On the Importance of Inherent Structural Properties for Learning in Markov Decision Processes

by

Saghar Adler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Vijay Subramanian, Chair
Associate Professor Asaf Cohen
Professor Mingyan Liu
Professor Lei Ying

Saghar Adler
ssaghar@umich.edu
ORCID iD: 0009-0004-8214-0222

Dedicated to my parents
Hamid Seyed Abbas Zadeh and Mahshid Abdollahi
and my husband
Arad Adler

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor Vijay Subramanian for his unwavering guidance and support throughout my doctoral journey. His mentorship allowed me to explore a new research area and provided invaluable assistance in navigating the challenges of a new field. I am also grateful to my committee members, Professor Mingyan Liu, Professor Lei Ying, and Professor Asaf Cohen, for their significant contributions to the development of my thesis and their insightful comments. My sincere appreciation extends to Professor Ehsan Afshari, whose teachings and mentorship enriched my intellectual journey, and I am incredibly thankful for his support throughout my PhD. I extend my sincere gratitude to my course instructors, Professor Demosthenis Tenekezis, Professor Mark Rudelson, and Professor David Barrett whose guidance and knowledge have been invaluable throughout my academic journey. I would also like to acknowledge the shared experiences with my fellow graduate students, Mehrdad Moharrami, Nouman Khan, Hossein Dabirian, Hossein Naghavi, Vahnood Pourahmad, and Ali Mostajeran, and others who have been sources of inspiration and collaboration.

To my family, especially my parents Mahshid and Hamid, my sister Sayeh, and my husband Arad, I owe an immeasurable debt of gratitude for your unwavering support, understanding, and encouragement throughout these challenging years. Last but not least, heartfelt thanks to my closest friends, Chris Najafi, Haniyeh Zamani, Saeed Kazemi, Banafsheh Pouyanfar, Ali Rashti, Melika Meraji, Peyman Sirous, Navid Nouri, Behnoush Rostami, Milad Mousavifar, Maryam Salim, Menglou Rao, Ramin Ansari, Amirhossein Tajedini, Najva Akbari, Sattar Vakili, and others for their constant encouragement, laughter, and understanding. This journey wouldn't have been the same without your support, and I am genuinely thankful for that.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF APPENDICES

APPENDIX

ABSTRACT

Recently, reinforcement learning methodologies have been applied to solve sequential decision-making problems in various fields, such as robotics and autonomous control, communication and networking, and resource allocation and scheduling. Despite great practical success, there has been less progress in developing theoretical performance guarantees for such complex systems. This dissertation aims to address the limitations of current theoretical frameworks and extend the applicability of learning-based control methods to more complex, real-life domains discussed above. This objective is achieved in two different settings using the inherent structural properties of the Markov decision processes used to model such systems. For admission control in systems modeled by the Erlang-B blocking model with unknown arrival and service rates, in the first setting, we use model knowledge to compensate for the lack of reward signals. Here, we propose a learning algorithm based on the self-tuning adaptive control and not only prove that our algorithm is asymptotically optimal but also provide finite-time regret guarantees. The second setting develops a framework to address the challenge of applying reinforcement learning methods to Markov decision processes with countably infinite state spaces and unbounded cost functions. An existing learning algorithm based on Thompson sampling with dynamically-sized episodes is extended to countably infinite state space using the ergodicity properties of Markov decision processes. We establish asymptotic optimality of our learning-based control policy by providing a sub-linear (in time-horizon) regret guarantee. Our framework is focused on models that arise in queueing system models of communication networks, computing systems, and processing networks. Hence, to demonstrate the applicability of our method, we also apply it to the problem of controlling two queueing systems with unknown dynamics.

CHAPTER 1

Introduction

1.1 Motivation

Recent advances in reinforcement learning (RL) have demonstrated its great potential for addressing complex real-world challenges [58, 82, 63, 74]. Despite the computational success, the current theoretical understanding of various reinforcement learning algorithms is limited, and the underlying reasons for their empirical success are not well understood. These theoretical limitations are particularly evident in the context of learning in Markov decision processes (MDPs) with infinite state or action spaces. The theorectical analysis of RL methods is mainly carried out for finite-state and finite-action MDPs (using tabular methods [80]) and linear MDPs (using functional approximation methods [110, 41]). Another limitation in current reinforcement learning methods is that these methods need dense reward signals to perform efficiently, which may not exist in many problems (heuristic solutions are employed in practice by constructing reward functions [33]).

Communication networks [62], computing systems [67, 95], supply chains [35, 83] and manufacturing systems [81] are important application domains with significant real-world impacts. In these application domains, admission control, rate control, server speed scaling, scheduling, resource allocation, and matching must be carried out effectively for efficient operation. Queueing models are commonly used to design and analyze algorithms employed for these tasks [20, 15, 26]. Knowing the underlying system parameters, queueing theoretic methods allow us to predict and optimize various performance measures, such as latency or loss, and determine the sensitivity of these measures to system parameters and algorithms. In practice, however, some or all of the system parameters may not be known, but efficient system operation is still required. As a result, it is essential to adapt existing learning algorithms or design new algorithms to address learning-based control in queueing systems when some or all system parameters are unknown. The unknown parameter assumption is also necessary for covering scenarios where it may not be possible to observe system parameters or where the parameters may (slowly) vary over time, such as in server processing times in large-scale server farms or treatment times in hospitals.

Developing learning algorithms applicable to real-world queueing models presents several challenges. This thesis focuses on overcoming two of these challenges, namely, limited information structure and large state spaces:

1. **Limited information structure:** In certain settings, reward functions are either not observed (neither perfectly nor directly) [104], costly to observe accurately [52], or too complex to characterize [46]. If the class of applicable models is known, then an alternative to the (dense) reward signals used in RL is the use of information rewards via likelihood values. We explore this direction to develop learning-based optimal control for a specific Markov decision process.

2. **Large state spaces:** Many queueing networks are modeled using infinite buffers and are naturally modeled using infinite state space MDPs. As a result, reinforcement learning schemes designed for learning unknown transition kernels in finite state spaces are not applicable. To tackle the challenge of learning for queueing systems with infinite buffers, we study the problem of learning within countable state space Markov decision processes with unbounded cost functions.

To overcome the aforementioned challenges, in this dissertation, I present two possible directions in which problem structure can be exploited to develop approximately optimal learning-based control methods. The first direction explores the utilization of model knowledge to augment the lack of dense reward signals and contrasts dense reward-based approaches versus methods using information signals based on model class knowledge. Motivated by the broad applicability of the Erlang-B blocking model, admission control for such a system is studied with unknown arrival and service rates with the aim of designing a dispatching policy that maximizes the long-term average reward by observing arrival times and system state at arrivals. The dispatcher observes neither service times nor departure epochs, precluding the use of reward-based reinforcement learning approaches. In contrast to the model-agnostic viewpoint in RL, the knowledge of the queueing dynamics is used to design an algorithm matched to our setting. We develop our learning-based dispatch scheme as a parametric learning problem a'la self-tuning adaptive control [55]. We prove that our proposed algorithm asymptotically converges to the optimal policy and present finite-time regret guarantees.

The second contribution of this thesis comes from utilizing the intrinsic ergodicity structure of specific Markov decision processes to extend learning schemes from finite state space settings to countably infinite state space with unbounded cost functions. Algorithmic and learning procedures developed to produce optimal policies mainly focus on finite state settings and do not directly apply to these models. We focus on a Bayesian framework and assume that the unknown transition probabilities are generated from a given prior distribution. In the countably infinite state-space

setting, it is crucial to establish certain assumptions regarding the class of models from which the unknown system is drawn to avoid many technical difficulties. To start with, the number of deterministic stationary policies is no longer finite. Moreover, in average cost optimal control problems, without stability assumptions or unbounded cost functions, the optimal policy may not exist or be stationary or deterministic [10]. With that in mind, we assume that for any state-action pair, the transition kernels in the model class are categorical and skip-free to the right, which is a common feature of many applications where an increase in some state corresponds to arrivals to the system. In addition, another set of assumptions ensures stability by assuming that the Markov process obtained by trying out different policies in the policy class is geometrically ergodic with some uniformity imposed over the entire parameter class. From these assumptions, moments on hitting times are derived in terms of another Lyapunov function that ensures polynomial ergodicity. The existence of the Lyapunov function for polynomial ergodicity is guaranteed by the assumed geometric ergodicity. However, to derive a useful bound for the moments of hitting times, we need the polynomial Lyapunov function to have certain properties, which we will discuss in Chapter 2. Using our assumptions, we show that a solution to the average cost optimality equation exists and provide a characterization of it.

To optimally control the unknown Markov decision process, we propose an algorithm based on Thompson sampling with dynamically-sized episodes. To evaluate the performance of our proposed algorithm, we utilize the metric of regret, which is defined as the expected total cost attained by a learning policy until time horizon T compared to the policy that achieves the optimal infinite-horizon average cost in a given policy class. Finally, using the solution of the average cost Bellman equation, we provide an upper bound on the Bayesian regret of our algorithm.

Collectively, these results help us improve the performance of existing reinforcement learning algorithms by using the structural properties inherent in specific Markov decision processes and establishing theoretical performance guarantees for these methodologies. Specifically, we introduce two distinct settings that use model knowledge to design learning algorithms matched to each setting. We further show theoretical performance guarantees for our proposed schemes and argue that the proposed learning algorithms attain asymptotically optimal performance.

1.2 Relevant literature

This section reviews three directions within the broader research domain of planning and learning in stochastic dynamical systems, outlined as follows:

1. **Stochastic control:** In this direction, the focus is on planning when the model is known.

2. **Adaptive control:** This approach involves learning within a parametric setting, where like-

lihood values function as informative rewards.

3. **Reinforcement learning:** This direction studies learning when the model is unknown by adopting a model-agnostic perspective, but where (dense) reward signals are available.

1.2.1 Stochastic control

This research direction studies a given stochastic dynamical system where the transition kernels are fully known and considers the problem of designing an admissible control strategy that optimizes a specified performance criterion, such as a cost function, within that framework [14, 11, 55]. In the finite-horizon setting, an optimal or ϵ-optimal policy for this optimization problem can be determined in a finite number of steps by the dynamic programming principle and solving the corresponding Bellman equation [13]. In the infinite-horizon version of this problem, assuming certain contraction assumptions hold, the optimal cost function is the solution to a fixed point equation, satisfies the Bellman equation, and a stationary optimal policy exists. Computationally, the optimal policy can be computed by successive application of the Bellman operator or other iterative methods, such as policy iteration. Under certain conditions, in the finite state and action space, it can be shown that these methods converge in a finite number of steps.

In this thesis, we consider a specific class of Markov decision processes with unknown transition kernels with the goal of simultaneous learning and control. At each stage of our learning algorithm, we estimate the unknown transition kernels, and assuming the estimate reflects the true dynamics, we apply the optimal policy according to this estimate. To derive this optimal policy, we utilize the existing stochastic control results and algorithms that provide us with the optimal or approximately optimal control law.

1.2.2 Adaptive control

The adaptive control literature studies the problem of asymptotic learning-based control of a stochastic dynamical system governed by an unknown parameter; for a detailed discussion, refer to Section 2.1.1. These classes of problems can be studied within two settings: a Bayesian setting, in which a prior distribution is given for the unknown parameter, or a non-Bayesian setting, in which the parameter is generated arbitrarily or there is no knowledge of the underlying prior distribution. We will focus on the latter setting, also referred to as self-tuning adaptive control [55]. The standard solution to this problem consists of two steps: first, the unknown parameter is estimated using an estimation method such as maximum likelihood estimation (MLE). Secondly, the optimal control law according to this parameter is applied. In [55], it is argued that in an MDP with finite state and action spaces, under certain identifiability conditions, the MLE converges to

the true parameter almost surely. To extend this result to the MDPs in which the restrictive identifiability conditions do not hold, forced exploration schemes are used, in which every admissible action is applied infinitely often. Another common solution is Reward-Biased MLE [53], wherein, instead of maximizing the likelihood function, a biased version that favors parameters with less optimal cost is maximized.

One challenge associated with implementing the discussed methods is that the theoretical results mainly hold when certain conditions, such as the mentioned identifiability conditions hold, or the probability distributions are uniformly bounded away from zero. However, these conditions may not hold in certain settings. In the seminal work on the self-tuning adaptive control, [66] shows that under identifiability, MLE converges to the true parameter under any control law, but [16] shows via an example that in the absence of identifiability such a conclusion may not hold. Additionally, in the context of queueing systems, the transition probabilities are not bounded away from zero; thus, the adaptive control literature results do not directly apply and modifications are needed [70]. In Chapter 3, we study the problem of learning-based control in such a queueing model, in which, at each arrival, the dispatcher can either accept (subject to availability) or reject the arrival. In the case of rejection in an empty queue, the queue will remain empty and states other than the empty state are visited with probability zero, which precludes the use of the methods introduced in the adaptive control literature. Another important point is that until recently the self-tuning adaptive control literature mainly focuses on asymptotic results. In contrast, we show finite-time performance guarantees for our proposed learning schemes in two different settings in Chapters 2 and 3.

1.2.3 Reinforcement learning

This line of research studies the problem of learning for sequential decision-making problems modeled by Markov decision processes with unknown models. The goal is to propose a model-agnostic learning algorithm to control the system such that a given measure of accumulated reward is maximized while receiving feedback via observed reward signals. One main category of reinforcement learning algorithms relies heavily on the reward signal and the associated Bellman equation. An instance of this general class of algorithms is the temporal-difference methodology, wherein a value function is learned to estimate the expected long-term reward for taking a specific action in a given state [92].

Another important category of learning algorithms in this context employs Bayesian learning, where a prior distribution is imposed on the unknown transition probabilities. These algorithms form and update a posterior distribution iteratively based on the received samples of the system. Using this posterior distribution, an estimate of the unknown transition probabilities is formed. It is

important to note that Thompson sampling schemes naturally balance the exploration/exploitation trade-off by maintaining a distribution of unknown parameters. Another advantage of these learning schemes is that the model knowledge can be encoded through the choice of prior distribution, and, as a result, the sample complexity of the learning algorithm can be improved. Characterizing the evolution of the posterior distribution during the learning process can be challenging. Hence, the learning problem is usually formulated and analyzed in an episodic framework. Another method to overcome the analysis challenges is that the learning problem is usually analyzed by assuming that the MDP has certain properties. As an example of the related work in this domain, [80] considers the problem of learning in an infinite-horizon MDP with finite state and action space. Using an algorithm based on Thompson sampling, called Thompson Sampling with Dynamically-sized Episodes (TSDE), they show an $\tilde{O}(S\sqrt{AT})$ Bayesian regret bound, where S and A are the sizes of state and action space, and T is the time-horizon. Similarly, [36], under recurrence assumptions, bounded likelihood ratios, and other technical conditions, establishes a sublinear frequentist regret bound in a finite state and action space MDP.

This thesis aims to extend the existing theoretical results to countable state space MDPs, particularly to achieve learning in unknown queueing systems with infinite buffers. To overcome the challenges inherent in countable state space MDPs with unbounded cost functions, we impose ergodicity assumptions uniformly throughout the parameter and policy classes. From these assumptions and other technical details, we prove a sublinear regret for our proposed algorithm based on Thompson sampling with dynamically-sized episodes algorithm.

1.3 Thesis overview

We next present an overview of our results and contributions.

1.3.1 Chapter 2: Self-tuning Adaptive Control for Admission Control in Erlang-B Systems

Motivated by applications of the Erlang-B blocking model beyond communication networks to sizing and pricing in production, messaging, and app-based parking systems, we study admission control for it with unknown arrival and service rates. In our model, a dispatcher assigns every arrival to an available server or blocks it. Every served job yields a fixed reward but incurs a per unit time holding cost. We aim to design a dispatching policy that maximizes the long-term average reward by observing arrival times and system state at arrivals, a realistic sampling of such systems. The dispatcher observes neither service times nor departure epochs, precluding the use of reward-based reinforcement learning approaches. We develop our learning-based dispatch scheme as a

parametric learning problem *a'la* self-tuning adaptive control. In our problem, certainty equivalent control switches between *always admit if room* (explore infinitely often) and *never admit* (terminate learning), so at judiciously chosen times, we avoid the never admit recommendation. We prove that our proposed policy asymptotically converges to the optimal policy and present finite-time regret guarantees. The extreme contrast in the control policies shows up in our regret bounds for different parameter regimes: constant in one versus logarithmic in another.

1.3.2 Chapter 3: Bayesian Learning in Countable State Space Markov Decision Processes

Models of many real-life applications, such as queueing models of communication networks or computing systems, have a countably infinite state-space. Algorithmic and learning procedures developed to produce optimal policies mainly focus on finite state settings and do not directly apply to these models. To overcome this lacuna, we study the problem of optimal control of a family of discrete-time countable state-space Markov decision processes governed by an unknown parameter $\theta \in \Theta$ and defined on a countably-infinite state-space $\mathcal{X} = \mathbb{Z}_+^d$, with finite action space $\mathcal{A}$ and an unbounded cost function. We take a Bayesian perspective with the random unknown parameter $\boldsymbol{\theta}^*$ generated via a given fixed prior distribution on Θ. To optimally control the unknown MDP, we propose an algorithm based on Thompson sampling with dynamically-sized episodes: at the beginning of each episode, the posterior distribution formed via Bayes' rule is used to produce a parameter estimate, which then decides the policy applied during the episode. To ensure the stability of the Markov chain obtained by following the policy chosen for each parameter, we impose ergodicity assumptions. From this condition and using the solution of the average cost Bellman equation, we establish an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ upper bound on the Bayesian regret of our algorithm, where T is the time-horizon, and h determines the skip-free to the right property.

Finally, to provide examples of our framework, we consider two different queueing models that meet our technical conditions and show that our algorithm can be applied to develop approximately optimal control algorithms even though the underlying dynamics are unknown. The first example is a continuous-time queueing system with two heterogeneous servers with unknown service rates and a common infinite buffer. In this setting, the optimal policy that minimizes the average waiting time is a threshold policy given in [54]. We verify our assumptions for a class of optimal policies corresponding to different service rates and conclude that our algorithm is well-suited for this setting. The second model is a two-server queueing system, each with separate infinite buffers. In this setting, the optimal policy to minimize the waiting time is unknown, so we aim to find the best policy among policies that assign the arrival to the queue with minimum weighted queue length. Similarly, we show that our assumptions hold for this model, and our algorithm can be used to

learn the best-in-class policy.

1.4 Summary of contributions

In summary, the major contributions of this thesis are as follows:

- **Chapter 2:** We study the problem of learning the unknown service rate of an $M/M/k/k$ queueing system. We design a dispatching policy based on maximum likelihood estimation and the certainty equivalent law coupled with forced exploration. We show the convergence of our proposed policy to the optimal policy, which is a threshold policy. Specifically, when the true service rate is above the threshold value, after a random finite time, all job arrivals are accepted (subject to availability). In contrast, when the true service rate is equal to or below the threshold value, all new arrivals are rejected after a random finite time. Moreover, in one parameter regime, we show a finite regret bound for an exploration function growing slower than exponential. Additionally, we prove $O(\log(n))$ regret for a specific exploration function in the other regime, where n is the number of arrivals to the system. Consequently, we conclude that by using model knowledge and designing a learning algorithm tailored to our specific problem structure, we can compensate for the absence of direct access to the reward function. Furthermore, model knowledge can be used to enhance learning performance, even compared to generic learning algorithms that observe reward signals.

- **Chapter 3:** We generalize the existing learning algorithm, Thompson sampling with dynamically-sized episodes [80], from finite state space setting to countably infinite state space setting. We further provide a finite-time performance guarantee by proving a sublinear regret for our algorithm. Specifically, we explore the following three cases:

    1. We demonstrate that our algorithm can be used to learn the optimal policy with an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ Bayesian regret.
    2. We illustrate that our algorithm learns the optimal policy within a specified policy class with an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ Bayesian regret.
    3. We argue that our proposed algorithm and finite-time regret guarantees can be extended to scenarios in which we do not have access to an oracle providing us with the optimal/best-in-class policy. Instead, in cases where we only have access to approximately optimal policies, we show that our learning algorithm can be employed under specific conditions to identify the best policy within a given policy class, and the same sublinear regret order carries through.

Moreover, we demonstrate that our learning algorithm can be applied to two different queueing models with infinite buffers with the goal of learning unknown service rates.

Notation of the Thesis: Notation varies across chapters. However, each chapter is self-contained, and the notation of each chapter is defined in that chapter. Appendices follow the notation of the corresponding chapter.

CHAPTER 2

Self-tuning Adaptive Control for Admission Control in Erlang-B Systems

2.1 Introduction

Queueing systems are widely applicable models used to study resource allocation problems in communication networks, distributed computing systems, semiconductor manufacturing, supply chains, and many other dynamic systems. Queueing models are analyzed under various system information settings, but a common assumption is that the core system parameters like arrival rates, service rates and distributions are available to the system designer (e.g., [87, 40]). However, there are various applications where these parameters are unknown, and the designer needs to learn them to be able to optimally assign jobs to the servers or block them. For example, the service rate of every server in large-scale server farms may be unknown, or the treatment times in hospitals may be unpredictable and time-varying.

The focus of this chapter is the Erlang-B system ([47, 87]). The traditional use of this system has been for sizing and analyzing voice and circuit-switched systems, i.e., loss systems. It is also used for sizing and analyzing call-center systems ([32]) with no waiting room and no reneging. Furthermore, it has been employed to study multiple-access schemes in wireless networks ([68]). More recently, these systems have been used to design and size production systems; for instance, by Amazon for its SimpleDB database service, Facebook for the back-end of its chat service, WhatsApp for its messaging servers, Motorola in call processing products used for public safety, etc. These applications motivate us to consider learning for the Erlang-B queuing system. Specifically, our problem formulation aligns with the call-center systems mentioned above, assuming the call center is operated by a third-party entity and the servers are homogeneous. It also extends to applications such as the pricing of parking lots in app-based parking systems or messaging systems implemented using third-party cloud servers.

Motivated by these applications and to highlight challenges in learning-based optimal control, we study optimal admission control in an Erlang-B queueing system with exponentially distributed

service times, and unknown arrival and service rates, denoted by λ and μ, with the goal of designing an optimal learning-based dispatching policy. At every arrival, the dispatcher can accept or block the arrival. Accepted jobs incur a service cost c per unit time, and yield a fixed reward R. Assuming that the service rate is known, the dispatcher can maximize its expected reward using a threshold policy: if the service rate exceeds c/R, all arrivals are admitted subject to availability; otherwise, all arrivals are rejected. When the service rate equals c/R, the dispatcher is indifferent between admitting or rejecting arrivals.

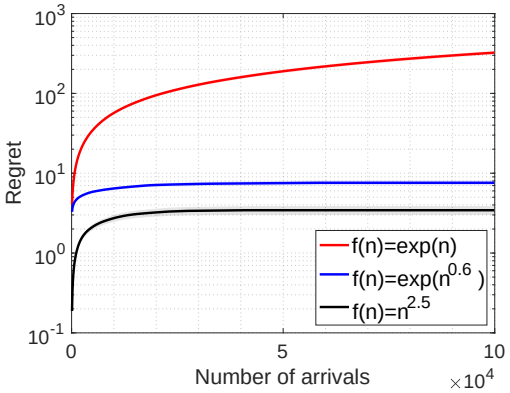
A key aspect of our problem setting is that the information available to the dispatcher consists only of the inter-arrival times and the number of busy servers at each arrival, as the system is sampled at arrivals. Contrarily, the service rate, departure times, and service times are not known to the dispatcher. Hence, the dispatcher cannot form a direct estimate of the service rate (e.g., by taking an empirical average of the observed service times) to then choose its policy, and instead has to use the queueing dynamics to estimate the service time for policy determination. This facet of the problem brings it closer to practice but also complicates the analysis. Based on this information structure, our focus is to design an optimal policy that maximizes the long-term average reward.

We study the problem of learning the service rate in the framework of parametric learning of a stochastic dynamical system. Specifically, consider a stochastic system governed by parameter θ:

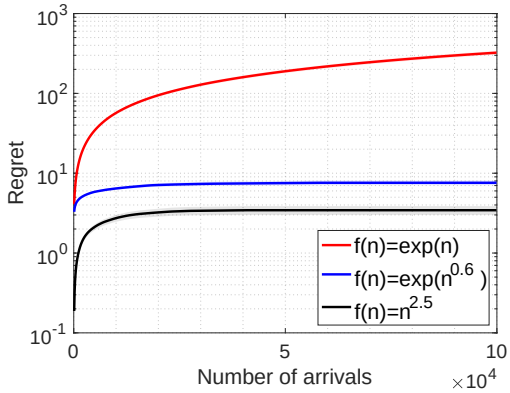
$$X_{t+1} = \mathcal{F}_t(X_t, U_t, W_t; \theta), \quad t = 0, 1, \dots \tag{2.1}$$

where $X_t \in \mathcal{X}$, $U_t \in \mathcal{U}$, $W_t \in \mathcal{W}$ are the state of the system, control input, and noise at time t and $\mathcal{F}_t$ is any measurable function. Further, $\theta \in \Theta$ is a fixed but unknown parameter, and the initial state and noise process are mutually independent. In line with the literature, we study a system where our controller *perfectly observes* the state X_t and uses its history of observations to choose the control U_t. For a specified reward function $r_t(x, u)$ for $(x, u) \in \mathcal{X} \times \mathcal{U}$, the objective is to maximize the long-term reward. We also assume that the optimal policy $\mathcal{G}^*(.; \theta)$ is *known* for each $\theta \in \Theta$.

To achieve the optimization objective whilst learning the unknown parameter θ, an adaptive control law is applied: using past observations $X_{1:t}$, an estimate $\hat{\theta}_{t+1}$ is formed, and then by certainty equivalent control law, the optimal policy according to $\hat{\theta}_{t+1}$, or $\mathcal{G}^*(.; \hat{\theta}_{t+1})$, is applied. One approach to form the estimate $\hat{\theta}_{t+1}$ is to use the maximum likelihood estimate (MLE). [66] prove that under identifiability, the MLE converges to the true parameter. When these conditions do not hold, to guarantee convergence, [53, 54] use reward bias-based exploration schemes to ensure asymptotic optimality. Our problem fits the above paradigm: the system state X_t is the number of busy servers at time t with the dispatcher observing the (continuous-time) system state at arrivals,



(a) $\mu = 2.05, \mu \in (c/R, +\infty)$.



(b) $\mu = 1.05, \mu \in (0, c/R)$.

Figure 2.1: Comparison of regret performance of Algorithm 1 for different functions $f(n)$ in a 5 server system with $\lambda = 5$ and $c/R = 1.3$. Different service rates are considered. The shaded region indicate the $\pm\sigma$ area of mean regret. In (a), the y axis is plotted on a logarithmic scale to display the differences clearly.

and the unknown parameter is the service rate μ, so $\Theta = \mathbb{R}_+$ [1]. Using an adaptive control law with forced exploration, we propose a dispatching policy to maximize the long-term average reward. Our main analysis-related contributions are:

1. **Asymptotic optimality.** We prove the convergence of our learning-based policy to the optimal policy. We first focus on a single-server Erlang-B queueing system; see Section 2.4.1. An underlying independence structure lets us establish asymptotic optimality using the strong law of large numbers. For the multi-server setting, the independence structure does not hold anymore, so in Section 2.5.1, a more intricate argument based on martingale sequences is used for the proof to prove convergence.

2. **Finite-time performance analysis.** In Section 2.4.2, we characterize the finite-time regret for the single-server system in two distinct service rate regimes. In the first regime, we show finite regret using independence and concentration inequalities. However, in the other regime, the exploration done by our policy leads to a regret upper bound that scales as $\log(n)$, where n is the number of arrivals. We also generalize our results to the multi-server setting to observe that the regret exhibits similar behavior as in the single-server setting; see Section 2.5.2. The analysis for the multi-server setting is based on Doob's decomposition and concentration inequalities for martingale sequences.

[1] More generally, we can take both the arrival and service rates, λ and μ, to be the unknown parameters.

