(p), n_m^r(p)\} \geq n_W$ **then**

- 7: Continue.

- 8: **end if**

- 9: **end for**

- 10: Compute and update $\check{d}_W(p), \check{g}_W(p), \check{h}_W(p), \hat{\ell}_{np}, \hat{\rho}_W(p), \forall p \in \mathcal{P}$.

- 11: Prune “unstable” prices $\{p \in \mathcal{P} : \hat{\rho}_W(p) > \rho_W\}$ and retain $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$.

- 12: Update $\alpha^{\lambda_p} = \alpha^{\lambda_p} + n_W(p), \beta^{\lambda_p} = \beta^{\lambda_p} + \sum_{i=1}^{n_W(p)} \hat{d}_i(p),$
 $\alpha^{\mu_p} = \alpha^{\mu_p} + n_W^s(p), \beta^{\mu_p} = \beta^{\mu_p} + \sum_{i=1}^{n_W^s(p)} \hat{g}_i(p),$
 $\alpha^{\nu_p} = \alpha^{\nu_p} + n_W^b(p), \beta^{\nu_p} = \beta^{\nu_p} + \sum_{i=1}^{n_W^b(p)} \hat{h}_i(p),$
 $\alpha^{\eta_p} = \alpha^{\eta_p} + n_m^c(p) - n_m^b(p),$
 $\beta^{\eta_p} = \beta^{\eta_p} + n_m^b(p), \forall p \in \mathcal{P}_L.$

13: **Learning phase:**

- 14: Compute the length of batch m by $\min\{I_m \tau, \text{the number of remaining periods}\}$.

- 15: **while** the number of remaining periods > 0 **do**

- 16: Sample $\hat{\lambda}_p \sim \text{Gamma}(\alpha^{\lambda_p}, \beta^{\lambda_p}), \hat{\mu}_p \sim \text{Gamma}(\alpha^{\mu_p}, \beta^{\mu_p}), \hat{\nu}_p \sim \text{Gamma}(\alpha^{\nu_p}, \beta^{\nu_p}), \hat{\eta}_p \sim \text{Beta}(\alpha^{\eta_p}, \beta^{\eta_p}),$ for $p \in \mathcal{P}$.

- 17: Compute $\hat{\ell}_p^{br}$ by plugging in $(\hat{\lambda}_p, \hat{\mu}_p, \hat{\nu}_p, \hat{\eta}_p)$ and $r^E(\hat{\lambda}_p, \hat{\mu}_p, \hat{\nu}_p, \hat{\eta}_p, p) = (\check{g}_W(p)(\check{d}_W(p) - (\check{h}_W(p) + \frac{n_W^b(p)}{n_W^s(p)} \check{d}_W(p)) \hat{\ell}_p^{br}) + \text{Rad}_W^E(p))p, \forall p \in \mathcal{P}_L.$

- 18: Choose $p_m = \arg \max_{p \in \mathcal{P}_L} r^E(\hat{\lambda}_p, \hat{\mu}_p, \hat{\nu}_p, \hat{\eta}_p, p)$ and offer p_m in batch m .

- 19: Record $\hat{d}_i(p_m)$ for arriving customers, $n_m^b(p)$ for balking customers, $\hat{h}_i(p)$ for reneging customers, $\hat{g}_i(p)$ for leaving customers, $\forall p \in \mathcal{P}_L$.

- 20: Compute and update $\check{d}_m(p), \check{g}_m(p), \check{h}_m(p), n_m(p), n_m^c(p), n_m^s(p), n_m^r(p), \forall p \in \mathcal{P}.$

- 21: Update $\alpha^{\lambda_p}, \beta^{\lambda_p}, \alpha^{\mu_p}, \beta^{\mu_p}, \alpha^{\nu_p}, \beta^{\nu_p}, \alpha^{\eta_p}, \beta^{\eta_p}, \forall p \in \mathcal{P}_L$ (via the same rules as in Step 12).
- 22: Update $m = m + 1$ and the number of remaining periods.
- 23: Compute the length of batch m by $\min\{I_m \tau, \text{the number of remaining periods}\}$.
- 24: **end while**

Similar to BTS, we use Bayesian regret as the performance measure of Extended BTS.

Definition 7. We define the relaxed Bayesian regret of the state-independent optimal policy by period T as

$$\overline{\text{BR}}^{\text{IND}}(\pi, T) = \mathbb{E}_{(\lambda, \mu, \nu, \eta) \sim \text{prior}}[\mathbb{E}[J^{\text{EXT}} - J^\pi \mid \mathcal{F}_T]].$$

Theorem 4. The T -period cumulative Bayesian ε_T -regret of the extended BTS algorithm (for systems with balking and reneging) is bounded by $\tilde{O}(\sqrt{PT})$.

Proof Sketch. The proof of Theorem 4 is similar to that of Theorem 2 and is thus omitted. The regret during the warm-up phase is same. The regret during the learning phase is also analyzed by decomposing it into four parts. The loss of instability and pruning errors is the same as those of BTS (based on Proposition 3). The analysis for the loss of nonstationarity is the same as that of the extended BUCB in Theorem 3 (based on Proposition 7). The results of the loss of suboptimality here are the same as those for BTS by substituting the new UCB equation in Definition 6. $\square$

6. Numerical Experiments

We present the numerical results of BUCB and BTS (without and with balking and reneging).

6.1. Experimental Setup

The total operation time horizon is 1,000 periods, and the capacity of the reusable resource is $c = 100$. We consider three balking and reneging cases: (i) no balking and reneging happens, which means that the system under any price can be reduced to a standard $M/M/c$ queue; (ii) the system is a loss system, where a newly arrived customer balks immediately when there is no available resource; and (iii) both balking and reneging may happen, where a newly arrived customer balks with probability 0.5 when there is no available resource, and the reneging rate of the queue is one, which means that one of the customers in the queue will get impatient and leave the queue after an uncertain amount of time following an exponential distribution with mean one. We choose the price from five candidate prices, of which the corresponding service rates are equal. We consider three scenarios of the arrival rates associated with candidate prices and thus the corresponding system dynamics. In scenario 1, the state-independent optimal price has a relatively low utilization factor. In

Table 1. Parameter Settings and Experimental Design (Discrete Price Set)

	Price	Arrival rate λ	Service rate μ	Utilization factor	Revenue rate under case (i)	Revenue rate under case (ii)	Revenue rate under case (iii)
Scenario 1	18	5.5	0.1	0.55	Instance 1 990.00	Instance 2 988.95	Instance 3 989.37
	16	6	0.1	0.60	960.00	957.10	959.26
	14	6.5	0.1	0.65	910.00	907.91	909.81
	12	8	0.1	0.80	960.00	958.12	959.78
	10	9.5	0.1	0.95	950.00	902.37	910.91
Scenario 2	18	5.2	0.1	0.52	Instance 4 936.00	Instance 5 935.20	Instance 6 935.39
	16	6	0.1	0.60	960.00	957.10	959.26
	14	7.5	0.1	0.75	1050.00	1049.47	1049.68
	12	8.5	0.1	0.85	1020.00	1009.27	1011.65
	10	9.5	0.1	0.95	950.00	902.37	910.91
Scenario 3	18	5	0.1	0.50	Instance 7 900.00	Instance 8 898.96	Instance 9 899.06
	16	5.5	0.1	0.55	880.00	878.55	879.65
	14	6	0.1	0.60	840.00	837.46	839.35
	12	7.5	0.1	0.75	900.00	898.97	899.58
	10	9.8	0.1	0.98	980.00	915.88	921.04

scenario 2, the state-independent optimal price has a relatively moderate utilization factor. In scenario 3, the state-independent optimal price has the highest utilization factor. To conclude, we have nine instances in total (i.e., three scenarios under three cases), and the corresponding details are presented in Table 1.

6.2. Benchmark

We adopt the blind network revenue management algorithm proposed in Besbes and Zeevi (2012) to our settings as the benchmark, referred to as the BZ benchmark for short. Algorithm 5 presents the algorithmic details of the BZ benchmark, which is implemented in this section. To conclude, for each instance, we implement five pricing algorithms: (Extended) BUCB, (Extended) BTS, and three benchmark policies with different values of the fraction of time allocated to the exploration phase, that is, θ in Algorithm 5. We set θ as 0.22, 0.1, and 0.05 in three BZ benchmarks, respectively (the same settings as in table 1 in Besbes and Zeevi (2012)).

Algorithm 5 (BZ Benchmark (Adopted from Besbes and Zeevi 2012))

- 1: **Input:** $T, \mathcal{P}, \theta$.
- 2: **for** $p \in \mathcal{P}$ **do**
- 3: Offer p for next $\lfloor (\theta \cdot T)/P \rfloor$ periods.
- 4: Collect revenue $\hat{J}_{p_m}$ from customers accepted with price p_m .
- 5: **end for**
- 6: Offer $p = \arg \max_{p \in \mathcal{P}} \hat{J}_p$ for the remaining periods until T .

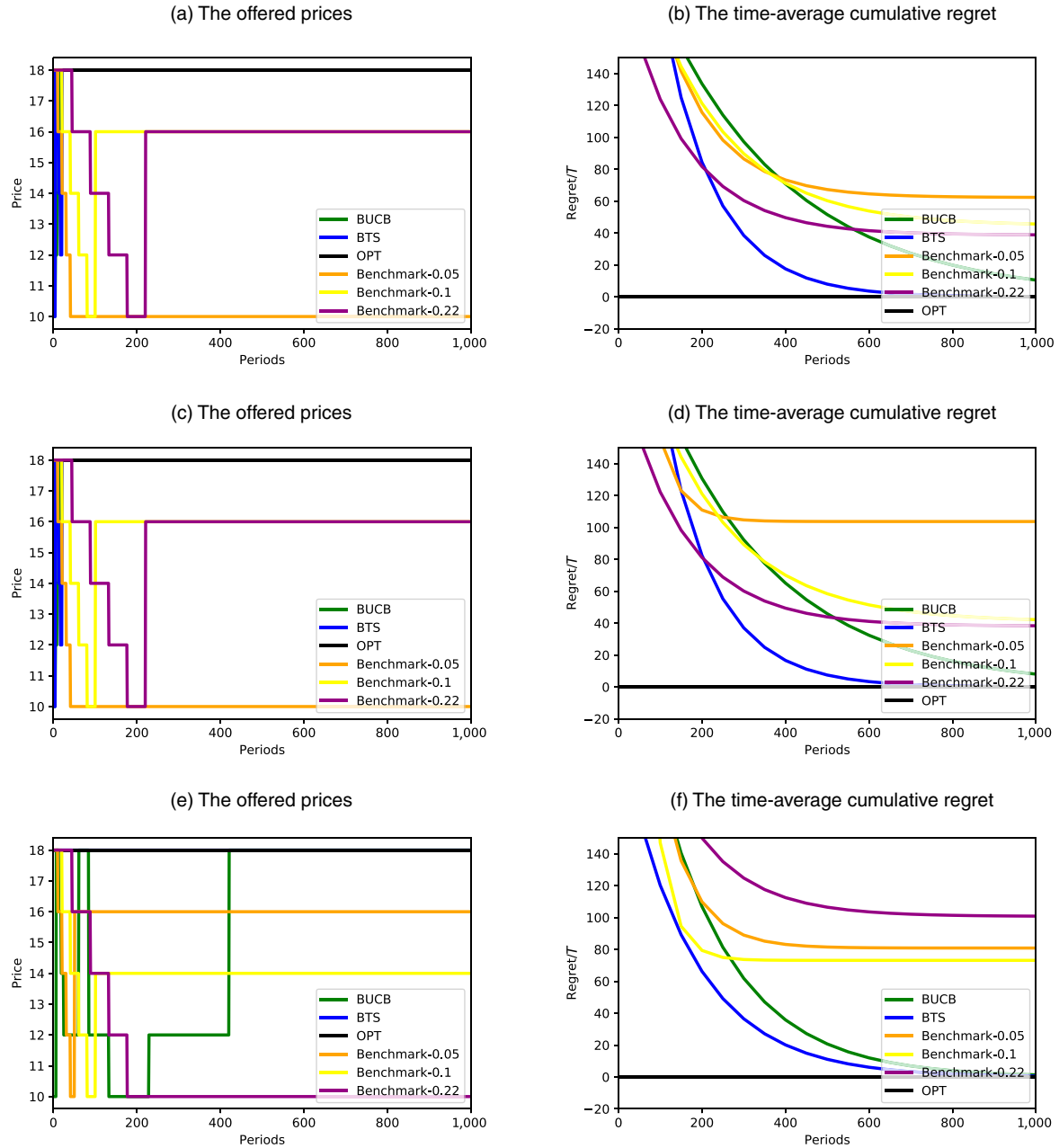
We compare the results of the above five pricing policies with the (static) optimal price (“OPT” in Figures 2–4). We present two figures for the results of each instance: the first shows the offered price over periods of each algorithm, $p_t, t = 1, \dots, T$; the second depicts the

cumulative time-average relaxed regret, including the relaxed regret of the state-dependent optimal policy for case (i) and the relaxed regret of the state-independent optimal policy for cases (ii) and (iii), computed as the ratio between the relaxed regret and the time periods, that is, $(\sum_{t'=1}^t J_{t'}^{LP} - \sum_{t'=1}^t J_{t'}^{\pi})/t$ for case (i), and $(\sum_{t'=1}^t J_{t'}^{\text{IND}} - \sum_{t'=1}^t J_{t'}^{\pi})/t$ for cases (ii) and (iii). For case (i), the relaxed regret of the state-dependent optimal policy is defined in Definition 2, and for cases (ii) and (iii), the relaxed regret of the state-independent optimal policy is defined in Definition 5. For each implemented policy, $\sum_{t'=1}^t J_{t'}^{\pi}$ is the revenue by the end of period t , which is collected from the simulation with the system starting with no customers at time zero. The results are presented in Figures 2–4.

If the policy finally chooses the state-independent optimal price $\tilde{p}$, then the average regret over periods will shrink to zero after a certain time, to offset the regret generated in the first several periods because of (i) choosing a suboptimal price and (ii) the nonstationary transient states of the service system. If the policy finally chooses a suboptimal price, then the average regret over periods will converge to a constant number, which theoretically is the difference between the stationary revenue rates of the chosen and the optimal state-independent price. In our theoretical proof, the regret analysis requires long time periods, that is, a large T in Assumption 1. Our numerical results show that the proposed BUCB and BTS have good practical performance even when T is small; for example, we set $T = 1,000$ in our experiments.

6.3. Performance Against BZ

From the numerical results, we can see that both BUCB and BTS perform very well for all instances. They end up with the state-independent optimal price, and the

Figure 2. (Color online) Offered Price and Time-Average Cumulative Regret of Policies in Instances 1 to 3 (Rate–Price Scenario 1 Under Three Balking and Reneging Cases)

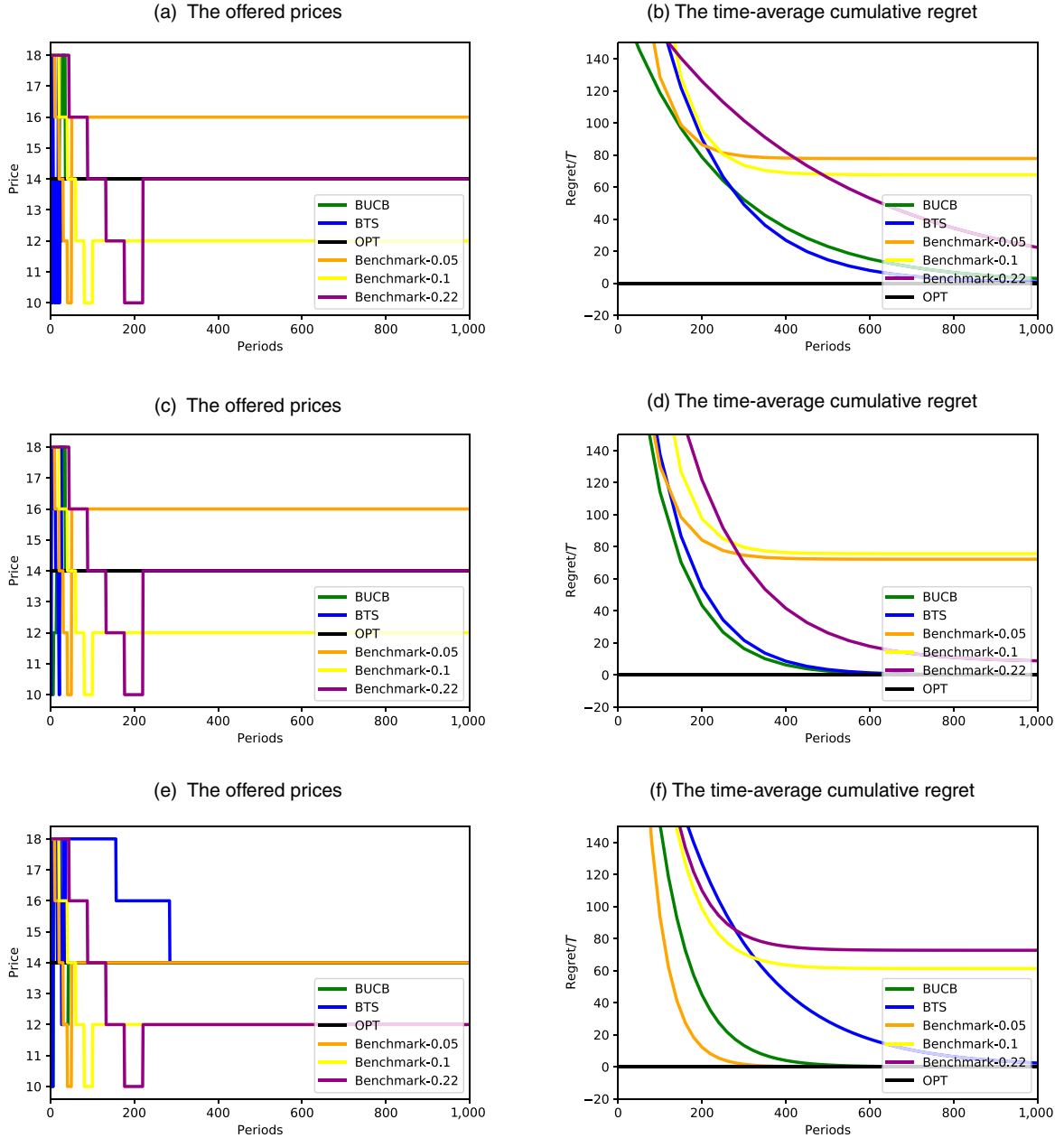
average regret converges to zero quickly. The performance of three BZ benchmarks, parametrized with different θ , varies among different instances. In general, the benchmark with a larger θ has the higher probability of identifying the state-independent optimal price in the exploitation phase, but the performance is not guaranteed. One merit of our proposed policies is that the exploration and exploitation are both motivated by received information, and thus even when the policy chooses a suboptimal price, it still has a chance to identify

the correct one afterward. For example, in Instances 3 and 6, BUCB and BTS choose a suboptimal price in the first several periods but identify and switch back to the state-independent optimal price later. However, the benchmarks cannot change the decision after entering the exploitation phase.

7. Conclusion

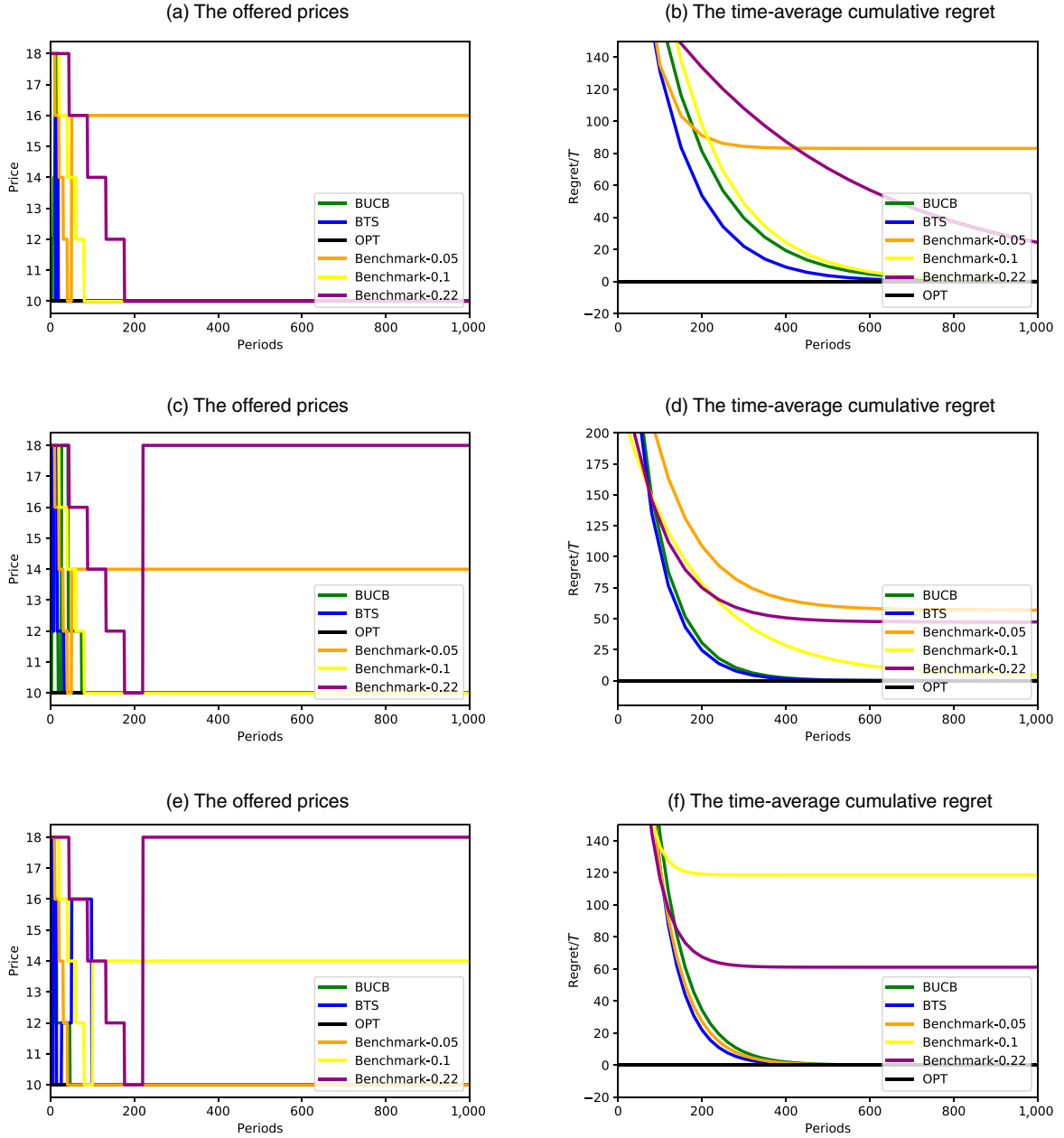
We consider a price-based revenue management problem with a single reusable resource over a finite time

Figure 3. (Color online) Offered Price and Time-Average Cumulative Regret of Policies in Instances 4 to 6 (Rate–Price Scenario 2 Under Three Balking and Reneging Cases)



horizon. We develop two online learning and dynamic pricing policies to (i) learn the unknown mappings between the system dynamics and prices from past sales and (ii) maximize the cumulative revenue. We propose a batch upper confidence bound algorithm and a batch Thompson sampling algorithm based on the multiarmed bandit-related algorithms. Our algorithms and regret analysis take into account the following salient challenges: (i) the unknown rate-and-price mapping information, (ii) the dynamic nature of reusable

resources being committed over time, (iii) the transient behavior of the service system when the price changes, and (iv) unbounded and heavy-tailed distributions of observable random variables. We prove that the cumulative regret is bounded from above by $\tilde{O}(\sqrt{PT})$ under the discrete price setting, where T is the total number of time periods, and P is the number of the discrete prices. We also develop a new coupling analysis for analyzing the time for a system to reach the steady state from an empty state or from a steady state under another set of

Figure 4. (Color online) Offered Price and Time-Average Cumulative Regret of Policies in Instances 7 to 9 (Rate–Price Scenario 3 Under Three Balking and Reneging Cases)

system parameters. Our numerical experiments demonstrate that the performance of the proposed BUCB and BTS algorithms is consistently excellent.

7.1. Continuous Price Space

One important direction to consider is how to incorporate continuous price space into our model. This is likely to be very challenging because it corresponds to continuum-armed bandits with heavy-tailed rewards and mixing times, which was not studied in prior literature. A promising model can be cast as follows. The

firm can choose any price between a fixed range $[p_L, p_U]$. Specifically, we consider that each price within this range is associate with a d_f -dimensional feature vector $\mathbf{x}_p$ and there exists a linear relationship between known feature vectors associated with prices, and the arrival/service times are linear on feature vectors:

$$\frac{1}{\lambda_p} = \theta_\lambda^T \mathbf{x}_p \text{ and } \frac{1}{\mu_p} = \theta_\mu^T \mathbf{x}_p, \quad (16)$$

where $\theta_\lambda \in \mathbb{R}^{d_f}$ and $\theta_\mu \in \mathbb{R}^{d_f}$ are unknown coefficient vectors.

One could adapt linear bandit approaches to tackle this problem. To address the mixing time and heavy-tailed reward challenges, the algorithmic design could follow the same batched design template but substitute the UCB part with LinUCB (Li et al. 2010, Han et al. 2020). In terms of regret analysis, Filippi et al. (2010) provides a good starting point for analyzing this compound linear UCB algorithm (because each “arm” of this RM problem has two types of linear dependent rewards, $1/\lambda$ and $1/\mu$). Then the analysis can be completed by further bounding the loss induced by the i.i.d. violation following the analysis of the self-normalized tail inequality in Abbasi-Yadkori et al. (2011). The loss due to suboptimality should be $\tilde{O}(d_f \sqrt{T})$. If the algorithm can reach a well-defined $\mathcal{P}$ relatively fast, for example, the length of the warm-up phase is at most $\tilde{O}(\sqrt{T})$. Then the cumulative regret (or ε_T -regret) is $\tilde{O}(d_f \sqrt{T})$, which matches the state-of-the-art regret lower bound of linear bandits $\Omega(d_f \sqrt{T})$ (Rusmevichientong and Tsitsiklis 2010).

7.2. Other Research Directions

There are several other important and plausible future research avenues. First, this paper considers the single-resource setting for revenue management problems

with reusable resources. Therefore, demand learning and dynamic pricing under a multiproduct and multire-source setting is an interesting research problem (such as the setting considered in Owen and Simchi-Levi 2017, Doan et al. 2020). Second, one may consider generalizing the current model to accommodate general arrival and service distributions. However, this would require different methods in analyzing the regret, especially the loss of nonstationarity, which we shall leave for future research. Third, this paper considers stationary demand and homogeneous personal activities in the system. There is also recent literature considering heterogeneous and nonstationary demand (see, e.g., Borgs et al. 2014, Besbes et al. 2015, Lei and Jasin 2020). Thus, it is worthwhile to investigate any nonstationary models.

Acknowledgments

The authors thank area editors Ramandeep Randhawa and Itai Gurvich, the associate editor, and the three anonymous referees for their very detailed and constructive comments, which helped significantly improve the content and exposition of this paper.

Appendix A. Summary of Major Notation

Table A.1. Summary of Major Notation for Problem Formulation

Notation	Description
$\mathcal{P}$	The set of candidate prices, $P = \mathcal{P} $
T	The number of periods in consideration
c	The capacity of reusable resources
n	The number of customers in the system
λ_p	The arrival rate under price $p \in \mathcal{P}$
μ_p	The service rate of one unit of resource under price $p \in \mathcal{P}$
k_p	The identical threshold length of the queue for balking under price $p \in \mathcal{P}$
η_p	The probability for customers to join the queue when balking happens under price $p \in \mathcal{P}$
ν_p	The reneging rate when reneging happens under price $p \in \mathcal{P}$
ℓ_{np}	The probability that the system has n customers in the steady state under price $p \in \mathcal{P}$
$r_{\max}$	A valid upper bound of the ratio between arrival and service rates of all candidate prices, $r_{\max} \geq \max_{p \in \mathcal{P}} \lambda_p / \mu_p$
$\pi(n, t)$	The pricing policy $\pi : \{(n, t) : n = 0, 1, \dots, \infty, t = 1, \dots, T\} \rightarrow \mathcal{P}$, where the firm selects price $\pi(n, t) \in \mathcal{P}$ when there are n customers in the system at the beginning of period t
J_t^π	The expected revenue under pricing policy π during period t , $\forall t = 1, \dots, T$,
J^π	The expected revenue over periods $\{1, \dots, T\}$, $J^\pi = \sum_{t=1}^T J_t^\pi$
π^*	The state-dependent optimal policy $\pi^* = \operatorname{argmax}_\pi J^\pi$
J^*	The state-dependent optimal expected revenue $J^* = J^{\pi^*}$
$r(\lambda_p, \mu_p, p)$	The stationary revenue rate of price $p \in \mathcal{P}$ with $\rho_p < 1$, $r(\lambda_p, \mu_p, p) = \frac{\lambda_p}{\mu_p} p$
ε	A small positive hyperparameter, specified by the algorithm user
ε_T	A parameter related to stability of the algorithm and $\varepsilon_T = \varepsilon / \log(T)$
$\mathcal{P}_{\varepsilon_T}$	The $(1 - \varepsilon_T)$ -stable price set, $\mathcal{P}_{\varepsilon_T} = \{p \in \mathcal{P} : \rho_p \leq 1 - \varepsilon_T\} \subseteq \mathcal{P}$
$\tilde{p}$	The (static) optimal $(1 - \varepsilon_T)$ -stable price, $\tilde{p} := \tilde{p}_{\varepsilon_T} = \operatorname{argmax}_{p \in \mathcal{P}_{\varepsilon_T}} \frac{\lambda_p}{\mu_p} p$.
$J_{\varepsilon_T}^{\text{LP}}$	The optimal objective value of the LP defined in (2), $J_{\varepsilon_T}^{\text{LP}} = \frac{J_{\tilde{p}}^{\text{LP}}}{\mu_{\tilde{p}} \tilde{p}}$

Table A.2. Summary of Major Notation for BUCB and BTS

Notation	Description
δ	A parameter used to control the length of the warm-up phase and the threshold for pruning “unstable” prices, $\delta = \bar{\epsilon} / (2(1 + r_{\max}/c) \log(T))$
N_W	The number of observations need to collect during the warm-up phase, $N_W = \max\{8 \log T, (\log T)/2\delta^2\}$
ρ_W	Threshold value for $\hat{\rho}$, $\rho_W = 1 - \delta(1 + r_{\max}/c)/(1 + \delta)$
$\hat{\rho}_W(p)$	The empirical utilization factor, $\hat{\rho}_W(p) = \bar{g}_W(p) \bar{d}_W(p)$
$\mathcal{P}_L$	The resulting price set after the warm-up phase, $\mathcal{P}_L = \{p \in \mathcal{P} : \hat{\rho}_W(p) \leq \rho_W\}$
M	The number of batches
τ	The value for computing the length of each batch, $\tau = (\log(T))^2$
I_m	Batch m , $m = 1, \dots, M$, contains $I_m \tau$ periods, where $I_m = 2^m$
$n_m(p)$	The number of arrived customers under price p by the end of batch m
$n_m^s(p)$	The number of customers have been successfully served under price p by the end of batch m
$\hat{d}_i(p)$	The observed time intervals between arrivals of customer i and customer $i - 1$ under price $p \in \mathcal{P}$
$\bar{d}_m(p)$	An empirical estimate of arrival rate by the end of batch m , $\bar{d}_m(p) = \frac{\sum_{i=1}^{n_m(p)} 1/\hat{d}_i(p)}{n_m(p)}$
$\hat{g}_i(p)$	The observed service time of customer i under price $p \in \mathcal{P}$
$\bar{g}_m(p)$	The empirical mean of service time by the end of batch m , $\bar{g}_m(p) = \frac{\sum_{i=1}^{n_m(p)} \hat{g}_i(p)}{n_m(p)}$
$n_m^c(p)$	The number of arrived customers when the system has more than $k_p + 1$ customers under price $p \in \mathcal{P}$
$n_m^b(p)$	The number of arrived customers who balk under price $p \in \mathcal{P}$
$n_m^r(p)$	The number of customers who join the queue but leave before getting served under price $p \in \mathcal{P}$
$\hat{h}_i(p)$	The observed waiting time that a customer get impatient and leave the queue under price $p \in \mathcal{P}$
$\bar{h}_m(p)$	An empirical estimate of reneging rate by the end of batch m , $\bar{h}_m(p) = \frac{\sum_{i=1}^{n_m^r(p)} 1/\hat{h}_i(p)}{n_m^r(p)}$
$\hat{\ell}_p^{br}$	The portion of the time when the number of customers in the system is greater than or equal to $k_p + 1$ after the first 2τ periods in each batch
p_m	The offered price in batch m by the algorithm
$J_m(p)$	The practical revenue collected in batch m with price p
$U_m(p)$	The upper confidence bound of expected stationary revenue rate of price p by the end of batch m
$L_m(p)$	The lower confidence bound of expected stationary revenue rate of price p by the end of batch m
$\mathcal{M}_p \subseteq \{1, \dots, M\}$	The batches choosing price p
$T_p = \sum_{m \in \mathcal{M}_p} I_m \tau$	The number of periods choosing price p
Gamma ($\alpha^{\lambda_p}, \beta^{\lambda_p}$)	The prior/posterior distribution of λ_p
Gamma ($\alpha^{\mu_p}, \beta^{\mu_p}$)	The prior/posterior distribution of μ_p
Gamma ($\alpha^{\nu_p}, \beta^{\nu_p}$)	The prior/posterior distribution of ν_p
Beta ($\alpha^{\eta_p}, \beta^{\eta_p}$)	The prior/posterior distribution of η_p
$\mathcal{C}_{2\tau}$	The event that the system reaches the steady state under the new price within the first 2τ periods
$\bar{\mathcal{C}}_{2\tau}$	The complementary event of $\mathcal{C}_{2\tau}$ that the system does not reach the steady state within 2τ periods