We end by contrasting our work with the literature on learning in stochastic dynamical systems. We study an example of a parametric learning problem for which we do not expect a single policy to achieve minimum regret in all regions of the parameter space. Whereas we don't have an explicit proof of such a claim, the contrasting behavior an optimal adaptive control scheme must exhibit in different parameter regimes—quickly converging to always admitting arrivals if room versus quickly rejecting all arrivals—gives credence to the claim. We discuss the above point in Figure 2.1, which depicts the performance of our algorithm for functions $f(n) \in \{n^{2.5}, \exp(n^{0.6}), \exp(n)\}$ where $1/f(n)$ is proportional to the (forced) exploration probability. For $f(n) = n^{2.5}$, exploration is employed aggressively, causing better performance for $\mu \in (c/R, +\infty)$, and higher regret in the other regime. Conversely, when $f(n) = \exp(n)$, aggressive exploitation is enforced, leading to the opposite behavior. For $\mu \in (c/R, +\infty)$, we show finite regret for $f(n) \in \{n^{2.5}, \exp(n^{0.6})\}$ in Section 2.5.2, but finite regret is not guaranteed for $f(n) = \exp(n)$ in our analysis. In Section 2.5.2, when $\mu \in (0, c/R)$, we establish an $O(\log^{5/3}(n))$ regret bound for $f(n) = \exp(n^{0.6})$. Similar arguments lead to a $O(\log(n))$ upper bound for $f(n) = \exp(n)$ in the same regime. From this discussion, we expect big differences in performance of any algorithm based on the parameter regime. Based on our numerical results, we also conjecture that for $\mu \in (0, c/R)$, there is an $\Omega(\log(n))$ regret lower bound. This is consistent with the lower bound on the asymptotic growth of regret from the literature on learning in unknown stochastic systems under the assumption that the transition kernels of the underlying controlled Markov chains are strictly bounded away from 0; see [5, 37].

Furthermore, our simulation results in Section 2.6 investigate other aspects that highlight the subtleties in designing learning schemes. For example, they provide evidence that depending on the relationship between the arrival rate and the service rate, sampling our continuous-time system at a faster rate than the arrivals could reduce the regret. We also show that subtle differences in variable updates in the learning scheme have a substantial impact on the regret achieved. Thus, the choice of the trade-off of regret between the different parameter regimes determines the learning scheme.

2.1.1 Related work

Adaptive control. The self-tuning adaptive control literature studies asymptotic learning in the parametric or non-parametric version of the problem described in (2.1), and the study was initiated by Mandl. [66] showed that the MLE converges to the true parameter under an identifiability condition. Since then, the adaptive control problem has been vastly studied in great generality; see [16, 53, 54, 5, 37, 36]. The work in [16] studied the adaptive control problem when the identifiability condition need not hold and proved that for a finite-state controlled Markov chain with finite

parameter space, the maximum likelihood estimate converges almost surely to a parameter with the same transition probabilities of the true parameter, if the transition probabilities are uniformly lower bounded. In [53, 54], a finite-state and finite-control Markov process with finite parameter space is considered, and an adaptive control law is presented that optimizes the long-term average cost using a combination of biased maximum likelihood estimation and certainty equivalent control law. Reference [5] view the problem of learning in an unknown controlled Markov chain with finite state, action, and parameter space as a multi-armed bandit problem and introduce a control scheme to minimize the rate of increase on the expected regret. In [37], a controlled Markov process is considered on a general state space and a compact parameter space, and an adaptive control law is presented that minimizes the expected regret in a particular class of control laws. Reference [56] develops adaptive control schemes in the non-parametric setting by working with a set of policies. Finally, [36] take a Bayesian viewpoint and develop expected regret bounds for Thompson sampling based schemes. Additionally, learning in queueing systems is one of the applications in this literature; see [56, 55].

A core assumption in the above literature is that the transition kernels of the underlying controlled Markov chains are strictly bounded away from 0 and 1, with the bound uniform in the parameter and the class of (optimal) policies. This core assumption does not hold in our problem: the controlled Markov chain found by sampling the queueing system at arrivals has drastically different behavior under the available class of policies—admit if room or never admit—, and thus the conclusions of this literature do not apply. Furthermore, in the above literature, most of the results are on asymptotic learning, and only recently, finite-time regret guarantees have been obtained. The existing finite-time regret guarantees are largely for certain discrete-time queueing systems with geometrically distributed service times and unknown parameters, which we will discuss below.

Queueing systems. There is a growing body of work on learning-based control in discrete-time queueing systems; see [103] for a recent survey. References [50, 51] studied a discrete-time multi-class, multi-server queueing system with unknown service rates. After imposing stability conditions on the problem parameters, [50] used a forced exploration-based learning scheme to prove finite regret compared to the $c\mu$ rule in a system with service rates known. In another work, [51] used UCB and Thompson sampling-based algorithms to prove a polylogarithmic regret bound. Reference [22] proved an $\tilde{O}(\sqrt{T})$ regret over time horizon T using a queue-length agnostic randomized-routing-based algorithm for a multi-server discrete-time queueing system. All of these works form empirical service rate estimates by observing and averaging service successes and failures.

Furthermore, [88] studied the problem of finding the optimum server for service in a discrete-time multi-server system with unknown service rates and a single queue and proves constant regret by sampling service rates during idle periods. In another work, [75] employed generative adver-

sarial networks to numerically learn the unknown service time distributions in a $G/G/\infty$ queuing system. In a recent work, [109] studied scheduling in a multi-class queue with abandonment with unknown arrival, service, and abandonment rates. By using service and patience times and forming estimates of the service and abandonment rates, logarithmic regret is shown against the $c\mu/\theta$ rule using an exploration-exploitation based scheme. Reference [108] studied social-welfare maximizing admission control in an $M/M/1$ queuing system with unknown service and arrival rates; with system parameters known a threshold-based admission control scheme is optimal. By observing the queueing system at all times, they propose a dispatching algorithm that achieves constant regret for one set of parameters, and $O(\log^{1+\epsilon}(n))$ regret for any $\epsilon > 0$ for another set of parameters (n is the number of arrivals).

In all these works, all completed service times or entire queueing processes are observed and used for learning. Such observations may not be feasible in real-world queueing systems due to increased computation and memory requirements: see [89, 40]. Multi-server settings introduce other complications: to correctly identify completed service times, server assignments need to be tracked from the entire process history (even for homogeneous servers). In our work, observations are the (minimal) Markov state of the system at each arrival, which despite being a nonlinear function of service times, aligns better with real-world systems. In Section 2.6, using simulations, we also show that sampling of such continuous-time systems needs careful design.

Learning-based decision-making has also been studied in inventory control and dynamic pricing. Reference [8] studies an inventory control problem with unknown demand distribution. The goal is to minimize the total cost associated with inventory holding and lost sales penalties over T periods by observing the minimum of demand and inventory. A learning algorithm is proposed based on the convexity of the average cost function under the benchmark base-stock policies, and a $O(\sqrt{T})$ regret is established. Reference [21] studies a dynamic pricing problem in a $GI/GI/1$ queue with the objective of determining the optimal service fee and service capacity that maximize the expected total profit. A gradient-based online-learning algorithm is proposed that estimates the gradient of the objective function from the history of arrivals, waiting times, and the server's busy times and a logarithmic regret bound in the total number of served customers is established. In another work, [45] study a price-based revenue management problem with finite reusable resources under price-dependent unknown arrival and service rates. The goal is to find the optimal pricing policy that maximizes the total expected revenue by observing the inter-arrival and service times. Two different online algorithms based on Thompson Sampling and Upper Confidence Bound are proposed, and a regret upper bound of $\tilde{O}(\sqrt{T})$ is proved, where T is the time-horizon.

Another related line of work focuses on the use of pricing strategies to regulate queue sizes and studies differences between individually optimal and socially optimal strategies (with model parameters known). Reference [72] studied regulating an $M/M/1$ queue with fixed reward and

linear holding cost, which was then generalized in [48] to an $M/M/k$ queuing model with fixed reward and nonlinear holding cost. In both, to ensure social optimality, customers are subject to a toll upon joining the queue to counteract the increased congestion when agents selfishly optimize. Similarly, [61] investigated a stochastic congestion system with random reward and linear holding cost and argued that individuals acting in self-interest over-congest a system relative to the socially optimal rule. Similar to [72, 48], a toll can be charged to induce customers to act in a socially optimal way. In these works, to ensure social optimality, customers are subject to a toll upon joining the queue to counteract the increased congestion when agents selfishly optimize.

Reinforcement learning (RL). Recently, RL methods have been applied to queueing problems with the goal of finding the average cost optimal policy, in both known model and cost parameter cases ([27]), and unknown parameter cases with available rewards ([69]). These methods do not apply to our setting as we neither observe the reward sequence nor know the expected rewards: the random reward is a linear function of the service times of accepted jobs which are not observed, and the expected reward is a function of the unknown arrival and service rates. We only observe the system state: a nonlinear and complex function of the reward. In contrast to the model-agnostic viewpoint in RL, we use the knowledge of the queueing dynamics to design an algorithm matched to our setting. Although RL methods do not apply to our setting, in Section 2.6, we consider a fictitious setup wherein the service times are observed ahead of time and implement an average reward RL algorithm, R-learning ([92]). Despite not observing the service times, our policy outperforms R-learning, providing evidence that model-class knowledge can be as effective as observing the reward signal; see Figure 2.4. In Figure 2.4, we also compare our algorithm to a Thompson sampling-based algorithm ([36]), showing that our algorithm using model-class knowledge is again as effective as Thompson sampling.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the problem and the learning objective. Section 2.3 discusses our learning-based dispatching policy and Section 2.4 shows the asymptotic optimality of our proposed policy in a single-server Erlang-B system. Moreover, we characterize the regret of our proposed policy compared to the system with knowledge of the service rate. Section 2.5 extends the results of Section 2.4 to the multi-server setting. In Section 2.6, we study the performance of our proposed policy through experiments and verify our theoretical analysis.

2.2 Problem formulation

We consider an $M/M/k/k$ queueing system with k identical servers. Arrivals to the system are according to a Poisson process with rate λ, and at each arrival, a dispatcher decides between admitting the arrival or blocking it. If admitted, the arrival is dispatched to the first available server

and serviced with exponentially distributed service times with parameter μ. Otherwise, if blocked, it leaves the system. Each time an arrival is accepted, the dispatcher receives a fixed reward R (after service completion), but incurs a cost of c per unit time service; we assume that rejecting an arrival has no penalty. In our setting, we assume that the dispatcher knows the parameters R and c but does not know either the service rate μ or the arrival rate λ. We also assume that the dispatcher observes the arrival times to the system and the system state upon arrivals. In contrast to the inter-arrival times, the service times of completed services are unknown.

Consider the queueing system sampled at arrival i for $i \in \{0, 1, \ldots\}$, and let A_i denote the action of the dispatcher to admit or block arrival i. If arrival i is blocked, $A_i = 0$; otherwise, if arrival i is admitted (when there's room), $A_i = 1$. We define N_i as the number of busy servers just before arrival i, and the system starts with empty servers, i.e., $N_0 = 0$. Let T_i be the inter-arrival time between arrival $i-1$ and i, and M_i be the number of departures during inter-arrival T_i. Notice that

$$N_{i-1} + A_{i-1} = M_i + N_i$$

and the value of M_i can be found with the knowledge of $\{N_{i-1}, N_i, A_{i-1}\}$. The dispatcher chooses A_i based on past observations up to arrival i, i.e., $\mathcal{H}_i = \{T_1, \ldots, T_i, A_0, A_1, \ldots, A_{i-1}, N_0, N_1, \ldots, N_i\}$. Using this history, the dispatcher's goal is to choose action sequence $\{A_n\}_{n=0}^{\infty}$ to maximize the expected average reward per unit time, which by PASTA ([87]) is

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}[K(A_i, \sigma_i)],$$

where σ_i is the service time of arrival i, and the reward function $K(\cdot, \cdot)$ is given by $K(a, s) = a(R - cs)$.

In a system with known service rate μ, the optimal policy of the dispatcher is to accept all arrivals if $\mu > c/R$ (subject to availability) and block all arrivals if $\mu < c/R$. The dispatcher is indifferent between accepting or rejecting when $\mu = c/R$. We evaluate the performance of a candidate policy with respect to the optimal policy, denoted by Π^*. In Section 2.3, we propose a dispatching policy that uses past observations to estimate the service rate μ, and in Sections 2.4.1 and 2.5.1, we show the asymptotic optimality of our policy by proving its convergence to Π^*. Further, in Sections 2.4.2 and 2.5.2, the finite-time performance of our policy is evaluated using the following definition.

Definition 1. *Set A_i^{Π} as the action taken at arrival i in a system that follows policy Π. The expected*

regret of a policy Π *with respect to the optimal policy* Π^* *after* n *arrivals is defined as*

$$\mathbb{E}\left[\mathcal{R}\left(n\right);\Pi\right] = \left|\,\mathbb{E}\left[\sum_{i=0}^{n-1}(A_i^{\Pi} - A_i^{\Pi^*})\right]\right|.$$

2.3 Proposed maximum likelihood estimate-based dispatching policy

In our problem setting, both the arrival and service rates, λ and μ, are unknown, but for the optimal dispatching policy it is sufficient to estimate the service rate. We would like a dispatching policy that (asymptotically) performs optimally, and further, (if possible) we want to minimize the regret of this system with respect to the system with known μ. As mentioned in Section 2.1, we take a self-tuning adaptive control viewpoint: we consider the system as being driven by parameter μ, and the learning problem as a parameter estimation problem using system measurements given by the sequence of policies chosen. Specifically, we use maximum likelihood (ML) estimation to estimate parameter μ, and then select the certainty equivalent control but with forced exploration.

2.3.1 Maximum likelihood estimate derivation

In this section, we derive the log-likelihood function and the corresponding MLE. The probability of m_i departures and n_i incomplete services at inter-arrival t_i given $m_i + n_i = N_{i-1} + A_{i-1}$ is

$$p\left(m_i, n_i, t_i; \mu\right) = \binom{n_i + m_i}{n_i} \left(1 - \exp\left(-\mu t_i\right)\right)^{m_i} \left(\exp\left(-\mu t_i\right)\right)^{n_i}. \tag{2.2}$$

The above equation follows from the exponential distribution of the service times. From (2.2), the conditional probability of observing sequences $\{m_i\}_{i=1}^n$ and $\{n_i\}_{i=1}^n$ for a fixed μ given the inter-arrival sequence $\{t_i\}_{i=1}^n$ is given by

$$\mathbb{P}\left(M_1 = m_1, \ldots, M_n = m_n, N_1 = n_1, \ldots, N_n = n_n \,\middle|\, \mu, \{t_i\}_{i=1}^n\right) = \prod_{i=1}^{n} p\left(m_i, n_i, t_i; \mu\right). \tag{2.3}$$

In our problem formulation, no prior distribution is assumed for μ, and thus, the posterior probability of a fixed μ given observations of $\{m_i\}_{i=1}^n$,$\{n_i\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$ is proportional to (2.3). From (2.2) and (2.3), we form the likelihood function of the past observations $\mathcal{H}_n$ under parameter μ as

$$L\left(\mathcal{H}_n; \mu\right) := c_b \prod_{i=1}^{n} \left(1 - \exp\left(-\mu T_i\right)\right)^{M_i} \left(\exp\left(-\mu T_i\right)\right)^{N_i}, \tag{2.4}$$

where c_b is the product of the binomial coefficients found in (2.2) and independent of μ. Maximization of likelihood function $L(\mathcal{H}_n;\mu)$ is equivalent to maximization of log-likelihood function $l(\mathcal{H}_n;\mu)$ defined as

$$l(\mathcal{H}_n;\mu) := \log L(\mathcal{H}_n;\mu) = \log c_b + \sum_{i=1}^{n} M_i \log\left(1-\exp\left(-\mu T_i\right)\right) - \mu \sum_{i=1}^{n} N_i T_i. \tag{2.5}$$

If $M_i = 0$ for all i, the maximum of $l(\mathcal{H}_n;\mu)$ in $[0,+\infty)$ is obtained for $\mu = 0$, and if $N_i = 0$ for all i, the maximum is reached at $+\infty$. Otherwise, from differentiability and strict concavity of the log-likelihood function, it follows that it has at most one maximizer, and as

$$\lim_{\mu\to 0} l(\mathcal{H}_n;\mu) = \lim_{\mu\to+\infty} l(\mathcal{H}_n;\mu) = -\infty,$$

there exists a unique $\hat{\mu}_n > 0$ that maximizes $l(\mathcal{H}_n;\mu)$, which can be found by taking the derivative with respect to μ and setting it equal to 0. The derivative of $l(\mathcal{H}_n;\mu)$ is given by

$$l'(\mathcal{H}_n;\mu) = \sum_{i=1}^{n} \frac{M_i T_i \exp\left(-\mu T_i\right)}{1-\exp\left(-\mu T_i\right)} - \sum_{i=1}^{n} N_i T_i. \tag{2.6}$$

From (2.6), the maximum likelihood estimate $\hat{\mu}_n$ is the solution to the following equation:

$$\sum_{i=1}^{n} g(T_i, M_i, \hat{\mu}_n) = \sum_{i=1}^{n} h(T_i, N_i, \hat{\mu}_n), \tag{2.7}$$

where

$$g(t,m,\mu) := \frac{mt\exp\left(-\mu t\right)}{1-\exp\left(-\mu t\right)}, \qquad\qquad h(t,n,\mu) := nt.$$

It is easy to verify that $\sum_{i=1}^{n} g(T_i, M_i, \mu)$ is a positive and decreasing function of μ. Moreover,

$$\lim_{\mu\to 0} \sum_{i=1}^{n} g(T_i, M_i, \mu) = +\infty, \qquad\qquad \lim_{\mu\to+\infty} \sum_{i=1}^{n} g(T_i, M_i, \mu) = 0.$$

Since $\sum_{i=1}^{n} h(T_i, N_i, \mu)$ is a positive constant independent of μ, Equation (2.7) has a unique positive solution $\hat{\mu}_n$. However, given the simple set of optimal policies for our problem, we do not need to solve this equation to determine our policy. For a given estimate $\hat{\mu}_n$, the optimal policy only requires a comparison of $\hat{\mu}_n$ and c/R, and, based on the properties of g and h, to compare $\hat{\mu}_n$ with c/R, it suffices to compare $\sum_{i=1}^{n} g(T_i, M_i, c/R)$ with $\sum_{i=1}^{n} h(T_i, N_i, c/R)$.

Algorithm 1 Proposed ML estimate-based Policy for Learning the Optimal Dispatching Policy

Input: $f : \mathbb{N} \cup \{0\} \to [1, \infty)$, increasing, and $\lim_{n\to+\infty} f(n) = +\infty$.
Initialize $N_0 = 0, \alpha_0 = 0$.
At arrival $n \geq 0$, **do**
Update α_n using (2.8), and find $S(n) = \max\{0 \leq i \leq n : N_i = 0\}$.
if $N_n = k$ **then**
    Block the arrival.
else if $N_n < k$ and $\sum_{i=1}^{S(n)} g(T_i, M_i, c/R) > \sum_{i=1}^{S(n)} h(T_i, N_i, c/R)$ **then**
    Admit the arrival.
else if $N_n < k$ and $\sum_{i=1}^{S(n)} g(T_i, M_i, c/R) \leq \sum_{i=1}^{S(n)} h(T_i, N_i, c/R)$ **then**
    Admit the arrival with probability $p_{\alpha_n} = 1/f(\alpha_n)$.
end if

2.3.2 The learning algorithm

The discussion at the end of the previous subsection leads to the following two cases:

1. $\sum_{i=1}^{n} g(T_i, M_i, c/R) > \sum_{i=1}^{n} h(T_i, N_i, c/R)$ implies that $\hat{\mu}_n > c/R$.
2. $\sum_{i=1}^{n} g(T_i, M_i, c/R) \leq \sum_{i=1}^{n} h(T_i, N_i, c/R)$ implies that $\hat{\mu}_n \leq c/R$.

In Case 1, the MLE indicates the *always admit if room* policy is optimal. In our proposed policy, we follow the MLE whenever Case 1 applies and admit the arrival (if there is a free server). In contrast to Case 1, the MLE in Case 2 suggests blocking all arrivals. However, if we follow the MLE in both cases, we may falsely identify the service rate and incur linear regret. Notably, using the optimal policy in Case 2 results in no arrivals and new system samples. Thus, to ensure learning, in Case 2, our policy will not use the certainty equivalent control with a small probability that converges to 0. Finally, we introduce Algorithm 1 for optimal dispatch in an Erlang-B system with unknown service rate.

We label the policy in Algorithm 1 as Π_{Alg1}. Then $S(n)$ is defined as the last arrival instance before or at arrival n when the system is empty, i.e., all servers are available. The probability of using the sub-optimal policy in Case 2 is equal to $p_{\alpha_n} = 1/f(\alpha_n)$, where a valid function $f : \mathbb{N} \cup \{0\} \to [1, \infty)$ is increasing and converges to infinity as α_n goes to infinity. Further, $\alpha_0 = 0$ and α_n is defined as below for $n \geq 1$

$$\alpha_n = \begin{cases} \alpha_{n-1} + 1, & \text{if } \sum_{i=1}^{n-1} g(T_i, M_i, c/R) \leq \sum_{i=1}^{n-1} h(T_i, N_i, c/R), A_{n-1} = 1, N_{n-1} = 0, \\ \alpha_{n-1}, & \text{otherwise.} \end{cases} \tag{2.8}$$

In other words, α_n is the number of accepted arrivals $0 \leq l < n$ such that $\sum_{i=1}^{l} g_i(c/R) \leq \sum_{i=1}^{l} h_i$ and the system is empty right before arrival l. In the following remark, we clarify the effect of function f on the performance of Π_{Alg1}.

Remark 1. *Any choice of $f \geq 1$ that increases to infinity leads to asymptotic optimality of Π_{Alg1} and convergence to Π^*, as proved in Sections 2.4.1 and 2.5.1. However, the class of admissible functions is restricted in Sections 2.4.2 and 2.5.2 to provide finite-time guarantees.*

The parameters of policy Π_{Alg1} are only updated when the system becomes empty, i.e., at busy period boundaries, rather than at all arrivals. The reason for this modification is that the busy period boundary is a regenerative epoch that provides sufficient independence needed in the analysis, whereas the regret of the policy that updates its parameters at all arrivals is hard to analyze. However, this alternate policy, called Π_{Alg3}, is also asymptotically optimal, and we empirically compare it to Π_{Alg1} in Section 2.6.

Remark 2. *In the single-server setting, the two policies Π_{Alg1} and Π_{Alg3} coincide. In other words, the proposed policy in Algorithm 1 is equivalent to the policy for which the parameters are updated at every arrival. In Sections 2.4 and 2.5, to provide intuition for the more general setting of the multi-server system, we initially discuss the case of the single-server system, and then extend our results to the multi-server setting.*

2.4 Single-server queueing model

We initially focus on the single-server Erlang-B queueing system to provide a simpler pathway to the multi-server setting. In Section 2.4.1, we first prove that asymptotic learning holds for the proposed policy Π_{Alg1} for any $\mu \in (0, +\infty)$ and valid function f. In Section 2.4.2, we evaluate the finite-time performance of our proposed policy in terms of the expected regret defined in Definition 1.

2.4.1 Asymptotic optimality

In the single-server setting, the policy Π_{Alg1} is equivalent to the policy Π_{Alg3} that updates at each arrival, so $S(n) = n$ in Algorithm 1. With this in place, we describe a stochastic process whose limiting behavior will determine the performance of our learning scheme. Define $\{\tilde{X}_n\}_{n=0}^{\infty}$ as

$$\tilde{X}_n = (X_n, N_n, \alpha_n) = \Big(\sum_{i=1}^{n} \left(g\left(T_i, M_i, c/R\right) - h\left(T_i, N_i, c/R\right) \right), N_n, \alpha_n \Big). \tag{2.9}$$

We note that the action at arrival n defined by Π_{Alg1} is uniquely determined by $\tilde{X}_n$: if the server is available and $X_n > 0$, the arrival will be accepted. Otherwise, if $X_n < 0$, the arrival will be admitted with probability p_{α_n}. To prove asymptotic optimality, we show that eventually, X_n will

always be positive for $\mu > c/R$, and negative for $\mu < c/R$. In the process $\{\tilde{X}_n\}_{n=0}^{\infty}$, X_n is updated as

$$X_n - X_{n-1} = g\left(T_n, M_n, c/R\right) - h\left(T_n, N_n, c/R\right). \tag{2.10}$$

In (2.10), random variables N_n and M_n only depend on the history through the previous state $\tilde{X}_{n-1}$ and α_n is updated from the previous state $\tilde{X}_{n-1}$ by (2.8). Thus, the stochastic process $\{\tilde{X}_n\}_{n=0}^{\infty}$ forms a Markov process (with a continuous-state component). Random variables $\{X_n - X_{n-1}\}_{n=1}^{\infty}$ are not independent since values of N_n and M_n depend on the previous state $\tilde{X}_{n-1}$. Hence, it is not straightforward to analyze the asymptotic behavior of the Markov process $\{\tilde{X}_n\}_{n=0}^{\infty}$. We will define a new stochastic process that will address this issue and establish convergence results for this process. Define $\{\beta_n\}_{n=0}^{\infty}$ as the sequence of the indices of accepted arrivals and $Y := X_{\beta_n}$. We down-sample the Markov process $\{\tilde{X}_n\}_{n=0}^{\infty}$ using sequence $\{\beta_n\}_{n=0}^{\infty}$ to get the stochastic process $\{\tilde{Y}_n\}_{n=0}^{\infty}$ given by

$$\tilde{Y}_n = \tilde{X}_{\beta_n} = \left(X_{\beta_n}, N_{\beta_n}, \alpha_{\beta_n}\right) =: \left(Y_n, 0, \alpha_{\beta_n}\right). \tag{2.11}$$

Note that $N_{\beta_n} = 0$ as the server is empty just before an arrival is accepted. To ensure process $\{\tilde{Y}_n\}_{n=0}^{\infty}$ is well-defined, in Lemma 17, we prove that the number of accepted arrivals following Π_{Alg1} is almost surely infinite; see Appendix A.1.1. Let E_n be the potential service time of arrival n; it is exponentially distributed with parameter μ and is independent of the inter-arrival times. We define l_n as the first arrival after β_n that the server is available, i.e., the service of the accepted arrival β_n is complete. Equivalently,

$$l_n = \min_m \left\{ m \geq 1 : \sum_{j=1}^{m} T_{\beta_n + j} \geq E_{\beta_n} \right\}. \tag{2.12}$$

In the following lemma, using the memoryless property of the exponentially-distributed service times and inter-arrival times, we investigate the behavior of l_n by constructing an alternate representation of the process $\{\tilde{Y}_n\}_{n=0}^{\infty}$.

Lemma 1. *Random variables $\{l_n\}_{n=0}^{\infty}$ defined in* (2.12) *are geometric, independent, and identically distributed.*

Proof of Lemma 1 is given in Appendix A.1.2. From the above observation, in Lemma 2 we prove that random variables $\{Y_n - Y_{n-1}\}_{n=1}^{\infty}$ are independent and identically distributed (*i.i.d.*).

Lemma 2. *Random variables $\{Y_n - Y_{n-1}\}_{n=1}^{\infty}$ are i.i.d..*

Proof of Lemma 2. Using (2.10), we can write $Y_n - Y_{n-1}$ as follows

$$Y_n - Y_{n-1} = \sum_{j=1}^{\beta_n - \beta_{n-1}} \left(X_{\beta_{n-1}+j} - X_{\beta_{n-1}+j-1}\right) = -\sum_{j=1}^{l_{n-1}-1} T_{\beta_{n-1}+j} + g\left(T_{\beta_{n-1}+l_{n-1}}, 1, c/R\right), \tag{2.13}$$

where we have used the fact that for $j < l_{n-1}$, $N_{\beta_{n-1}+j} = 1$ (the server is busy); otherwise, $N_{\beta_{n-1}+j} = 0$. Also, $M_{\beta_{n-1}+j} = 1$ for $j = l_{n-1}$, and 0 otherwise, as the arrival departs in inter-arrival $T_{\beta_{n-1}+l_{n-1}}$. As the server remains empty until an arrival is accepted, $M_{\beta_{n-1}+j}$ and $N_{\beta_{n-1}+j}$ are both equal to 0 for $l_{n-1}+1 \leq j \leq \beta_n - \beta_{n-1}$. Finally, since $\{l_n\}_{n=0}^{\infty}$ and $\{T_n\}_{n=0}^{\infty}$ are *i.i.d.*, Lemma 2 follows. □

Notice that $\{Y_n\}_{n=0}^{\infty}$ is the partial sums process of *i.i.d.* random variables. As a result, by the strong law of large numbers (SLLN), we observe that Y_n converges to infinity with the sign depending on $\mathbb{E}[Y_n - Y_{n-1}]$. We now present the main result of this subsection in Theorem 1, which proves the asymptotic optimality of policy Π_{Alg1} for any $\mu > 0$ in the single-server setting and argues that convergence of Y_n results in the convergence of Π_{Alg1} to the optimal policy.

Theorem 1. *Consider a single-server Erlang-B queueing system with service rate μ. For any $\mu \in (0, +\infty)$, policy Π_{Alg1} converges to the true optimal policy Π^*. Specifically, for $\mu \in (c/R, +\infty)$, we have $\lim_{n\to+\infty} Y_n = +\infty$ a.s., and the proposed policy admits all arrivals if room after a random finite time. Similarly, for $\mu \in (0, c/R)$, we have $\lim_{n\to+\infty} Y_n = -\infty$ a.s., and after a random finite time, an arrival is only accepted with a probability that converges to 0 as $n \to +\infty$.*

Proof of Theorem 1. We assume that $\mu > c/R$, and prove Theorem 1. The proof when $\mu < c/R$ follows similarly. We first find $\mathbb{E}[Y_{i+1} - Y_i]$ using (2.13) as below

$$\mathbb{E}[Y_{i+1} - Y_i] = \sum_{m=1}^{\infty} \mathbb{P}(l_i = m)\, \mathbb{E}\Big[- \sum_{j=1}^{l_i - 1} T_{\beta_i + j} + g\left(T_{\beta_i + l_i}, 1, \frac{c}{R}\right) \Big| \, l_i = m \Big]. \tag{2.14}$$

We have

$$\begin{aligned}
&\mathbb{E}\Big[g\Big(T_{\beta_i+l_i}, 1, \frac{c}{R}\Big) \Big| \, l_i = m \Big] \\
&= \frac{\mu+\lambda}{\mu} \int_{t=0}^{+\infty} \int_{x=0}^{t} g\left(t, 1, \frac{c}{R}\right) \mu \exp(-\mu x)\, \lambda \exp(-\lambda t)\, dx dt \\
&= \frac{\mu+\lambda}{\mu} \int_{t=0}^{+\infty} \lambda t \exp\left(-t\left(\lambda + \frac{c}{R}\right)\right) (1 - \exp(-\mu t)) \Big(\sum_{s=0}^{+\infty} \exp\left(-st\frac{c}{R}\right) \Big) dt \\
&= \frac{\mu+\lambda}{\mu} \Big(\sum_{s=1}^{\infty} \frac{\lambda}{\left(\lambda + s\frac{c}{R}\right)^2} - \sum_{s=1}^{\infty} \frac{\lambda}{\left(\lambda + \mu + s\frac{c}{R}\right)^2} \Big),
\end{aligned} \tag{2.15}$$

where the second line follows by $\frac{1}{1-\exp\left(-t\frac{c}{R}\right)} = \sum_{s=0}^{+\infty} \exp\left(-st\frac{c}{R}\right)$. Furthermore,

$$\mathbb{E}\left[T_{\beta_i+j} \,\middle|\, l_i = m, j < l_i \right] = \frac{\mu+\lambda}{\lambda} \int_{t=0}^{+\infty} \int_{x=t}^{+\infty} t\mu \exp(-\mu x)\, \lambda \exp(-\lambda t)\, dx\, dt = \frac{1}{\lambda+\mu}.$$

As $\mu > c/R$, using (2.14)

$$\begin{aligned}
&\mathbb{E}\left[Y_{i+1} - Y_i\right] \\
&= \frac{-1}{\lambda+\mu}\sum_{m=1}^{\infty}\mathbb{P}\left(l_i = m\right)(m-1) + \frac{\mu+\lambda}{\mu}\Big(\sum_{s=1}^{\infty}\frac{\lambda}{\left(\lambda + s\frac{c}{R}\right)^2} - \sum_{s=1}^{\infty}\frac{\lambda}{\left(\lambda+\mu+s\frac{c}{R}\right)^2}\Big) \\
&= \frac{-1}{\lambda+\mu}\frac{\lambda}{\mu} + \frac{\mu+\lambda}{\mu}\Big(\sum_{s=1}^{\infty}\frac{\lambda}{\left(\lambda + s\frac{c}{R}\right)^2} - \sum_{s=1}^{\infty}\frac{\lambda}{\left(\lambda+\mu+s\frac{c}{R}\right)^2}\Big) \\
&> \frac{-\lambda}{\mu\left(\lambda+\mu\right)} + \frac{\mu+\lambda}{\mu}\frac{\lambda}{\left(\lambda+\frac{c}{R}\right)^2} \\
&> \frac{-\lambda}{\mu\left(\lambda+\mu\right)} + \frac{\lambda}{\mu\left(\lambda+\frac{c}{R}\right)} \quad > \quad 0. \qquad (2.16)
\end{aligned}$$

From (2.16) and Lemma 2, $\{Y_{i+1} - Y_i\}_{i=0}^{\infty}$ are *i.i.d.* with $\mathbb{E}[Y_{i+1} - Y_i] > 0$. Thus, by the SLLN,

$$\lim_{n\to+\infty} Y_n = \lim_{n\to+\infty}\sum_{i=0}^{n-1}\left(Y_{i+1} - Y_i\right) = +\infty \qquad \textit{a.s.}$$

Consider the process $\{X_i\}_{i=0}^{\infty}$ from $i = \beta_n$ to $i = \beta_{n+1}$. For $l_n + 1 \le j \le \beta_{n+1} - \beta_n$, we have

$$N_{\beta_n+j} = M_{\beta_n+j} = 0$$

and in the inter-arrivals after the departure of arrival β_n, the server remains empty. Hence, for $l_n + 1 \le j \le \beta_{n+1} - \beta_n$,

$$X_{\beta_n+j} = X_{\beta_n+l_n} + \sum_{i=1}^{j}\left(g\Big(T_{\beta_n+j}, 0, \frac{c}{R}\Big) - h\Big(T_{\beta_n+j}, 0, \frac{c}{R}\Big)\right) = X_{\beta_n+l_n}.$$

Specifically, for $j = \beta_{n+1} - \beta_n$, we have

$$X_{\beta_{n+1}} = X_{\beta_n+l_n} = Y_{n+1},$$

meaning that from arrival $\beta_n + l_n$ at which the system is empty for the first time after β_n, the decision to accept or reject the following arrivals is made based on the sign of Y_{n+1}, which is eventually always positive. Thus, after a (random) finite time, the arrival is accepted whenever the server is available. □

2.4.2 Finite-time performance analysis

To study the finite-time performance of Π_{Alg1}, we characterize the regret in terms of the processes $\{\tilde{X}_n\}_{n=0}^{\infty}$ and $\{\tilde{Y}_n\}_{n=0}^{\infty}$. As the sign of $\{Y_n\}_{n=0}^{\infty}$ determines the acceptance law, we would like to upper bound the expected number of times Y_n has an undesirable sign—specifically, being non-positive when $\mu > c/R$ and non-negative in the other regime. In Lemma 2, we showed that random variables $\{Y_{n+1} - Y_n\}_{n=0}^{\infty}$ are *i.i.d.* We further show $Y_{n+1} - Y_n$ is sub-exponentially distributed in Lemma 3. The definition of a sub-exponential random variable and its properties are also given in Appendix A.1.3.

Lemma 3. *Random variables $\{Y_{n+1} - Y_n\}_{n=0}^{\infty}$ are sub-exponentially distributed.*

The intuition behind Lemma 3 is the following: from (2.13), random variable $Y_{n+1} - Y_n$ can be written as the sum of exponential random variables and a bounded random variable; see proof in Appendix A.1.3. The results of Lemmas 2 and 4 allow us to use Bernstein's concentration inequality for the sum of independent sub-exponential random variables and establish an exponentially decaying upper bound for the probability of a suboptimal action resulting from the value of Y_n.

Lemma 4. *Consider a single-server Erlang-B queueing system with service rate μ following policy Π_{Alg1}. For $\mu \in (c/R, +\infty)$, there exists a positive problem-dependent constant c_1 such that the process $\{Y_n\}_{n=0}^{\infty}$ satisfies*

$$\mathbb{P}\left(Y_n \leq 0\right) \leq \exp\left(-c_1 n\right),$$

and for any $\mu \in (0, c/R)$, for a positive problem-dependent constant c_2, the following holds

$$\mathbb{P}\left(Y_n \geq 0\right) \leq \exp\left(-c_2 n\right).$$

Lemma 4 is proved in Appendix A.1.4. We first give an upper bound for the expected regret when $\mu > c/R$. In this regime, when Y_n is positive, Π_{Alg1} follows the optimal policy Π^* and admit the arrivals (as long as there is an available server). However, for non-positive Y_n, the arrival is only admitted with a given probability. We quantify the impact of the arrival instances for which Y_n is non-positive using the exponentially decaying probability established in Lemma 4. Finally, in Theorem 2, we prove that for $\mu \in (c/R, +\infty)$, the expected regret is finite.

Theorem 2. *Consider a single-server Erlang-B queueing system with service rate μ. For any $\mu \in (c/R, +\infty)$ and (valid) function f such that $\log(f) = o(n)$, the expected regret $\mathbb{E}\left[\mathcal{R}\left(n\right); \Pi_{\text{Alg1}}\right]$ under policy Π_{Alg1} is upper bounded by a constant independent of n.*

Proof of Theorem 2. We define H_n as the number of times an arrival is rejected between arrival β_n and β_{n+1} when the server is available. Consider the system that accepts all arrivals subject to

availability, i.e., follows the optimal policy for $\mu > c/R$; call this system $Q^{(0)}$. We couple $Q^{(0)}$ with our system from the first arrival so that we can ensure whenever our system is busy, $Q^{(0)}$ is also busy and rejects the arrival. Thus, we have the following upper bound for the expected regret:

$$\mathbb{E}\left[\mathcal{R}\left(n\right);\Pi_{\text{Alg1}}\right] \leq \mathbb{E}\Big[\sum_{i=0}^{\infty} H_i\Big] = \sum_{i=0}^{\infty} \mathbb{E}\left[H_i \,\middle|\, Y_{i+1} < 0\right] \mathbb{P}\left(Y_{i+1} \leq 0\right),$$

which holds because when $Y_{i+1} > 0$, the number of rejected arrivals, H_i, is zero. Conditioned on the event $\{Y_{i+1} \leq 0\}$, H_i is geometric with parameter $1/f(\alpha_{\beta_i+l_i})$, where $\alpha_{\beta_i+l_i}$ is less than or equal to the number of admitted arrivals up to β_i+l_i, which is equal to $i+1$. Thus, using Lemma 4,

$$\mathbb{E}\left[\mathcal{R}\left(n\right);\Pi_{\text{Alg1}}\right] \leq \sum_{i=0}^{\infty} f\left(i+1\right) \mathbb{P}\left(Y_{i+1} \leq 0\right) \leq \sum_{i=0}^{\infty} f\left(i+1\right) \exp\left(-c_1\left(i+1\right)\right).$$

The above summation converges if f grows slower than the exponential function, and we conclude that the expected regret is bounded by a constant independent of n. □

Next, we present the finite-time performance guarantee for a single-server system with service rate $\mu < c/R$. In this regime, the expected regret consists of two terms. The first term arises from the arrivals for which the corresponding $Y_n > 0$, and we use the exponentially decaying probability of Lemma 4 to bound this term. The second term results from the arrivals accepted with a given probability when $Y_n \leq 0$. We will use Lemma 5 presented below to address this term; proof is given in Appendix A.1.5. In conclusion, Theorem 3 proves a polynomial in $\log(n)$ upper bound for the expected regret in the case of $\mu \in (0, c/R)$.

Lemma 5. *Let $f(n) = \exp(n^{1-\epsilon})$ and $d = \lceil 3(\log^{\frac{1}{1-\epsilon}}(n+1))\rceil$ for a fixed $\epsilon \in (0,1)$. Then, for independent geometric random variables $\{y_i\}_{i=1}^n$ with corresponding success probabilities $\{f(i)^{-1}\}_{i=1}^n$, the sum $\sum_{i=d}^{n-1} i\,\mathbb{P}(y_1+\cdots+y_i < n, y_1+\cdots+y_{i+1} \geq n)$ is bounded by a constant determined by ϵ.*

Theorem 3. *Consider a single-server Erlang-B queueing system with service rate $\mu \in (0, c/R)$. For $f(n) = \exp\left(n^{1-\epsilon}\right)$, the expected regret under policy Π_{Alg1} is $\mathbb{E}\left[\mathcal{R}\left(n\right);\Pi_{\text{Alg1}}\right] = O\big(\log^{\frac{1}{1-\epsilon}}(n)\big)$.*

Proof of Theorem 3. For $\mu \in (0, c/R)$, the optimal policy rejects all arrivals, and thus, the expected regret of the proposed policy is equal to the expected number of accepted arrivals up to arrival n,

or

$$\mathbb{E}\left[\mathcal{R}\left(n\right);\Pi_{\text{Alg1}}\right] = \mathbb{E}\Big[\sum_{i=0}^{n-1}\mathbb{1}\left\{A_i=1\right\}\Big]$$
$$= \mathbb{E}\Big[\sum_{i=0}^{n-1}\mathbb{1}\left\{A_i=1,X_i>0\right\}\Big] + \mathbb{E}\Big[\sum_{i=0}^{n-1}\mathbb{1}\left\{A_i=1,X_i\leq 0\right\}\Big],$$

where X_i is the first component of the state of the Markov chain defined in (2.9). Using the sampled process $\{\tilde{Y}_n\}_{n=0}^{\infty}$ and Lemma 4, we simplify the first term on the RHS of the above equation as,

$$\mathbb{E}\Big[\sum_{i=0}^{n-1}\mathbb{1}\left\{A_i=1,X_i>0\right\}\Big] \leq \sum_{i=0}^{+\infty}\mathbb{P}\left(Y_i>0\right) \leq \sum_{i=1}^{+\infty}\exp\left(-c_2 i\right) < \infty. \tag{2.17}$$

We next upper bound the expected number of arrivals accepted when $X_i \leq 0$. We consider a system that has infinite servers (to avoid rejecting arrivals) and regardless of the sign of X_i, accepts with probability $1/f(i)$, if i arrivals have already been accepted (the acceptance rule is compatible with the original system when $X_i \leq 0$). By coupling this system with the system following Algorithm 1, taking $d = \lceil 3\log^{\frac{1}{1-\epsilon}}(n)\rceil$ and $\{y_i\}_{i=1}^{n}$ as defined in Lemma 5,

$$\mathbb{E}\Big[\sum_{i=0}^{n-1}\mathbb{1}\left\{A_i=1,X_i\leq 0\right\}\Big] \leq \mathbb{E}\Big[\sum_{i=0}^{d-1}\mathbb{1}\left\{A_i=1,X_i\leq 0\right\}\Big] + \mathbb{E}\Big[\sum_{i=d}^{n-1}\mathbb{1}\left\{A_i=1,X_i\leq 0\right\}\Big]$$
$$\leq d + \sum_{i=d}^{n-1} i\,\mathbb{P}\left(y_1+\cdots+y_i<n, y_1+\cdots+y_{i+1}\geq n\right).$$

Finally, By Lemma 5 and (2.17), for $\mu \in (0, c/R)$, the expected regret is bounded by a polylogarithmic function. □

Remark 3. *There is an exploration-exploitation trade-off in selecting $f(n)$ on the two sides of $\mu = c/R$. When admitting is optimal, we want $f(n)$ to increase to infinity as slow as possible. Also, based on the proof of Theorem 2, for our current bound, we cannot take $f(n)$ to grow exponentially fast since its exponent needs to depend on unknown μ to ensure constant regret. Conversely, when blocking all arrivals is optimal, we need $f(n)$ to converge to infinity as fast as possible. As the learning algorithm needs to be agnostic about the parameter regime, $f(n) = \exp\left(n^{1-\epsilon}\right)$ is a good choice: it ensures constant regret in one regime and polynomial regret in $\log(n)$ in the other.*

We next consider a decreasing sequence of ϵ values by choosing $\epsilon_n := \frac{\varepsilon}{\sqrt{1+\log(n+1)}}$ for $n \geq 1$, where $\varepsilon \in (0,1)$. The algorithm corresponding to the exploration function $f(n) = \exp\left(n^{1-\epsilon_n}\right)$ is asymptotically optimal from Theorem 1. To determine the regret when $\mu > c/R$, we observe that

$\log(f) = o(n)$ and the regret in this regime remains finite. For the case of $\mu < c/R$, we are able to reduce the order of regret further to $\log(n)$, as shown in Corollary 1 with proof in Appendix A.1.6.

Corollary 1. *Consider a single-server Erlang-B queueing system with service rate $\mu \in (0, c/R)$. For $f(n) = \exp\left(n^{1-\epsilon_n}\right)$ where $\epsilon_n = \frac{\varepsilon}{\sqrt{1+\log(n+1)}}$ for all $n \geq 1$ and $\varepsilon \in (0,1)$, the expected regret under policy Π_{Alg1} is $\mathbb{E}\left[\mathcal{R}\left(n\right); \Pi_{\text{Alg1}}\right] = O\big(\log(n)\big)$.*

Remark 4. *For some parameters, our problem setting overlaps with the setting of [108]: when $\mu \leq c/R$, our setting can be viewed as learning in an $M/M/1$ system with the optimal admission threshold of 0, and when $c/R < \mu \leq h(\lambda, c/R) < +\infty$ (for a function h), our setting corresponds to an $M/M/1$ system with an optimal threshold of 1. However, our work samples the system only at arrivals, in contrast to [108] which samples the system at all times (so service times of departed jobs are known). Despite observing less information, our proposed policy exhibits the same regret behavior as [108] as shown in Corollary 1 and Theorem 2.*

2.5 Multi-server queueing model

In this section, we extend the results of Section 2.4 to a multi-server setting. In Section 2.5.1, the convergence of Π_{Alg1} to the optimal policy is shown by a martingale-based analysis coupled with SLLN for martingale sequences. Moreover, in Section 2.5.2, we prove that the regret bounds of the single-server model extend to the multi-server setting using martingale concentration inequalities.

2.5.1 Asymptotic optimality

First, we extend the processes defined in Section 2.4.1 to the multi-server setting. We define the stochastic process $\{\tilde{X}_n\}_{n=0}^{\infty}$ as

$$\tilde{X}_n = (X_n, N_n, \alpha_n) = \Big(\sum_{i=1}^{n} \Big(g\left(T_i, M_i, c/R\right) - h\left(T_i, N_i, c/R\right) \Big), N_n, \alpha_n \Big). \tag{2.18}$$

As in the single-server case, $\{\tilde{X}_n\}_{n=0}^{\infty}$ forms a Markov process (with a continuous-state component). Our goal is to down-sample the Markov process $\{\tilde{X}_n\}_{n=0}^{\infty}$ at arrival acceptances for which the system is empty and establish convergence results for the resulting process. Similar to the single-server case, we first argue that these instances happen infinitely often (almost surely) in Appendix A.2.1. Let $\{\beta_n\}_{n=0}^{\infty}$ be the sequence of the indices of accepted arrivals when the system is empty. We down-sample the Markov process $\{\tilde{X}_n\}_{n=0}^{\infty}$ using the sequence $\{\beta_n\}_{n=0}^{\infty}$ to get process

$\{\tilde{Y}_n\}_{n=0}^{\infty}$ where Y_n is defined as $Y_n := X_{\beta_n}$, and

$$\tilde{Y}_n := \tilde{X}_{\beta_n} := (X_{\beta_n}, N_{\beta_n}, \alpha_{\beta_n}) = (Y_n, 0, \alpha_{\beta_n})\,. \tag{2.19}$$

In the single-server setting, the processes depicted in (2.18) and (2.19) are equivalent to the processes defined in (2.9) and (2.11), as the single-server system is empty whenever an arrival is accepted. In contrast to the single-server case, random variables $\{Y_n - Y_{n-1}\}_{n=1}^{\infty}$ are not independent in the multi-server setting, as here, unlike the single-server setting, $Y_n - Y_{n-1}$ depends on the acceptance probabilities. We will argue that process $\{Y_n\}_{n=0}^{\infty}$ is a submartingale (or supermartingale), and using this result, we will analyze its convergence. We define random variable D_i as the change in X_i at inter-arrival T_i, i.e., $D_i := X_i - X_{i-1}$. Next, for any $n \geq 0$, we define process $\{W_{n,m}\}_{m=0}^{\infty}$ as

$$W_{n,m} = Y_n + \sum_{i=1}^{m} D_{\beta_n+i} = X_{\beta_n+m}. \tag{2.20}$$

We define the random variable τ_n as the index of the first arrival after β_n that finds the system empty, i.e.,

$$\tau_n = \min\{i \geq 1 : N_{\beta_n+i} = 0\}\,.$$

Note that by (2.20), $W_{n,\tau_n} = X_{\beta_n+\tau_n}$. We claim that process $\{X_n\}_{n=0}^{\infty}$ at the first arrival acceptance after τ_n, i.e., $X_{\beta_{n+1}}$, is equal to W_{n,τ_n}. Indeed, process $\{X_n\}_{n=0}^{\infty}$ does not change when there are no departures or ongoing services. Hence,

$$W_{n,0} = Y_n, \qquad\qquad W_{n,\tau_n} = X_{\beta_{n+1}} = Y_{n+1}, \tag{2.21}$$

and random variable Y_{n+1} is equivalent to the process $\{W_{n,m}\}_{m=0}^{\infty}$ stopped at τ_n. Thus, to analyze the convergence of process $\{Y_n\}_{n=0}^{\infty}$, we study the properties of process $\{W_{n,m}\}_{m=0}^{\infty}$ and random variable τ_n for $n \geq 1$. We determine the behavior of τ_n by coupling the system that runs Algorithm 1 with a system that accepts all arrivals (subject to availability) as follows.

Let $Q^{(n)}$ denote the system that accepts all arrivals as long as it has at least one available server. We also define random variable ζ_n as the first arrival after arrival β_n that finds $Q^{(n)}$ empty, starting from an empty state. Starting from arrival β_n, we couple this system with the system that follows Algorithm 1 such that at each arrival, the number of busy servers in $Q^{(n)}$ is greater than or equal to our system. We couple the arrival sequences in both systems such that the inter-arrival times are equal. Moreover, when an arrival is accepted in both systems, we assume that its service time is identical in both. System $Q^{(n)}$ will accept all arrivals unless none of its servers are available. Suppose all of the servers of $Q^{(n)}$ are busy, and our system accepts an arrival. In this case, we assume that the service time of the accepted arrival in our system equals the remaining service

time of the k^{th} server in $Q^{(n)}$, which has an exponential distribution with parameter μ due to the memoryless property. Using this coupling, we verify that all moments of τ_n are finite in Lemma 6.

Lemma 6. *All moments of random variable τ_n are bounded by a constant independent of n.*

Proof of Lemma 6. By the above coupling of $Q^{(n)}$ with the system that follows our proposed policy, we ensure that at each arrival, the number of busy servers in $Q^{(n)}$ is greater than or equal to our system. Hence, the moments of τ_n are bounded by the moments of ζ_n. In system $Q^{(n)}$, the number of busy servers just before each arrival forms a finite-state irreducible Markov chain, and random variable ζ_n is the first passage time of the state zero starting from zero, and has moments bounded by a constant which only depends on λ, μ and the number of servers. □

After characterizing the behavior of τ_n, in Lemma 7, we show that the process $\{W_{n,m}\}_{m=0}^{\infty}$ is a submartingale or supermartingale depending on the sign of $\mu - c/R$.

Lemma 7. *Fix $n \geq 0$. For $\mu \in (c/R, +\infty)$, the stochastic process $\{W_{n,m}\}_{m=0}^{\infty}$ forms a submartingale sequence with respect to the filtration $\{\mathcal{G}_{n,m}\}_{m=0}^{\infty}$, wherein the σ-algebra $\mathcal{G}_{n,m}$ is defined as*

$$\mathcal{G}_{n,m} := \sigma\left(T_{\beta_n+1}, \ldots, T_{\beta_n+m}, N_{\beta_n+1}, \ldots, N_{\beta_n+m}, \alpha_{\beta_n}, \ldots, \alpha_{\beta_n+m}, A_{\beta_n+1}, \ldots, A_{\beta_n+m}, Y_n\right).$$

For $\mu \in (0, c/R)$, the process $\{W_{n,m}\}_{m=0}^{\infty}$ is a supermartingale with respect to filtration $\{\mathcal{G}_{n,m}\}_{m=0}^{\infty}$.

Proof of Lemma 7. We show the proof for the case of $\mu > c/R$. The other region follows similarly. To prove $\{W_{n,m}\}_{m=0}^{\infty}$ is a submartingale sequence, we first show $\mathbb{E}\left[|W_{n,m}|\right] < \infty$. From (2.20),

$$\begin{aligned}
\mathbb{E}\left[|W_{n,m}|\right] &\leq \mathbb{E}\Big[\,|Y_n| + \sum_{i=1}^{m} |D_{\beta_n+i}|\Big] \\
&\leq \mathbb{E}\Big[|Y_n| + \sum_{i=1}^{m} \Big|g\big(T_{\beta_n+i}, M_{\beta_n+i}, \frac{c}{R}\big) - h\big(T_{\beta_n+i}, N_{\beta_n+i}, \frac{c}{R}\big)\Big|\Big] \\
&\leq \mathbb{E}\left[|Y_n|\right] + k\sum_{i=1}^{m}\Big(\mathbb{E}\left[g\big(T_{\beta_n+i}, 1, \frac{c}{R}\big)\right] + \mathbb{E}\left[T_{\beta_n+i}\right]\Big), \qquad (2.22)
\end{aligned}$$

where (2.22) holds as $0 \leq M_{\beta_n+i}, N_{\beta_n+i} \leq k$. For $t > 0$, we have $g\left(t, 1, x\right) \leq \frac{1}{x}$, and thus, the summation in (2.22) is finite. To show that $\mathbb{E}\left[|Y_n|\right] < \infty$, it suffices to show $\mathbb{E}\left[|Y_{n+1} - Y_n|\right]$ is

finite for all n:

$$\begin{aligned}
\mathbb{E}\left[|Y_{n+1} - Y_n|\right] &= \mathbb{E}\left[|W_{n,\tau_n} - Y_n|\right] \\
&= \mathbb{E}\Big[\Big|\sum_{i=1}^{\tau_n} D_{\beta_n+i}\Big|\Big] \\
&\leq k\mathbb{E}\Big[\sum_{i=1}^{\tau_n}\Big(T_{\beta_n+i} + g\Big(T_{\beta_n+i}, 1, \frac{c}{R}\Big)\Big)\Big] \qquad (2.23) \\
&\leq k\mathbb{E}\Big[\sum_{i=1}^{\zeta_n}\Big(T_{\beta_n+i} + g\Big(T_{\beta_n+i}, 1, \frac{c}{R}\Big)\Big)\Big] \\
&= k\mathbb{E}\left[\zeta_n\right]\mathbb{E}\left[T_{\beta_n+1} + g\Big(T_{\beta_n+1}, 1, \frac{c}{R}\Big)\right], \qquad (2.24)
\end{aligned}$$

where (2.23) is derived similar to (2.22) and (2.24) follows from coupling $Q^{(n)}$ with the system that runs Algorithm 1. Hitting time ζ_n is a stopping time for the finite-state irreducible Markov chain found by sampling $Q^{(n)}$ at arrivals and $\mathbb{E}[\zeta_n] < \infty$. Also, $\{T_{\beta_n+i}\}_{i=1}^{\infty}$'s are independent and identically distributed. Hence, (2.24) follows from Wald's equation ([28]), and $\mathbb{E}\left[|Y_{n+1} - Y_n|\right] < \infty$, which implies that $\mathbb{E}\left[|Y_n|\right] < \infty$, and by (2.22), $\mathbb{E}\left[|W_{n,m}|\right] < \infty$. We next verify the submartingale property of $\{W_{n,m}\}_{m=0}^{\infty}$. From the Markov property of $\{\tilde{X}_n\}_{n=0}^{\infty}$,

$$\mathbb{E}\left[W_{n,m+1} - W_{n,m}\,\middle|\,\mathcal{G}_{n,m}\right] = \mathbb{E}\left[X_{\beta_n+m+1} - X_{\beta_n+m}\,\middle|\,X_{\beta_n+m}, N_{\beta_n+m}, \alpha_{\beta_n+m}, A_{\beta_n+m}\right], \quad (2.25)$$

which is equal to the expected change in X_i during inter-arrival T_{β_n+m+1}. To show $\mathbb{E}[W_{n,m+1} - W_{n,m}\,\big|\,\mathcal{G}_{n,m}] \geq 0$, we argue that $\mathbb{E}[X_{i+1} - X_i\,\big|\,X_i, N_i, \alpha_i, A_i]$ is non-negative for all i as follows,

$$\begin{aligned}
&\mathbb{E}\big[X_{i+1} - X_i\big|X_i, N_i, \alpha_i, A_i\big] \\
&= \mathbb{E}\big[g\big(T_{i+1}, N_i + A_i - N_{i+1}, \frac{c}{R}\big) - h\big(T_{i+1}, N_{i+1}, \frac{c}{R}\big)\big|N_i, A_i\big] \\
&= \mathbb{E}\big[\big(N_i + A_i - N_{i+1}\big)g\big(T_{i+1}, 1, \frac{c}{R}\big)\big|N_i, A_i\big] - \mathbb{E}\left[T_{i+1}N_{i+1}\,\middle|\,N_i, A_i\right] \\
&= \big(N_i + A_i\big)\mathbb{E}\big[g\big(T_{i+1}, 1, \frac{c}{R}\big)\big] - \mathbb{E}\big[N_{i+1}g\big(T_{i+1}, 1, \frac{c}{R}\big)\big|N_i, A_i\big] - \big(N_i + A_i\big)\mathbb{E}\left[T_{i+1}\mathbb{1}_A\right], \\
&\qquad (2.26)
\end{aligned}$$

where A is the event that a fixed server from the $N_i + A_i$ busy servers remains busy during inter-

arrival T_{i+1}. The second term of (2.26) can be simplified as follows

$$\begin{aligned}\mathbb{E}\big[N_{i+1}g\big(T_{i+1},1,\frac{c}{R}\big)\,\Big|\,N_i,A_i\big] &= (N_i+A_i)\,\mathbb{E}\big[g\big(T_{i+1},1,\frac{c}{R}\big)\mathbb{1}_A\big]\\ &= (N_i+A_i)\int_{t=0}^{+\infty}\frac{t\exp\left(-t\frac{c}{R}\right)}{1-\exp\left(-t\frac{c}{R}\right)}\lambda\exp\left(-\lambda t\right)\exp\left(-\mu t\right)dt\\ &= (N_i+A_i)\sum_{j=0}^{\infty}\frac{\lambda}{\left(\lambda+\mu+(j+1)\,\frac{c}{R}\right)^2},\end{aligned} \tag{2.27}$$