Table A.3. Summary of Major Notation for Coupling Analysis

Notation	Description
S_t	The number of customers in the target system at time t , starting from the empty state or from the steady state under another price
S_∞	The number of customers in the system in steady state
$\hat{S}_t$	An auxiliary stochastic process starting with the steady-state distribution and remaining thereafter, $\hat{S}_t = S_\infty, t \geq 0$
A_n	The busy period between n and $n - 1$ customers, $A_n = \min\{t: n \text{ customers in the system at time } 0^+ \text{ and } n - 1 \text{ customers in the system at time } t\}$, $\forall n \geq 1$; A_1 and σ_1 are the mean and the standard deviation of A_1 .
τ	The coupling time, $\tau = (\log(T))^2$
$\mathcal{A}_t$	The event that system S_t and system $\hat{S}_t$ couple after time t , $\mathcal{A}_t = \{S_t = \hat{S}_t \text{ after } t\}$
$\mathcal{B}_n^a$	The event where a new customer arrives before any current customers leaves starting from n customers in the system
$\mathcal{B}_n^l$	The event where a current customer leaves before any new customer arrives, starting from n customers in the system
$\mathcal{C}_t$	The event where S reaches the steady state under the new price before t , $t \in \{\tau, 2\tau\}$
$\bar{\mathcal{C}}_t$	The complementary events of $\mathcal{C}_t$, $t \in \{\tau, 2\tau\}$
$\hat{S}$	The virtual system that has a coupling speed slower than or equal to that of the target system S
$\bar{\mathcal{C}}_t$	The event where the virtual system $\hat{S}$ reaches the steady state under the new price before t , $t \in \{\tau, 2\tau\}$

Appendix B. Preliminaries of UCB and TS

Consider that in MAB problems, the player aims to find the arm resulting in the highest expected reward. The stochastic MAB problems consider i.i.d. rewards, and the players have no prior information about uncertain rewards. Motivated by adaptive exploration, the concepts of confidence bounds first appeared in Lai and Robbins (1985), and the first version of UCB algorithm was proposed in Lai (1987). The UCB algorithm defines an upper confidence bound of the expected reward of each arm based on the past observations, where the actual expected reward is below this bound with a high probability. Therefore, this upper confidence bound can be viewed as an optimistic estimate. In a simple UCB, that is, UCB1 in Auer et al. (2002) and Slivkins (2019), it chooses the arm with the highest upper confidence bound, that is, the best arm based on the optimistic estimate, in each round. There are also many UCB variants (see, e.g., Garivier and Moulines 2011), and our BUCB is based on UCB1 in Auer et al. (2002) and Slivkins (2019). The UCB algorithms achieve better performance compared with nonadaptive exploration-based algorithms. We refer to chapter 7 of Lattimore and Szepesvári (2020) and chapter 3 of Slivkins (2019) for more technical details and concrete summaries of UCB algorithms.

As a randomized Bayesian algorithm, TS was first proposed by Thompson (1933, 1935). Consider that in MAB problems, the player also aims to find the arm resulting in the highest expected reward. The basic idea of TS is that, at each round, random samples are generated by the posterior distribution of the reward for each arm, and then the arm with the highest sampled reward is chosen. Different from stochastic bandits, the player holds other information about the reward associated with each arm; that is, the problem instance is drawn from a known prior distribution. Therefore, the remarkable difference of TS from UCB lies in that TS holds priors for the distributions of rewards. The MAB problems under this Bayesian assumption are referred to as Bayesian MAB problems (see, e.g., Slivkins 2019, chapter 3). TS is also known as the “probability matching” property because the probability of an arm being chosen matches the posterior probability that this arm has the highest expected reward. TS gradually becomes a representative algorithm for solving Bayesian MAB problems, and it has been adapted to various Bayesian bandit problem settings (Russo and Van Roy 2014). Compared with other popular MAB algorithms, TS achieves similar theoretical performance guarantees and often better empirical performance (see Chapelle and Li 2011, Kaufmann et al. 2012). We refer interested readers to Russo et al. (2018) and Agrawal and Goyal (2013) and two books, by Slivkins (2019) and Lattimore and Szepesvári (2020), for technical details of TS. A work similar to our paper is by Ferreira et al. (2018), where the authors consider a price-based network revenue management problem (with perishable resources) and propose a pricing algorithm based on TS to address the challenge of balancing the exploration–exploitation trade-off.

Appendix C. Proofs of Section 2.1

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 (2), and the expected revenue of this admissible policy is exactly the corresponding objective value in (2).

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 (2b) are immediately satisfied. Variable $\hat{\pi}_{nt}^p = 1$ if the policy $\hat{\pi}$ chooses price p at state (n, t) , and otherwise, $\hat{\pi}_{nt}^p = 0$, for $n = 0, \dots, \infty, t = 1, \dots, T, p \in \mathcal{P}$. As a result, constraints (2c) hold because on the left-hand side, we have one variable equal to one and others all zero. Variable $\hat{J}_{nt}$ takes the value of the 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 , and let $\hat{p}$ denote the price chosen at state (n, t) by the policy $\hat{\pi}$. Then, by plugging $\hat{\pi}_{nt}^{\hat{p}} = 1$ in the right-hand side of constraints (2d), 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 (assuming all arrived customers are served successfully). We call $\frac{\lambda_{\hat{p}}}{\mu_{\hat{p}}} \hat{p}$ the nominal revenue rate of price $\hat{p}$. Consequently, we can conclude that the right-hand side of (2d) 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 (2d) are 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 linear programming objective value at solution $(\hat{\pi}, \hat{a}, \hat{J})$ is $J^{\hat{\pi}}$. Therefore, we have $J^{\text{LP}} \geq J^{\hat{\pi}}$, for any admissible policy π , and thus also for the optimal state-dependent policy π^* . Q.E.D.

Proof of Corollary 1. This corollary follows directly from Propositions 1 and 4. Based on the above constructed optimal solution $(\hat{\pi}, \hat{a}, \hat{J})$, one can compute that the difference between the revenue upper bound $J_{\varepsilon_T}^{\text{LP}}$ and the collected revenue under the static pricing policy $J^{\pi^{\hat{p}}}$ 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 $T \rightarrow \infty$, $J_{\varepsilon_T}^{\text{LP}} - J^{\pi^{\hat{p}}} \leq O((\log(T))^2)$, which suggests that $\pi^{\hat{p}}$ is asymptotically $(1 - \varepsilon_T)$ -stable optimal. Q.E.D.

Discussion over state variable definitions. In LP (2), we use (n, t) as the state variable, where n reflects the number of customers currently in the system and t is the current time stamp. Proposition 1 asserts that $J^{\text{LP}} = T \frac{\lambda_{\hat{p}}}{\mu_{\hat{p}}} \hat{p}$

is a valid upper bound of the state-dependent optimal policy with (n, t) being the state variable, i.e., $J^* \leq J^{LP}$. One may consider a different and much more refined definition for the state variable. Consider $(\mathbf{n}, t)$ as the new state variable, where the vector $\mathbf{n}$ keeps track of the full information of each customer currently in the system (exact location, service elapsed, price committed). We denote the optimal state-dependent policy under this new state variable definition by π^{**} and the corresponding expected revenue by J^{**} . Then we have two results. First, it is evident that the optimal state-dependent policy under the original state variable definition is a feasible policy for the new state variable definition and thus $J^{**} \geq J^*$. Second, we still have $J^{**} \leq J^{LP}$. To prove this, we can replace the state variables in decision variables in LP (2) to obtain a new linear program, denoted by LP2. Then following the same analysis as in Proposition 1 we can readily show that the optimal objective value of LP2 is a valid upper bound of J^{**} and $J^{LP2} = J^{LP} = T \frac{\lambda_p}{\mu_p} \tilde{p}$. Therefore, we have $J^{**} \leq J^{LP2} = J^{LP}$. In fact, under any valid state variable definition, J^{LP} provides an upper bound of the state-dependent optimal policy.

Appendix D. Proofs of Section 3.1

Proof of Proposition 2. The proof is based on the concentration inequality for subexponential random variables. We show the intermediate results in Lemmas D.1 to D.3. By definitions of $n(p)$ and $n^s(p)$, one has $n(p) \geq n^s(p)$, and thus

$$\begin{aligned}
 & \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}(p)\check{d}(p)\right| \leq 4\left(r_{\max} + \bar{g}(p)\check{d}(p)\right)\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \\
 &= \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}(p)\check{d}(p)\right| \leq 4r_{\max}\sqrt{\frac{2\log(T)}{n^s(p)}} + 4\bar{g}(p)\check{d}(p)\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \\
 &\geq \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}(p)\check{d}(p)\right| \leq 4r_{\max}\sqrt{\frac{2\log(T)}{n^s(p)}} + 4\bar{g}(p)\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right) \\
 &= \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \lambda_p\bar{g}(p) + \lambda_p\bar{g}(p) - \bar{g}(p)\check{d}(p)\right| \leq 4r_{\max}\sqrt{\frac{2\log(T)}{n^s(p)}}\right. \\
 &\quad \left.+ 4\bar{g}(p)\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right) \\
 &= \mathbb{P}\left(\left|\lambda_p\left(\frac{1}{\mu_p} - \bar{g}(p)\right) + \bar{g}(p)(\lambda_p - \check{d}(p))\right| \leq 4r_{\max}\sqrt{\frac{2\log(T)}{n^s(p)}}\right. \\
 &\quad \left.+ 4\bar{g}(p)\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right) \\
 &\geq \mathbb{P}\left(\left|\lambda_p\left(\frac{1}{\mu_p} - \bar{g}(p)\right)\right| \leq r_{\max}\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \\
 &\quad \cdot \mathbb{P}\left(\left|\bar{g}(p)(\lambda_p - \check{d}(p))\right| \leq 4\bar{g}(p)\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right) \\
 &= \mathbb{P}\left(\left|\frac{1}{\mu_p} - \bar{g}(p)\right| \leq \frac{4r_{\max}}{\lambda_p}\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \\
 &\quad \cdot \mathbb{P}\left(\left|\lambda_p - \check{d}(p)\right| \leq 4\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right). \tag{D.1}
 \end{aligned}$$

Given $n(p) \geq n^s(p) \geq 8\log(T)$ and $r_{\max} \geq \max_{p \in \mathcal{P}} \lambda_p / \mu_p$, by Lemmas D.2 and D.3, one has

$$\begin{aligned}
 & \mathbb{P}\left(\left|\frac{1}{\mu_p} - \bar{g}(p)\right| \leq \frac{4r_{\max}}{\lambda_p}\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \geq \mathbb{P}\left(\left|\frac{1}{\mu_p} - \bar{g}(p)\right| \leq \frac{4}{\mu_p}\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \\
 &\geq 1 - \frac{2}{T^4}, \tag{D.2}
 \end{aligned}$$

and

$$\mathbb{P}\left(\left|\lambda_p - \check{d}(p)\right| \leq 4\check{d}(p)\sqrt{\frac{2\log(T)}{n(p)}}\right) \geq 1 - \frac{2}{T^4}. \tag{D.3}$$

Therefore, one can derive the result that

$$\begin{aligned}
 & \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}(p)\check{d}(p)\right| \leq 4\left(r_{\max} + \bar{g}(p)\check{d}(p)\right)\sqrt{\frac{2\log(T)}{n^s(p)}}\right) \geq \left(1 - \frac{2}{T^4}\right)^2 \\
 &\geq 1 - \frac{4}{T^4}. \quad \text{Q.E.D.}
 \end{aligned}$$

Proof of Proposition 3. The proof here is very similar to the proof of Proposition 2. Before proving the two statements in Proposition 3, we first derive the following results. Based on Lemmas D.2 and D.3, and intermediate results in (D.2) and (D.3) (from the proof of Proposition 2), we have two direct results:

$$\mathbb{P}\left(\left|\frac{1}{\mu_p} - \bar{g}_W(p)\right| \leq \frac{r_{\max}}{\lambda_p}\sqrt{\frac{\log(T)}{2n_W^s(p)}}\right) \geq 1 - \frac{2}{\sqrt{T}}, \tag{D.4}$$

and

$$\mathbb{P}\left(\left|\lambda_p - \check{d}_W(p)\right| \leq \check{d}_W(p)\sqrt{\frac{\log(T)}{2n_W(p)}}\right) \geq 1 - \frac{2}{\sqrt{T}}. \tag{D.5}$$

Thus, we have

$$\mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \bar{g}_W(p)\check{d}_W(p)\right| \leq \left(r_{\max} + \bar{g}_W(p)\check{d}_W(p)\right)\sqrt{\frac{\log(T)}{2n_W^s(p)}}\right) \geq 1 - \frac{4}{\sqrt{T}}. \tag{D.6}$$

We first prove that for price $p \in \mathcal{P}_L$, $\mathbb{P}(\rho_p \geq 1) \leq \frac{4}{\sqrt{T}}$. For an arbitrary price $p \in \mathcal{P}_L$, we have $\hat{\rho}_W(p) \leq 1 - \delta(1 + r_{\max}/c)/(1 + \delta)$. For the notation simplicity, let $\rho_{\max} = r_{\max}/c$. Thus, we have

$$\hat{\rho}_W(p) \leq 1 - \frac{\delta(1 + \rho_{\max})}{1 + \delta} \iff \hat{\rho}_W(p) \leq 1 - \delta(\rho_{\max} + \hat{\rho}_W(p)).$$

Therefore, we can derive the probability bound and finish the proof:

$$\mathbb{P}(\rho_p \geq 1) \leq \mathbb{P}(\rho_p \geq \hat{\rho}_W(p) + \delta(\rho_{\max} + \hat{\rho}_W(p))) \tag{D.7a}$$

$$\leq \mathbb{P}(|\rho_p - \hat{\rho}_W(p)| \geq \delta(\rho_{\max} + \hat{\rho}_W(p))) \tag{D.7b}$$

$$\leq \mathbb{P}\left(|\rho_p - \hat{\rho}_W(p)| \geq (\rho_{\max} + \hat{\rho}_W(p))\sqrt{\frac{\log(T)}{2n_W^s(p)}}\right) \tag{D.7c}$$

$$\leq \frac{4}{\sqrt{T}}, \tag{D.7d}$$

where (D.7c) is by the algorithmic step that $n_W^s(p) \geq N_W = \max\{8\log T, (\log T)/2\delta^2\}$, and (D.7d) is a direct result of (D.6).

In the following analysis, we prove that for an arbitrary price $p \in \mathcal{P}$ whose $\rho_p \leq 1 - \varepsilon$, the probability that it will stay in $\mathcal{P}_L$ is larger than $1 - 4/\sqrt{T}$. By the pruning rule, we have

$$\mathbb{P}(p \notin \mathcal{P}_L) = \mathbb{P}\left(\hat{\rho}_W(p) \geq 1 - \frac{\delta(1 + \rho_{\max})}{1 - \delta}\right).$$

We also have the following equivalence:

$$\hat{\rho}_W(p) \geq \rho_p + \delta(\hat{\rho}_W(p) + \rho_{\max}) \Leftrightarrow \hat{\rho}_W(p) \geq \frac{\rho_p + \delta\rho_{\max}}{1 - \delta}.$$

By plugging in $\rho_p \leq 1 - \varepsilon = 1 - 2\delta(1 + \rho_{\max})$, we have

$$\mathbb{P}\left(\hat{\rho}_W(p) \geq \frac{\rho_p + \delta\rho_{\max}}{1 - \delta}\right) \geq \mathbb{P}\left(\hat{\rho}_W(p) \geq 1 - \frac{\delta(1 + \rho_{\max})}{1 - \delta}\right).$$

With the above results, we are ready to derive that

$$\mathbb{P}(p \notin \mathcal{P}_L) = \mathbb{P}\left(\hat{\rho}_W(p) \geq 1 - \frac{\delta(1 + \rho_{\max})}{1 - \delta}\right) \quad (\text{D.8a})$$

$$\leq \mathbb{P}\left(\hat{\rho}_W(p) \geq 1 - \frac{\delta(1 + \rho_{\max})}{1 - \delta}\right) \quad (\text{D.8b})$$

$$\leq \mathbb{P}\left(\hat{\rho}_W(p) \geq \frac{\rho_p + \delta\rho_{\max}}{1 - \delta}\right) \quad (\text{D.8c})$$

$$= \mathbb{P}\left(\hat{\rho}_W(p) \geq \rho_p + \delta(\hat{\rho}_W(p) + \rho_{\max})\right) \quad (\text{D.8d})$$

$$\leq \mathbb{P}\left(|\rho_p - \hat{\rho}_W(p)| \geq \delta(\rho_{\max} + \hat{\rho}_W(p))\right) \quad (\text{D.8e})$$

$$\leq \mathbb{P}\left(|\rho_p - \hat{\rho}_W(p)| \geq (\rho_{\max} + \hat{\rho}_W(p))\sqrt{\frac{\log(T)}{2n_W^s(p)}}\right) \quad (\text{D.8f})$$

$$\leq \frac{4}{\sqrt{T}}, \quad (\text{D.8g})$$

where (D.8f) and (D.8g) are exactly the same as (D.7c) and (D.7d). To conclude, we can prove that

$$\mathbb{P}(p \in \mathcal{P}_L) = 1 - \mathbb{P}(p \notin \mathcal{P}_L) \geq 1 - \frac{4}{\sqrt{T}}.$$

Q.E.D.