We derive $\mathbb{E}\left[g\left(T_{i+1},1,c/R\right)\right]$ using the same calculations as in (2.15),

$$\mathbb{E}\big[g\big(T_{i+1},1,\frac{c}{R}\big)\big] = \int_{t=0}^{+\infty}\frac{t\exp\left(-t\frac{c}{R}\right)}{1-\exp\left(-t\frac{c}{R}\right)}\lambda\exp\left(-\lambda t\right)dt = \sum_{j=0}^{\infty}\frac{\lambda}{\left(\lambda+(j+1)\,\frac{c}{R}\right)^2}. \tag{2.28}$$

Next, we simplify the third term of (2.26):

$$\begin{aligned}(N_i+A_i)\,\mathbb{E}\left[T_{i+1}\mathbb{1}_A\right] &= (N_i+A_i)\int_{t=0}^{+\infty}\int_{x=t}^{+\infty}t\mu\exp\left(-\mu x\right)\lambda\exp\left(-\lambda t\right)dxdt\\ &= (N_i+A_i)\,\frac{\lambda}{\left(\lambda+\mu\right)^2}.\end{aligned} \tag{2.29}$$

Substituting the terms found in the above equation, (2.28), and (2.27), in Equation (2.26), we have

$$\mathbb{E}\big[X_{i+1}-X_i\,\Big|\,X_i,N_i,\alpha_i,A_i\big] = \tilde{\delta}\left(N_i+A_i\right), \tag{2.30}$$

where

$$\tilde{\delta} := -\frac{\lambda}{\left(\lambda+\mu\right)^2}+\sum_{j=0}^{\infty}\frac{\lambda}{\left(\lambda+(j+1)\,\frac{c}{R}\right)^2}-\frac{\lambda}{\left(\lambda+\mu+(j+1)\,\frac{c}{R}\right)^2}.$$

and is positive for $\mu\in(c/R,+\infty)$. Hence, from (2.25),

$$\mathbb{E}\left[W_{n,m+1}-W_{n,m}\,\Big|\,\mathcal{G}_{n,m}\right] = \tilde{\delta}\left(N_{\beta_n+m}+A_{\beta_n+m}\right)\geq 0, \tag{2.31}$$

and we conclude that $\{W_{n,m}\}_{m=0}^{\infty}$ is a submartingale sequence with respect to $\{\mathcal{G}_{n,m}\}_{m=0}^{\infty}$. □

Next, in Proposition 1 we argue that the stopped sequence $\{W_{n,\tau_n}\}_{n=0}^{\infty}$ or $\{Y_n\}_{n=0}^{\infty}$ also forms a submartingale or supermartingale sequence depending on the problem parameters.

Proposition 1. *Sequence $\{Y_n\}_{n=0}^{\infty}$ forms a submartingale or supermartingale (depending on the sign of $\mu-c/R$) with respect to filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$ defined as*

$$\mathcal{F}_n = \sigma\left(Y_0,\ldots,Y_n,\alpha_{\beta_0},\ldots,\alpha_{\beta_n}\right).$$

Specifically, $\{Y_n\}_{n=0}^{\infty}$ is a submartingale sequence if $\mu > c/R$ and a supermartingale sequence otherwise.

Proof of Proposition 1. We show the proof for the case of $\mu > \frac{c}{R}$, and the other regime follows similarly. Note that Y_{n+1} is equal to submartingale $\{W_{n,m}\}_{m=0}^{\infty}$ stopped at τ_n; in other words,

$$Y_{n+1} = W_{n,\tau_n} = Y_n + \sum_{i=1}^{\tau_n} D_{\beta_n+i}.$$

In Lemma 6, we argued that $\mathbb{E}[\tau_n] < \infty$. Moreover,

$$\begin{aligned}\mathbb{E}\left[|W_{n,m+1} - W_{n,m}| \,\middle|\, \mathcal{G}_{n,m}\right] &= \mathbb{E}\left[|D_{\beta_n+m+1}| \,\middle|\, \mathcal{G}_{n,m}\right] \\ &\leq k\mathbb{E}\Big[g\Big(T_{\beta_n+1}, 1, \frac{c}{R}\Big)\Big] + k\mathbb{E}\left[T_{\beta_n+1}\right]. \end{aligned} \tag{2.32}$$

As g is bounded, the RHS of (2.32) is also finite. Hence, we can use Doob's optional stopping theorem [28, Theorem 4.8.5] for submartingale $\{W_{n,m}\}_{m=0}^{\infty}$ and stopping time τ_n with a finite expected value to get

$$\mathbb{E}\big[Y_{n+1} \,\big|\, \mathcal{G}_{n,0}\big] = \mathbb{E}\big[W_{n,\tau_n} \,\big|\, \mathcal{G}_{n,0}\big] \geq \mathbb{E}\big[W_{n,0} \,\big|\, \mathcal{G}_{n,0}\big] = Y_n.$$

Thus, we have

$$\mathbb{E}\big[Y_{n+1} - Y_n \,\big|\, \mathcal{G}_{n,0}\big] = \mathbb{E}\big[Y_{n+1} - Y_n \,\big|\, \mathcal{F}_n\big] \geq 0.$$

As $\mathbb{E}[|Y_n|]$ is finite, $\{Y_n\}_{n=0}^{\infty}$ is a submartingale sequence with respect to $\{\mathcal{F}_n\}_{n=0}^{\infty}$. □

Now that we proved the submartingale (or supermartingale) property of $\{Y_n\}_{n=0}^{\infty}$, we can examine the convergence of this process. We will achieve this using Doob's decomposition and studying the convergence of the derived martingale and the predictable sequence. From Proposition 1 and Doob's decomposition of $\{Y_n\}_{n=0}^{\infty}$, we have

$$Y_n = Y_n^A + Y_n^M, \tag{2.33}$$

where Y_n^M is a martingale sequence, and Y_n^A is a predictable and almost surely increasing (or decreasing) sequence with $Y_0^A = 0$. In Lemmas 8 and 9, we examine the limiting behavior of sequences $\{Y_n^A\}_{n=0}^{\infty}$ and $\{Y_n^M\}_{n=0}^{\infty}$. The basic idea is to show that $\{Y_n^A\}_{n=0}^{\infty}$ converges to infinity, and $\{Y_n^M\}_{n=0}^{\infty}$ is well-behaved in a way that their sum, $\{Y_n\}_{n=0}^{\infty}$, converges to infinity.

Lemma 8. *For $\mu \in (c/R, +\infty)$, there exists a positive problem-dependent constant $\tilde{\delta}_1$ such that*

the predictable increasing process $\{Y_n^A\}_{n=0}^{\infty}$ from Doob's decomposition of $\{Y_n\}_{n=0}^{\infty}$ satisfies

$$Y_n^A \geq \tilde{\delta}_1 n \quad \text{a.s.}$$

Similarly, for $\mu \in (0, c/R)$, there exists a negative problem-dependent constant $\tilde{\delta}_2$ such that the predictable decreasing process $\{Y_n^A\}_{n=0}^{\infty}$ satisfies

$$Y_n^A \leq \tilde{\delta}_2 n \quad \text{a.s.}$$

Proof of Lemma 8. WLOG, we assume $\mu \in (c/R, +\infty)$. By Proposition 1, sequence $\{Y_n\}_{n=0}^{\infty}$ is a submartingale with respect to filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$. Hence, the increasing sequence is given as below

$$Y_n^A = \sum_{m=0}^{n-1} \mathbb{E}\left[Y_{m+1} - Y_m \,\middle|\, \mathcal{F}_m\right] = \sum_{m=0}^{n-1} \left(\mathbb{E}\left[W_{m,\tau_m} \,\middle|\, \mathcal{F}_m\right] - Y_m\right). \tag{2.34}$$

In Lemma 7, we argued $\{W_{n,m}\}_{m=0}^{\infty}$ is a submartingale with respect to $\{\mathcal{G}_{n,m}\}_{m=0}^{\infty}$. From Doob's decomposition, we get

$$W_{n,m} = W_{n,m}^A + W_{n,m}^M.$$

For the predictable process $\{W_{n,m}^A\}_{m=0}^{\infty}$, from (2.31),

$$W_{n,m}^A = \sum_{i=0}^{m-1} \mathbb{E}\left[W_{n,i+1} - W_{n,i} \,\middle|\, \mathcal{G}_{n,i}\right] = \sum_{i=0}^{m-1} \tilde{\delta}\left(N_{\beta_n+i} + A_{\beta_n+i}\right). \tag{2.35}$$

Next, we use Doob's optional stopping theorem for the martingale sequence $\{W_{n,m}^M\}_{m=0}^{\infty}$ to find $\mathbb{E}\left[W_{n,\tau_n}^M \,\middle|\, \mathcal{F}_n\right]$. The stopping time τ_n has finite expectation as argued in Lemma 6, and

$$\begin{aligned}
\mathbb{E}\left[\left|W_{n,i+1}^M - W_{n,i}^M\right| \,\middle|\, \mathcal{G}_{n,i}\right] &= \mathbb{E}\left[\left|W_{n,i+1} - W_{n,i} - \left(W_{n,i+1}^A - W_{n,i}^A\right)\right| \,\middle|\, \mathcal{G}_{n,i}\right] \\
&= \mathbb{E}\left[\left|D_{\beta_n+i+1} - \mathbb{E}\left[D_{\beta_n+i+1} \,\middle|\, \mathcal{G}_{n,i}\right]\right] \,\middle|\, \mathcal{G}_{n,i}\right] \\
&\leq \mathbb{E}\left[\left|2D_{\beta_n+i+1}\right| \,\middle|\, \mathcal{G}_{n,i}\right], \qquad (2.36)
\end{aligned}$$

where (2.36) is bounded by a constant, as argued in (2.32). After verifying the conditions of the optional stopping theorem, we are able to use this theorem to get

$$\mathbb{E}\big[W_{n,\tau_n}^M \,\big|\, \mathcal{F}_n\big] = \mathbb{E}\big[W_{n,0}^M \,\big|\, \mathcal{F}_n\big] = Y_n. \tag{2.37}$$

From (2.34) and (2.35), we can find Y_n^A as follows

$$Y_n^A = \tilde{\delta} \sum_{m=0}^{n-1} \mathbb{E}\Big[\sum_{i=0}^{\tau_m - 1} (N_{\beta_m+i} + A_{\beta_m+i}) \,\Big|\, \mathcal{F}_m \Big]. \tag{2.38}$$

Note that $A_{\beta_m} = 1$, as arrival β_n is accepted by the definition of the sampling times $\{\beta_n\}_{n=0}^{\infty}$. Hence, $\mathbb{E}\Big[\sum_{i=0}^{\tau_m-1} (N_{\beta_m+i} + A_{\beta_m+i}) \,\Big|\, \mathcal{F}_m \Big] \geq 1$, which gives $Y_n^A \geq \tilde{\delta} n$. □

We next state the strong law of large numbers for martingale sequences in Theorem 4 and then, using this result, prove Lemma 9.

Theorem 4. *[86, Corollary 7.3.2] let $\{M_n\}_{n=0}^{\infty}$ be a martingale sequence with $M_0 = 0$ and $\mathbb{E}\big[\, |M_n|^{2r} \,\big] < \infty$ for some $r \geq 1$, and it satisfies $\sum_{n=1}^{\infty} n^{-(1+r)} \mathbb{E}\left[|M_n - M_{n-1}|^{2r} \right] < \infty$. Then, we have the strong law of large numbers for martingales, which states that*

$$\lim_{n\to\infty} \frac{M_n}{n} = 0. \quad \text{a.s.}$$

Lemma 9. *The martingale process $\{Y_n^M\}_{n=0}^{\infty}$ found by Doob's decomposition of $\{Y_n\}_{n=0}^{\infty}$ satisfies*

$$\lim_{n\to\infty} \frac{Y_n^M}{n} = 0. \quad \text{a.s.}$$

Proof of Lemma 9. We prove Lemma 9 for $\mu > c/R$. The result in the other region is proved similarly. We first derive upper and lower bounds for the martingale difference sequence $Y_{n+1}^M - Y_n^M$. We have

$$\begin{aligned}
&Y_{n+1}^M - Y_n^M \\
&= Y_{n+1} - Y_n - \big(Y_{n+1}^A - Y_n^A\big) \\
&= \sum_{i=1}^{\tau_n} D_{\beta_n+i} - \mathbb{E}\Big[\tilde{\delta} \sum_{i=0}^{\tau_n - 1} (N_{\beta_n+i} + A_{\beta_n+i}) \,\Big|\, \mathcal{F}_n \Big] && (2.39) \\
&= \sum_{i=1}^{\tau_n} \Big(g\Big(T_{\beta_n+i}, M_{\beta_n+i}, \frac{c}{R}\Big) - h\Big(T_{\beta_n+i}, N_{\beta_n+i}, \frac{c}{R}\Big)\Big) - \mathbb{E}\Big[\tilde{\delta} \sum_{i=0}^{\tau_n-1} (N_{\beta_n+i} + A_{\beta_n+i}) \,\Big|\, \mathcal{F}_n \Big], && (2.40)
\end{aligned}$$

where (2.39) is true by (2.38), and (2.40) follows from the definition of D_i. To derive an upper bound for the martingale difference sequence, we only consider the non-negative terms in (2.40) as below

$$Y_{n+1}^M - Y_n^M \leq \sum_{i=1}^{\tau_n} g\Big(T_{\beta_n+i}, M_{\beta_n+i}, \frac{c}{R}\Big) \leq k \frac{R}{c} \tau_n, \tag{2.41}$$

which holds as for $t > 0$, we have $g(t, 1, x) \leq \frac{1}{x}$. To find a lower bound, using the non-positive terms,

$$\begin{aligned} Y_{n+1}^M - Y_n^M &\geq -\sum_{i=1}^{\tau_n} h\Big(T_{\beta_n+i}, N_{\beta_n+i}, \frac{c}{R}\Big) - \mathbb{E}\Big[\tilde{\delta} \sum_{i=0}^{\tau_n - 1} (N_{\beta_n+i} + A_{\beta_n+i}) \,\Big|\, \mathcal{F}_n\Big] \\ &\geq -k \sum_{i=1}^{\tau_n} T_{\beta_n+i} - \tilde{\delta} k \mathbb{E}\left[\tau_n \,\middle|\, \mathcal{F}_n\right], \end{aligned} \tag{2.42}$$

where we have used the definition of function h. From Lemma 6, $\tilde{\delta} k \mathbb{E}\left[\tau_n \,\middle|\, \mathcal{F}_n\right]$ is bounded by a constant, which we call $c_{\tilde{\delta}}$. By (2.41) and (2.42), we have

$$-k \sum_{i=1}^{\tau_n} T_{\beta_n+i} - c_{\tilde{\delta}} \leq Y_{n+1}^M - Y_n^M \leq k \frac{R}{c} \tau_n. \tag{2.43}$$

We next verify the conditions of Theorem 4 for the martingale sequence Y_n^M with $r = 1$. From (2.43),

$$\mathbb{E}\big[\,\big(Y_{n+1}^M - Y_n^M\big)^2\,\big] \leq k^2 \frac{R^2}{c^2} \mathbb{E}\left[\tau_n^2\right] + k^2 \mathbb{E}\Big[\Big(\sum_{i=1}^{\tau_n} T_{\beta_n+i}\Big)^2\Big] + 2kc_{\tilde{\delta}} \mathbb{E}\Big[\sum_{i=1}^{\tau_n} T_{\beta_n+i}\Big] + c_{\tilde{\delta}}^2. \tag{2.44}$$

We aim to show the right-hand side of (2.44) is bounded by a constant independent of n. From Wald's equation [28, Theorem 4.8.6], we have that $\mathbb{E}\left[\sum_{i=1}^{\tau_n} T_{\beta_n+i}\right]$ is bounded by a constant. For the second term, we use Wald's second equation [28, Exercise 4.8.4] for *i.i.d.* random variables $\{\tilde{T}_i\}_{i=1}^n$ defined as $\tilde{T}_i := T_{\beta_n+i} - \frac{1}{\lambda}$, with $\mathbb{E}[\tilde{T}_i] = 0$ for all i. We take $\tilde{S}_n := \sum_{i=1}^n \tilde{T}_i$. From Wald's second equation, for stopping time τ_n with finite expectation,

$$\mathbb{E}\big[\tilde{S}_{\tau_n}^2\big] = \frac{1}{\lambda^2} \mathbb{E}\left[\tau_n\right].$$

In addition, from the definition of $\tilde{S}_n$, we have

$$\mathbb{E}\big[\tilde{S}_{\tau_n}^2\big] = \mathbb{E}\big[\big(\sum_{i=1}^{\tau_n} T_{\beta_n+i} - \frac{\tau_n}{\lambda}\big)^2\big].$$

Finally, we bound the second term on the right-hand side of (2.44) with a constant as below

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{i=1}^{\tau_n} T_{\beta_n+i}\Big)^2\Big] \\
&= \frac{1}{\lambda^2}\mathbb{E}[\tau_n] + \frac{2}{\lambda}\mathbb{E}\Big[\tau_n \sum_{i=1}^{\tau_n} T_{\beta_n+i}\Big] - \frac{1}{\lambda^2}\mathbb{E}\Big[\tau_n^2\Big] \\
&\leq \frac{1}{\lambda^2}\mathbb{E}[\tau_n] + \frac{1}{\lambda}\mathbb{E}\Big[\sum_{i=1}^{\tau_n} 2\tau_n T_{\beta_n+i}\Big] \leq \frac{1}{\lambda^2}\mathbb{E}[\tau_n] + \frac{1}{\lambda}\mathbb{E}\Big[\sum_{i=1}^{\tau_n} T_{\beta_n+i}^2\Big] + \frac{1}{\lambda}\mathbb{E}\Big[\tau_n^3\Big]. \qquad (2.45)
\end{aligned}$$

The last line uses inequality $2xy \leq x^2 + y^2$. We argued that the moments of τ_n are bounded by the moments of the first hitting time to 0 of a finite-state irreducible Markov chain found by sampling system $Q^{(n)}$, or ζ_n, and thus, are finite. Hence, the first and third terms of (2.45) are bounded by a constant. By Wald's equation, the second term is also bounded by a constant. In conclusion, (2.45) is bounded by a constant independent of n. Similarly, the first term on the right-hand side of (2.44) is also bounded by a constant. Now, we verify the condition of Theorem 4 as follows

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}\big[\big(Y_n^M - Y_{n-1}^M\big)^2\big]}{n^2} \leq c_5 \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

and the conditions of Theorem 4 are satisfied. Thus, by Theorem 4,

$$\lim_{n\to+\infty} \frac{Y_n^M}{n} = 0 \quad \textit{a.s.}$$

□

We now state Theorem 5, which generalizes Theorem 1 to the multi-server setting and proves the asymptotic optimality of our proposed policy for the multi-server queueing system. The proof of this theorem is based on the submartingale (or supermartingale) property of the sequence $\{Y_n\}_{n=0}^{\infty}$.

Theorem 5. *Consider the multi-server Erlang-B queueing system with k servers and service rate μ. For any $\mu \in (0, +\infty)$, policy Π_{Alg1} converges to the true optimal policy Π^*. Specifically, for $\mu \in (c/R, +\infty)$, Y_n converges to $+\infty$* a.s. *and the proposed policy admits all arrivals after a random finite time subject to availability. Similarly, for $\mu \in (0, c/R)$, Y_n converges to $-\infty$* a.s.*, and after a random finite time, an arrival is only accepted with a probability that converges to 0 as $n \to +\infty$.*

Proof of Theorem 5. For $\mu \in (c/R, +\infty)$, by Doob's decomposition for submartingale $\{Y_n\}_{n=0}^{\infty}$

and Lemmas 8 and 9,

$$\lim_{n\to+\infty} Y_n = +\infty \quad \textit{a.s.}$$

In Algorithm 1, $X_{S(n)}$ determines the acceptance rule, and between arrival β_n and β_{n+1}, $X_{S(n)}$ is either equal to $X_{\beta_n} = Y_n$ or $X_{\beta_{n+1}} = Y_{n+1}$. Hence, the sign of Y_n and Y_{n+1} determines the acceptance rule between arrival β_n and β_{n+1}. Thus, after a finite time, as long as there is an available server, the arrival is accepted, and the proposed policy converges to the optimal policy Π^*. The same arguments apply for the regime of $\mu \in (0, c/R)$. □

2.5.2 Finite-time performance analysis

To extend the finite-time results of the single-server queueing system to the more general setting of the multi-server system, we characterize the regret in terms of the submartingale (or supermartingale) sequence $\{Y_n\}_{n=0}^{\infty}$ and processes $\{Y_n^A\}_{n=0}^{\infty}$ and $\{Y_n^M\}_{n=0}^{\infty}$ found from Doob's decomposition. As the sign of $\{Y_n\}_{n=0}^{\infty}$ determines the acceptance rule, we provide an upper bound for the probability of the event that Y_n has an undesirable sign. Without loss of generality, in describing the methodology we assume that $\mu \in (c/R, +\infty)$ and from Doob's decomposition and Lemma 8,

$$\mathbb{P}\left(Y_n \le 0\right) = \mathbb{P}\left(Y_n^A + Y_n^M \le 0\right) \le \mathbb{P}\left(Y_n^M \le -\tilde{\delta}_1 n\right) \text{ for some } \tilde{\delta}_1 > 0. \tag{2.46}$$

Thus, it suffices to bound $\mathbb{P}\left(Y_n^M \le -\tilde{\delta}_1 n\right)$, as done in Lemma 10. The proof of Lemma 10 given in Appendix A.2.2, verifies a conditional sub-exponential property for the martingale difference sequence $\{Y_{n+1}^M - Y_n^M\}_{n=0}^{\infty}$, and utilizes a Bernstein-type bound for martingale difference sequences.

Lemma 10. *Consider a multi-server Erlang-B queueing system with service rate μ following policy Π_{Alg1}. For $\mu \in (c/R, +\infty)$, there exists a problem-dependent constant c_3 such that the martingale process $\{Y_n^M\}_{n=0}^{\infty}$ satisfies*

$$\mathbb{P}\left(Y_n^M \le -\tilde{\delta}_1 n\right) \le \exp\left(-c_3 n\right),$$

and for any $\mu \in (0, c/R)$, there exists a positive problem-dependent constant c_4 such that the following holds

$$\mathbb{P}\left(Y_n^M \ge -\tilde{\delta}_2 n\right) \le \exp\left(-c_4 n\right).$$

We begin with stating Theorem 6 that extends Theorem 2 to the multi-server setting, and argues that for the multi-server queueing system with $\mu \in (c/R, +\infty)$ and any (valid) function $f(n)$ such that $\log(f) = o(n)$, finite regret is achieved. The proof is similar to Theorem 2 and bounds the

expected number of arrivals for which $Y_n \leq 0$ using the exponentially decaying probability shown in Lemma 10.

Theorem 6. *Consider the multi-server Erlang-B queueing system with k servers and service rate μ. For any $\mu \in (c/R, +\infty)$ and (valid) function f such that $\log(f) = o(n)$, the expected regret $\mathbb{E}\left[\mathcal{R}; \Pi_{\text{Alg1}}(n)\right]$ under policy Π_{Alg1} is upper bounded by a constant independent of n.*

Proof of Theorem 6. Let K_n be the number of arrivals rejected after or at $\beta_n + \tau_n$ and before the first acceptance, β_{n+1}, i.e.,

$$K_n = \min\left\{i \geq 0 : A_{\beta_n+\tau_n+i} = 1\right\} = \beta_{n+1} - \beta_n - \tau_n.$$

Note that if $Y_n > 0$, the proposed policy will accept all arrivals from $\beta_{n-1} + \tau_{n-1}$ up to $\beta_n + \tau_n$ (subject to availability). In this case, $\beta_{n-1} + \tau_{n-1} = \beta_n$. But, if $Y_n \leq 0$, the arrivals are accepted with a certain probability and can contribute to the expected regret. Thus, we upper bound the regret as below

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}\left(n\right); \Pi_{\text{Alg1}}\right] &\leq \mathbb{E}\left[\tau_0\right] + \mathbb{E}\Big[\sum_{i=1}^{\infty}(\tau_i + K_{i-1})\mathbb{1}\{Y_i \leq 0\}\Big] \\
&= \sum_{i=0}^{\infty}\mathbb{E}\left[\tau_i \mathbb{1}\left\{Y_i \leq 0\right\}\right] + \sum_{i=1}^{\infty}\mathbb{E}\left[K_{i-1}\mathbb{1}\left\{Y_i \leq 0\right\}\right] \\
&\leq \sum_{i=0}^{\infty}\mathbb{E}\left[\tau_i \,\middle|\, Y_i \leq 0\right]\mathbb{P}\left(Y_i \leq 0\right) + \sum_{i=1}^{\infty} f\left(i\right)\mathbb{P}\left(Y_i \leq 0\right) \\
&\leq \sum_{i=0}^{\infty}\mathbb{E}\left[\tau_i \,\middle|\, Y_i \leq 0\right]\exp\left(-c_3 i\right) + \sum_{i=1}^{\infty} f\left(i\right)\exp\left(-c_3 i\right).
\end{aligned}$$

In the second line, we used the fact that given $Y_i \leq 0$, K_i is geometric with $\mathbb{E}[K_i] \leq f(i)$. The last inequality follows from (2.46) and Lemma 10. In Lemma 6, we argued that $\mathbb{E}[\tau_i | Y_{i-1} \leq 0]$ is bounded by a constant. Hence, for any function f with $\log(f) = o(n)$, the expected regret is finite. □

Lastly, we argue that the expected regret for a multi-server queueing system with $\mu \in (0, c/R)$ grows polylogarithmically in n. Analogous to Theorem 3, we bound the expected number of arrivals wherein $Y_n > 0$ using Lemma 10. Moreover, we capture the effect of the arrivals being accepted with a given probability by Lemma 5, leading to a polynomial bound in $\log(n)$. Further, in Corollary 2, following the same ideas as in Corollary 1, we improve the regret to achieve a $O(\log(n))$ regret.