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

Proof of Lemma D.1. According to the definition of subexponential random variables (e.g., in Boucheron et al. 2013), 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 $s \leq \frac{\lambda}{2}$,

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

Let $t = \frac{s}{\lambda}$; then we have

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

Q.E.D.

Lemma D.2 (Subexponential Tail Bound for Exponential Distribution). *For random variables $X_i, i = 1, \dots, n$, following i.i.d. exponential distributions with mean $1/\lambda$, if $n \geq 8\log(T)$, then one has*

$$\mathbb{P}\left[\left|\frac{1}{\lambda} - \frac{1}{n} \sum_{i=1}^n X_i\right| \geq \frac{4}{\lambda} \sqrt{\frac{2\log(T)}{n}}\right] \leq \frac{2}{T^4}. \quad (\text{D.9})$$

Proof of Lemma D.2. By Lemma D.1, according to the concentration inequality of subexponential random variables, one has

$$\mathbb{P}\left[\left|\mathbb{E}[X_i] - \frac{1}{n} \sum_{i=1}^n X_i\right| \geq t\right] \leq \begin{cases} 2\exp\left(-\frac{nt^2}{2\frac{4}{\lambda^2}}\right) & \text{if } 0 \leq t \leq \frac{2}{\lambda}, \\ 2\exp\left(-\frac{nt}{2\frac{2}{\lambda}}\right) & \text{if } t > \frac{2}{\lambda}. \end{cases} \quad (\text{D.10})$$

Plug in $t = \frac{4}{\lambda} \sqrt{\frac{2\log(T)}{n}}$ and $n \geq 8\log(T)$, one has $\frac{4}{\lambda} \sqrt{\frac{2\log(T)}{n}} \leq \frac{2}{\lambda}$, and thus, by (D.10),

$$\begin{aligned} \mathbb{P}\left[\left|\frac{1}{\lambda} - \frac{1}{n} \sum_{i=1}^n X_i\right| \geq \frac{4}{\lambda} \sqrt{\frac{2\log(T)}{n}}\right] &\leq 2\exp\left(-\frac{n \frac{16 \cdot 2\log(T)}{\lambda^2}}{2\frac{4}{\lambda^2}}\right) \\ &= 2\exp(-4\log(T)) \\ &= \frac{2}{T^4}. \quad \text{Q.E.D.} \end{aligned}$$

Lemma D.3 (Subexponential Tail Bound for a Poisson Process). *For random variables $X_i, i = 1, \dots, n$, following i.i.d. exponential distributions with mean $1/\lambda$, if $n \geq 8\log(T)$, then one has*

$$\mathbb{P}\left[\left|\lambda - \frac{n}{\sum_{i=1}^n X_i}\right| \geq \frac{4}{\frac{1}{n} \sum_{i=1}^n X_i} \sqrt{\frac{2\log(T)}{n}}\right] \leq \frac{2}{T^4}. \quad (\text{D.11})$$

Proof of Lemma D.3. We can view $X_i, i = 1, \dots, n$, as time intervals between events in a Poisson process with rate λ . Therefore, $\frac{n}{\sum_{i=1}^n X_i}$ is an empirical estimate of λ . Let $g = \frac{n}{\sum_{i=1}^n X_i}$ and $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$; then, one has

$$\begin{aligned} \mathbb{P}\left[\left|\lambda - \frac{n}{\sum_{i=1}^n X_i}\right| \geq \frac{4}{\frac{1}{n} \sum_{i=1}^n X_i} \sqrt{\frac{2\log(T)}{n}}\right] \\ &= \mathbb{P}\left[\left|\frac{\lambda - g}{\lambda g}\right| \geq \frac{4}{\bar{X} \lambda g} \sqrt{\frac{2\log(T)}{n}}\right] \\ &= \mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{\lambda}\right| \geq \frac{4}{\lambda} \sqrt{\frac{2\log(T)}{n}}\right] \\ &\leq \frac{2}{T^4}, \end{aligned}$$

where the last inequality is the direct result of Lemma D.2. Q.E.D.

Appendix E. Proofs of Section 4.1

Proof of Proposition 4. 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 $\hat{S}_\tau = S_\tau$, the two processes have an equal number of customers at any time thereafter, and thus the two processes couple. Define $\bar{S} = \lfloor \frac{(\log(T))^{3/2}}{-\log \rho} \rfloor$. Then we have

$$\begin{aligned} \mathbb{P}(\hat{S}_\tau = S_\tau) &\geq \mathbb{P}(\hat{S}_\tau = S_\tau | \hat{S}_0 \leq \bar{S}) \mathbb{P}(\hat{S}_0 \leq \bar{S}) \\ &\geq \mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_n \leq \tau\right) \mathbb{P}(\hat{S}_0 \leq \bar{S}). \end{aligned} \quad (\text{E.1})$$

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

$$\mathbb{P}(\hat{S}_0 \leq \bar{S}) = 1 - \frac{\ell_0 c^c}{c!} \rho^{\bar{S}+1} \geq 1 - \frac{1}{T^2}.$$

By Lemma E.1, 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 and Marathe 1978, Daley and Servi 1998), $\bar{A}_1$ can be derived by $\bar{A}_c$, where $\bar{A}_c = \frac{1}{c\mu - \lambda}$. By Lemma E.2, we have

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

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

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

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

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

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

$$\geq \left(1 - e^{-(\log(T))^{3/2}}\right) \left(1 - \frac{1}{T^2}\right) \quad (\text{E.2g})$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \geq 1 - \frac{2}{T^2}, \quad (\text{E.2h})$$

where C_3 and C_4 are positive constants. Q.E.D.

Proof of Proposition 5. Without loss of generality, we consider two cases: (1) the price changes from a lower price p_l to a higher price p_h , and (2) the price changes

from a higher price p_h to a lower price p_l . For each case, we further consider two subcases: (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. Let S denote the actual system, and let $\tilde{S}$ denote the virtual system. For $t \in \{\tau, 2\tau\}$, define the following events:

$$C_t = \{S \text{ reaches the steady state under the new price before } t\}, \quad (\text{E.3})$$

$$\check{C}_t = \{\tilde{S} \text{ reaches the steady state under the new price before } t\}. \quad (\text{E.4})$$

Case 1(i) (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 $\tilde{S}$ that starts with S_0 customers. The future customers arrive by the same process as the actual system, that is, 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 of 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

$$\mathbb{P}(\check{C}_{2\tau}) \quad (\text{E.5a})$$

$$\geq \mathbb{P}(\check{C}_{2\tau} | S_0 \leq 2\bar{S}_h) \cdot \mathbb{P}(S_0 \leq 2\bar{S}_h) \quad (\text{E.5b})$$

$$\geq \mathbb{P}\left(\sum_{n=1}^{2\bar{S}_h} \bar{A}_{1h} \leq 2\tau\right) \cdot \mathbb{P}(S_0 \leq 2\bar{S}_h) \quad (\text{E.5c})$$

$$\geq \left(\mathbb{P}\left(\sum_{n=1}^{\bar{S}_h} \bar{A}_{1h} \leq \tau\right)\right)^2 \cdot \mathbb{P}(S_0 \leq 2\bar{S}_h) \quad (\text{E.5d})$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \left(1 - \frac{\ell_{0l} c^c}{c!} \rho_l^{2\bar{S}_h+1}\right) \quad (\text{E.5e})$$

$$\geq \left(1 - \frac{1}{T^2}\right)^2 \left(1 - \rho_l^{\frac{2}{\log \rho_h} (-\log(T))^{3/2}}\right) \quad (\text{E.5f})$$

$$\geq \left(1 - \frac{1}{T^2}\right)^3 \quad (\text{E.5g})$$

$$\geq 1 - \frac{3}{T^2}, \quad (\text{E.5h})$$

where $\bar{S}_h = \frac{(\log(T))^{3/2}}{-\log \rho_h}$, and ℓ_{0l} is the steady-state probability of the system with zero customers under price p_l . The inequality (E.5g) holds because of Assumption 2, part 3.

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

$$\mathbb{P}(C_{2\tau}) \geq \mathbb{P}(\check{C}_{2\tau}) \geq 1 - \frac{3}{T^2}. \quad (\text{E.6})$$

Case 1(ii) (From p_l to p_h and $\mu_{p_l} \leq \mu_{p_h}$). Consider a virtual system $\tilde{S}$ that starts with S_0 customers. The future customers arrive as the process under p_l , that is, 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 by $\hat{t}$. After $\hat{t}$, the further customers that 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 $\hat{t}$. Therefore, the probability of the virtual system reaching the steady state under p_l before 2τ is the probability that $\hat{t} + \hat{t} \leq 2\tau$. Therefore, we can compute the probability as

$$\mathbb{P}(\check{C}_{2\tau}) \quad (E.7a)$$

$$= \mathbb{P}(\hat{t} + \hat{t} \leq 2\tau) \quad (E.7b)$$

$$\geq \mathbb{P}(\hat{t} + \hat{t} \leq 2\tau | S_0 \leq 2\bar{S}_l) \cdot \mathbb{P}(S_0 \leq 2\bar{S}_l) \quad (E.7c)$$

$$\geq \mathbb{P}(\hat{t} \leq \tau) \cdot \mathbb{P}(\hat{t} \leq \tau | S_0 \leq 2\bar{S}_l) \cdot \mathbb{P}(S_0 \leq 2\bar{S}_l) \quad (E.7d)$$

$$\geq \left(1 - \frac{2}{T^2}\right)^2 \quad (E.7e)$$

$$\geq 1 - \frac{4}{T^2}, \quad (E.7f)$$

where (E.7e) is by Proposition 4. The virtual system considers a longer service time for the existing customers that arrived under a lower price in the previous batch. Therefore, in probabilistic sense, the virtual system takes a longer time to reach steady state:

$$\mathbb{P}(C_{2\tau}) \geq \mathbb{P}(\check{C}_{2\tau}) \geq 1 - \frac{4}{T^2}. \quad (E.8)$$

Case 2(i) (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, part 1, and $\Delta_\lambda \leq \frac{\rho_{p_l}}{-3\log p_l}$ by Assumption 2, part 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 in the actual system, that is, λ_{p_h} and Δ_λ . The customers arriving 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 $\check{S}_t^{\text{buffer}}$. 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 $\check{S}_t^{\text{buffer}}$. 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:

$$\mathbb{P}(\hat{t} \leq \tau) \geq 1 - \frac{2}{T^2}. \quad (E.9)$$

Similarly, by Proposition 4, if the starting length of the system $\check{S}_t^{\text{buffer}}$ 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

$$\mathbb{P}(\check{C}_{2\tau}) \quad (E.10a)$$

$$\geq \mathbb{P}(\check{C}_{2\tau} | \hat{t} \leq \tau) \cdot \mathbb{P}(\hat{t} \leq \tau) \quad (E.10b)$$

$$\geq \mathbb{P}\left(\check{C}_{2\tau} | \hat{t} \leq \tau, \check{S}_t^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right) \quad (E.10c)$$

$$\cdot \mathbb{P}\left(\check{S}_t^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}} \middle| \hat{t} \leq \tau\right) \cdot \mathbb{P}(\hat{t} \leq \tau) \quad (E.10d)$$

$$\geq \left(1 - \frac{3}{T^2}\right) \cdot \mathbb{P}\left(\check{S}_t^{\text{buffer}} \leq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right). \quad (E.10e)$$

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

$$\mathbb{P}\left(\check{S}_t^{\text{buffer}} \geq \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}\right) \quad (E.11a)$$

$$\leq \min_a e^{\tau \Delta_{p_l} e^a - \tau \Delta_{p_l} - a \frac{(\log(T))^{3/2}}{-\log \rho_{p_l}}} \quad (E.11b)$$

$$\leq e^{-(\log(T))^{3/2} \left(\frac{1 + \log(-\log \rho_{p_l} \Delta_\lambda \sqrt{\log(T)})}{\log \rho_{p_l}} \right)} \quad (E.11c)$$

$$\leq e^{-(\log(T))^{3/2}} \quad (E.11d)$$

$$\leq \frac{1}{T^2}, \quad (E.11e)$$

where (E.11b) is by Chernoff bound. The inequality (E.11c) holds because $\frac{1 + \log(-\log \rho_{p_l} \Delta_\lambda \sqrt{\log(T)})}{\log \rho_{p_l}} \geq 1$ by Assumption 2, part

3. Therefore, we can future derive that

$$\mathbb{P}(\check{C}_{2\tau}) \geq \left(1 - \frac{3}{T^2}\right) \left(1 - \frac{1}{T^2}\right) \geq 1 - \frac{4}{T^2}. \quad (E.12)$$

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 if there are available resources before $\hat{t}$. Therefore, in probabilistic sense, the virtual system takes a longer time to reach steady state:

$$\mathbb{P}(C_{2\tau}) \geq \mathbb{P}(\check{C}_{2\tau}) \geq 1 - \frac{4}{T^2}. \quad (E.13)$$

Case 2(ii) (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 $\check{S}$ that 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 bound the coupling speed of the virtual queue, we consider the speed conditional on $S_0 \leq \bar{S}_l$. Consider another virtual system $\hat{S}$ that 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 $\check{S}$ reach the steady state equals the probability that

these two systems couple with the virtual system $\hat{S}$. We have shown that once they reach the 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 the following events:

$$\check{C}_\tau^0 = \{\check{S} \text{ reaches the empty state before } \tau\}, \quad (\text{E.14})$$

$$\hat{C}_\tau^0 = \{\hat{S} \text{ reaches the empty state before } \tau\}, \quad (\text{E.15})$$

$$C_\tau^0 = \{S \text{ reaches the empty state before } \tau\}. \quad (\text{E.16})$$

Based on the above analysis, we can derive

$$\mathbb{P}(C_\tau) \quad (\text{E.17a})$$

$$\geq \mathbb{P}(C_\tau | S_0 \leq \bar{S}_l) \cdot \mathbb{P}(S_0 \leq \bar{S}_l) \quad (\text{E.17b})$$

$$\geq \mathbb{P}(\check{C}_\tau^0 | S_0 \leq \bar{S}_l) \cdot \mathbb{P}(S_0 \leq \bar{S}_l) \cdot \mathbb{P}(\check{C}_\tau^0) \quad (\text{E.17c})$$

$$\geq \mathbb{P}(\check{C}_\tau^0 | S_0 \leq \bar{S}_l) \cdot \mathbb{P}(S_0 \leq \bar{S}_l) \cdot \mathbb{P}(\hat{C}_\tau^0) \quad (\text{E.17d})$$

$$\geq \mathbb{P}(\check{C}_\tau^0 | S_0 \leq \bar{S}_l) \cdot \mathbb{P}(S_0 \leq \bar{S}_l) \cdot \mathbb{P}(\hat{C}_\tau^0 | \hat{S}_0 \leq \bar{S}_l) \cdot \mathbb{P}(\hat{S}_0 \leq \bar{S}_l) \quad (\text{E.17e})$$

$$\geq \mathbb{P}(\check{C}_\tau^0 | S_0 \leq \bar{S}_h) \cdot \mathbb{P}(S_0 \leq \bar{S}_h) \cdot \mathbb{P}(\hat{C}_\tau^0 | \hat{S}_0 \leq \bar{S}_l) \cdot \mathbb{P}(\hat{S}_0 \leq \bar{S}_l) \quad (\text{E.17f})$$

$$\geq \left(1 - \frac{1}{T^2}\right)^4 \quad (\text{E.17g})$$

$$\geq 1 - \frac{4}{T^2}, \quad (\text{E.17h})$$

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 (E.17f) is by Assumption 2, part 1, and $\mu_{p_l} < \mu_{p_h}$. The inequality (E.17g) is by the intermediate results we derive in Proposition 4. Q.E.D.

The analysis also considers the first-order stochastic dominance between two random variables defined in definition 1 in Hadar and Russell (1969) and in Seth and Yalonetzky (2014).

Lemma E.1 (First-Order Stochastic Dominance Between A_i and A_j in an $M/M/c$ Queue). For random variables A_i and A_j , where $i < j \leq c$, A_i first-order dominates A_j , that is, $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$ for any $x > 0$.