Theorem 7. *Consider the multi-server Erlang-B queueing system with k servers and service rate $\mu \in (0, c/R)$. For $f(n) = \exp(n^{1-\epsilon})$, the expected regret under policy Π_{Alg1} is $\mathbb{E}[\mathcal{R}(n); \Pi_{\text{Alg1}}] = O\big(\log^{\frac{1}{1-\epsilon}}(n)\big)$.*

Proof of Theorem 7. In this case, the expected regret up to arrival n equals the expected number of arrivals accepted from the first n arrivals. Hence, we have

$$\begin{aligned}\mathbb{E}\Big[\mathcal{R}(n); \Pi_{\text{Alg1}}\Big] &= \mathbb{E}\Big[\sum_{i=0}^{n-1} \mathbb{1}\{A_i = 1\}\Big] \\ &= \mathbb{E}\Big[\sum_{i=0}^{n-1} \mathbb{1}\big\{A_i = 1, X_{S(i)} > 0\big\}\Big] + \mathbb{E}\Big[\sum_{i=0}^{n-1} \mathbb{1}\big\{A_i = 1, X_{S(i)} \le 0\big\}\Big]. \quad (2.47)\end{aligned}$$

We first upper bound the first term using (2.46) and Lemma 10 as follows

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=0}^{n-1} \mathbb{1}\big\{A_i = 1, X_{S(i)} > 0\big\}\Big] &\le \sum_{i=0}^{\infty} \mathbb{E}\Big[\mathbb{1}\{Y_i > 0\}\tau_i\Big] \\ &\le \sum_{i=0}^{\infty} \mathbb{E}\Big[\tau_i \,\Big|\, Y_i > 0\Big] \exp(-c_4 i). \quad (2.48)\end{aligned}$$

By Lemma 6, the above summation is bounded by a constant c_p. Next, we upper bound the second term of (2.47). As defined before, τ_i is the first $j > \beta_i$ such that $N_{\beta_i+j} = 0$ and K_i is equal to $\beta_{i+1} - \beta_i - \tau_i$, i.e., the number of rejected arrivals before arrival β_{i+1} and after or at $\beta_i + \tau_i$. If $X_{\beta_i+\tau_i} \le 0$, then K_i is geometric with parameter $1/\alpha_{\beta_i+\tau_i}$. We define $G(i)$ as the index of the first accepted arrival after $i-1$ arrivals, or

$$G(i) := \min_m \big\{m \ge 0 : \sum_{j=0}^{m} (\tau_j + K_j) \ge i\big\}.$$

We also take $F(i)$ to be the smallest m such that the sum of the first $m+1$ geometric trials exceeds $i-1$, i.e.,

$$F(i) := \min_m \big\{m \ge 0 : \sum_{j \in B_m} (K_j + 1) \ge i\big\},$$

where $B_m = \{j : 0 \le j \le m, X_{\beta_j+\tau_j} \le 0\}$. From these definitions, it follows that $G(i) \le F(i)$. The second term of (2.47) is less than or equal to the expected number of times an arrival $i < n$

with $X_{S(i)} \leq 0$ is accepted until arrival $\beta_{G(n)+1}$. Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{n-1} \mathbb{1}\left\{A_i = 1, X_{S(i)} \leq 0\right\}\Big] \\
&\leq \mathbb{E}\Big[\sum_{i=0}^{G(n)} \tau_i \mathbb{1}\left\{X_{\beta_i} \leq 0\right\}\Big] \\
&\leq \mathbb{E}\Big[\sum_{i=0}^{F(n)} \tau_i \mathbb{1}\left\{X_{\beta_i} \leq 0\right\}\Big] \\
&\leq \sum_{j=0}^{n-1} \mathbb{E}\Big[\sum_{i=0}^{F(n)} \tau_i \,\Big|\, F\left(n\right) = j\Big] \mathbb{P}\left(F\left(n\right) = j\right) \\
&\leq c_\tau \sum_{j=0}^{d} (j+1)\, \mathbb{P}\left(F\left(n\right) = j\right) + c_\tau \sum_{j=d+1}^{n-1} (j+1)\, \mathbb{P}\Big(\sum_{i=1}^{j-1} y_i < n, \sum_{i=1}^{j} y_i \geq n\Big) \qquad (2.49) \\
&\leq c_\tau\, \mathbb{E}\left[(F(n)+1)\mathbb{1}\{F(n) \leq d\}\right] + c_\tau \sum_{j=d}^{n-2} (j+2)\, \mathbb{P}\Big(\sum_{i=1}^{j} y_i < n, \sum_{i=1}^{j+1} y_i \geq n\Big), \qquad (2.50)
\end{aligned}
$$

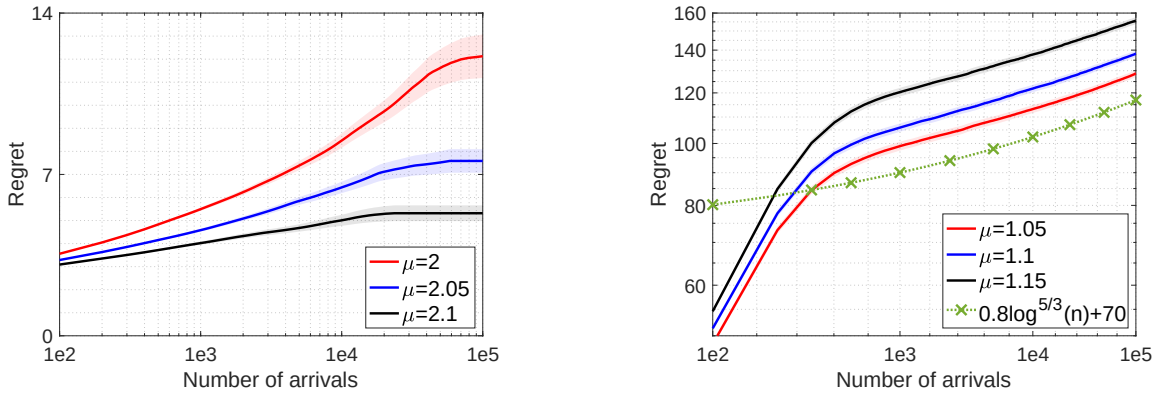
where $\{y_i\}_{i=1}^{n}$ are defined in Lemma 5, $d = \lceil 3(\log^{\frac{1}{1-\epsilon}}(n+1))\rceil$, c_τ is found using Lemma 6 and is proportional to $\sum_{j=0}^{k} \frac{\lambda^j}{\mu^j j!}$. Furthermore, (2.49) follows from the fact that the event $\{F(n) = j\}$ is equivalent to the event $\{\sum_{i=1}^{j-1} y_i < n, \sum_{i=1}^{j} y_i \geq n\}$. From Lemma 5, (2.50) is bounded by $c_\tau(d + 3 + c_\epsilon)$, where c_ϵ is a constant determined by ϵ. Finally, from (2.48) and (2.50), Theorem 7 follows. □

Corollary 2. *Consider the multi-server Erlang-B queueing system with k servers and service rate $\mu \in (0, c/R)$. For $f(n) = \exp\left(n^{1-\epsilon_n}\right)$ where $\epsilon_n = \frac{\varepsilon}{\sqrt{1+\log(n+1)}}$ for all $n \geq 1$ and $\varepsilon \in (0,1)$, the expected regret under policy Π_{Alg1} is $\mathbb{E}\left[\mathcal{R}\left(n\right); \Pi_{\text{Alg1}}\right] = O\big(\log(n)\big)$.*

We conclude by noting that the finite-time performance guarantees of the single-server setting hold for the general setting of multiple servers. Particularly, for $\mu \in (0, c/R)$, the expected regret is upper bounded as shown in Theorem 7 and Corollary 2; whereas, in the case of $\mu \in (c/R, +\infty)$, a finite regret bound is achieved following our proposed policy in Algorithm 1 as argued in Theorem 6.

2.6 Simulation-based numerical results

In this section, we empirically evaluate the performance of policy Π_{Alg1}. We calculate the regret by finding the difference in the number of sub-optimal actions taken by Π_{Alg1} compared to the optimal



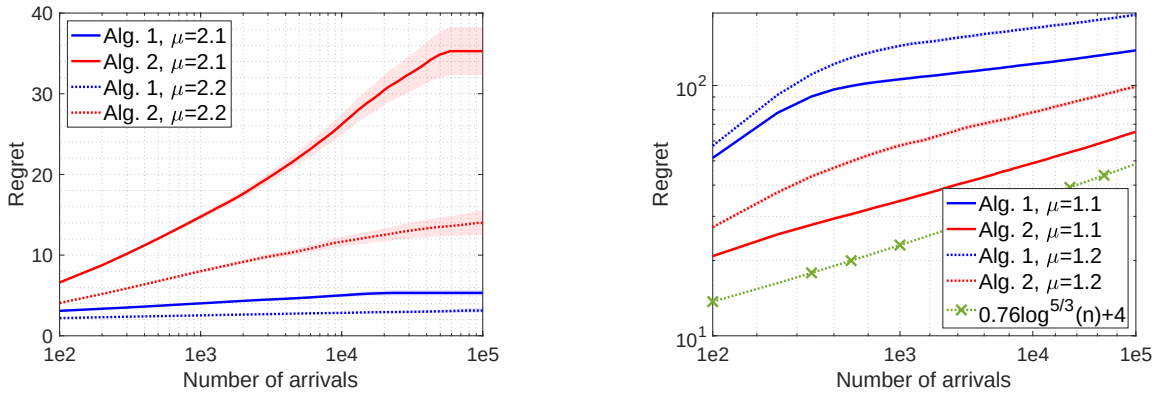
(a) $\mu \in (c/R, +\infty)$ (b) $\mu \in (0, c/R)$

Figure 2.2: Variations of regret for different service rates in a 5 server system with $\lambda = 5$, $c/R = 1.3$, $\epsilon = 0.4$, $\frac{1}{1-\epsilon} = 5/3$, and $f(n) = \exp\left(n^{1-\epsilon}\right)$ following Algorithm 1.

policy with the knowledge of the true service rate. The regret is averaged over 2500 simulation runs and plotted versus the number of incoming jobs. From our simulations, it can be observed that the proposed policy achieves finite regret for $\mu > c/R$, as predicted by our analysis. Further, the finite-time performance in the other regime corroborates our theoretical bound. We demonstrate the finite-time performance under various service rates and compare the performance of Π_{Alg1} against the dispatching scheme that updates the acceptance rule at every arrival. Furthermore, we compare the performance of Algorithm 1 with two RL algorithms: R-learning and Thompson sampling. In the plots of this section, we use a logarithmic scale for the x-axis when $\mu > c/R$ to display the variations clearly. Moreover, when $\mu < c/R$, we plot $\log\log(x)$ versus $\log(y)$ as the regret is bounded by a polynomial in $\log(n)$, where n is the number of arrivals, and this axes scaling provides a clearer depiction of the regret. Furthermore, the shaded regions in all plots indicate the $\pm\sigma$ area of the mean regret.

Figure 2.2 shows the regret performance for different service rates in a system with 5 servers, $\lambda = 5$, $c/R = 1.3$, and $f(n) = \exp\left(n^{0.6}\right)$. We can see that the regret grows as the service rate approaches the boundary value c/R (from either direction). In addition, as the gap between the service rate and the boundary value narrows, the regret converges more slowly to its final value when $\mu > c/R$. The results of Figures 2.2a and 2.2b corroborate the theoretical bounds of Theorems 6 and 7.

In Figure 2.3, we compare the performance of Algorithm 1 with an algorithm that updates the policy parameters at every arrival, called Algorithm 3. The problem parameters $\lambda, k, c, R, \epsilon$ are the same as the setting of Figure 2.2. In Algorithm 3, the admission probability decays faster than Algorithm 1, resulting in less exploration and better regret performance when $\mu < c/R$. From



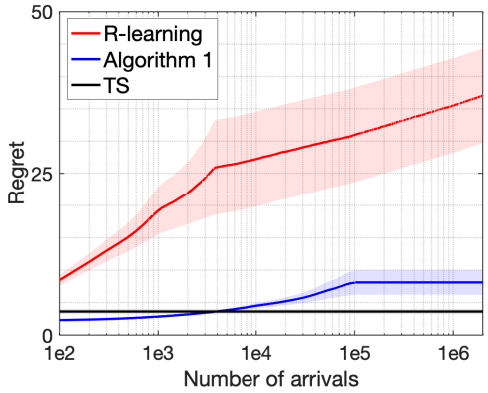
(a) $\mu \in (c/R, +\infty)$ (b) $\mu \in (0, c/R)$

Figure 2.3: Comparison of regret performance of Algorithm 1 against Algorithm 3 in a 5 server system with $\lambda = 5$, $c/R = 1.3$, $\epsilon = 0.4$, $\frac{1}{1-\epsilon} = 5/3$, and $f(n) = \exp\left(n^{1-\epsilon}\right)$.

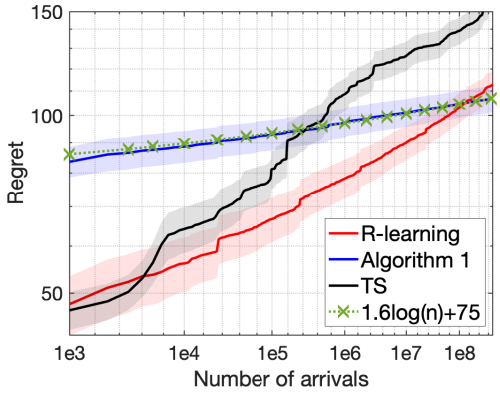
Figure 2.3a, Algorithm 1 outperforms Algorithm 3 for $\mu > c/R$ due to its slower decaying admission probability and the greater number of arrivals accepted. Another intuitive justification is that Algorithm 1 updates the policy parameters after observing a collection of arrivals, not prematurely after one sample, and the resulting averaging (and variance reduction) is useful in this regime. Conversely, in the case of $\mu \in (0, c/R)$, as suggested in Figure 2.3b, the additional exploration of Algorithm 1 leads to a worse regret performance.

In Figure 2.4, we compare the performance of Algorithm 1 with two other algorithms: R-learning ([92]) and Thompson sampling ([36]). We consider a system with $k = 5$, $\lambda = 5$, and $c/R = 1.3$. We also assume $f(n) = \exp\left(n^{1-\epsilon_n}\right)$ with $\epsilon_n = \frac{\varepsilon}{\sqrt{1+\log(n+1)}}$ and $\varepsilon = 0.2$. As noted in Section 2.1, the R-learning algorithm assumes that the service times are known ahead of the time when an arrival is accepted. Despite not observing the service times, Figure 2.4 depicts that Algorithm 1 outperforms R-learning in both regimes. Furthermore, empirically R-learning seems to have growing regret in both regimes. To implement the Thompson sampling algorithm, we use a uniform prior distribution defined on the two-point support $\{\mu_1, \mu_2\}$, where $\mu_1 = \frac{c}{2R} < \frac{c}{R}$ and $\mu_2 = \frac{3c}{2R} > \frac{c}{R}$, and update the posterior using (2.4) upon every arrival. As shown in Figure 2.4a, when $\mu > c/R$, the Thompson sampling algorithm has a better final regret value compared to our algorithm, but both algorithms have constant regret. However, when $\mu < c/R$, Algorithm 1 outperforms Thompson sampling; empirically, the asymptotic behavior of regret of both algorithms seem similar. We end by noting that theoretical analysis characterizing the regret performance for R-learning and Thompson sampling algorithms is not available in the literature.

In Figure 2.5, we compare the performance of Algorithm 1 in a $5-$server system with $\lambda = 5$ and $c/R = 1.3$ for two different exploration functions $f(n) = \exp\left(n^{1-\epsilon}\right)$ and $f(n) = \exp\left(n^{1-\epsilon_n}\right)$,



(a) $\mu = 2.5, \mu \in (c/R, +\infty)$

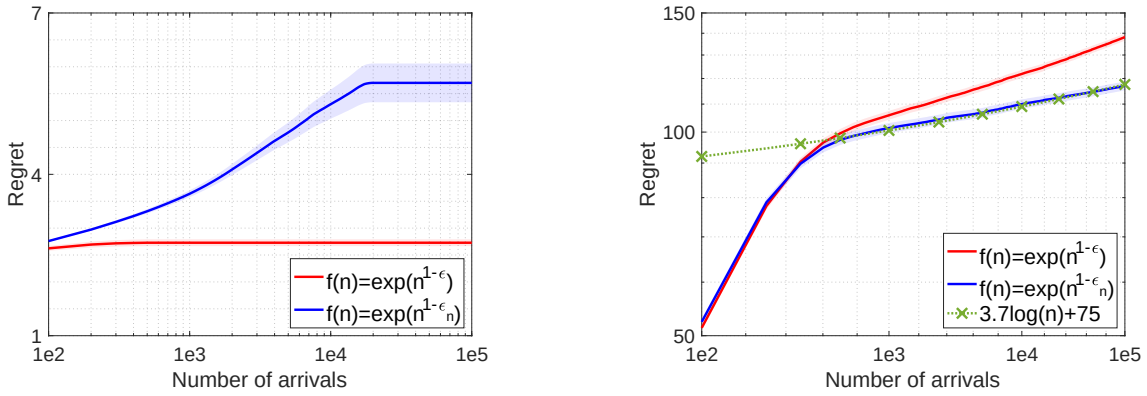


(b) $\mu = 1.05, \mu \in (0, c/R)$

Figure 2.4: Comparison of regret performance of Algorithm 1 against RL algorithms in a 5 server system with $\lambda = 5$, $c/R = 1.3$, $\varepsilon = 0.2$, and $f(n) = \exp\left(n^{1-\epsilon_n}\right)$.

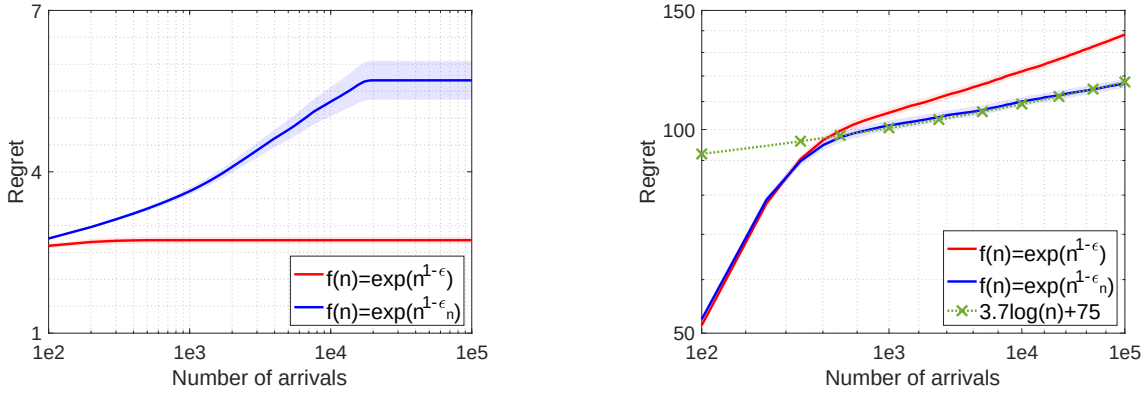
where $\epsilon_n = \frac{\varepsilon}{\sqrt{1+\log(n+1)}}$ and $\epsilon = \varepsilon = 0.55$. In Corollary 2, employing $f(n) = \exp\left(n^{1-\epsilon_n}\right)$ allows us to improve the order of the expected regret from $O(\log^{\frac{1}{1-\epsilon}}(n))$ to $O(\log(n))$. This improvement is shown in the numerical results of Figure 2.5b. Since ϵ_n decreases with n, the arrival acceptance due to exploration decreases faster, leading to slightly inferior performance when $\mu > c/R$, as shown in Figure 2.5a.

We next discuss a variant of our setting in which we can sample the system at other instances rather than only at the arrivals. One feasible approach is to modify the learning process as follows. Set a fixed sampling duration d. At each sampling time t, update functions g and h and the admittance probability accordingly. From any sampling time t, if an arrival occurs before d units of time, sample the system at the arrival and decide admission according to updated parameters. Otherwise, if d units of time pass without an arrival, sample the system at $t + d$. After a new sampling is done, repeat the previous steps. Note that (as a rule of thumb) for sampling to contribute to the learning, sampling duration d should be less than $1/\lambda$; setting $d = +\infty$ corresponds to policy Π_{Alg1}. In Figure 2.6, in a $2-$server system with $\lambda = 2$, $c/R = 1.5$, $f(n) = \exp\left(n^{1-\epsilon}\right)$, and $\epsilon = 0.4$, we depict the performance of the sampling scheme. When $\mu > \lambda$, the performance of Algorithm 1 can be improved by sampling; see Figure 2.6a. However, as shown in Figure 2.6b, when sampling according to the arrival rate is fast enough, performance does not improve with additional sampling. Moreover, Figure 2.6 suggests that an adaptive sampling scheme might achieve the best trade-off.



(a) $\mu = 2.2, \mu \in (c/R, +\infty)$ (b) $\mu = 1.1, \mu \in (0, c/R)$

Figure 2.5: Comparison of regret performance of Algorithm 1 for different functions $f(n)$ in a 5 server system with $\lambda = 5$, $c/R = 1.3$, and $\epsilon = \varepsilon = 0.55$.



(a) $\mu = 2.6, \mu \in (c/R, +\infty)$ (b) $\mu = 0.6, \mu \in (0, c/R)$

Figure 2.6: Regret performance for different sampling durations in a 2 server system with $\lambda = 2$, $c/R = 1.5$, $\epsilon = 0.4$, $\frac{1}{1-\epsilon} = \frac{5}{3}$, and $f(n) = \exp\left(n^{1-\epsilon}\right)$.

CHAPTER 3

Bayesian Learning in Countable State Space Markov Decision Processes

3.1 Introduction

Many real-life applications, such as communication networks, supply chains, and computing systems, are modeled using queueing models with countably infinite state-space. In the existing analysis of these systems, the models are assumed to be known, but despite this, developing optimal control schemes is hard, with only a few examples worked out [55, 10, 90]. However, knowing the model, algorithmic procedures exist to produce approximately optimal policies [55] (such as value iteration and linear programming). Given the success of data-driven optimal control design, in particular Reinforcement Learning (RL), we explore the use of such methods for the countable state-space controlled Markov processes. However, current RL methods that focus on finite-state settings do not apply to the mentioned queueing models. With the model unknown, our goal is to develop a meta-learning scheme that is RL-based but obtains good performance by utilizing algorithms developed when models are known. Specifically, we study the problem of optimal control of a family of discrete-time countable state-space MDPs governed by an unknown parameter θ from a general space Θ with each MDP evolving on the countable state-space $\mathcal{X} = \mathbb{Z}_+^d$ and finite action space $\mathcal{A}$. The cost function is unbounded and polynomially dependent on the state, following the examples of minimizing waiting times in queueing systems. Taking a Bayesian view, we assume the model is governed by an unknown parameter $\boldsymbol{\theta}^* \in \Theta$ generated from a fixed and known prior distribution. We aim to learn a policy π that minimizes the optimal infinite-horizon average cost over a given class of policies Π with low Bayesian regret with respect to the (parameter-dependent) optimal policy in Π.

To avoid many technical difficulties in countably infinite state-space settings, it is crucial to establish certain assumptions regarding the class of models from which the unknown system is drawn; some examples are: i) the number of deterministic stationary policies is not finite; and ii) in average cost optimal control problems, without stability/ergodicity assumptions, an optimal

policy may not exist [64], and when it exists, it may not be stationary or deterministic [30]. With these in mind, we assume that for any state-action pair, the transition kernels in the model class are categorical and skip-free to the right, i.e., with finite support with a bound depending on the state only in an additive manner; both are common features of queueing models where an increase in state is due to arrivals (with only a finite number of arrivals possible at any arrival instance). A second set of assumptions ensure stability by assuming that the Markov chains obtained by using different policies in Π are geometrically ergodic with uniformity across Θ. From these assumptions, moments on hitting times are derived in terms of Lyapunov functions for polynomial ergodicity, which exists due to geometric ergodicity. These assumptions also yield a solution to the average cost optimality equation (ACOE) [10] and provide a characterization of this solution.

3.1.1 Contributions

To optimally control the unknown MDP, we propose an algorithm based on Thompson sampling with dynamically-sized episodes; posterior sampling is used based on its broad applicability and computational efficiency [77, 78]. At the beginning of each episode, a posterior distribution is formed using Bayes' rule, and an estimate is realized from this distribution which then decides the policy used throughout the episode. To evaluate the performance of our proposed algorithm, we use the metric of Bayesian regret, which compares the expected total cost achieved by a learning policy π_L until time horizon T with the policy achieving the optimal infinite-horizon average cost in the policy class Π. We consider regret guarantees in three different settings as follows:

1. In Theorem 8, for Π being the set of all policies and assuming that we have oracle access to the optimal policy for each parameter, we establish an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ upper bound on the Bayesian regret of this algorithm compared to the optimal policy, where T is the time-horizon, d is the dimension of the state space, and $\tilde{O}$ hides logarithmic factors inproblem parameters.

2. In Corollary 3, where class Π is a subset of all stationary policies and where we know the best policy within this subset for each parameter via an oracle, we prove an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ upper bound on the Bayesian regret of our proposed algorithm, relative to the best-in-class policy.

3. In Theorem 9, we explore a scenario where we have access to an approximately optimal policy, rather than the optimal policy in set Π (which are all assumed to be stationary policies). When the approximately optimal policies satisfy Assumptions 3-4, we prove an $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$ regret bound, relative to the optimal policy in set Π.

Finally, to provide examples of our framework, we consider two different queueing models that meet our technical conditions, showing the applicability of our algorithm in developing approxi-

mately optimal control algorithms for stochastic systems with unknown dynamics. The first example is a continuous-time queueing system with two heterogeneous servers with unknown service rates and a common infinite buffer with the decision being the use of the slower server. Here, the optimal policy that minimizes the average waiting time is a threshold policy [59] which yields a queue-length after which the slower server is always used. The second model is a two-server queueing system, each with separate infinite buffers, to one of which a dispatcher routes an incoming arrival. Here, the optimal policy minimizing the waiting time is a switching curve [39] with the specifics unknown for general parameter values, so we aim to find the best policy within a commonly used set of switching-curve policies (Max-Weight policies [96, 97]), and assign the arrival to the queue with minimum weighted queue length. For both models, we verify our assumptions for the class of optimal/best-in-class policies corresponding to different service rates and conclude that our proposed algorithm can be used to learn the optimal/best-in-class policy.

3.1.2 Related work

Thompson sampling [100], or posterior sampling, has been applied to RL in many contexts of unknown MDPs [91, 76] and partially observed MDPs [43]; see tutorials [34, 84] for a comprehensive survey. It has been used in the parametric learning context [7] to minimize either Bayesian [77, 78, 80, 1, 98, 99] or frequentist [6, 36] regret. The bulk of the literature, including [6, 36, 80], analyzes finite-state and finite-action models but with different parameterizations such that a general dependence of the models on the parameters is allowed. The work in [99] studies general state-space MDPs but with a scalar parameterization with a Lipschitz dependence of the underlying models. Our problem formulation specifically considers countable state-space models with the models related via ergodicity, which we believe is a natural choice. Our focus on parametric learning is also connected to older work in adaptive control [3, 37] which studies asymptotically optimal learning for general parameter settings but with either a finite or countably infinite number of policies. Learning-based asymptotically optimal control in queues has a long history [56, 55] but recently there is increased work that also characterizes finite-time regret performance with respect to a well-known good policy or the optimal policy; see [103] for a survey. A series of work has studied learning with Max-Weight policies to get stability and linear regret [73, 49] or just stability [106]. A recent related work [27] considers learning optimal paramterized policies in queueing networks when the MDP is known. In a finite or countable state-space setting of specific queueing models where the parameters can be estimated, several works [50, 88, 51, 23, 31, 2, 25] have used forced exploration type schemes to obtain either regret that is constant or scaling logarithmically in the time-horizon.

Another line of work studies the problem of learning the optimal policy in an undiscounted

finite-horizon MDP with a bounded reward function. Reference [107] uses a Thompson sampling-based learning algorithm with linear value function approximation to study an MDP with a bounded reward function in a finite-horizon setting. Reference [24] considers an episodic finite-horizon MDP with known bounded rewards but unknown transition kernels modeled using linearly parameterized exponential families with unknown parameters. A maximum likelihood (ML) based algorithm coupled with exploration done by constructing high probability confidence sets around the ML estimate is used to learn the unknown parameters. In another work, [79] extends the problem setting of [24] to an episodic finite-horizon MDP with unknown rewards and transitions modeled using parametric bilinear exponential families. To learn the unknown parameters, they use a ML based algorithm with exploration done with explicit perturbation. To compare these works with our problem, we note that all mentioned works consider a finite-horizon problem. In contrast, our work considers an average cost problem, an infinite-horizon setting, and provides finite-time performance guarantees. In addition, these works focus on an MDP with a bounded reward function. Our focus, however, is learning in MDPs with unbounded rewards with the goal of covering practical queueing examples. We note that the parameterization of transitions used in [79, 24] can be used within our framework. However, similar to our work, additional stability assumptions are necessary to guarantee asymptotic learning and sub-linear regret. Another issue with exponential transition families is that they do not allow for 0 entries, which limits their applicability in queueing models such as our examples.