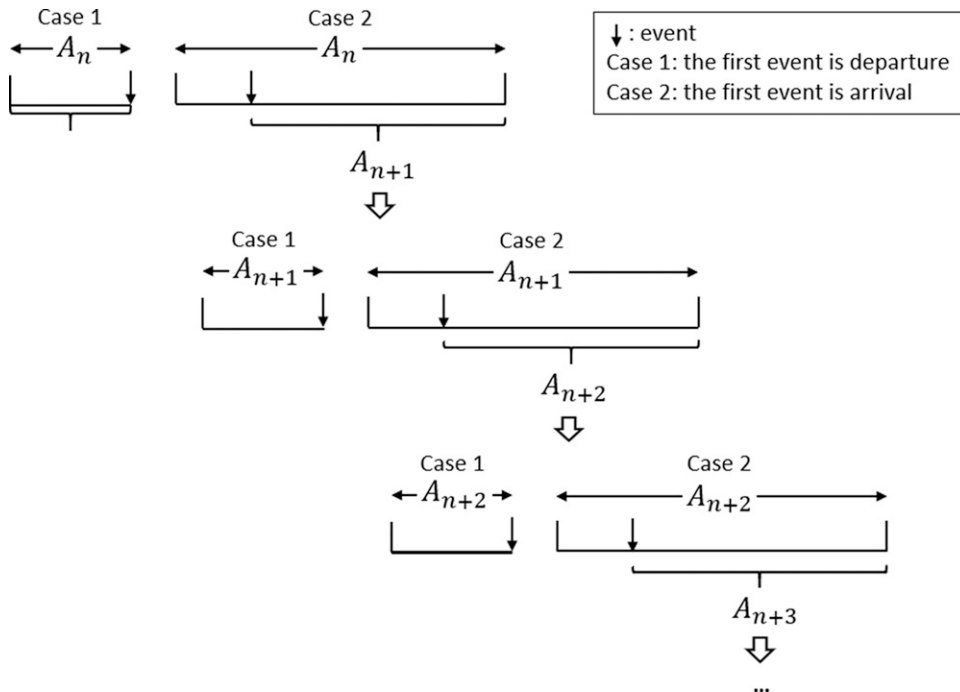
Proof of Lemma E.1. To compare random variables A_i and A_j , where $i < j$ of an $M/M/c$ queue process Q , we construct an auxiliary process $\hat{Q}$, which is the same as Q except that when the system has n customers with $j \leq n \leq c + j - i$, the service rate of $\hat{Q}$ is $n - j + i$. Similarly, we define $\hat{A}_n$ for $n = 1, \dots, \infty$ of $\hat{Q}$. Omahen and Marathe (1978, lemma 1) and Daley and Servi (1998) showed the recursion between A_n , for $n = 1, \dots, \infty$. The term A_n can be decomposed into intervals shown in Figure E.1. 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:

$$\bar{A}_n = \mathbb{P}(\mathcal{B}_n^a) \cdot (\mathbb{E}[t_n^a] + \bar{A}_{n+1} + \bar{A}_n) + \mathbb{P}(\mathcal{B}_n^l) \cdot \mathbb{E}[t_n^l], \quad (\text{E.18})$$

where $\bar{A}_n$ is the expected value of the random variable A_n . 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 E.1;

Figure E.1. Intervals Within Busy Period A_n



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 the same (when $n \geq c + j - i$) departure rate in the intervals in Figure E.1.

Therefore, $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$. By definition 1 in Seth and Yalonetzky (2014), one further has $\mathbb{E}_{A_i}[u(x)] < \mathbb{E}_{A_j}[u(x)]$ for all strictly increasing, continuous functions $u(x) : [0, \infty] \rightarrow \mathbb{R}$. Q.E.D.

Lemma E.2 (Concentration Inequality for Independent Samples). *This lemma is similar to lemma 11 in Chen and Shi (2019). Let ξ_i be i.i.d. random variables with mean zero and standard deviation σ . If the moment generating function of ξ_i around zero is finite, that is, 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 $\bar{S} = \left\lceil \frac{(\log(T))^{3/2}}{-\log \rho} \right\rceil$.

Proof of Lemma E.2. For s in $[-\delta, \delta]$, define

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

For $x > 0$ and $s \in [0, \delta]$, 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 zero; then 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 $x = 2\sigma\sqrt{-\log \rho}$, we have

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

Appendix F. Proofs of Section 4.2

Proof of Theorem 1. We bound the relaxed regret during the warm-up and learning phases separately. We further decompose the loss during the learning phase into four parts. The first is the loss of the instability, which happens if any retained price $p \in \mathcal{P}_L$ has a utilization factor larger than one. The second is the loss of nonstationarity, which represents the revenue loss during the transient state when the system changes price. The third is the loss of pruning error, which is induced by mistakenly prune the $(1 - \varepsilon_T)$ -stable state-independent price $\tilde{p}$ from $\mathcal{P}_L$. The fourth is the loss of suboptimality, which is induced by not offering the state-independent optimal price from the learning price set $\mathcal{P}_L$ in certain time periods.

Regret of the Warm-Up Phase. The difference between $J_{\text{Warm-up}}^*$ and $J_{\text{Warm-up}}^{\text{BUCB}}$ is at most linear on the length of the warm-up phase, which is $O(P(\log(T))^2)$. Therefore, we can

directly have

$$\overline{\text{Regret}}_{\text{Warm-up}}(\pi_{\text{BUCB}}, \varepsilon_T, T) = O(P(\log(T))^3). \quad (\text{F.1})$$

Loss of Instability During the Learning Phase. Define event $\mathcal{R}$ as when any retained price has a utilization factor larger than or equal to one. Therefore, we can define the complementary event $\bar{\mathcal{R}}$ as when all retained prices have utilization factors less than one. Based on Proposition 3,

$$\begin{aligned} \mathbb{P}(\bar{\mathcal{R}}) &= \mathbb{P}(\mathcal{P}_L \subseteq \mathcal{P}_1) = \prod_{p \in \mathcal{P}_L} (\rho_p < 1) \geq \left(1 - \frac{2}{\sqrt{T}}\right)^{|\mathcal{P}_L|} \geq 1 - \frac{2|\mathcal{P}_L|}{\sqrt{T}} \\ &\geq 1 - \frac{2P}{\sqrt{T}}. \end{aligned} \quad (\text{F.2})$$

Therefore, the regret in the learning phase can be expressed as

$$\overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T) \quad (\text{F.3a})$$

$$= \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}})\mathbb{P}(\bar{\mathcal{R}}) + \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\mathcal{R})\mathbb{P}(\mathcal{R}) \quad (\text{F.3b})$$

$$\leq \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}}) \left(1 - \frac{2P}{\sqrt{T}}\right) + O\left(T \cdot \frac{P}{\sqrt{T}}\right) \quad (\text{F.3c})$$

$$= \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}}) \left(1 - \frac{2P}{\sqrt{T}}\right) + O(P\sqrt{T}) \quad (\text{F.3d})$$

$$\leq \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}}) + O(P\sqrt{T}). \quad (\text{F.3e})$$

Below, we analyze $\overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}})$.

Loss of Nonstationarity During the Learning Phase. We denote the offered price of batch m during the learning phase of BUCB by p_m . According to the above analysis, we conduct the following analysis conditional on $\rho_p < 1$, $\forall p \in \mathcal{P}_L$. Therefore, the expected revenue collected in batch m is

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

where $J_{m,0-2\tau}(p_m)$ denotes the expected revenue collected in the first 2τ periods, and $J_{m,2\tau-I_m\tau}(p_m)$ denotes the expected revenue collected from the $(2\tau + 1)$ th period to the end of batch m of BUCB. 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 $J_{m,2\tau-I_m\tau}(p_m)$ 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

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

Loss of Pruning Error During the Learning Phase. For batch m , we can compute the difference of the expected revenue rates between the optimal $(1 - \varepsilon_T)$ -stable linear programming

benchmark and selected price p_m as

$$\Delta_{\tilde{p}}(p_m) = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - r(\lambda_{p_m}, \mu_{p_m}, p_m).$$

Define $p_L^* = \arg \max_{p \in \mathcal{P}_L} p \lambda_p / \mu_p$ as the learning upper-bound price, and correspondingly define $\Delta_L(p_m) = r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$. By Proposition 3, we have

$$\mathbb{P}(\tilde{p} \in \mathcal{P}_L) \geq 1 - \frac{4}{\sqrt{T}},$$

and thus

$$\mathbb{P}(\Delta_{\tilde{p}}(p_m) \leq \Delta_L(p_m)) \geq 1 - \frac{4}{\sqrt{T}}.$$

Loss of Suboptimality During the Learning Phase. We further have the following results for $\Delta_L(p_m)$. Recall the upper and lower confidence bounds of the revenue rate defined in Definition 3. It is clear that

$$U_m(p) - L_m(p) = 2\text{Rad}_m(p) = 8(r_{\max} + \bar{g}_m(p)\check{d}_m(p))\sqrt{\frac{2\log(T)}{n_m^s(p)}}p.$$

For batch m , we can compute the difference in the revenue rates between the optimal price and selected price p_m as

$$\Delta_L(p_m) = r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - r(\lambda_{p_m}, \mu_{p_m}, p_m) \quad (\text{F.5a})$$

$$= (r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_m)) + (U_m(p_m) - L_m(p_m)) \\ + (L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)) \quad (\text{F.5b})$$

$$\leq (r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)) + (U_m(p_m) - L_m(p_m)) \\ + (L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)), \quad (\text{F.5c})$$

where the last inequality holds because $U_m(p_m) \geq U_m(p_L^*)$ by the algorithm construction.

Thus, the cumulative regret during the learning phase is (by incorporating the analysis on the loss of instability, nonstationarity, and pruning error)

$$\overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T) \quad (\text{F.6a})$$

$$\leq \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}}) + O(P\sqrt{T}) \quad (\text{F.6b})$$

$$= \sum_{m=1}^M (r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p})I_m\tau - J_m(p_m)) + O(P\sqrt{T}) \quad (\text{F.6c})$$

$$= \sum_{m=1}^M \mathbb{E}[\Delta_{\tilde{p}}(p_m)I_m\tau + O(\tau)] + O(P\sqrt{T}) \quad (\text{F.6d})$$

$$\leq \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau + O(\tau)] + O(\sqrt{T}) + O(P\sqrt{T}) \quad (\text{F.6e})$$

$$= \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau] + \sum_{m=1}^M O(\tau) + O(P\sqrt{T}). \quad (\text{F.6f})$$

By plugging in the results in (F.5), we have

$$\sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau] = \sum_{m=1}^M \mathbb{E}[(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*))I_m\tau] \\ + \sum_{m=1}^M \mathbb{E}[(U_m(p_m) - L_m(p_m))I_m\tau] \\ + \sum_{m=1}^M \mathbb{E}[(L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m))I_m\tau]. \quad (\text{F.7})$$

We first analyze the middle term of (F.7). Let $\mathcal{M}_p$ denote the batches choosing price p , and let T_p denote the number of periods choosing price p , that is, $T_p = \sum_{m \in \mathcal{M}_p} I_m\tau$. Then,

$$\sum_{m=1}^M \mathbb{E}[(U_m(p_m) - L_m(p_m))I_m\tau] \\ = \sum_{p \in \mathcal{P}_L} \sum_{m \in \mathcal{M}_p} \mathbb{E}[U_m(p) - L_m(p)]I_m\tau \quad (\text{F.8a})$$

$$= \sum_{p \in \mathcal{P}} \sum_{m \in \mathcal{M}_p} \mathbb{E}[U_m(p) - L_m(p)]I_m\tau \quad (\text{F.8b})$$

$$= \sum_{p \in \mathcal{P}} \sum_{m \in \mathcal{M}_p} 2\mathbb{E}[\text{Rad}_m(p)]I_m\tau \quad (\text{F.8c})$$

$$\leq \sum_{p \in \mathcal{P}} O(\log(T)\sqrt{T_p}) \quad (\text{F.8d})$$

$$= O(\log(T)) \sum_{p \in \mathcal{P}} \sqrt{T_p} \quad (\text{F.8e})$$

$$\leq O(\log T \sqrt{PT}), \quad (\text{F.8f})$$

where (F.8d) holds by Lemma F.1 (see Appendix F), and (F.8f) follows Jensen's inequality.

Below, we analyze the first term of (F.7), and the same logic applies to the third term, of which we omit the details here:

$$\sum_{m=1}^M \mathbb{E}[(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*))I_m\tau] \quad (\text{F.9a})$$

$$\leq \sum_{m=1}^M \mathbb{E}[(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)) \mathbb{1}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*))I_m\tau] \quad (\text{F.9b})$$

$$\leq \sum_{m=1}^M \mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) \mathbb{1}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*)) \\ > U_m(p_L^*)) - U_m(p_L^*) \mathbb{1}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*))I_m\tau] \quad (\text{F.9c})$$

$$= \sum_{m=1}^M r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) \mathbb{P}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*))I_m\tau \\ - \sum_{m=1}^M \mathbb{E}[U_m(p_L^*) \mathbb{1}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*))I_m\tau] \quad (\text{F.9d})$$

$$\leq \sum_{m=1}^M r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) \mathbb{P}(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) > U_m(p_L^*))I_m\tau \quad (\text{F.9e})$$

$$\leq O\left(\frac{1}{T^3}\right), \quad (\text{F.9f})$$

where (F.9e) is because $U_m(p_L^*) > 0$ by Definition 3, and (F.9f) follows Proposition 2.

Regret of the Learning Phase. Given that $O(n_M(p)) = O(T_p)$, and the total number of batches is $\log_2\left(1 + \frac{T - PN_W}{2(\log(T))^2}\right) = O(\log(T))$, we have $\sum_{m=1}^M O(\tau) = O((\log(T))^3)$. Combining the three terms together,

$$\begin{aligned}
 \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T) &\leq \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m) I_m \tau] \\
 &\quad + \sum_{m=1}^M O(\tau) + O(P\sqrt{T}) \\
 &= O(\log(T)\sqrt{PT}) + O\left(\frac{1}{T^3}\right) \\
 &\quad + O((\log(T))^3) + O(P\sqrt{T}) \\
 &= O(\log(T)\sqrt{PT}). \tag{F.10}
 \end{aligned}$$

Total Regret of BUCB. By adding up the relaxed regret during the warm-up and learning phases, we have the regret of BUCB as

$$\overline{\text{Regret}}(\pi_{\text{BUCB}}, \varepsilon_T, T) \tag{F.11a}$$

$$= \overline{\text{Regret}}_{\text{Warm-up}}(\pi_{\text{BUCB}}, \varepsilon_T, T) + \overline{\text{Regret}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T) \tag{F.11b}$$

$$= O(P(\log(T))^3) + O(\log(T)\sqrt{PT}) \tag{F.11c}$$

$$= O(\log(T)\sqrt{PT}) \tag{F.11d}$$

$$= \tilde{O}(\sqrt{PT}). \tag{F.11e}$$

Q.E.D.

Lemma F.1. For any price $p \in \mathcal{P}$,

$$\sum_{m \in \mathcal{M}_p} I_m \tau \mathbb{E}[\text{Rad}_m(p)] = O\left(\sqrt{T_p \log(T)}\right). \tag{F.12}$$

Proof of Lemma F.1. Consider that for any price p , the number of new observations (denoted by n) collected in a certain time is proportional to the length of this time duration (denoted by t_n). Thus, there exists a constant $\mathcal{Y}$ such that $t_n \leq \mathcal{Y}n$. We further denote $n_{m,m-1}(p)$ as the new observations collected during batch m , and thus $n_{m,m-1}(p) \leq n_m(p)$ and $I_m \tau \leq \mathcal{Y}n_{m,m-1}(p)$. Plugging in the definition of $\text{Rad}_m(p)$, $\forall m = 1, \dots, M$, one has

$$\sum_{m \in \mathcal{M}_p} I_m \tau \mathbb{E}[\text{Rad}_m(p)] \leq O\left(\sqrt{\log(T)}\right) \sum_{m \in \mathcal{M}_p} I_m \tau \frac{1}{\sqrt{n_{m,m-1}(p)}} \tag{F.13a}$$

$$\leq O\left(\sqrt{\log(T)}\right) \sum_{m \in \mathcal{M}_p} I_m \tau \frac{1}{\sqrt{n_{m,m-1}(p)}} \tag{F.13b}$$

$$\leq O\left(\sqrt{\log(T)}\right) \cdot \sqrt{\mathcal{Y}} \sum_{m \in \mathcal{M}_p} \sqrt{I_m \tau} \tag{F.13c}$$

$$\leq O\left(\sqrt{\log(T)} \cdot \sqrt{|\mathcal{M}_p| T_p}\right) \tag{F.13d}$$

$$= O(\log(T)\sqrt{T_p}), \tag{F.13e}$$

where (F.13d) is by Jensen's inequality, and (F.13e) is by $|\mathcal{M}_p| \leq \sum_{m \in \mathcal{M}} |\mathcal{M}_p| = O(\log(T))$. $\square$

Appendix G. Proofs of Section 4.3

Proof of Theorem 2. The analysis here follows the same structure as the proof of Theorem 1.

Bayesian Regret in the Warm-Up Phase. We can bound the Bayesian regret in the warm-up phase in the same order as the length of the warm-up phase:

$$\overline{\text{BR}}_{\text{Warm-up}}(\pi_{\text{BTS}}, \varepsilon_T, T) = O(P \log(T)^3). \tag{G.1}$$

Bayesian Regret in the Learning Phase. We also decompose the regret loss during the learning phase into four parts: instability, nonstationarity, pruning error, and suboptimality. We analyze them below, respectively.

Loss of Instability During the Learning Phase. We keep using the definition of event $\mathcal{R}$ as when any retained price has a utilization factor larger than or equal to one, and correspondingly the complementary event $\bar{\mathcal{R}}$. The probability bound in (F.2) is still valid, and we have

$$\overline{\text{BR}}_{\text{Learning}}(\pi_{\text{BTS}}, \varepsilon_T, T) \leq \overline{\text{BR}}_{\text{Learning}}(\pi_{\text{BTS}}, \varepsilon_T, T|\bar{\mathcal{R}}) + O(P\sqrt{T}). \tag{G.2}$$

The intermediate steps are similar to (F.3). Below we analyze $\overline{\text{BR}}_{\text{Learning}}(\pi_{\text{BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}})$.