In another work, [85] studies discounted MDPs with unknown dynamics, and unbounded state-space, but with bounded rewards, and learns an online policy that satisfies a specific notion of stability. It is also assumed that a Lyapunov function ensuring stability for the optimal policy exists. We note that [85] ignores optimality and focuses on finding a stable policy, which contrasts with our work that evaluates performance relative to the optimal policy. Secondly, [85] considers a discounted reward problem, essentially a finite-time horizon problem (given the geometrically distributed lifetime). Average cost problems, such as ours, are infinite-time horizon problems, so connections to discounted problems can only be made in the limit of the discount parameter going to 1 and after normalizing the total discounted reward by 1 minus the discount parameter. Moreover, [85] considers a bounded reward function, simplifying their analysis but not practical for many queueing examples. Further, the assumption of a stable optimal policy with a Lyapunov function (as in [85]) is highly restrictive for bounded reward settings with discounting. For example, if the rewards increase to a bounded value as the state goes to infinity, then the stationary optimal policy (if it exists) will likely be unstable as the goal will be to increase the state as much as possible. Additionally, average cost problems with bounded costs need strong state-independent recurrence conditions for the existence of (stationary) optimal solutions, which many queueing examples don't satisfy; see [18]. Further complications can also arise with bounded costs: e.g., [30]

shows that a stationary average cost optimal policy may not exist.

3.2 Problem formulation

We consider a family of discrete-time Markov Decision Processes (MDPs) governed by parameter $\theta \in \Theta$ with the MDP for parameter θ described by $(\mathcal{X}, \mathcal{A}, c, P_\theta)$. For exposition purposes, we assume that all the MDPs are on (a common) countably infinite state-space $\mathcal{X} = \mathbb{Z}_+^d$. We denote the finite action space by $\mathcal{A}$, the transition kernel by $P_\theta : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, and the cost function by $c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$. As mentioned earlier, we will take a Bayesian view of the problem and assume that the model is generated using an unknown parameter $\boldsymbol{\theta}^* \in \Theta$, which is generated from a given fixed prior distribution $\nu(\cdot)$ on Θ. Our goal is to find a policy $\pi : \mathcal{X} \to \mathcal{A}$ that tries to achieve Bayesian optimal performance in policy class Π, i.e., minimizes the expected regret with $\boldsymbol{\theta}^*$ chosen from the prior distribution $\nu(\cdot)$. For each value $\theta \in \Theta$, the minimum infinite-horizon average cost is defined as

$$J(\theta) = \inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\big[\sum_{t=1}^{T} c(\boldsymbol{X}(t), A(t))\big], \tag{3.1}$$

where we optimize over a given class of policies Π and $\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t)) \in \mathcal{X}$ and $A(t) \in \mathcal{A}$ are the state and action at $t \in \mathbb{N}$. Typically, we set this class to be all (causal) policies, but it is also possible to consider Π to be a proper subset of all policies as we will explore in our results. For a learning policy π_L that aims to select the optimal control without model knowledge but with knowledge of Θ and the prior ν, the Bayesian regret until time horizon $T \geq 2$ is defined as

$$R(T, \pi_L) = \mathbb{E}\big[\sum_{t=1}^{T} \big[c(\boldsymbol{X}(t), A(t)) - J(\boldsymbol{\theta}^*)\big]\big], \tag{3.2}$$

where the expectation is taken over $\boldsymbol{\theta}^* \sim \nu$ and the dynamics induced by π_L. Owing to underlying challenges in countable state-space MDPs, we require the below assumptions on the cost function. The assumptions listed below are one potential general means to address the challenges.

Assumption 1. *The cost function $c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ is assumed to satisfy the following two conditions:*

1. ***(Coercive)** For every number $z \geq 0$ and for every valid action a in state $\boldsymbol{x}$, we assume that the cost function $c(\boldsymbol{x}, a)$ is greater than or equal to z outside a finite subset of $\mathcal{X}$.*

2. ***(Polynomially bounded growth-rate)** The cost function is upper-bounded by a multivariate polynomial $f_c : \mathbb{Z}_+^d \to \mathbb{R}_+$ which is increasing in every component on $\boldsymbol{x} \in \mathbb{Z}_+^d$ and has*

maximum degree of r (≥ 1) *in any dimension. We can assume that* $f_c(\boldsymbol{x}) = K \sum_{i=1}^{d} (x_i)^r$ *for some* $K > 0$*, where* $\boldsymbol{x} = (x_1, \ldots, x_d)$.

Thus, the cost function increases without bound (in the state) at a polynomial rate. Many examples of cost functions used in practice, say in queueing models of communication networks or manufacturing systems, depend polynomially on the state and fall under this setting; we will discuss a few in our evaluation section. To avoid technical issues the infinite state-space setting also necessitates some assumptions on the class from which the unknown model is drawn. For instance, irreducibility of Markov chains on such state-spaces (by using Markov or stationary policies) does not ensure positive recurrence (and ergodicity); thus, positive recurrence needs to be ensured using additional conditions. Moreover, for average cost optimal control problems, without stability even the existence of an optimal policy is not guaranteed, and we need more conditions. Other issues can also arise: an optimal control policy may not exist [64], and when it exists, it may not be stationary or deterministic [30], the average cost optimality equation (ACOE) may not have a solution [10, Section 5], and many others. The following assumption ensures a skip-free behaviour for transitions, which holds in many queueing models, where an increase in state corresponds to (new) arrivals (either external or internal), and this increment being bounded is thus reasonable. More generally, we can encapsulate this property as a maximum bound on the distance of the transitions instead of just being in one direction.

Assumption 2. ***(Skip-free to the right)*** *From any state-action pair* $(\boldsymbol{x}, a)$*, we assume that the transition is to a finite number of states; in essence, each such distribution is assumed to be a categorical distribution. We also assume that all transition kernels are skip-free to the right: for some* $h \geq 1$ *which is independent of* $\theta \in \Theta$ *and* $(\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A}$*, we have* $P_\theta(\boldsymbol{x}'; \boldsymbol{x}, a) = 0$ *for all* $\boldsymbol{x}' \in \{\tilde{\boldsymbol{x}} \in \mathbb{Z}_+^d : \|\tilde{\boldsymbol{x}}\|_1 > \|\boldsymbol{x}\|_1 + h\}$.

Learning necessitates some commonalities within the class of models so that using a policy well-suited to one model provides information on other models too. For us, these are in the form of constraints on the transition kernels of the models and stability assumptions for the policies that will be used; these stability assumptions will also ensure the existence of moments of certain functionals. As simple union bound arguments don't work in the countably infinite state-space setting, we will use the stability assumptions instead. In our setting, we consider a class of models, each with a policy being well-suited to at least one model in the class, and use the set of policies to search within. Using a reduced set of policies is necessary as the number of deterministic stationary policies is infinite. To learn correctly while restricting attention to this subset policy class, requires some regularity assumptions when a policy well-suited to one model is tried on a different model. Our ergodicity assumptions are one convenient choice; see Appendix B.1.1 for

details. These assumptions let us characterize the distributions of the first passage times or hitting times of the Markov processes via stability conditions; see Lemmas 23 and 24.

Assumption 3. ***(Geometric ergodicity)*** *For any MDP $(\mathcal{X}, \mathcal{A}, c, P_\theta)$ with parameter $\theta \in \Theta$, there exists a unique optimal policy π^*_θ that minimizes the infinite-horizon average cost within the class of policies Π. Furthermore, for any $\theta_1, \theta_2 \in \Theta$, the Markov process with transition kernel $P_{\theta_1}^{\pi^*_{\theta_2}}$ obtained from the MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy $\pi^*_{\theta_2}$ is irreducible, aperiodic, and geometrically ergodic with geometric ergodicity coefficient $\gamma^g_{\theta_1,\theta_2} \in (0,1)$ and stationary distribution μ_{θ_1,θ_2}. This is equivalent to the existence of finite set $C^g_{\theta_1,\theta_2}$ and Lyapunov function $V^g_{\theta_1,\theta_2} : \mathcal{X} \to [1, +\infty)$ satisfying*

$$\Delta V^g_{\theta_1,\theta_2}(\boldsymbol{x}) \le -\big(1-\gamma^g_{\theta_1,\theta_2}\big)V^g_{\theta_1,\theta_2}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X} \setminus C^g_{\theta_1,\theta_2} \text{ and } P_{\theta_1}^{\pi^*_{\theta_2}} V^g_{\theta_1,\theta_2}(\boldsymbol{x}) < +\infty,\ \boldsymbol{x} \in C^g_{\theta_1,\theta_2},$$

where $\Delta V^g_{\theta_1,\theta_2}(\boldsymbol{x}) := P_{\theta_1}^{\pi^*_{\theta_2}} V^g_{\theta_1,\theta_2}(\boldsymbol{x}) - V^g_{\theta_1,\theta_2}(\boldsymbol{x})$. Setting $b^g_{\theta_1,\theta_2} := \max_{\boldsymbol{x} \in C^g_{\theta_1,\theta_2}} P_{\theta_1}^{\pi^*_{\theta_2}} V^g_{\theta_1,\theta_2}(\boldsymbol{x}) + V^g_{\theta_1,\theta_2}(\boldsymbol{x})$ yields

$$\Delta V^g_{\theta_1,\theta_2}(\boldsymbol{x}) \le -\big(1-\gamma^g_{\theta_1,\theta_2}\big)V^g_{\theta_1,\theta_2}(\boldsymbol{x}) + b^g_{\theta_1,\theta_2}\mathbb{I}_{C^g_{\theta_1,\theta_2}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}. \tag{3.3}$$

Then, we have the following assumptions relating all the models in Θ:

1. *The geometric ergodicity coefficient is uniformly bounded below 1: $\gamma^g_* := \sup_{\theta_1,\theta_2 \in \Theta} \gamma^g_{\theta_1,\theta_2} < 1$.*

2. *We assume that $\{0^d\} \subseteq \cap_{\theta_1,\theta_2 \in \Theta} C^g_{\theta_1,\theta_2}$, that is, 0^d is a common state to all $C^g_{\theta_1,\theta_2}$. Furthermore, $C^g_* = \cup_{\theta_1,\theta_2 \in \Theta} C^g_{\theta_1,\theta_2}$ is a finite set. We further assume that $b^g_* := \sup_{\theta_1,\theta_2} b^g_{\theta_1,\theta_2} < +\infty$.*

Remark 5. *An implication of the assumptions above is that for every $\theta_1, \theta_2 \in \Theta$*

$$\mu_{\theta_1,\theta_2}(V^g_{\theta_1,\theta_2}) \le \frac{b^g_{\theta_1,\theta_2}}{1-\gamma^g_{\theta_1,\theta_2}} \le \frac{\sup_{\theta_1,\theta_2} b^g_{\theta_1,\theta_2}}{1-\gamma^g_*} < +\infty.$$

Remark 6. *The uniqueness of the optimal policy is not essential for the validity of our results, provided that all optimal policies satisfy our assumptions. When this condition is not met, we need to select an optimal policy that is geometrically ergodic for all $\theta \in \Theta$. This could entail searching over all optimal policies when non-uniqueness holds. This issue can be avoided by using a smaller subset of policies for which ergodicity can be shown, such as Max-Weight policies for scheduling, for which the ergodicity can be established for all policies.*

Geometric ergodicity implies that all moments of the hitting time of state 0^d, say τ_{0^d}, from any initial state $\boldsymbol{x} \neq 0^d$ are finite as $\mathbb{E}_{\boldsymbol{x}}[\kappa^{\tau_{0^d}}] \le c_1 V^g(\boldsymbol{x})$ (for specific $\kappa > 1$ and c_1), and so,

$\mathbb{E}_{\boldsymbol{x}}[\tau_{0^d}^k] \leq c_1 V^g(\boldsymbol{x}) k! / \log^k(\kappa) < +\infty$ for all $k \in \mathbb{N}$; see Appendix B.1.2. Function V^g is typically exponential in some norm of the state and yields an exponential bound for moments of hitting times, and a poor regret bound. To improve the regret bound, we need a different drift equation with function V^p with polynomial dependence on a norm of the state that bounds certain polynomial moments of τ_{0^d}.

Assumption 4. ***(Polynomial ergodicity)*** *Given $\theta_1, \theta_2 \in \Theta$, Markov process with transition kernel $P_{\theta_1}^{\pi^*_{\theta_2}}$ obtained from MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy $\pi^*_{\theta_2}$ is irreducible, aperiodic, and polynomially ergodic (with stationary distribution μ_{θ_1,θ_2}) through the Foster-Lyapunov criteria: there exists a finite set $C^p_{\theta_1,\theta_2}$, constants $\beta^p_{\theta_1,\theta_2}$, $b^p_{\theta_1,\theta_2} > 0$, $\alpha^p_{\theta_1,\theta_2} \in [\frac{r}{r+1}, 1)$, and function $V^p_{\theta_1,\theta_2} : \mathcal{X} \to [1, +\infty)$ satisfying (r is defined in Assumption 1)*

$$\Delta V^p_{\theta_1,\theta_2}(\boldsymbol{x}) \leq -\beta^p_{\theta_1,\theta_2} \big(V^p_{\theta_1,\theta_2}(\boldsymbol{x})\big)^{\alpha^p_{\theta_1,\theta_2}} + b^p_{\theta_1,\theta_2} \mathbb{I}_{C^p_{\theta_1,\theta_2}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}. \tag{3.4}$$

Then, we have the following assumptions relating all the models in Θ:

1. *$V^p_{\theta_1,\theta_2}$ is a polynomial with positive coefficients, maximum degree (in any dimension) $r^p_{\theta_1,\theta_2}$, and sum of coefficients $s^p_{\theta_1,\theta_2}$. We assume $r^p_* = \sup_{\theta_1,\theta_2} r^p_{\theta_1,\theta_2} < \infty$ and $s^p_* = \sup_{\theta_1,\theta_2} s^p_{\theta_1,\theta_2} < \infty$.*

2. *We assume that $\{0^d\} \subseteq \cap_{\theta_1,\theta_2 \in \Theta} C^p_{\theta_1,\theta_2}$, that is, 0^d is common to all $C^p_{\theta_1,\theta_2}$. Furthermore, $C^p_* = \cup_{\theta_1,\theta_2 \in \Theta} C^p_{\theta_1,\theta_2}$ is a finite set. We further assume that $\beta^p_* := \inf_{\theta_1,\theta_2} \beta^p_{\theta_1,\theta_2} > 0$ and $b^p_* := \sup_{\theta_1,\theta_2} b^p_{\theta_1,\theta_2} < \infty$.*

3. *Let $K_{\theta_1,\theta_2}(\boldsymbol{x}) := \sum_{n=0}^{\infty} 2^{-n-2} \big(P_{\theta_1}^{\pi^*_{\theta_2}}\big)^n (\boldsymbol{x}, 0^d)$, which is positive for any pair $\theta_1, \theta_2 \in \Theta$ by irreducibility. We assume that it is strictly positive in Θ: $K_* := \inf_{\theta_1,\theta_2} \min_{\boldsymbol{x} \in C^p_*} K_{\theta_1,\theta_2}(\boldsymbol{x}) > 0$.*

Remark 7. *An implication of the assumptions above is that for every $\theta_1, \theta_2 \in \Theta$*

$$\mu_{\theta_1,\theta_2}\left((V^p_{\theta_1,\theta_2})^{\alpha^p_{\theta_1,\theta_2}}\right) \leq \frac{b^p_{\theta_1,\theta_2}}{\beta^p_{\theta_1,\theta_2}} \leq \frac{\sup_{\theta_1,\theta_2} b^p_{\theta_1,\theta_2}}{\beta^p_*} < +\infty. \tag{3.5}$$

Note that $V^g_{\theta_1,\theta_2}$ satisfies the Foster-Lyapunov criterion in Assumption 4 for every $\alpha^p_{\theta_1,\theta_2} \in (0, 1)$. Assumptions 3-4 hold in many models of interest; see Appendix B.5. As average cost optimality is our design criterion, we need to ensure the existence of solutions to ACOE when Π is the set of all policies, or Poisson equation when Π is a subset of all policies. We discuss these two cases separately.

Case 1: Π ***is the set of all policies.*** For any parameter $\theta \in \Theta$, the MDP $(\mathcal{X}, \mathcal{A}, c, P_\theta)$ is said to

satisfy the ACOE if there exists a constant $J(\theta)$ and a unique function $v(\cdot;\theta) : \mathcal{X} \to \mathbb{R}$ such that

$$J(\theta) + v(\boldsymbol{x};\theta) = \min_{a\in\mathcal{A}} \big\{c(\boldsymbol{x},a) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_\theta(\boldsymbol{y}|\boldsymbol{x},a)v(\boldsymbol{y};\theta)\big\} \text{ with } v(0^d;\theta) = 0.$$

From [19] if the following conditions hold, ACOE has a solution, J_θ is the optimal infinite-horizon average cost, and there is an optimal stationary policy with ACOE becoming (3.6):

1. for every $(\boldsymbol{x}, a)$ and $z \geq 0$, cost function $c(\boldsymbol{x}, a) \geq z$ outside a finite subset of $\mathcal{X}$;
2. there is a stationary policy with an irreducible and aperiodic Markov process with finite average cost;
3. from every $(\boldsymbol{x}, a)$ transition to a finite number of states is possible.

From Assumptions 1-3, the above conditions hold, there exists an average cost optimal stationary policy, and the ACOE has a solution.

Case 2: Π ***is a proper subset of all policies.*** Here, we posit that for every $\theta \in \Theta$ and its best in-class policy π^*_θ, there exists a constant $J(\theta)$, the average cost of π^*_θ, and a function $v(\cdot;\theta) : \mathcal{X} \to \mathbb{R}$ with

$$J(\theta) + v(\boldsymbol{x};\theta) = c(\boldsymbol{x}, \pi^*_\theta(\boldsymbol{x})) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_\theta(\boldsymbol{y}|\boldsymbol{x}, \pi^*_\theta(\boldsymbol{x}))v(\boldsymbol{y};\theta). \tag{3.6}$$

This holds by the solution of the Poisson equation with the appropriate forcing function. For a Markov process $\boldsymbol{X}$ on the space $\mathcal{X}$ with time-homogeneous transition kernel P and cost function $\bar{c}(\cdot)$ (which will be the forcing function below), a solution to the Poisson equation [65] is a scalar J and function $v(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ such that

$$J + v = \bar{c} + Pv, \tag{3.7}$$

where $v(\boldsymbol{z}) = 0$ for some $\boldsymbol{z} \in \mathcal{X}$. Just like for the ACOE, if (v, J) is a solution to the Poisson equation, then so is $(v + b, J)$ for any scalar b. Hence, it is common to seek solutions such that $v(\boldsymbol{z}) = 0$ for some specific $\boldsymbol{z} \in \mathcal{X}$. In our setting using [65, Sections 9.6-9.8], for a model governed by $\theta \in \Theta$ following policy π^*_θ, we show a solution to the Poisson equation exists and is given by $v^{\pi^*_\theta}(0^d) = 0$ and

$$J(\theta) = \bar{C}^{\pi^*_\theta}(0^d)\big(\mathbb{E}^{\pi^*_\theta}_{0^d}[\tau_{0^d}]\big)^{-1} \text{ and } v^{\pi^*_\theta}(\boldsymbol{x}) = \bar{C}^{\pi^*_\theta}(\boldsymbol{x}) - J(\theta)\mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}[\tau_{0^d}], \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{3.8}$$

where $\bar{C}^{\pi^*_\theta}(\boldsymbol{x}) = \mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}\big[\sum_{i=0}^{\tau_{0^d}-1} c(\boldsymbol{X}(i), \pi^*_\theta(\boldsymbol{X}(i)))\big]$, and expectation is over trajectories of Markov chain $\boldsymbol{X}$ with transition kernel $P^{\pi^*_\theta}_\theta$ starting in state $\boldsymbol{x}$. In Appendix B.1.3, we present related definitions and show that from Assumptions 3-4, the requirements for the existence and finiteness of

Algorithm 2 Thompson Sampling with Dynamically-sized Episodes (TSDE)

Input: ν_0
Initialization: $\boldsymbol{X}(1) = 0^d$, $t \leftarrow 1$
for episodes $k = 1, 2, ...$ **do**
  $t_k \leftarrow t$
  Generate $\theta_k \sim \nu_{t_k}$
  while $t \leq t_k + \tilde{T}_{k-1}$ and $N_t(\boldsymbol{x}, a) \leq 2N_{t_k}(\boldsymbol{x}, a)$ for all $(\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A}$ **do**
    Apply action $A(t) = \pi^*_{\theta_k}(\boldsymbol{X}(t))$
    $N_t(\boldsymbol{X}(t), A(t)) \leftarrow N_t(\boldsymbol{X}(t), A(t)) + 1$
    Observe new state $\boldsymbol{X}(t+1)$
    Update ν_{t+1} according to (3.9)
    $t \leftarrow t + 1$
  end while
  $\tilde{T}_k \leftarrow t - t_k$
  while $\boldsymbol{X}(t) \neq 0^d$ **do**
    Apply action $A(t) = \pi^*_{\theta_k}(\boldsymbol{X}(t))$
    Observe new state $\boldsymbol{X}(t+1)$
  end while
  $T_k \leftarrow t - t_k$
end for

the solutions to Poisson equation are satisfied. Finally, we assume $\sup_{\theta\in\Theta} J(\theta)$ is finite, which typically holds as a result of the boundedness assumptions over all models in Θ stated in Asumptions 3 or 4, along with Assumption 1; this will be clear in our evaluation examples, but we mention it separately for completeness.

Remark 8. *In Assumption 4 we can use any other policy π_{θ_2} such that the Markov process obtained from MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy π_{θ_2} is irreducible and polynomially ergodic via the Foster-Lyapunov criteria with the uniformity discussed. Irreducibility is important as the policy will be used at times when the state is not known in advance, specifically at Steps 14-17 in Algorithm 2.*

Assumption 5. *We assume that $J^* := \sup_{\theta\in\Theta} J(\theta) < +\infty$.*

3.3 Learning algorithm: Thompson sampling with dynamically-sized episodes

We will use the learning algorithm Thompson sampling with dynamically-sized episodes from [80] to learn the unknown parameter $\boldsymbol{\theta}^* \in \Theta$ and the corresponding policy, $\pi^*_{\boldsymbol{\theta}^*}$, but suitably modify it for our countable state-space setting. Consider the prior distribution $\nu_0 = \nu$ defined on Θ from

which $\boldsymbol{\theta}^*$ is sampled. At each time $t \in \mathbb{N}$, the posterior distribution ν_t is updated according to Bayes' rule as

$$\nu_{t+1}(d\theta) = \frac{\mathbb{P}_\theta\left(\boldsymbol{X}\left(t+1\right)|\boldsymbol{X}\left(t\right), A\left(t\right)\right)\nu_t(d\theta)}{\int_{\theta' \in \Theta}\mathbb{P}_{\theta'}\left(\boldsymbol{X}\left(t+1\right)|\boldsymbol{X}\left(t\right), A\left(t\right)\right)\nu_t(d\theta')}, \tag{3.9}$$

and the posterior estimate θ_{t+1}, if generated, is from the posterior distribution ν_{t+1}. The modified Thompson-sampling with dynamically-sized episodes algorithm (TSDE) is presented in Algorithm 2. The TSDE algorithm operates in episodes: at the beginning of each episode k, parameter θ_k is sampled from the posterior distribution ν_{t_k} and during episode k, actions are generated from the stationary policy according to θ_k, i.e., $\pi^*_{\theta_k}$. Notice that $\pi^*_{\theta_k}$ is the optimal policy that minimizes the average expected cost of (3.1) in MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_k})$ either over all policies or a given set of policies. Let t_k be the time the k-th episode begins. Define $\tilde{t}_{k+1}$ as the first time after t_k that the conditions of Line 6 of Algorithm 2 is triggered and t_{k+1} as the first time at or after $\tilde{t}_{k+1}$ where state 0^d is visited; for the last episode started before or at T, we ensure that t_k and $\tilde{t}_k$ are less than or equal $T+1$. Explicitly, $t_1 = 1$ and for $k > 1$,

$$t_k = \min\{t \geq \tilde{t}_k : \boldsymbol{X}\left(t\right) = 0^d \text{ or } t > T\}.$$

Let $T_k = t_{k+1} - t_k$ be the length of the k-th episode and set $\tilde{T}_k = \tilde{t}_{k+1} - t_k$ with the convention $\tilde{T}_0 = 1$. The length of each episode k is determined in Line 6 of Algorithm 2 and is not fixed as it depends on the evolution of the Markov process determined by the true parameter $\boldsymbol{\theta}^*$ and the policy $\pi^*_{\theta_k}$ being used. For any state-action pair $(\boldsymbol{x}, a)$, we define $N_1(\boldsymbol{x}, a) = 0$ and for $t > 1$,

$$N_t(\boldsymbol{x}, a) = \left|\{t_k \leq i < \tilde{t}_{k+1} \leq t \text{ for some } k \geq 1 : (\boldsymbol{X}(i), A(i)) = (\boldsymbol{x}, a)\}\right|.$$

Notice that for all state-action pairs $(\boldsymbol{x}, a)$ and $\tilde{t}_{k+1} \leq t \leq t_{k+1}$, we have $N_t(\boldsymbol{x}, a) = N_{\tilde{t}_{k+1}}(\boldsymbol{x}, a)$. We denote K_T as the number of episodes started by or at time T, or $K_T = \max\{k : t_k \leq T\}$. The length of episode $k < K_T$ is not fixed and is determined according to two stopping criteria: (1) $t > t_k + \tilde{T}_{k-1}$, (2) $N_t(\boldsymbol{x}, a) > 2N_{t_k}(\boldsymbol{x}, a)$ for some state-action pair $(\boldsymbol{x}, a)$. After either criterion is met, the system will still follow policy $\pi^*_{\theta_k}$ until the first time at which state 0^d is visited; see Line 14 and Figure 3.1. We use this settling period to 0^d because the system state can be arbitrary when the first stopping criterion is met. As the countable state-space setting precludes a simple union-bound argument to overcome this uncertainty (as in the literature for finite state settings), we let the system reach the special state 0^d. Another (essentially equivalent) option is to wait until the state hits the finite set C^g_* or C^p_* and then use a union bound argument for all states in either set. For analytical convenience, we only use the state samples observed before arrival $\tilde{t}_{k+1}$ to update the posterior distribution, and not the samples of the system after time $\tilde{t}_{k+1}$ and before the beginning of episode $k+1$, i.e., t_{k+1}. The posterior update is halted during the settling period to 0^d as we have

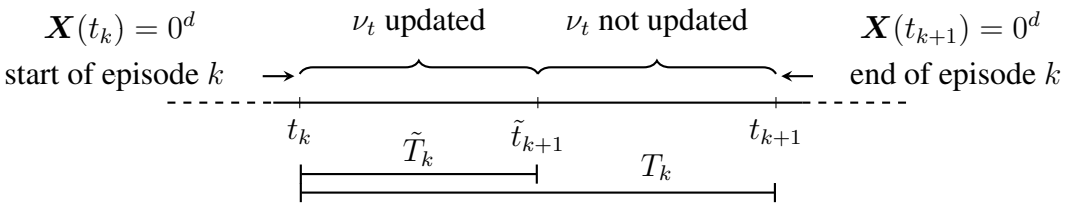


Figure 3.1: MDP evolution in episode $k < K_T$.

no control on the states visited during it, despite it being finite in duration (by our assumptions).