Loss of Nonstationarity During the Learning Phase. In order to compute the expected revenue under a price during a batch, we need to compute both the revenue collected during the former transient state and during the later steady state. We keep using the definition of event $\mathcal{C}_{2\tau}$ as that the system reaches the steady state under the new price within the first 2τ periods, and its complementary event $\bar{\mathcal{C}}_{2\tau}$. Thus, we have the expected revenue during batch m as (same as (F.4))

$$J_m(p_m) = r(\lambda_{p_m}, \mu_{p_m}, p_m) I_m \tau - O(\tau). \tag{G.3}$$

Loss of Pruning Error During the Learning Phase. Based on the above regret analysis of instability and nonstationarity, the key of the regret is to compute $\Delta_{\tilde{p}}(p_m) = r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$. By letting $p_L^* = \arg \max_{p \in \mathcal{P}_L} p \lambda_p / \mu_p$ be the learning upper-bound price, we correspondingly have $\Delta_L(p_m) = r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - r(\lambda_{p_m}, \mu_{p_m}, p_m)$. By Proposition 3, we have $\mathbb{P}(\tilde{p} \in \mathcal{P}_L) \geq 1 - \frac{4}{\sqrt{T}}$, and thus $\mathbb{P}(\Delta_{\tilde{p}}(p_m) \leq \Delta_L(p_m)) \geq 1 - \frac{4}{\sqrt{T}}$.

Loss of Suboptimality During the Learning Phase. The Bayesian regret in batch m is

$$\bar{\text{BR}}_m = \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, \tilde{p}) I_m \tau - J_m(p_m) | \mathcal{F}_m]] \tag{G.4a}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - J_m(p_m) | \mathcal{F}_m]] + I_m \tau \\
 &\quad \cdot O\left(\frac{4}{\sqrt{T}}\right). \tag{G.4b}
 \end{aligned}$$

The following analysis mainly relies on the distribution matching property of the Thompson sampling algorithm. In batch m , the chosen price p_m and the independent optimal price benchmark of during the learning phase p_L^* are identically distributed conditional on $\mathcal{F}_m$, which indicates that $\mathbb{E}[U_m(p_m) - U_m(p_L^*) | \mathcal{F}_m] = 0$ (e.g., see the proof of lemma 3.7 in Slivkins (2019)). Therefore, we can further bound the first term in the above regret as

$$\mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - J_m(p_m) | \mathcal{F}_m]] \tag{G.5a}$$

$$= \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[I_m \tau U_m(p_m) - J_m(p_m) + r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - I_m \tau U_m(p_L^*) | \mathcal{F}_m]] \quad (\text{G.5b})$$

$$= \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[I_m \tau U_m(p_m) - J_m(p_m) | \mathcal{F}_m]] + \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - I_m \tau U_m(p_L^*) | \mathcal{F}_m]] \quad (\text{G.5c})$$

$$= \mathbb{E}[I_m \tau U_m(p_m) - J_m(p_m)] + \mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - I_m \tau U_m(p_L^*)] \quad (\text{G.5d})$$

$$\leq \mathbb{E}[I_m \tau U_m(p_m) - I_m \tau \cdot r(\lambda_{p_m}, \mu_{p_m}, p_m) + O(\tau)] + \mathbb{E}[I_m \tau \cdot r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - I_m \tau U_m(p_L^*)] \quad (\text{G.5e})$$

$$= I_m \tau (\mathbb{E}[U_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)] + \mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)]) + O(\tau). \quad (\text{G.5f})$$

Following two properties of the upper and lower bounds of the stationary revenue rate in (8) and (9) (derived using Proposition 2), we have

$$\begin{aligned} \mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)] &\leq \mathbb{E}[(r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*))^+] \\ &\leq \sum_{p \in \mathcal{P}} \mathbb{E}[(r(\lambda_p, \mu_p, p) - U_m(p))^+] \\ &\leq \frac{\gamma}{T}, \end{aligned} \quad (\text{G.6})$$

and similarly,

$$\begin{aligned} \mathbb{E}[U_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)] &= \mathbb{E}[2 \cdot \mathbf{Rad}_m(p_m) + L_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)] \\ &\leq \mathbb{E}[2 \cdot \mathbf{Rad}_m(p_m)] + \mathbb{E}[(L_m(p) - r(\lambda_{p_m}, \mu_{p_m}, p_m))^+] \\ &\leq \mathbb{E}[2 \cdot \mathbf{Rad}_m(p_m)] + \sum_{p \in \mathcal{P}} \mathbb{E}[(L_m(p) - r(\lambda_{p_m}, \mu_{p_m}, p_m))^+] \\ &\leq \mathbb{E}[2 \cdot \mathbf{Rad}_m(p_m)] + \frac{\gamma}{T}. \end{aligned} \quad (\text{G.7})$$

Summing over all batches m , $\forall m = 1, \dots, M$, we have

$$\sum_{m=1}^M \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - J_m(p_m) | \mathcal{F}_m]] \quad (\text{G.8a})$$

$$= \sum_{m=1}^M (I_m \tau (\mathbb{E}[U_m(p_m) - r(\lambda_{p_m}, \mu_{p_m}, p_m)] + \mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) - U_m(p_L^*)]) + O(\tau)) \quad (\text{G.8b})$$

$$\leq \sum_{m=1}^M (I_m \tau (2 \frac{\gamma}{T} + 2 \mathbb{E}[\mathbf{Rad}_m(p_m)])) + O(\tau \log(T)) \quad (\text{G.8c})$$

$$\leq 2 \sum_{m=1}^M I_m \tau \frac{\gamma}{T} + 2 \sum_{p \in \mathcal{P}} \sum_{m \in \mathcal{M}_p} I_m \tau \mathbb{E}[\mathbf{Rad}_m(p)] + O((\log(T))^3) \quad (\text{G.8d})$$

$$= 2\gamma + \sum_{p \in \mathcal{P}} O(\log(T) \sqrt{T_p}) + O((\log(T))^3) \quad (\text{G.8e})$$

$$\leq O(\log(T) \sqrt{PT}) + O((\log(T))^3) \quad (\text{G.8f})$$

$$= O(\log(T) \sqrt{PT}), \quad (\text{G.8g})$$

where (G.8e) is by Lemma F.1 (see Appendix F) and (G.8f) is by Jensen's inequality. By plugging this result back in, we have the cumulative relaxed Bayesian regret by the end of the learning phase as

$$\overline{\text{BR}}_{\text{Learning}}(\pi_{\text{BTS}}, \varepsilon_T, T | \bar{\mathcal{R}}) \quad (\text{G.9a})$$

$$= \sum_{m=1}^M \overline{\text{BR}}_m \quad (\text{G.9b})$$

$$\leq \sum_{m=1}^M \mathbb{E}_{(\lambda, \mu) \sim \text{prior}} [\mathbb{E}[r(\lambda_{p_L^*}, \mu_{p_L^*}, p_L^*) I_m \tau - J_m(p_m) | \mathcal{F}_m]] + \sum_{m=1}^M I_m \tau \cdot O\left(\frac{4}{\sqrt{T}}\right) \quad (\text{G.9c})$$

$$= O(\log(T) \sqrt{PT}) + O(\sqrt{T}) \quad (\text{G.9d})$$

$$= O(\log(T) \sqrt{PT}). \quad (\text{G.9e})$$

Regret of the BTS Algorithm. By adding up the $(1 - \varepsilon_T)$ relaxed Bayesian regret during the warm-up and learning phases, we have the regret of BTS algorithm as

$$\overline{\text{BR}}(\pi_{\text{BTS}}, \varepsilon_T, T) \quad (\text{G.10a})$$

$$= \overline{\text{BR}}_{\text{Warm-up}}(\pi_{\text{BTS}}, \varepsilon_T, T) + \overline{\text{BR}}_{\text{Learning}}(\pi_{\text{BTS}}, \varepsilon_T, T | \bar{\mathcal{R}}) + O(P\sqrt{T}) \quad (\text{G.10b})$$

$$= O(P \log(T)^3) + O(\log(T) \sqrt{PT}) + O(P\sqrt{T}) \quad (\text{G.10c})$$

$$= O(\log(T) \sqrt{PT}) \quad (\text{G.10d})$$

$$= \tilde{O}(\sqrt{PT}). \quad (\text{G.10e})$$

Q.E.D.

Appendix H. Proofs of Section 5.2

Proof of Proposition 6. The proof is similar to the proof of Proposition 2. We present the intermediate results in Lemmas H.1 to H.3.

We have

$$\begin{aligned} &\mathbb{P}\left(\left|\left(\frac{1}{\mu_p} \left(\lambda_p - (v_p + (1 - \eta_p) \lambda_p) \ell_p^{br}\right)\right) - \left(\bar{g}_m(p) \left(\check{d}_m(p) - \left(\check{h}_m(p) + \frac{n_m^b(p)}{n_m^c(p)} \check{d}_m(p)\right) \check{\ell}_p^{br}\right)\right)\right|\right) \\ &= \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{v_p}{\mu_p} \ell_p^{br} - (1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \check{d}_m(p) \bar{g}_m(p) + \check{h}_m(p) \bar{g}_m(p) \check{\ell}_p^{br} + \frac{n_m^b(p)}{n_m^c(p)} \check{\ell}_p^{br} \check{d}_m(p) \bar{g}_m(p)\right|\right) \leq \Gamma. \end{aligned} \quad (\text{H.1})$$

By plugging in the value of

$$\Gamma = \frac{4(y_{\max} + r_{\max})}{T^3} + (8r_{\max} + 4y_{\max} + 9\check{d}_m(p)\bar{g}_m(p) + 4\check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p), n_m^r(p)\}'}}$$

we have

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \frac{v_p}{\mu_p} \ell_p^{br} - (1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \check{d}_m(p)\bar{g}_m(p) + \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br} + \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p)\bar{g}_m(p)\right| \leq \Gamma\right) \\ & \geq \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \check{d}_m(p)\bar{g}_m(p)\right| \leq 4(r_{\max} + \check{d}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{n_m^s(p)}}\right) \\ & \quad \cdot \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br}\right| \leq \frac{4y_{\max}}{T^3} + 4(y_{\max} + \check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}'}}\right) \\ & \quad \cdot \mathbb{P}\left(\left|(1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p)\bar{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + (4r_{\max} + 5\check{d}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p)\}'}}\right). \end{aligned} \quad (\text{H.2})$$

By Proposition 2, we have

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{\lambda_p}{\mu_p} - \check{d}_m(p)\bar{g}_m(p)\right| \leq 4(r_{\max} + \check{d}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{n_m^s(p)}}\right) \\ & \leq 1 - \frac{4}{T^4}. \end{aligned}$$

By Lemma H.1, we have

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br}\right| \leq \frac{4y_{\max}}{T^3} + 4(y_{\max} + \check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}'}}\right) \\ & \geq 1 - \frac{3}{T^3}. \end{aligned}$$

By Lemma H.3, we can further derive that

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{1}{c\mu_p} (\lambda_p - (v_p + (1 - \eta_p)\lambda_p) \ell_p^{br}) - \frac{\bar{g}_m(p)}{c} (\check{d}_m(p) - (\check{h}_m(p) + \frac{n_m^b(p)}{n_m^c(p)} \check{d}_m(p)) \hat{\ell}_p^{br})\right| \leq \Gamma\right) \\ & \geq \left(1 - \frac{4}{T^4}\right) \left(1 - \frac{3}{T^3}\right) \left(1 - \frac{4}{T^3}\right) \\ & \geq 1 - \frac{8}{T^3}. \end{aligned} \quad (\text{H.3})$$

Q.E.D.

Lemma H.1. If the number of the observed service time intervals $n_m^s(p) \geq 8\log(T)$ and the number of observed reneging time intervals $n_m^r(p) \geq 8\log(T)$, then one has

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br}\right| \leq \frac{4y_{\max}}{T^3} + 4(y_{\max} + \check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}'}}\right) \\ & \geq 1 - \frac{3}{T^3}, \end{aligned}$$

where $y_{\max} \geq \max_{p \in \mathcal{P}} v_p / \mu_p$.

Proof of Lemma H.1. First, we have

$$\begin{aligned} & \left|\frac{v_p}{\mu_p} \ell_p^{br} - \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br}\right| \\ & = \left|\frac{v_p}{\mu_p} \ell_p^{br} - \frac{v_p}{\mu_p} \hat{\ell}_p^{br} + \frac{v_p}{\mu_p} \hat{\ell}_p^{br} - \check{h}_m(p)\bar{g}_m(p) \hat{\ell}_p^{br}\right| \\ & \leq \frac{v_p}{\mu_p} \left|\ell_p^{br} - \hat{\ell}_p^{br}\right| + \hat{\ell}_p^{br} \left|\frac{v_p}{\mu_p} - \check{h}_m(p)\bar{g}_m(p)\right|. \end{aligned} \quad (\text{H.4})$$

Define event $\mathcal{G}$ as the queue being in steady state after the first 2τ time periods, and define the complementary event as $\bar{\mathcal{G}}$. Plug in $y_{\max} \geq \max_{p \in \mathcal{P}} v_p / \mu_p$. By Proposition 4 and Lemma I.2, we have

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4y_{\max}}{T^3}\right) \\ & \geq \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4v_p}{\mu_p T^3}\right) \\ & = \mathbb{P}\left(\left|\ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4}{T^3} \mid \mathcal{G}\right) \cdot \mathbb{P}(\mathcal{G}) + \mathbb{P}\left(\left|\ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4}{T^3} \mid \bar{\mathcal{G}}\right) \cdot \mathbb{P}(\bar{\mathcal{G}}) \\ & \geq \mathbb{P}\left(\left|\ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4}{T^3} \mid \mathcal{G}\right) \cdot \mathbb{P}(\mathcal{G}) \\ & \geq 1 - \frac{2}{T^3}. \end{aligned} \quad (\text{H.5})$$

We also have

$$\mathbb{P}\left(\left|\hat{\ell}_p^{br} \frac{v_p}{\mu_p} - \check{h}_m(p)\bar{g}_m(p)\right| \leq 4(y_{\max} + \check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}'}}\right) \quad (\text{H.6a})$$

$$\geq \mathbb{P}\left(\left|\frac{v_p}{\mu_p} - \check{h}_m(p)\bar{g}_m(p)\right| \leq 4(y_{\max} + \check{h}_m(p)\bar{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}'}}\right) \quad (\text{H.6b})$$

$$\geq 1 - \frac{4}{T^4}, \quad (\text{H.6c})$$

where the inequality (H.6b) is by $0 \leq \hat{\ell}_p^{br} \leq 1$, and the inequality (H.6c) is by Proposition 2.

To conclude, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \tilde{h}_m(p) \tilde{g}_m(p) \ell_p^{br}\right| \leq \frac{4y_{\max}}{T^3} + 4(y_{\max} + \tilde{h}_m(p) \tilde{g}_m(p))\right. \\
& \quad \left.\sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}}}\right) \\
&= \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \hat{\ell}_p^{br}\right| + \left|\hat{\ell}_p^{br} - \tilde{h}_m(p) \tilde{g}_m(p)\right| \leq \frac{4y_{\max}}{T^3}\right) \\
&\leq \frac{4y_{\max}}{T^3} + 4(y_{\max} + \tilde{h}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}}} \\
&\geq \mathbb{P}\left(\left|\frac{v_p}{\mu_p} \ell_p^{br} - \hat{\ell}_p^{br}\right| \leq \frac{4y_{\max}}{T^3}\right) \\
&\quad \cdot \mathbb{P}\left(\left|\hat{\ell}_p^{br} - \tilde{h}_m(p) \tilde{g}_m(p)\right| \leq 4(y_{\max} + \tilde{h}_m(p) \tilde{g}_m(p))\right) \\
&\quad \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^r(p)\}}} \\
&\geq \left(1 - \frac{2}{T^3}\right) \left(1 - \frac{4}{T^4}\right) \\
&\geq 1 - \frac{3}{T^3}. \quad \text{Q.E.D.}
\end{aligned}$$

Lemma H.2 (Concentration Inequality of Balking Probability).

$$\mathbb{P}\left(\left|\frac{n_m^b(p)}{n_m^c(p)} - (1 - \eta_p)\right| \leq \sqrt{\frac{2\log(T)}{n_m^c(p)}}\right) \geq 1 - \frac{2}{T^4}.$$

Proof of Lemma H.2. Define $X_1, X_2, \dots, X_{n_m^c}$ as random variables such that $X_i = 1$ if the customer balks and $X_i = 0$ otherwise. Therefore, by Hoeffding's inequality, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n_m^c(p)} X_i - \mathbb{E}\left[\sum_{i=1}^{n_m^c(p)} X_i\right]\right| \leq a\right) \geq 1 - 2\exp\left(-\frac{2a^2}{n_m^c(p)}\right).$$

By letting $a = \sqrt{2\log(T)n_m^c(p)}$, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|\frac{\sum_{i=1}^{n_m^c(p)} X_i}{n_m^c(p)} - \frac{\mathbb{E}[\sum_{i=1}^{n_m^c(p)} X_i]}{n_m^c(p)}\right| \leq \frac{a}{n_m^c(p)}\right) \geq 1 - 2\exp\left(-\frac{2a^2}{n_m^c(p)}\right) \\
&\Rightarrow \mathbb{P}\left(\left|\frac{n_m^b(p)}{n_m^c(p)} - (1 - \eta_p)\right| \leq \frac{a}{n_m^c(p)}\right) \geq 1 - 2\exp\left(-\frac{2a^2}{n_m^c(p)}\right) \\
&\Rightarrow \mathbb{P}\left(\left|\frac{n_m^b(p)}{n_m^c(p)} - (1 - \eta_p)\right| \leq \sqrt{\frac{2\log(T)}{n_m^c(p)}}\right) \geq 1 - \frac{2}{T^4}. \quad \text{Q.E.D.}
\end{aligned}$$

Lemma H.3. If $n_m^s(p) \geq 8\log(T)$ and $n_m^c(p) \geq 8\log(T)$, then one has

$$\begin{aligned}
& \mathbb{P}\left(\left|(1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + (4r_{\max} + 5\check{d}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p)\}}}\right) \\
&\geq 1 - \frac{4}{T^3}.
\end{aligned}$$

Proof of Lemma H.3. First, we can derive that

$$\begin{aligned}
& \left|(1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \\
&= \left|(1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - (1 - \eta_p) \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right. \\
&\quad \left.+ (1 - \eta_p) \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p) - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \\
&\leq (1 - \eta_p) \\
&\quad \cdot \left|\ell_p^{br} \frac{\lambda_p}{\mu_p} - \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| + \left|\hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p) - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \\
&\leq \left|\ell_p^{br} \frac{\lambda_p}{\mu_p} - \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| + \left|\check{d}_m(p) \tilde{g}_m(p) - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right|. \quad (\text{H.7})
\end{aligned}$$

Similar to Lemma H.1, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|\ell_p^{br} \frac{\lambda_p}{\mu_p} - \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + 4(r_{\max} + \check{d}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{n_m^s(p)}}\right) \geq 1 - \frac{3}{T^3}. \quad (\text{H.8})
\end{aligned}$$

By Lemma H.2, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|\check{d}_m(p) \tilde{g}_m(p) - \left|(1 - \eta_p) - \frac{n_m^b(p)}{n_m^c(p)}\right| \leq \check{d}_m(p) \tilde{g}_m(p) \sqrt{\frac{2\log(T)}{n_m^c(p)}}\right) \\
&= \mathbb{P}\left(\left|(1 - \eta_p) - \frac{n_m^b(p)}{n_m^c(p)}\right| \leq \sqrt{\frac{2\log(T)}{n_m^c(p)}}\right) \\
&\geq 1 - \frac{2}{T^4}. \quad (\text{H.9})
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|(1 - \eta_p) \ell_p^{br} \frac{\lambda_p}{\mu_p} - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + (4r_{\max} + 5\check{d}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p)\}}}\right) \\
&= \mathbb{P}\left(\left|\ell_p^{br} \frac{\lambda_p}{\mu_p} - \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| + \left|\check{d}_m(p) \tilde{g}_m(p) - \frac{n_m^b(p)}{n_m^c(p)} \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + (4r_{\max} + 5\check{d}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{\min\{n_m^s(p), n_m^c(p)\}}}\right) \\
&\geq \mathbb{P}\left(\left|\ell_p^{br} \frac{\lambda_p}{\mu_p} - \hat{\ell}_p^{br} \check{d}_m(p) \tilde{g}_m(p)\right| \leq \frac{4r_{\max}}{T^3} + 4(r_{\max} + \check{d}_m(p) \tilde{g}_m(p)) \sqrt{\frac{2\log(T)}{n_m^s(p)}}\right. \\
&\quad \left.+ \mathbb{P}\left(\left|\check{d}_m(p) \tilde{g}_m(p) - \left|(1 - \eta_p) - \frac{n_m^b(p)}{n_m^c(p)}\right| \leq \check{d}_m(p) \tilde{g}_m(p) \sqrt{\frac{2\log(T)}{n_m^c(p)}}\right)\right) \\
&\geq \left(1 - \frac{3}{T^3}\right) \left(1 - \frac{2}{T^4}\right) \\
&\geq 1 - \frac{4}{T^3}. \quad \text{Q.E.D.}
\end{aligned}$$

Appendix I. Proofs of Section 5.3

Proposition 1.1 (Coupling Probability of $M/M/c$ Queue with Balking and Reneging Starting from the Empty State). *Assuming that Assumption 4 holds, the probability that the system with balking and reneging reaches the steady state within τ starting from an empty state is bounded from below by $1 - \frac{2}{T^2}$, where $\tau = (\log(T))^2$.*

Proof of Proposition 1.1. For any $t \geq \tau$, one has

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

where the equality holds because once $\hat{S}_\tau = S_\tau$, the two processes couple. Define $\bar{S} = \left\lfloor \frac{\log(T)\sqrt{\log(T)}}{-\log \rho} \right\rfloor$. By Lemma 1.1, 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 $\mathbb{P}(\sum_{n=1}^{\bar{S}} \hat{S}_n \leq \tau)$. Therefore, we have

$$\mathbb{P}(\hat{S}_\tau = S_\tau) \quad (I.1a)$$

$$\geq \mathbb{P}(\hat{S}_\tau = S_\tau | \hat{S}_0 \leq \bar{S}) \mathbb{P}(\hat{S}_0 \leq \bar{S}) \quad (I.1b)$$

$$= \mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_n \leq \tau\right) \mathbb{P}(\hat{S}_0 \leq \bar{S}) \quad (I.1c)$$

$$\geq \mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_1 \leq \tau\right) \mathbb{P}(\hat{S}_0 \leq \bar{S}). \quad (I.1d)$$

By Lemmas 1.5 and 1.4, (I.1d) can be bounded by

$$\mathbb{P}\left(\sum_{n=1}^{\bar{S}} A_1 \leq \tau\right) \mathbb{P}(S_0 \leq \bar{S}) \quad (I.2a)$$

$$\geq \mathbb{P}\left(\sum_{n=1}^{\bar{S}} \tilde{A}_1 \leq \tau\right) \mathbb{P}(\tilde{N} \leq \bar{S}). \quad (I.2b)$$

In the proof of Proposition 4, we have shown that (I.2b) is bounded from below by $1 - \frac{2}{T^2}$. Q.E.D.

Proof of Proposition 7. The proof is similar to the proof of Proposition 5 and we omit the details here. $\square$

Proof of Theorem 3. Similar to Theorem 1, we bound the relaxed regret during the warm-up and learning phases separately. We can also decompose the loss during the learning phase into four parts: loss due to instability, non-stationarity, pruning error, and suboptimality.

Regret of the Warm-Up Phase. The analysis here is the same as that in (F.1). Therefore, we have

$$\overline{\text{Regret}}_{\text{Warm-up}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, T) = O(\text{Plog}(T)^3). \quad (I.3)$$

Loss of Instability During the Learning Phase. For Extended BUCB, Proposition 3 is also valid, and thus the analysis here is the same as for the loss of instability of BUCB in (F.3).

Therefore, the regret in the learning phase can be expressed as

$$\begin{aligned} & \overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T) \\ & \leq \overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T | \tilde{\mathcal{R}}) + O(P\sqrt{T}). \end{aligned} \quad (I.4)$$

Below we analyze $\overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T | \tilde{\mathcal{R}})$.

Loss of Nonstationarity During the Learning Phase. Denote the offered price of batch m in Extended BUCB by p_m . The analysis here is similar to that for the loss of nonstationarity of BUCB in (F.4). For this extended system, we can apply the mixing time result in Proposition 7. Therefore, we have

$$\begin{aligned} \mathbb{E}[J_m(p_m)] &= \mathbb{E}[J_{m,0-2\tau}(p_m) + J_{m,\mathcal{C}_{2\tau}}(p_m) | \mathbb{P}(\mathcal{C}_{2\tau}) \\ & \quad + J_{m,\bar{\mathcal{C}}_{2\tau}}(p_m)(1 - \mathbb{P}(\mathcal{C}_{2\tau}))] \\ &\geq O(\tau) + \left(\frac{p_m}{\mu_{p_m}} (\lambda_{p_m} - (v_{p_m} + (1 - \eta_{p_m})\lambda_{p_m})\ell_{p_m}^{\text{br}}) \right) \\ & \quad (I_m\tau - 2\tau) \left(1 - \frac{4}{T^2} \right) + O\left(\frac{I_m\tau - 2\tau}{T^2} \right) \\ &= \left(\frac{p_m}{\mu_{p_m}} (\lambda_{p_m} - (v_{p_m} + (1 - \eta_{p_m})\lambda_{p_m})\ell_{p_m}^{\text{br}}) \right) I_m\tau - O(\tau) \\ &= r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m) I_m\tau - O(\tau). \end{aligned}$$

Loss of Pruning Error During the Learning Phase. The analysis here is similar to the analysis on the loss of instability of BUCB. For this extended system, we define

$$\Delta_{\tilde{p}}(p_m) = r^E(\lambda_{\tilde{p}}, \mu_{\tilde{p}}, v_{\tilde{p}}, \eta_{\tilde{p}}, \tilde{p}) - r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m).$$

Let $p_L^* = \arg \max_{p \in \mathcal{P}_L} r^E(\lambda_p, \mu_p, v_p, \eta_p, p)$ be the learning upper-bound price, and correspondingly define

$$\Delta_L(p_m) = r^E(\lambda_{p_L^*}, \mu_{p_L^*}, v_{p_L^*}, \eta_{p_L^*}, p_L^*) - r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m).$$

Thus, following Proposition 3, we have

$$\mathbb{P}(\tilde{p} \in \mathcal{P}_L) \geq 1 - \frac{4}{\sqrt{T}} \text{ and } \mathbb{P}(\Delta_{\tilde{p}}(p_m) \leq \Delta_L(p_m)) \geq 1 - \frac{4}{\sqrt{T}}.$$

Loss of Suboptimality During the Learning Phase. The new UCB defined in Definition 6 indicates same properties of the new $\text{Rad}_m^E(p)$ of Extended BUCB and $\text{Rad}_m(p)$ in BUCB. Therefore, we can follow the same analysis structure as that of BUCB. We further derive that

$$\Delta_L(p_m) = r^E(\lambda_{p_L^*}, \mu_{p_L^*}, v_{p_L^*}, \eta_{p_L^*}, p_L^*) - r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m) \quad (I.5a)$$

$$\begin{aligned} &= \left(r^E(\lambda_{p_L^*}, \mu_{p_L^*}, v_{p_L^*}, \eta_{p_L^*}, p_L^*) - U_m^E(p) \right) + (U_m^E(p_m) - L_m^E(p_m)) \\ & \quad + (L_m^E(p_m) - r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m)) \end{aligned} \quad (I.5b)$$

$$\begin{aligned} &\leq \left(r^E(\lambda_{p_L^*}, \mu_{p_L^*}, v_{p_L^*}, \eta_{p_L^*}, p_L^*) - U_m^E(p_L^*) \right) + (U_m^E(p_m) - L_m^E(p_m)) \\ & \quad + (L_m^E(p_m) - r^E(\lambda_{p_m}, \mu_{p_m}, v_{p_m}, \eta_{p_m}, p_m)). \end{aligned} \quad (I.5c)$$

Then, the cumulative regret during the learning phase is (by incorporating the analysis on the loss of instability,

nonstationarity, and pruning error)

$$\overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T) \quad (\text{I.6a})$$

$$\leq \overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T|\bar{\mathcal{R}}) + O(P\sqrt{T}) \quad (\text{I.6b})$$

$$= \sum_{m=1}^M \mathbb{E}[\Delta_{\bar{p}}(p_m)I_m\tau + O(\tau)] + O(P\sqrt{T}) \quad (\text{I.6c})$$

$$\leq \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau + O(\tau)] + O(\sqrt{T}) + O(P\sqrt{T}) \quad (\text{I.6d})$$

$$= \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau] + \sum_{m=1}^M O(\tau) + O(P\sqrt{T}). \quad (\text{I.6e})$$

Then we can plug $\Delta_L(p_m)$ in (I.5c) into (I.6e) and analyze the summation of these three terms separately. Following exactly the same steps in (F.8) and (F.9), we also achieve same results for the extended BUCB:

$$\begin{aligned} & \sum_{m=1}^M \mathbb{E}[\Delta_L(p_m)I_m\tau] \\ &= \sum_{m=1}^M \mathbb{E}[(r^E(\lambda_{p_m^*}, \mu_{p_m^*}, \nu_{p_m^*}, \eta_{p_m^*}, p_m^*) - U_m^E(p_m^*))I_m\tau] \\ &+ \sum_{m=1}^M \mathbb{E}[(U_m^E(p_m) - L_m^E(p_m))I_m\tau] \\ &+ \sum_{m=1}^M \mathbb{E}[(L_m^E(p_m) - r(\lambda_{p_m}, \mu_{p_m}, \nu_{p_m}, \eta_{p_m}, p_m))I_m\tau] \\ &= O(\log(T)\sqrt{PT}) + O\left(\frac{1}{T^3}\right) \\ &= O(\log(T)\sqrt{PT}). \end{aligned} \quad (\text{I.7})$$

Regret of the Learning Phase. By substituting (I.7) back into (I.6), we reach the regret bound of the learning phase:

$$\begin{aligned} \overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, \varepsilon_T, T) &= O(\log(T)\sqrt{PT}) + \sum_{m=1}^M O(\tau) \\ &+ O(P\sqrt{T}) \\ &= O(\log(T)\sqrt{PT}). \end{aligned}$$

Total Regret of the Extended BUCB Algorithm. By adding up the relaxed regret during the warm-up and learning phases, we have the regret of the extend BUCB algorithm as

$$\begin{aligned} \overline{\text{Regret}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, T) &= \overline{\text{Regret}}_{\text{Warm-up}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, T) \\ &+ \overline{\text{Regret}}_{\text{Learning}}^{\text{EXT}}(\pi_{\text{EXT-BUCB}}, T) \end{aligned} \quad (\text{I.8a})$$

$$= O(P \log(T)^3) + O(\log(T)\sqrt{PT}) \quad (\text{I.8b})$$

$$= \tilde{O}(\sqrt{PT}). \quad (\text{I.8c})$$

Q.E.D.

Lemma I.1 (First-Order Stochastic Dominance Between A_i and A_j in an $M/M/c$ Queue with Balking and Reneging). *If Assumption 4 holds, then for random variables A_i and A_j , where $i < j$, A_i first-order dominates A_j , that is, $\mathbb{P}(A_j \leq x) \geq \mathbb{P}(A_i \leq x)$.*

Proof of Lemma I.1. Based on the proof of Lemma E.1, we skip the step for constructing an auxiliary queue and directly analyze the service and arrival rates of each interval. Please note that balking and reneging can be included as part of both the arrival and departure rates. In the following proofs, we include them in the arrival rate.

Consider an $M/M/c$ queue Q with balking and reneging, where parts 1 and 2 of Assumption 4 hold. Similarly, we can analyze the intervals included in A_i and A_j :

1. All intervals in A_i have the same or a smaller service rate compared with the intervals that sit in the same position in the recursion illustration of A_j , because the system has c servers, and thus the former always has fewer or the same amount of customers under service.

2. All intervals in A_i have the same or a larger average arrival rate compared with the intervals that sit in the same position in the recursion illustration of A_j , according to Assumption 4, parts 1 and 2.

Therefore, $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$, and A_i first-order dominates A_j . Q.E.D.

Lemma I.2 (Total Variation of the $M/M/c$ Queue with Balking and Reneging). *For $t \geq \tau$, where $\tau = \log(T)(\log \log(T))^2$, the total variation of S_t and S_∞ is bounded as*

$$\sup \{|\mathbb{P}(S_t \in \Xi) - \mathbb{P}(S_\infty \in \Xi)|, \text{measurable set } \Xi \subset [0, \infty)\} \leq \frac{4}{T^2}.$$

Proof of Lemma I.2. Based on Proposition I.1, we have, for any $t \geq \tau$ and any measurable set $\Xi \subset [0, \infty)$,

$$\begin{aligned} |\mathbb{P}(S_t \in \Xi) - \mathbb{P}(S_\infty \in \Xi)| &= |\mathbb{P}(\hat{S}_t \in \Xi) - \mathbb{P}(S_t \in \Xi)| \leq \mathbb{P}(\hat{S}_t = S_t) \\ &\times 0 + \mathbb{P}(\hat{S}_t \neq S_t) \times 2 \leq \frac{4}{T^2}. \end{aligned} \quad \text{Q.E.D.}$$

Lemma I.3 (First-Order Stochastic Dominance Between A_i and A_j in an $M/M/c$ Queue with Balking). *If Assumption 4, part 1, holds, then for random variables A_i and A_j , where $i < j$, A_i first-order dominates A_j . Similar to Lemma E.1, one has $\mathbb{P}(A_j \leq x) \geq \mathbb{P}(A_i \leq x)$ and $\mathbb{E}_{A_j}[u(x)] \leq \mathbb{E}_{A_i}[u(x)]$ for all strictly increasing, continuous utility functions $u(x) : [0, \infty) \rightarrow \mathbb{R}$.*

Proof of Lemma I.3. Consider an $M/M/c$ queue Q with balking where part 1 of Assumption 4 holds. The following analysis is also based on the arguments in lemma 1 in Omahen and Marathe (1978), and this recursion relationship is illustrated in Figure E.1. We can analyze the intervals included in A_i and A_j :

1. All intervals in A_i have the same or a smaller service rate compared with the intervals that sit in the same position in the recursion illustration of A_j , because the system has c servers, and thus the former always has fewer or the same amount of customers under service.