3.4 Regret analysis of Algorithm 2

The performance of any learning policy π_L is evaluated using the metric of expected regret compared to the optimal expected average cost of true parameter $\boldsymbol{\theta}^*$, namely, $J(\boldsymbol{\theta}^*)$. In this section, we evaluate the performance of Algorithm 2 and derive an upper bound for $R(T, \pi_{TSDE})$, its expected regret up to time T. In Section 3.2, we argued that at time t in episode k ($t_k \leq t < t_{k+1}$), there exist a constant $J(\theta_k)$ and a unique function $v(\cdot;\theta_k) : \mathcal{X} \to \mathbb{R}$ such that $v\left(0^d;\theta_k\right) = 0$ and

$$J(\theta_k) + v(\boldsymbol{X}(t);\theta_k) = c(\boldsymbol{X}(t), \pi^*_{\theta_k}(\boldsymbol{X}(t))) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}(\boldsymbol{y}|\boldsymbol{X}(t), \pi^*_{\theta_k}(\boldsymbol{X}(t)))v(\boldsymbol{y};\theta_k), \tag{3.10}$$

in which $\pi^*_{\theta_k}$ is the optimal or best-in-class policy (depending on the context) according to parameter θ_k and $J(\theta_k)$ is the average cost for the Markov process obtained from MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_k})$ by following $\pi^*_{\theta_k}$. We derive a bound for the expected regret $R(T, \pi_{TSDE})$ following the proof steps of [80] while extending it to the countable state-space setting of our problem. Using (3.10), the regret is decomposed into three terms and each term is bounded separately:

$$R(T, \pi_{TSDE}) = \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} c(\boldsymbol{X}(t), \pi^*_{\theta_k}(\boldsymbol{X}(t)))\right] - T\,\mathbb{E}\left[J\left(\boldsymbol{\theta}^*\right)\right] = R_0 + R_1 + R_2, \tag{3.11}$$

$$\text{with } R_0 = \mathbb{E}\left[\sum_{k=1}^{K_T} T_k J(\theta_k)\right] - T\,\mathbb{E}[J(\boldsymbol{\theta}^*)], \tag{3.12}$$

$$R_1 = \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left[v(\boldsymbol{X}(t);\theta_k) - v(\boldsymbol{X}(t+1);\theta_k)\right]\right], \tag{3.13}$$

$$R_2 = \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left[v(\boldsymbol{X}(t+1);\theta_k) - \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}(\boldsymbol{y}|\boldsymbol{X}(t), \pi^*_{\theta_k}(\boldsymbol{X}(t)))v(\boldsymbol{y};\theta_k)\right]\right]. \tag{3.14}$$

Before bounding the above regret terms, we address the complexities arising from the countable state-space setting. Firstly, we need to study the maximum state (with respect to the ℓ_∞-norm) visited up to time T in the MDP $(\mathcal{X}, \mathcal{A}, c, P_{\boldsymbol{\theta}^*})$ following Algorithm 2; we denote this maximum state by $M^T_{\boldsymbol{\theta}^*}$. We state the results that characterize the maximum l_∞-norm of the state vector achieved up until and including time T, and the resulting bounds on the number of episodes executed until time T. The results are listed as below:

1. In Lemma 11, we bound the moments of the maximum length of recurrence times of state 0^d, or $\max_{1\leq i\leq T} \tau^{(i)}_{0^d}$, using the ergodicity assumptions 3 and 4. This, along with the skip-free property, allows us to prove that the p-th moment of $\max_{1\leq i\leq T} \tau^{(i)}_{0^d}$ and $M^T_{\boldsymbol{\theta}^*}$ are both of order $O(\log^p T)$.

2. In Lemma 12, we find an upper bound for the number of episodes in which the second stopping criterion is met or there exists a state-action pair for which $N_t(\boldsymbol{x}, a)$ has increased more than twice in terms of random variable $M^T_{\boldsymbol{\theta}^*}$ and other problem-dependent constants.

3. In Lemma 13, we bound the total number of episodes K_T by time T by bounding the number of episodes triggered by the first stopping criterion, using the fact that in such episodes, $\tilde{T}_k = \tilde{T}_{k-1}+1$. Moreover, to account for the settling time of each episode, we use geometric ergodicity and Lemma 11. It follows that the expected value of the number of episodes K_T is of the order $\tilde{O}(h^d\sqrt{|\mathcal{A}|T})$.

Another challenge in analyzing the regret is that the relative value function $v(\boldsymbol{x};\theta)$ is unlikely to be bounded in the countable state-space setting. Hence, in (3.16) and (3.17), we find bounds for the relative value function in terms of hitting time τ_{0^d} from the initial state $\boldsymbol{x}$. Based on these results, we provide an upper bound for the regret of Algorithm 2 in Theorem 8.

3.4.1 Maximum state norm under polynomial and geometric ergodicity

We start with deriving upper bounds on the hitting times of state 0^d using the ergodicity conditions of Assumptions 3 and 4. Previous works [38, 42, 44] have already established bounds on hitting times in geometrically and polynomially ergodic chains in terms of their corresponding Lyapunov function. However, our objective is to provide a precise characterization of all constants included in these bounds in terms of the constants of the drift equations 3.3 and 3.4. This characterization allows us to derive uniform bounds across the model class. In Appendix B.3.1, using the polynomial Lyapunov function provided in Assumption 4, we establish upper bounds on the i-th moment of hitting time of state 0^d from any state $\boldsymbol{x} \in \mathcal{X}$ and for $1 \leq i \leq r+1$. Importantly, the derived bound is polynomial in terms of any component of the state x_i. Additionally, in Appendix B.3.2,

we characterize the tail probabilities of the return time to state 0^d starting from 0^d in terms of the geometric Lyapunov function of Assumption 3. The derived tail bounds will be used in Lemma 11 to derive upper bounds for all moments of hitting times in the model class. These bounds, along with the skip-free behavior of the model, allow us to study the maximum state (with respect to ℓ_∞-norm) achieved up to time T in MDP $(\mathcal{X}, \mathcal{A}, c, P_{\boldsymbol{\theta}^*})$ following Algorithm 2 as follows.

Lemma 11. *For $p \in \mathbb{N}$, the p-th moment of $\max_{1\le i\le T} \tau_{0^d}^{(i)}$ and $M_{\boldsymbol{\theta}^*}^T$, that is the maximum ℓ_∞-norm of the state vector achieved up until and including time T is $O(\log^p T)$.*

In the proof of Lemma 11 given in Appendix B.2.1, we make use of geometric ergodicity of the chain and the fact that hitting times have geometric tails to find an upper bound for moments of $M_{\boldsymbol{\theta}^*}^T$. Using this, we aim to bound the number of episodes started before or at T, denoted by K_T. We first find an upper bound for the number of episodes in which the second stopping criterion is met or there exists a state-action pair for which $N_t(\boldsymbol{x}, a)$ has increased more than twice. In the following lemma, we bound the number of such episodes, which we denote by K_M, in terms of random variable $M_{\boldsymbol{\theta}^*}^T$ and other problem-dependent constants. Proof of Lemma 12 is given in Appendix B.2.2.

Lemma 12. *The number of episodes triggered by the second stopping criterion and started before or at time T, denoted by K_M, satisfies $K_M \le 2|\mathcal{A}|(M_{\boldsymbol{\theta}^*}^T + 1)^d \log_2 T$ a.s.*

We next bound the total number of episodes K_T by bounding the number of episodes triggered by the first stopping criterion, using the fact that in such episodes, $\tilde{T}_k = \tilde{T}_{k-1} + 1$. Moreover, to address the settling time of each episode k, shown by $E_k = T_k - \tilde{T}_k$, we use the geometric ergodicity property and Lemma 11. Finally, the proof of Lemma 13 is given in Appendix B.2.3.

Lemma 13. *The number of episodes started by T satisfies $K_T \le 2\sqrt{|\mathcal{A}|(M_{\boldsymbol{\theta}^*}^T + 1)^d T \log_2 T}$ a.s.*

From Lemma 13, the upper bound given in Lemma 11 for moments of $M_{\boldsymbol{\theta}^*}^T$, and Cauchy–Schwarz inequality, it follows that the expected value of the number of episodes K_T is of the order $\tilde{O}(h^d\sqrt{|\mathcal{A}|T})$. This term has a crucial role in determining the overall order of the total regret up to time T.

Remark 9. *The skip-free to the right property in Assumption 2 yields a polynomially-sized subset of the underlying state-space that can be explored as a function of T. This polynomially-sized subset can be viewed as the effective finite-size of the system in the worst-case, and then, directly applying finite-state problem bounds [80] would result in a regret of order $\tilde{O}(T^{d+0.5})$; since $d \ge 1$, such a coarse bound is not helpful even for asserting asymptotic optimality! However, to achieve a regret of $\tilde{O}(\sqrt{T})$, it is essential to carefully understand and characterize the distribution of $M_{\boldsymbol{\theta}^*}^T$ and then its moments, as demonstrated in Lemma 11.*

Remark 10. *The derived regret bound can be extended to a larger class of MDPs which consist of transient states in addition to the single irreducible class. Specifically, for any $\theta_1, \theta_2 \in \Theta$, the Markov process with transition kernel $P_{\theta_1}^{\pi^*_{\theta_2}}$ obtained from the MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy $\pi^*_{\theta_2}$ has a single irreducible class I_{θ_1,θ_2} and a set of transient states T_{θ_1,θ_2}. Furthermore, Assumptions 3 and 4 hold for the single irreducible class. The reasoning behind the proof remains true in this case using the following argument: each episode k starts at 0^d which is in the irreducible set for the chosen policy $\pi^*_{\theta_k}$, hence, throughout the episode the algorithm remains in the irreducible set that is positive recurrent and never visits any transient states. In other words, episodes starting and ending at 0^d with a fixed episode dependent policy implies that reachable set of 0^d is all that can be explored, which is positive recurrent by our assumptions. As a result, we can restrict our proof derivations to the subset that is reachable from 0^d in each episode and follow the same analysis. The Lyapunov function based bounds apply to the positive recurrent states, and hence, restricting attention to states reachable from 0^d within each episode, we can use these bounds for our assessment of regret using norms of the state. Thereafter, the coarse bounds on the norms of the state can be applied as carried out in our proof.*

Remark 11. *By problem-dependent parameters, we refer to the parameters that characterize the complexity or size of the model class Θ. These parameters are not just a function of the size of the state-space and diameter of the MDP (as mentioned in the literature on finite-size problems[6, 36, 80]), as stability needs to be accounted for in the countable state-space setting. The dependence is, thus, more complex and requires the inclusion of stability parameters, such as Lyapunov functions, petite sets, and ergodicity coefficients that are discussed in Assumptions 1-4.*

3.4.2 Regret analysis

We first note a key property of Thompson sampling from [80], which states that for any episode k, measurable function f, and $\mathcal{H}_{t_k}-$measurable random variable Y, we have

$$\mathbb{E}\left[f(\theta_k, Y)\right] = \mathbb{E}\left[f(\boldsymbol{\theta}^*, Y)\right], \tag{3.15}$$

where $\mathcal{H}_t := \sigma\left(\boldsymbol{X}\left(1\right), \ldots, \boldsymbol{X}\left(t\right), A\left(1\right), \ldots, A\left(t-1\right)\right)$ for all $t \in \mathbb{N}$. Next, we bound regret terms R_0, R_1 and R_2 using the approach of [80] along with additional arguments to extend their result to a countably infinite state-space. We consider the relative value function $v(\boldsymbol{x}; \theta)$ of policy π^*_θ introduced for the optimal policy in ACOE or for the best in-class policy in the Poisson equation. In either of these cases, policy π^*_θ satisfies (3.6), which is the corresponding Poisson equation with forcing function $c(\boldsymbol{x}, \pi^*_\theta(\boldsymbol{x}))$ in a Markov chain with transition matrix $P_\theta^{\pi^*_\theta}$. In (3.8), we presented the solution (J, v) to the Poisson equation, which yields the following upper bound for the relative

value function, as argued in Appendix B.1.3:

$$v(\boldsymbol{x};\theta) \leq \bar{C}^{\pi^*_\theta}(\boldsymbol{x}) \leq \mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}\left[Kd\left(\|\boldsymbol{x}\|_\infty + h\tau_{0^d}\right)^r \tau_{0^d}\right]. \tag{3.16}$$

We can similarly lower bound the relative value function using Assumption 5 as

$$v(\boldsymbol{x};\theta) \geq -J(\theta)\mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}[\tau_{0^d}] \geq -J^*\mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}[\tau_{0^d}]. \tag{3.17}$$

From Assumption 3, all moments of τ_{0^d} and thus, the derived bounds are finite. Also, in Lemma 23 we bound the moments of τ_{0^d} of order $i \leq r+1$ using the polynomial Lyapunov function $V^p_{\theta_1,\theta_2}$, which is then used to bound the expected regret. We next bound the first regret term R_0 from the first stopping criterion in terms of the number of episodes K_T and the settling time of each episode k.

Lemma 14. *The first regret term R_0 satisfies $R_0 \leq J^*\,\mathbb{E}[K_T(\max_{1\leq i\leq T}\tau^{(i)}_{0^d}+1)]$.*

Proof of Lemma 14 is given in Appendix B.2.4. From Lemma 11, all moments of $\max_{1\leq i\leq T}\tau^{(i)}_{0^d}$ are bounded by a polylogarithmic function. Futhermore, as a result of Lemma 13, expected value of the number of episodes K_T is of the order $\tilde{O}(h^d\sqrt{|\mathcal{A}|T})$, which leads to a $\tilde{O}(h^d\sqrt{|\mathcal{A}|T})$ regret term R_0. Next, an upper bound on R_1 defined in (3.13) is derived. In the proof of Lemma 15 we argue that as the relative value function is equal to 0 at all time instances t_k for $k \leq K_T$, the only term that contributes to the regret is the value function at the end of time horizon T. We use the lower bound derived in (3.17) to show that the second regret term R_1 is $\tilde{O}(1)$; the proof is given in Appendix B.2.5.

Lemma 15. *The second regret term R_1 satisfies $R_1 \leq c_2\,\mathbb{E}[(M^T_{\boldsymbol{\theta}^*})^{r^p_*}] + c_3$, where $c_2 = J^*2^{r^p_*}s^p_*(\beta^p_*)^{-1}$ and $c_3 = J^*(\beta^p_*)^{-1}\big(s^p_*\,(2h)^{r^p_*} + b^p_*(K_*)^{-1}\big)$.*

From Lemma 11, $\mathbb{E}[(M^T_{\boldsymbol{\theta}^*})^{r^p_*}]$ is $O(\log^{r^p_*}T)$; hence, R_1 is upper bounded by a polylogarithmic function of the order r^p_*. Finally, in Lemma 16, we derive an upper bound for the third regret term R_2 defined in (3.14) using the bound derived for the relative value function in (3.16). To bound R_2, we characterize it in terms of the difference between the empirical and true unknown transition kernel and following the concentration method used in [105, 12, 80, 9], we argue that with high probability the total variation distance between the two distributions is small; for proof, see Appendix B.2.6.

Lemma 16. *For problem-dependent constant c_{p_3} and polynomial $Q(T) = c_{p_3}(Th)^{r+r^p_*}/48$, the second regret term R_2 satisfies*

$$R_2 \leq (\log(hT+h)+1)^d + c_{p_3}\sqrt{|\mathcal{A}|T}\log_2\left(2|\mathcal{A}|T^2Q(T)\right)\mathbb{E}\left[(M^T_{\boldsymbol{\theta}^*}+h)^{d+r+r^p_*}\big(\max_{1\leq i\leq T}\tau^{(i)}_{0^d}\big)\right].$$

The above Lemma results in a $\tilde{O}(KrdJ^*h^{d+2r+r_*^p}\sqrt{|\mathcal{A}|T})$ regret term as a result of Lemma 11, where h is the skip-free parameter defined in Assumption 2, d is the dimension of the state-space, K and r are the cost function parameters defined in Assumption 1, J^* is the supremum on the optimal cost, r_*^p is defined in Assumption 4, and where $\tilde{O}$ hides logarithmic factors in problem parameters one of which is $\log^{d+r+r_*^p+2}(T)$. For simplicity, we have not included the Lyapunov functions related parameters in the regret. Finally, from Lemmas 14, 15, 16, along with the Cauchy-Schwarz inequality, we conclude that the regret of Algorithm 2 $R(T, \pi_{TSDE})(= R_0 + R_1 + R_2)$ is $\tilde{O}(KrdJ^*h^{d+2r+r_*^p}\sqrt{|\mathcal{A}|T})$; for brevity, we will state that regret is of the order $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$.

Theorem 8. *Under Assumptions 1-5, the regret of Algorithm 2, $R(T, \pi_{TSDE})$, is $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$.*

Theorem 8 can be extended to the problem of finding the best policy within a sub-class of policies in set Π, which may or may not contain the optimal policy. In Section 3.2, we stated that Assumptions 3 and 4 hold for policies in Π and we used this to argue that the Poisson equation has a solution given in (3.8). As a result, repeating the same arguments as in Theorem 8 with the modification that π_θ^* is the best in-class policy of the MDP governed by parameter θ, yields the following corollary.

Corollary 3. *Under Assumptions 1 through 5, the regret of Algorithm 2 when using the best in-class policy is $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$.*

3.4.3 Requirement of an optimal policy oracle

To implement our algorithm, we need to find the optimal policy for each model sampled by the algorithm—optimal policy for Theorem 8 and optimal policy within policy class Π for Corollary 3. In the finite state-space setting, [80] provides a schedule of ϵ values and selects ϵ-optimal policies to obtain $\tilde{O}(\sqrt{T})$ regret guarantees. The issue with extending the analysis of [80] to the countable state-space setting is that we need to ensure (uniform) ergodicity for the chosen ϵ-optimal policies. In other words, we must verify ergodicity assumptions for a potentially large set of close-to-optimal algorithms whose structure is undetermined. Another issue is that, to the best of our knowledge, there isn't a general structural characterization of all ϵ-optimal stationary policies for countable state-space MDPs or even a characterization of the policy within this set that is selected by any computational procedure in the literature; current results only discuss characterization of the stationary optimal policy. In the absence of such results, stability assumptions with the same uniformity across models as in our submission will be needed, which are likely too strong to be useful. However, if we could verify the stability requirements of Assumptions 3 and 4 for a subset of policies, the optimal oracle is not needed, and instead, by choosing approximately optimal policies within this subset, we can follow the same proof steps as [80] to guarantee regret performance similar to Corollary 3 (without knowledge of model parameters). Thus, in Theorem 9 we

extend the previous regret guarantees to the algorithm employing ϵ-optimal policy; proof is given in Appendix B.2.8.

Theorem 9. *Consider a non-negative sequence $\{\epsilon_k\}_{k=1}^{\infty}$ such that for every $k \in \mathbb{N}$, ϵ_k is bounded above by $\frac{1}{k+1}$ and an ϵ_k-optimal policy satisfying Assumptions 3 and 4 is given. The regret incurred by Algorithm 2 while using the ϵ_k-optimal policy during any episode k is $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$.*

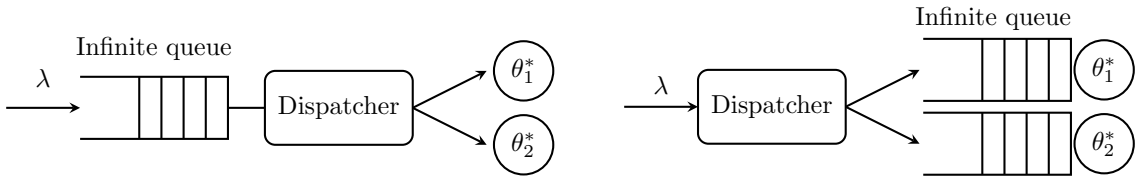
3.5 Evaluation: Application of Algorithm 2 to queueing models

Next, we present an evaluation of our algorithm. We study two different queueing models shown in Figure 3.2, each with Poisson arrivals at rate λ, and two heterogeneous servers with exponentially distributed service times with unknown service rate vector $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$. Vector $\boldsymbol{\theta}^*$ is sampled from the prior distribution ν defined on the space Θ given as

$$\Theta = \Big\{(\theta_1, \theta_2) \in \mathbb{R}_+^2 : \frac{\lambda}{\theta_1 + \theta_2} \leq \frac{1-\delta}{1+\delta}, 1 \leq \frac{\theta_1}{\theta_2} \leq R\Big\},$$

for fixed $R \geq 1$ and $\delta \in (0, 0.5)$. The first condition ensures the stability of the queueing models, while the second guarantees the compactness of the parameter space of the parameterized policies. In both systems, the goal of the dispatcher is to minimize the expected sojourn time of jobs, which by Little's law [87] is equivalent to minimizing the average number of jobs in the system. After verifying Assumptions 1-5 in Appendix B.5 for the cost function $c(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, Theorem 8 yields a Bayesian regret of order $\tilde{O}(\sqrt{|\mathcal{A}|T})$ for Algorithm 2.

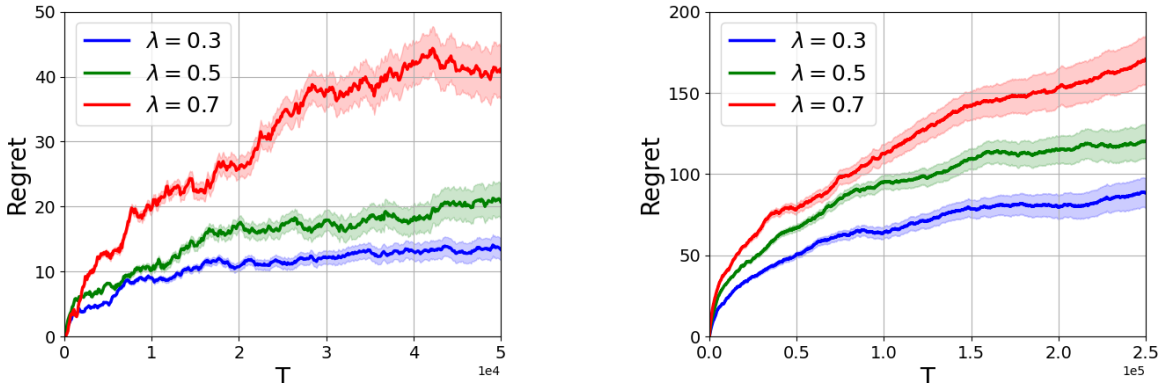
Model 1. Two-server queueing system with a common buffer. We consider the continuous-time queueing system of Figure 3.2a, where the countable state-space is $\mathcal{X} = \{\boldsymbol{x} = (x_0, x_1, x_2) \in \mathbb{Z}_+ \times \{0,1\}^2\}$, where x_0 is the queue length, and x_i, $i = 1, 2$ equal 1 if server i is busy. The action space is $\mathcal{A} = \{h, b, 1, 2\}$, where h means no action, b sends a job to both servers, and $i = 1, 2$ assigns a job to server i. In [59], when the system parameters are known, it is shown that by uniformization [60] and sampling the continuous-time Markov process at rate $\lambda + \theta_1^* + \theta_2^*$, a discrete-time Markov chain is obtained, which converts the original continuous-time problem to an equivalent discrete-time problem where we need to minimize $\limsup_{T\to\infty} T^{-1} \sum_{t=0}^{T-1} \|\boldsymbol{X}(t)\|_1$. Further, [59] shows that the optimal policy achieving the infimum average number of jobs is a threshold policy $\pi_{t(\boldsymbol{\theta}^*)}$ with optimal finite threshold $t(\boldsymbol{\theta}^*) \in \mathbb{N}$: always assign a job to the faster (first) server when free, and to the second server if it is free and $\|\boldsymbol{x}\|_1 > t(\boldsymbol{\theta}^*)$, and take no action otherwise. In Appendix B.5.1, we argue that the discrete-time Markov process governed by $\theta \in \Theta$ and following threshold policy π_t for any threshold t belonging to a compact set satisfies Assumptions 1-5.



(a) Queueing system with a common buffer.

(b) Two parallel queues.

Figure 3.2: Two-server queueing systems with heterogeneous service rates.



(a) Queueing system of Figure 3.2a.

(b) Queueing system of Figure 3.2b.

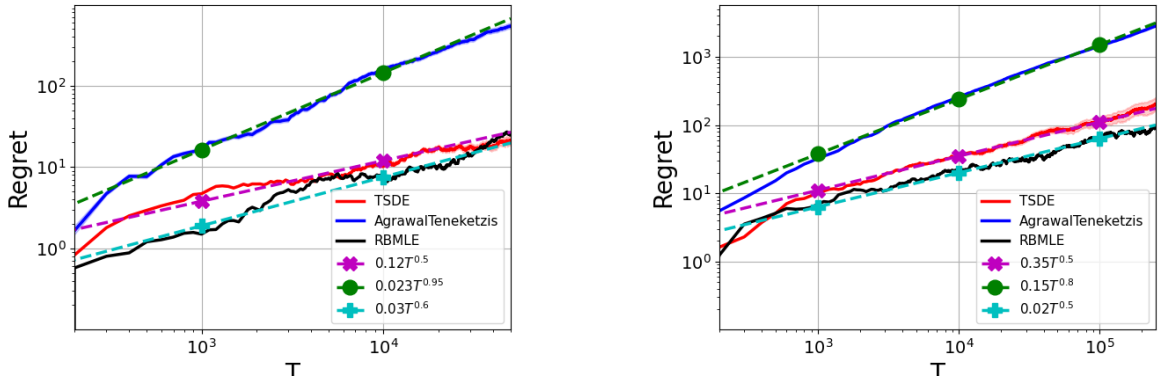
Figure 3.3: Regret performance for $\lambda = 0.3, 0.5, 0.7$. Shaded region shows the $\pm\sigma$ area of mean regret.

Model 2. Two heterogeneous parallel queues. We consider the continuous-time queueing system of Figure 3.2b with countable state-space $\mathcal{X} = \{\boldsymbol{x} = (x_1, x_2) \in \mathbb{Z}_+^2\}$, where x_i is the number of jobs in the server-queue pair i. The action space is $\mathcal{A} = \{1, 2\}$, where action i sends the arrival to queue i. We obtain the discrete-time MDP by sampling the queueing system at the arrivals, and then aim to find the average cost minimizing policy within the class $\Pi = \{\pi_\omega; \omega \in [(c_R R)^{-1}, c_R R]\}$, $c_R \geq 1$. Policy $\pi_\omega : \mathcal{X} \to \mathcal{A}$ routes arrivals based on the weighted queue lengths: $\pi_\omega(\boldsymbol{x}) = \arg\min(1 + x_1, \omega(1 + x_2))$ with ties broken for 1. Even with the transition kernel fully specified (by the values of arrival and service rates), the optimal policy in Π is not known except when $\theta_1 = \theta_2$ where the optimal value is $\omega = 1$, and so, to learn it, we will use Proximal Policy Optimization for countable state-space MDPs [27]. Note that [27] requires full model knowledge, which holds in our scheme as we use parameters sampled from the posterior for choosing the policy at the beginning of each episode. In Appendix B.5.2, we argue that the discrete-time Markov process governed by parameter $\theta \in \Theta$ and following policy π_ω for $\omega \in [(c_R R)^{-1}, c_R R]$ satisfies Assumptions 1-5.

Next, we report the numerical results of Algorithm 2 in the two queueing models of Figure 3.2 and calculate regret using (3.2). The regret is averaged over 2000 simulation runs and plotted against the number of transitions in the sampled discrete-time Markov process. Figure 3.3 shows the behavior of the regret of the two queueing models for three different arrival rates and service rates distributed according to a Dirichlet prior over $[0.5, 1.9]^2$. We observe that the regret is sub-linear in time and grows as the arrival rate increases. For the queueing model of Figure 3.2a, the minimum average cost $J(\theta)$ and optimal policy π^*_θ are known explicitly [59] for every $\theta \in \Theta$, which are used in Algorithm 2 and for regret calculation. Conversely, for the second queueing model, $J(\theta)$ and π^*_θ are not known. The PPO algorithm [27] is used to empirically find both the optimal weight and the policy's average cost. As expected from our theoretical guarantees, we observe that the regret is sub-linear in time. Furthermore, it grows as the arrival rate increases and the normalized load on the system converges to 1, which is expected since the system gets closer to the stability boundary. As discussed in Section 3.4, our bound on the expected regret is linearly dependent on J^* and, thus, will increase with the arrival rate. Additional details of the simulations and more plots are presented in Section 3.5.2.

3.5.1 Comparison of Algorithm 2 with other learning algorithms

We first note that due to the countably infinite state-space setting of our problem, we are unable to directly compare our algorithm to other learning algorithms proposed in the literature. One potential candidate algorithm uses the reward biased maximum likelihood estimation (RBMLE) [53, 54, 17, 70], which estimates the unknown model parameter with the likelihood perturbed a vanishing bias towards parameters with a larger long-term average reward (i.e., optimal value). This scheme also uses the principle of "optimism in the face of uncertainty" in how it perturbs the maximum likelihood estimate. The naive version of the RMBLE algorithm does not apply to our examples due the following key assumption: over all parameters (and the control policies used for them), the transition probabilities are assumed to be mutually absolutely continuous; this is critical for the proofs and also allows the use of log-likelihood functions for computations. Similarly, naive use of the algorithms in [56] and [37] is not possible, again due to a similar absolutely continuity assumption which is critical for the proofs. Our posterior computations avoid such issues as the true parameter always has non-zero mass during the execution of the algorithm: episode k always starts in state 0^d which is positive recurrent for the Markov chain with true parameter $\boldsymbol{\theta}^*$ and policy used $\pi^*_{\theta_k}$. The RBMLE algorithm has yet another issue in that it requires knowledge of the optimal value function, and hence, for our examples, it may only apply to Model 1 for which the value function is known analytically. Finally, whereas we do get to observe inter-arrival times for both model, we never directly observe completed service times owing to the sampling employed, and



(a) Queueing system of Figure 3.2a.