2. All intervals in A_i have the same or a larger average arrival rate compared with the intervals that sit in the same position in the recursion illustration of A_j , according to Assumption 4, part 1.

Therefore, $\mathbb{P}(A_j \leq x) > \mathbb{P}(A_i \leq x)$, and A_i first-order dominates A_j . Q.E.D.

We further establish the first-order stochastic dominance between the standard $M/M/c$ queue and the $M/M/c$ queue with balking and reneging under Assumption 4, parts 1 and 2. Consider the standard $M/M/c$ queue and the $M/M/c$ queue with balking and reneging under the same customer arrival and service rates. Denote the number of customers of $M/M/c$ queue with balking and reneging in steady state by N , and denote that of standard $M/M/c$ queue by $\tilde{N}$. Denote the busy period of the standard $M/M/c$ queue by $\tilde{A}_i$.

Lemma I.4 (First-Order Stochastic Dominance Between $\tilde{A}_i$ and A_i). *We have that $\tilde{A}_i$ first-order dominates A_i . That is, $\mathbb{P}(A_i \leq x) \geq \mathbb{P}(\tilde{A}_i \leq x)$ and $\mathbb{E}_{A_i}[u(n)] \leq \mathbb{E}_{\tilde{A}_i}[u(x)]$ for all strictly increasing, continuous utility functions $u(x) : [0, \infty] \rightarrow \mathbb{R}$.*

Proof of Lemma I.4. We can analyze the intervals included in A_i and $\tilde{A}_i$ similarly:

1. All intervals in A_i have the same service rate compared with those of $\tilde{A}_i$.
2. All intervals in A_i have the same or a smaller average arrival rate compared with those of $\tilde{A}_i$.

Therefore, $\mathbb{P}(A_i \leq x) \geq \mathbb{P}(\tilde{A}_i \leq x)$, and $\tilde{A}_i$ first-order dominates A_i . Q.E.D.

Lemma I.5 (First-Order Stochastic Dominance Between $\tilde{N}$ and N). *We have that $\tilde{N}$ first-order dominates N . That is, $\mathbb{P}(N \leq n) \geq \mathbb{P}(\tilde{N} \leq n)$ and $\mathbb{E}_N[u(n)] \leq \mathbb{E}_{\tilde{N}}[u(x)]$ for all strictly increasing, continuous utility functions $u(n) : [0, 1, 2, \dots] \rightarrow \mathbb{R}$.*

Proof of Lemma I.5. The proof here directly originates from the overall arrival and service rates of these two queues. The queue with balking and reneging always has a smaller arrival rate, and thus $\mathbb{P}(N \leq n) \geq \mathbb{P}(\tilde{N} \leq n)$. From another perspective, to avoid the effects of the number of customers in the system on the service rate, we can view balking and reneging as parts of the departure rate. Thus, the queue with balking and reneging has the same arrival rate but a larger departure rate, and furthermore, one has $\mathbb{P}(N \leq n) \geq \mathbb{P}(\tilde{N} \leq n)$. Q.E.D.

References

- Abate J, Whitt W (1987) Transient behavior of the $M/M/1$ queue: Starting at the origin. *Queueing Systems* 2(1):41–65.
- Abate J, Whitt W (1988) Transient behavior of the $M/M/1$ queue via Laplace transforms. *Adv. Appl. Probab.* 20(1):145–178.
- Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K, eds. *Advances in Neural Information Processing Systems*, vol. 24 (Curran Associates Inc., Red Hook, NY), 2312–2320.
- Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. Carvalho CM, Ravikumar P, eds. *Proc. 16th Internat. Conf. Artificial Intelligence Statist.* Proceedings of Machine Learning Research, vol. 31 (Microtome Publishing, Brookline, MA), 99–107.
- Al-Seedy RO, El-Sherbiny A, El-Shehawly S, Ammar S (2009) Transient solution of the $M/M/c$ queue with balking and reneging. *Comput. Math. Appl.* 57(8):1280–1285.
- Ammar SI, Helan MM, Al Amri FT (2013) The busy period of an $M/M/1$ queue with balking and reneging. *Appl. Math. Model.* 37(22):9223–9229.
- Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5):1169–1188.
- Auer P, Ortner R (2010) UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* 61(1–2):55–65.
- Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2):235–256.
- Badanidiyuru A, Kleinberg R, Slivkins A (2018) Bandits with knapsacks. *J. ACM* 65(3):1–55.
- Bailey NT (1954) A continuous time treatment of a simple queue using generating functions. *J. Royal Statist. Soc. B* 16(2):288–291.
- Banerjee S, Johari R, Riquelme C (2015) Pricing in ride-sharing platforms: A queueing-theoretic approach. Roughgarden T, ed. *Proc. 16th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 639–641.
- Bartlett PL, Tewari A (2009) Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. McAllester D, ed. *Proc. 25th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA), 35–42.
- Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.
- Besbes O, Zeevi A (2012) Blind network revenue management. *Oper. Res.* 60(6):1537–1550.
- Besbes O, Elmachtoub AN, Sun Y (2022) Static pricing: Universal guarantees for reusable resources. *Oper. Res.* 70(2):1143–1152.
- Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.
- Bimpikis K, Candogan O, Saban D (2019) Spatial pricing in ride-sharing networks. *Oper. Res.* 67(3):744–769.
- Borgs C, Candogan O, Chayes J, Lobel I, Nazerzadeh H (2014) Optimal multiperiod pricing with service guarantees. *Management Sci.* 60(7):1792–1811.
- Boucheron S, Lugosi G, Massart P (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, UK).
- Bourel H, Maillard OA, Talebi MS (2020) Tightening exploration in upper confidence reinforcement learning. *Proc. 37th Internat. Conf. Machine Learn.* Proceedings of Machine Learning Research, vol. 119 (Microtome Publishing, Brookline, MA), 1056–1066.
- Bubeck S, Cesa-Bianchi N, Lugosi G (2013) Bandits with heavy tail. *IEEE Trans. Inform. Theory* 59(11):7711–7717.
- Cesa-Bianchi N, Dekel O, Shamir O (2013) Online learning with switching costs and other adaptive adversaries. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates Inc., Red Hook, NY), 1160–1168.
- Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 24 (Curran Associates Inc., Red Hook, NY), 2249–2257.
- Chen B, Chao X (2019) Parametric demand learning with limited price explorations in a backlog stochastic inventory system. *IIE Trans.* 51(6):605–613.
- Chen Y, Shi C (2017) Optimal pricing policy for service systems with reusable resources and forward-looking customers. Technical report, University of Michigan, Ann Arbor, MI.
- Chen B, Shi C (2019) Tailored base-surge policies in dual-sourcing inventory systems with demand learning. Technical report, University of Michigan, Ann Arbor, MI.
- Chen Y, Shi C (2022) Network revenue management with online inverse batch gradient descent method. Technical report, University of Michigan, Ann Arbor, MI.
- Chen B, Chao X, Wang Y (2020a) Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Oper. Res.* 68(5):1445–1456.
- Chen Y, Levi R, Shi C (2017) Revenue management of reusable resources with advanced reservations. *Production Oper. Management* 26(5):836–859.

- Chen X, Liu Y, Hong G (2020b) An online learning approach to dynamic pricing and capacity sizing in service systems. Technical report, Chinese University of Hong Kong, Shenzhen, China.
- Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Oper. Res.* 65(6):1722–1731.
- Dai JG, Gluzman M (2022) Queueing network controls via deep reinforcement learning. *Stochastic Systems* 12(1):30–67.
- Daley DJ, Servi L (1998) Idle and busy periods in stable M/M/k queues. *J. Appl. Probab.* 35(4):950–962.
- den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys Oper. Res. Management Sci.* 20(1):1–18.
- Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. *Ann. Statist.* 7(2):269–281.
- Doan XV, Lei X, Shen S (2020) Pricing of reusable resources under ambiguous distributions of demand and service time with emerging applications. *Eur. J. Oper. Res.* 282(1):235–251.
- Dubey A, Pentland AS (2019) Thompson sampling on symmetric alpha-stable bandits. Kraus S, ed. *Proc. 28th Internat. Joint Conf. Artificial Intelligence* (International Joint Conferences on Artificial Intelligence Organization, The Cloud), 5715–5721.
- Dubey A, Pentland A (2020) Cooperative multi-agent bandits with heavy tails. Daumé III H, Singh A, eds. *Proc. 37th Internat. Conf. Machine Learn.* Proceedings of Machine Learning Research, vol. 119 (Microtome Publishing, Brookline, MA), 2730–2739.
- Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using Thompson sampling. *Oper. Res.* 66(6):1586–1602.
- Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 23 (Curran Associates Inc., Red Hook, NY), 586–594.
- Fruit R, Pirotta M, Lazaric A (2020) Improved analysis of UCRL2 with empirical Bernstein inequality. Preprint, submitted July 10, <https://arxiv.org/abs/2007.05456>.
- Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* 40(8):999–1020.
- Gans N, Savin S (2007) Pricing and capacity rationing for rentals with uncertain durations. *Management Sci.* 53(3):390–407.
- Gao Z, Han Y, Ren Z, Zhou Z (2019) Batched multi-armed bandits problem. Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates Inc., Red Hook, NY), 503–513.
- Garivier A, Moulines E (2011) On upper-confidence bound policies for switching bandit problems. Dasgupta S, Haghtalab N, eds., *Proc. 22nd Internat. Conf. Algorithmic Learn. Theory* (Springer-Verlag, Berlin), 174–188.
- Hadar J, Russell WR (1969) Rules for ordering uncertain prospects. *Amer. Econom. Rev.* 59(1):25–34.
- Han Y, Zhou Z, Zhou Z, Blanchet J, Glynn PW, Ye Y (2020) Sequential batch learning in finite-action linear contextual bandits. Preprint, submitted April 14, <https://arxiv.org/abs/2004.06321>.
- Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *J. Machine Learn. Res.* 11(51):1563–1600.
- Jia H, Shi C, Shen S (2021) Multi-armed bandit with sub-exponential rewards. *Oper. Res. Lett.* 49(5):728–733.
- Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, eds. *Proc. 32nd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 4868–4878.
- Kaewpuang R, Niyato D, Wang P, Hossain E (2013) A framework for cooperative resource management in mobile cloud computing. *IEEE J. Selected Areas Comm.* 31(12):2685–2700.
- Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An asymptotically optimal finite-time analysis. Bshouty NH, Stoltz G, Vayatis N, Zeugmann T, eds. *Internat. Conf. Algorithmic Learn. Theory*. Lecture Notes in Computer Science, vol. 7568 (Springer, Berlin), 199–213.
- Kelton WD, Law AM (1985) The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Oper. Res.* 33(2):378–396.
- Korda N, Kaufmann E, Munos R (2013) Thompson sampling for 1-dimensional exponential family bandits. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates Inc., Red Hook, NY), 1448–1456.
- Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* 15(3):1091–1114.
- Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.
- Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press, Cambridge, UK).
- Ledermann W, Reuter GEH (1954) Spectral theory for the differential equations of simple birth and death processes. *Philos. Trans. Roy. Soc. Lond. A* 246(914):321–369.
- Ledoux M (2001) *The Concentration of Measure Phenomenon* (American Mathematical Society, Providence, RI).
- Lei Y, Jasin S (2020) Real-time dynamic pricing for revenue management with reusable resources, advance reservation, and deterministic service time requirements. *Oper. Res.* 68(3):676–685.
- Levi R, Radovanović A (2010) Provably near-optimal LP-based policies for revenue management in systems with reusable resources. *Oper. Res.* 58(2):503–507.
- Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. Rappa M, Jones P, eds. *Proc. 19th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 661–670.
- Liu Y, Li Y (2017) Pricing scheme design of ridesharing program in morning commute problem. *Transportation Res. Part C: Emerging Tech.* 79:156–177.
- Liu B, Xie Q, Modiano E (2019) Reinforcement learning for optimal control of queueing systems. Liberzon D, Dominguez-Garcia A, eds. *Proc. 57th Annual Allerton Conf. Comm., Control, Comput.* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 663–670.
- Maglaras C (2006) Revenue management for a multiclass single-server queue via a fluid model analysis. *Oper. Res.* 54(5):914–932.
- Minton P, Raiffa H, Schlaifer R (1962) Applied statistical decision theory. *Amer. Math. Monthly* 69(1):72–73.
- Morisaku T (1976) Techniques for data-truncation in digital computer simulation. Unpublished doctoral thesis, University of Southern California, Los Angeles.
- Oliveira BB, Carravilla MA, Oliveira JF (2018) Integrating pricing and capacity decisions in car rental: A matheuristic approach. *Oper. Res. Perspect.* 5:334–356.
- Omahen K, Marathe V (1978) Analysis and applications of the delay cycle for the M/M/c queueing system. *J. ACM* 25(2):283–303.
- Owen Z, Simchi-Levi D (2017) Price and assortment optimization for reusable resources. Preprint, submitted November 16, <http://dx.doi.org/10.2139/ssrn.3070625>.
- Özer Ö, Phillips R (2012) *The Oxford Handbook of Pricing Management* (Oxford University Press, Oxford, UK).
- Parthasarathy P, Sharafali M (1989) Transient solution to the many-server Poisson queue: A simple approach. *J. Appl. Probab.* 26(3):584–594.
- Perchet V, Rigollet P, Chassang S, Snowberg E (2016) Batched bandit problems. *Ann. Statist.* 44(2):660–681.
- Püschel T, Schryen G, Hristova D, Neumann D (2015) Revenue management for cloud computing providers: Decision models

- for service admission control under non-probabilistic uncertainty. *Eur. J. Oper. Res.* 244(2):637–647.
- Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Math. Oper. Res.* 35(2):395–411.
- Rusmevichientong P, Sumida M, Topaloglu H (2020) Dynamic assortment optimization for reusable products with random usage durations. *Management Sci.* 66(7):2820–2844.
- Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.
- Russo DJ, Van Roy BV, Kazerouni A, Osband I, Wen Z (2018) *A Tutorial on Thompson Sampling*. Foundations and Trends in Machine Learning, vol. 11 (now Publishers, Hanover, MA).
- Savin SV, Cohen MA, Gans N, Katalan Z (2005) Capacity management in rental businesses with two customer bases. *Oper. Res.* 53(4):617–631.
- Seth S, Yalowitzky G (2014) Stochastic dominance with parametric distributions. Technical report, Oxford University, Oxford, UK.
- Shah D, Xie Q, Xu Z (2020) Stable reinforcement learning with unbounded state space. Bayen AM, Jadbabaie A, Pappas G, Parrilo PA, Recht B, Tomlin C, Zeilinger M, eds. *Proc. Second Conf. Learn. Dynamics Control*. Proceedings of Machine Learning Research, vol. 120 (Microtome Publishing, Brookline, MA), 581.
- Shalev-Shwartz S (2011) *Online Learning and Online Convex Optimization*. Foundations and Trends in Machine Learning, vol. 4 (now Publishers, Hanover, MA), 107–194.
- Simchi-Levi D, Xu Y (2019) Phase transitions and cyclic phenomena in bandits with switching constraints. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Red Hook, NY), 7523–7532.
- Slivkins A (2019) *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning, vol. 12 (now Publishers, Hanover, MA).
- Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
- Szepesvári C (2010) *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning (Morgan & Claypool, San Rafael, CA).
- Talluri KT, Van Ryzin GJ (2006) *The Theory and Practice of Revenue Management* (Springer, New York).
- Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.
- Thompson WR (1935) On the theory of apportionment. *Amer. J. Math.* 57(2):450–456.
- Vakili S, Liu K, Zhao Q (2013) Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE J. Selected Topics Signal Processing* 7(5):759–767.
- Wang K, Li N, Jiang Z (2010) Queueing system with impatient customers: A review. *Proc. 2010 IEEE Internat. Conf. Service Oper. Logist. Informatics* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 82–87.
- Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.
- Ward AR, Glynn PW (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43(1–2):103–128.
- Ward AR, Glynn PW (2005) A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* 50(4):371–400.
- Xia Y, Qin T, Ma W, Yu N, Liu TY (2016) Budgeted multi-armed bandits with multiple plays. Kambhampati S, ed. *Proc. 25th Internat. Joint Conf. Artificial Intelligence* (AAAI Press, Palo Alto, CA), 2210–2216.
- Xu H, Li B (2013) Dynamic cloud pricing for revenue maximization. *IEEE Trans. Cloud Comput.* 1(2):158–171.
- Zhang Y, Yue D, Yue W (2005) Analysis of an M/M/1/N queue with balking, reneging and server vacations. Eduardo J, ed. *Proc. Fifth Internat. Sympos. OR Its Appl.* (Springer, New York), 37–47.

Huiwen Jia is an applied scientist at Amazon. She received her PhD degree in Industrial and Operations Engineering from the University of Michigan at Ann Arbor in 2022. Her research interests include stochastic optimization and online learning algorithms with applications in transportation and revenue management.

Cong Shi is an associate professor in the Department of Industrial and Operations Engineering, University of Michigan at Ann Arbor. His recent research interests lie in the design and analysis of online learning algorithms with applications to revenue management and supply chain management.

Siqian Shen is an associate professor in the Department of Industrial and Operations Engineering, University of Michigan at Ann Arbor. Her research interests include stochastic programming, network optimization, and integer programming. Applications of her work include transportation and energy.