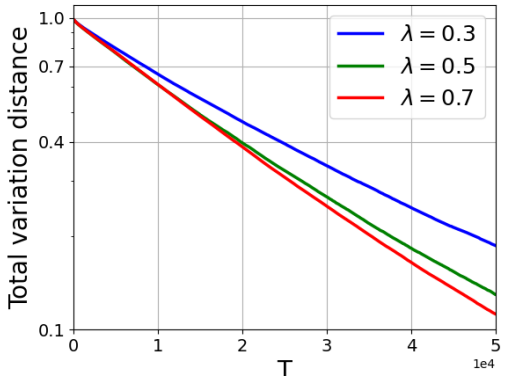
(b) Queueing system of Figure 3.2b.

Figure 3.4: Comparison of the regret performance of Algorithm 2 (referred to as TSDE) with the algorithm proposed by [4] (denoted as AgrawalTeneketzis) and the algorithm proposed by [53] (denoted as RBMLE) for the queueing models of Figure 3.2.

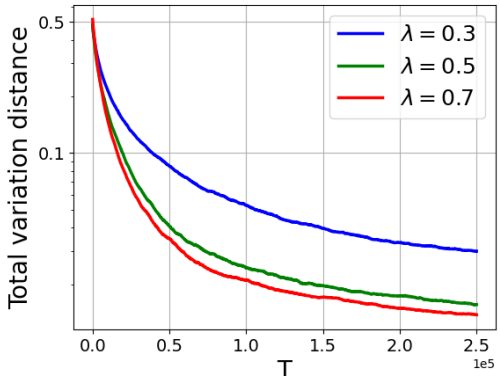
this precludes the direct use of Upper-Confidence-Bound based parameter estimation followed by certainty equivalent control algorithms. Owing to these issues, at this point in time, we're unable to perform empirical comparisons of Algorithm 2 to other candidate algorithms with theoretical performance guarantees in a countable state setting.

As discussed in the previous paragraph, learning algorithms with theoretical performance guarantees are established in the finite state setting. One such algorithm is the certainty equivalence control with forcing, which is proposed and discussed in detail in [4]. To assess the finite-time performance of our algorithm, in Figure 3.4, we compare the performance of our proposed learning algorithm, denoted as TSDE, with the algorithm introduced in [4], referred to as AgrawalTeneketzis. Reference [4] proposes a certainty equivalence control law with forced exploration, which operates in episodes with increasing lengths and a priori fixed sequences of forcing times. Specifically, at the beginning of each episode, all possible stationary control laws are explored for one recurrence interval of state $(0, 0)$. Subsequently, based on this exploration, an empirical estimate of the average collected reward is formed, and the control law resulting in the maximum average reward is implemented for the remainder of the episode. The length of the episodes are determined according to sequence $\{a_i\}_{i=0}^{\infty}$ defined as following:

$$
\begin{aligned}
a_0 &= 0, \\
a_i &= \sum_{k=1}^{i} b_k + ip, \qquad \text{for } i \geq 1,
\end{aligned}
$$



(a) Queueing system of Figure 3.2a.



(b) Queueing system of Figure 3.2b.

Figure 3.5: Total variation distance between the posterior and real distribution for $\lambda = 0.3, 0.5, 0.7$. The y axis is plotted on a logarithmic scale to display the differences clearly.

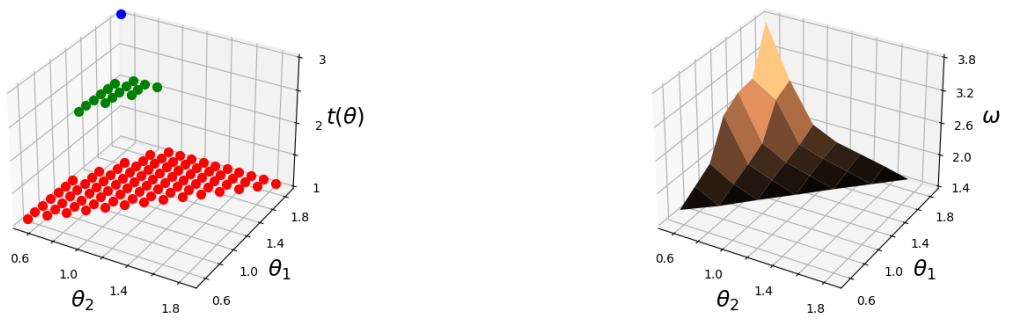
where p is the number of possible stationary control laws and $b_i = \left\lfloor \exp\left(i^{\frac{1}{1+\delta}}\right) \right\rfloor$ for any $\delta > 0$. Specifically, episode i terminates after completing additional $a_i - a_{i-1}$ recurrence intervals to state $(0, 0)$.

Another algorithm implemented in Figure 3.4 is Reward Biased MLE (RBMLE), which biases the maximum likelihood estimate towards the parameter with a smaller optimal average cost. In our setting, at each arrival t, we choose the estimate for unknown parameter θ as follows:

$$\theta_t \in \arg\max_{\Theta} \sum_{i=1}^{t-1} \log\left(P_\theta\left(\boldsymbol{X}(i+1)|\boldsymbol{X}(i), \pi^*_{\theta_i}(\boldsymbol{X}(i))\right)\right) - \alpha J(\theta)\log(t),$$

where α is a positive constant. A closed-form expression for the optimal average cost $J(\theta)$ is not available in the second model; instead, we rely on the estimated average cost obtained through the PPO algorithm (refer to Table 3.1).

Both algorithms are implemented in the two queueing systems of Figure 3.2, where the arrival rate is $\lambda = 0.5$ and service rates are distributed according to a Dirichlet prior over $[0.5, 1.9]^2$. In Figures 3.4a and 3.4b, we set $\delta = 3.5$ and $\delta = 3$, respectively, and $\alpha = 0.5$. These parameters are chosen to optimize the performance of the corresponding algorithms. Moreover, in Figure 3.4b, the goal is to find the optimal weight w in the set $\{1.5, 2, 2.5, 3, 3.5\}$. The results in Figure 3.4 show that both algorithms exhibit a sublinear regret performance. Specifically, Algorithm 2, TSDE, achieves an $\tilde{O}(\sqrt{T})$ as predicted in our theoretical results of Theorem 8 and Corollary 3. Furthermore, in both queueing models, our proposed algorithm either outperforms the other algorithms (AgrawalTeneketzis and RBMLE) in terms of regret order or attains the same regret order.



(a) Queueing system of Figure 3.2a.

(b) Queueing system of Figure 3.2b.

Figure 3.6: Optimal policy parameters for different service rate vectors in the two exemplary queuing systems in Model 1 and Model 2 with $\lambda = 0.5$.

3.5.2 Additional simulation details and discussion

Model 1: Two-server queueing system with a common buffer. Figure 3.3b illustrates the behavior of the regret of Model 1 for three different arrival rate values and averaged over 2000 simulation runs. In these simulations, the parameter space is selected as

$$\Theta = \left\{(\theta_1, \theta_2) \in [0.5, 0.6, \ldots, 1.9]^2 : \lambda < \theta_1 + \theta_2, \theta_2 < \theta_1\right\},$$

which results in a prior size of 105. As depicted in Figure 3.3a, the regret has a sub-linear behavior and increases with the arrival rate. The total variation distance between the posterior and real distribution, a point-mass on the random $\boldsymbol{\theta}^*$, are plotted in Figure 3.5a. As expected, the distance diminishes towards 0, indicating the learning of the true parameter. As mentioned in Appendix B.5.1, the optimal policy minimizing the average number of jobs in a system with parameter θ, is a threshold policy $\pi_{t(\theta)}$ with optimal finite threshold $t(\theta) \in \mathbb{N}$, which can be numerically determined as the smallest $i \in \mathbb{N}$ for which $J^i(\theta) < J^{i+1}(\theta)$, calculated in [59]. We compute the optimal threshold $t(\theta)$ for every $\theta \in \Theta$ and present the results in Figure 3.6a. We can see that the threshold increases as the ratio of the service rates grows. Specifically, this is why in Section 3.5.2, we imposed conditions on Θ to ensure that the ratio between the service rates is both upper and lower bounded.

Model 2: Two heterogeneous parallel queues Figure 3.3b illustrates the behavior of the regret of Model 2 for three different arrival rate values and averaged over 2000 simulation runs. We note that the regret is sub-linear and increases with higher arrival rates. In these simulations, the

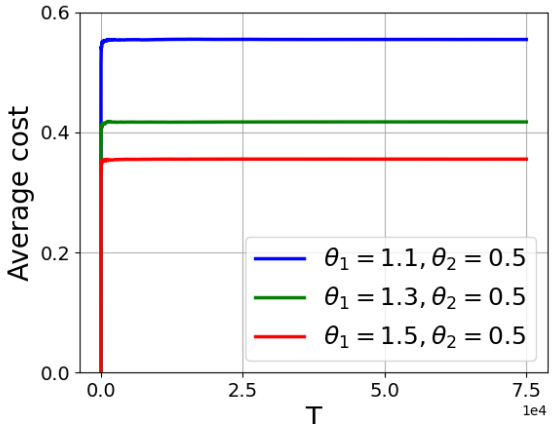


Figure 3.7: Estimated average cost of Model 2 for three different service rate vectors.

parameter space is selected as

$$\Theta = \left\{(\theta_1, \theta_2) \in [0.5, 0.7, \ldots, 1.9]^2 : \lambda < \theta_1 + \theta_2, \theta_2 < \theta_1\right\},$$

which results in a prior size of 28. As discussed earlier, our goal is to find the average cost minimizing policy within the class of policies $\Pi = \{\pi_\omega; \omega \in [(c_R R)^{-1}, c_R R]\}$, $c_R \geq 1$, where $\pi_\omega(\boldsymbol{x}) = \arg\min(1 + x_1, \omega(1 + x_2))$ with ties broken for 1. As discussed before, even with the transition kernel fully specified (by the values of arrival and service rates), the optimal policy in Π is not known except when $\theta_1 = \theta_2$ where the optimal value is $\omega = 1$, and so, to learn it, we will use Proximal Policy Optimization with approximating martingale-process (AMP) method for countable state-space MDPs [27]. We run the algorithm for 200 policy iterations, using 20 actors for each iteration. We take the state $(0, 0)$ as a regeneration state and simulate 1500 independent regenerative cycles per actor in each algorithm iteration. To approximate the value function, we employ a fully connected feed-forward neural network with one hidden layer consisting of 10×10 units and ReLU activation functions. The AMP method is also employed for variance reduction in value function estimation. The optimal ω for every $\theta \in \Theta$ is shown in Figure 3.6b, indicating that ω increases as the ratio of the service rates grows. Therefore, it is necessary to ensure that the ratio between the service rates is bounded from above and below. Furthermore, to evaluate the regret numerically, the value of $J(\theta)$ is required for every $\theta \in \Theta$, which is not known. Thus, after finding the optimal ω using the PPO algorithm, we perform a separate simulation to approximate the optimal average cost. In Figure 3.7, we plot the estimated average cost for three different service rate vectors, demonstrating that the optimal average cost decreases as the service rates increase. In Figure 3.5b we also depict the total variation distance between the posterior and real distribution, which is a point-mass on the random $\boldsymbol{\theta}^*$, and observe that the distance is converging to zero.

θ_1^*	θ_2^*	ω	$J(\boldsymbol{\theta}^*)$
0.7	0.5	1.5	1.04
0.9	0.5	1.5	0.82
1.1	0.5	2	0.67
1.3	0.5	2.5	0.56
1.5	0.5	2.5	0.47
1.7	0.5	3.5	0.41
1.9	0.5	3.5	0.35
0.9	0.7	1.5	0.70
1.1	0.7	1.5	0.59
1.3	0.7	2	0.51
1.5	0.7	2	0.44
1.7	0.7	2.5	0.39
1.9	0.7	2.5	0.34
1.1	0.9	1.5	0.54
1.3	0.9	1.5	0.47
1.5	0.9	1.5	0.42
1.7	0.9	2	0.37
1.9	0.9	2	0.33
1.3	1.1	1.5	0.44
1.5	1.1	1.5	0.39
1.7	1.1	1.5	0.35
1.9	1.1	2	0.32
1.5	1.3	1.5	0.37
1.7	1.3	1.5	0.33
1.9	1.3	1.5	0.30
1.7	1.5	1.5	0.32
1.9	1.5	1.5	0.29
1.9	1.7	1.5	0.28

Table 3.1: Optimal values of weight w in set $\{1.5, 2, 2.5, 3, 3.5\}$ and the corresponding average cost $J(\boldsymbol{\theta}^*)$ for different service rate values $(\theta_1^*, \theta_2^*) \in [0.5, 0.7, \ldots, 1.9]^2$.

CHAPTER 4

Conclusion

Throughout this thesis, we developed and analyzed two different learning schemes that leverage the inherent structure of a Markov decision process to design efficient learning algorithms with provable performance guarantees. We now proceed to present a few remarks on high-level takeaways and long-term future research directions.

4.1 High-level takeaways

The main ideas developed in this thesis are as follows:

- **Incorporating model knowledge can facilitate learning in the absence of direct observation of received rewards.**
  In Chapter 2, we studied the problem of learning-based optimal admission control of an Erlang-B blocking system with an unknown service rate. We showed that the extreme contrast in the optimal control schemes in different parameter regimes—quickly converging to always admitting arrivals if room versus quickly rejecting all arrivals—makes learning challenging. With the system being sampled only at arrivals, we designed a dispatching policy based on the maximum likelihood estimate of the unknown service rate, followed by using the certainty equivalent law with forced exploration. We proved the convergence of our proposed policy to the optimal policy in a system where the service rate is known and established finite-time guarantees for specific parameter settings: constant regret when $\mu > c/R$ and logarithmic regret when $\mu < c/R$. Through simulations, we also showed that our policy achieves a good trade-off of the regret over all parameter regimes. In addition to showing the difficulty of obtaining universally optimal learning algorithms for stochastic dynamic systems, we highlighted the complexity of obtaining lower bounds on the performance for continuous-time systems owing to issues such as sampling.

- **Exploiting inherent properties of certain Markov decision processes enables the adaptation of existing learning algorithms from finite state space to countably infinite state space.**

We studied the problem of learning optimal policies in countable state space Markov decision processes governed by unknown parameters. We proposed a learning policy based on Thompson sampling with dynamically-sized episodes and established finite-time performance guarantees for the Bayesian regret. We highlighted the practicality of our proposed algorithm by considering two different queueing models with unknown service rates and showing that our algorithm can be applied to develop optimal control policies. Specifically, we argued that for two different queueing models, the ergodicity assumptions of our algorithm are satisfied, and we further numerically investigated the regret performance of our algorithm.

4.2 Future directions

We conclude by exploring potential future directions related to the two chapters presented in this thesis.

- **Chapter 2.** The following questions naturally follow as future research topics. First, we proved a $\log(n)$ upper bound for the regret when $\mu < c/R$. One direction is to explore lower bounds in this regime; we conjecture that it is $\Omega(\log(n))$. Another direction is to allow for different sampling and update schemes (including by an independent Poisson process) and theoretically analyze the regret. Yet another direction is to extend our results to other service-time distributions, as the optimal admission control policy is unchanged due to the insensitivity ([47, 87]) of the Erlang-B system. Lastly, a broader theory is needed to study problems like ours where the problem structure changes non-smoothly across parameter choices.

- **Chapter 3.** For future work, it is worth studying how to extend the applicability of our algorithm to a broader class of problem settings and generalize it to consider policies that do not necessarily ensure stability. Additionally, in future work, it is also worth studying how to simplify our proposed algorithm by incorporating ideas from [99, 94].

In a broader context, we can conclude that the performance of existing learning algorithms can be enhanced by understanding and using the MDP's model and policy class knowledge. This observation points to the value of exploring this direction further and creating more comprehensive frameworks for utilizing model knowledge to improve the performance of the existing learning algorithms. Specifically, Chapter 2 highlights the promise of using information rewards; however, further exploration is required to demonstrate this more broadly, including within the general context assumed in the adaptive control literature. Additionally, within the general area of learning in queueing systems, the problem of learning-based control can be studied where the model class is known, but the observed samples from the system are sparse in some given sense; for instance,

queue lengths are partially observed. In this context, the stability of the queueing system is a central issue, and studying learning methodologies in the presence of potentially unstable policies is an interesting future direction.

Furthermore, recent advancements in the theoretical analysis of RL algorithms have mainly focused on finite state and action spaces. With the aim of generalizing existing results, Chapter 3 explored a learning problem in a countable state setting but with a finite action space. To further extend these findings, a potential future direction would be to explore learning in MDPs with infinite action spaces. This formulation is particularly relevant in many real-world control scenarios characterized by smooth and continuous actions, as exemplified in tasks like autonomous driving. Moreover, there exists potential for studying more complex problem settings (within the domain of queueing systems and beyond) to understand the implications of imperfectly observing reward functions during the learning process. Another framework within this direction is to investigate learning in MDPs where the rewards are not observable unless a query cost is paid. Studying these directions will provide us with novel tools and frameworks to address the complexity of learning-based control in real-world systems.

APPENDIX A

Appendix of Chapter 2

A.1 Analysis of the Single-server Erlang-B Queueing System

A.1.1 Lemma 17

Lemma 17. *In a single-server Erlang-B queueing system, the number of accepted arrivals following policy* Π_{Alg1} *is almost surely infinite.*

Proof of Lemma 17. Let A be the event that the system stops accepting new arrivals after some finite arrival, A_1 the event that the server is always busy after some finite arrival, A_2 the event that the server is available after some finite arrival but rejects all subsequent arrivals according to Line 10 of Algorithm 1, and $A_{2,m}$ as the event that for the first time at arrival m, the server is available but rejects all arrivals. We have

$$\mathbb{P}(A) = \mathbb{P}(A_1) + \mathbb{P}(A_2) = \mathbb{P}(A_2) = \sum_{m=0}^{\infty} \mathbb{P}(A_{2,m}) \leq \sum_{m=0}^{\infty} \lim_{n\to+\infty} \left(1 - \frac{1}{f(m)}\right)^n = 0, \quad \text{(A.1)}$$

where the inequality follows from the fact that for $n \geq m$, we have $\alpha_n = \alpha_m \leq m$, which means the acceptance probability is fixed after arrival m, as no other arrivals are accepted. From (A.1), we conclude that almost surely an infinite number of arrivals are accepted following Algorithm 1. □

A.1.2 Proof of Lemma 1

Proof of Lemma 1. Consider the queueing system sampled at sequence $\{\beta_n\}_{n=0}^{\infty}$. In the original representation, at state $\tilde{Y}_n$, service time E_{β_n} is realized when arrival β_n is accepted. Sequence $\{T_{\beta_n+j}\}_{j=1}^{\beta_{n+1}-\beta_n}$ is also realized until the next accepted arrival β_{n+1}. Based on the definition of l_n, arrival β_n departs during $T_{\beta_n+l_n}$. In the alternate process, instead of realizing E_{β_n} all at once, at each arrival, we generate two independent exponential random variables T'_{β_n+j} and E'_{β_n+j}, with parameters λ and μ. Let l'_n be the first arrival such that $l'_n = \min\{m \geq 1 : T'_{\beta_n+m} \geq E'_{\beta_n+m}\}$.

For $j < l'_n$, the minimum of T'_{β_n+j} and E'_{β_n+j} equals T'_{β_n+j}, and we assume the server is busy. This event occurs with probability $\frac{\lambda}{\lambda+\mu}$, and T'_{β_n+j} indicates the inter-arrival time between arrival $\beta_n + j - 1$ and $\beta_n + j$. At $j = l'_n$, E'_{β_n+j} is less than T'_{β_n+j} for the first time, and we assume the service of arrival β_n is complete. For the rest of the process, we only generate the inter-arrivals T'_{β_n+j} until an arrival is accepted. Note that l'_n is geometric with parameter $\frac{\mu}{\lambda+\mu}$. The equivalence of the process defined using T'_{β_n+j} and E'_{β_n+j} and the original process follows from the memoryless property of the exponential distribution and we deduce that $\{l_n\}_{n=0}^{+\infty}$ are *i.i.d.* and geometric random variables. □

A.1.3 Proof of Lemma 3

Proof of Lemma 3. Instead of directly verifying that the tail decay of random variables $\{Y_{n+1} - Y_n\}_{n=0}^{\infty}$ is at least as fast as an exponential distribution, we argue that an equivalent condition holds [101, Proposition 2.7.1]: there exists a constant $b > 0$ such that $\mathbb{E}\left[\exp\left(b\left|Y_{n+1} - Y_n\right|\right)\right] \leq 2$. From (2.13),

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(b\left|Y_{n+1} - Y_n\right|\right)\right] \\
&\leq \mathbb{E}\Big[\exp\Big(b\sum_{j=1}^{l_n-1} T_{\beta_n+j} + bg\left(T_{\beta_n+l_n}, 1, \frac{c}{R}\right)\Big)\Big] \\
&= \sum_{s=1}^{\infty} \mathbb{P}\left(l_n = s\right) \mathbb{E}\Big[\exp\Big(b\sum_{j=1}^{l_n-1} T_{\beta_n+j} + bg\left(T_{\beta_n+l_n}, 1, \frac{c}{R}\right)\Big) \,\Big|\, l_n = s\Big] \\
&= \sum_{s=1}^{\infty} \mathbb{P}\left(l_n = s\right) \Big(\mathbb{E}\Big[\exp\left(bT_{\beta_n+1}\right) \,\Big|\, l_n = s\Big]\Big)^{s-1} \mathbb{E}\Big[\exp\Big(bg\Big(T_{\beta_n+l_n}, 1, \frac{c}{R}\Big)\Big) \,\Big|\, l_n = s\Big]. \quad \text{(A.2)}
\end{aligned}$$

For $s > 1$ and $b < \lambda + \mu$, we simplify the first expectation to get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(bT_{\beta_n+1}\right) \,\middle|\, l_n = s\right] &= \frac{\mu+\lambda}{\lambda} \int_{t=0}^{+\infty} \int_{x=t}^{+\infty} \exp\left(bt\right) \mu \exp\left(-\mu x\right) \lambda \exp\left(-\lambda t\right) dxdt \\
&= \frac{\lambda+\mu}{\lambda+\mu-b}.
\end{aligned}$$

As $g(t,1,c/R) \leq R/c$ for all $t>0$, the second expectation term in (A.2) is bounded by $\exp(bR/c)$. Thus,

$$\begin{aligned}\mathbb{E}\left[\exp\left(b\left|Y_{n+1}-Y_n\right|\right)\right] &\leq \sum_{s=1}^{\infty}\mathbb{P}\left(l_n=s\right)\left(\frac{\lambda+\mu}{\lambda+\mu-b}\right)^{s-1}\exp\left(b\frac{R}{c}\right)\\ &=\sum_{s=1}^{\infty}\left(\frac{\lambda}{\lambda+\mu}\right)^{s-1}\frac{\mu}{\lambda+\mu}\left(\frac{\lambda+\mu}{\lambda+\mu-b}\right)^{s-1}\exp\left(b\frac{R}{c}\right)\\ &=\frac{\mu\exp\left(b\frac{R}{c}\right)}{\lambda+\mu}\sum_{s=1}^{\infty}\left(\frac{\lambda}{\lambda+\mu-b}\right)^{s-1}\end{aligned}$$

For $b<\mu$, the above sum converges, and $\mathbb{E}\left[\exp\left(b\left|Y_{n+1}-Y_n\right|\right)\right] \leq \frac{\mu}{\lambda+\mu}\frac{\lambda+\mu-b}{\mu-b}\exp\left(b\frac{R}{c}\right)$, which is less than 2 for small enough b. Thus, the sub-exponential property is proved. □

A.1.4 Proof of Lemma 4

Proof of Lemma 4. Without loss of generality, we take $\mu > c/R$. Note that $\mathbb{P}(Y_n \leq 0) = \mathbb{P}(\sum_{i=0}^{n-1}(Y_{i+1}-Y_i) \leq 0)$. In Lemmas 2 and 4, we showed that $\{Y_i - Y_{i-1}\}_{i=0}^{n-1}$ are *i.i.d* and sub-exponential; thus, the centered random variables $\{Y_{i+1}-Y_i-\mathbb{E}\left[Y_{i+1}-Y_i\right]\}_{i=0}^{n-1}$ are sub-exponential. We showed $\mathbb{E}\left[Y_{i+1}-Y_i\right]>0$ in (2.16); define $\mathbb{E}\left[Y_{i+1}-Y_i\right]=\delta$. From Bernstein's concentration inequality [101, Theorem 2.8.2],

$$\begin{aligned}\mathbb{P}\left(\sum_{i=0}^{n-1}(Y_{i+1}-Y_i)<0\right) &= \mathbb{P}\left(\sum_{i=0}^{n-1}(Y_{i+1}-Y_i)-n\delta<-n\delta\right)\\ &\leq \exp\left(-c_B\min\left(\frac{n^2\delta^2}{n},n\delta\right)\right)\\ &=\exp(-c_1 n).\end{aligned}$$

□

A.1.5 Proof of Lemma 5

Proof of Lemma 5. We first bound the probability term $\mathbb{P}\left(\sum_{j=1}^{i}y_j<n,\sum_{j=1}^{i+1}y_j\geq n\right)$ using the probability of the first event. We take $p_i=1-q_i=\exp\left(-i^{1-\epsilon}\right)$ and then use the Chernoff bound

to get

$$\mathbb{P}\left(y_1+\cdots+y_i<n, y_1+\cdots+y_{i+1}\geq n\right)\leq \mathbb{P}\left(y_1+\cdots+y_i\leq n\right)$$
$$\leq \min_{t\geq 0} e^{tn}\prod_{j=1}^{i}\frac{p_j}{e^t-(1-p_j)}. \tag{A.3}$$

Take $b=\lceil(\log(n+1))^{\frac{1}{1-\epsilon}}\rceil$ and $t\geq 0$ such that $e^t=\frac{n+1}{n}q_i$. From (A.3), for $i\geq d\geq b$ we have

$$\begin{aligned}\mathbb{P}\left(y_1+\cdots+y_i\leq n\right)&\leq \left(\frac{n+1}{n}\right)^n q_i^n\prod_{j=1}^{i}\frac{p_j}{\frac{1}{n}(1-p_i)+(p_j-p_i)}\\&\leq \left(\frac{n+1}{n}\right)^n q_i^n\prod_{j=1}^{i}p_j\prod_{j=1}^{b}\frac{1}{p_j-p_i}\prod_{j=b+1}^{i}\frac{n}{1-p_i}\\&\leq \left(\frac{n+1}{n}\right)^n q_i^{n-(i-b)}n^{i-b}\prod_{j=b+1}^{i}p_j\prod_{j=1}^{b}\frac{1}{1-\exp\left(-\left(i^{1-\epsilon}-j^{1-\epsilon}\right)\right)}.\end{aligned} \tag{A.4}$$

Since $q_i\leq 1$ and $n\geq i-b$, we have $\left(\frac{n+1}{n}\right)^n q_i^{n-(i-b)}\leq e$. By concavity and gradient inequality, for $1\leq j\leq i$, we have

$$i^{1-\epsilon}-j^{1-\epsilon}\geq\frac{1-\epsilon}{i^\epsilon}(i-j).$$

Using this inequality and setting $\kappa:=\lceil i^\epsilon/(1-\epsilon)\rceil$, we have

$$\begin{aligned}\prod_{j=1}^{b}\frac{1}{1-\exp\left(-\left(i^{1-\epsilon}-j^{1-\epsilon}\right)\right)}&\leq\prod_{j=1}^{b}\frac{1}{1-\exp\left(-\frac{1-\epsilon}{i^\epsilon}(i-j)\right)}\\&\leq\prod_{t=1}^{\infty}\frac{1}{1-\exp\left(-\left(\frac{1-\epsilon}{i^\epsilon}\right)t\right)}\\&\leq\prod_{t=1}^{\kappa-1}\frac{1}{1-\exp\left(-\left(\frac{1-\epsilon}{i^\epsilon}\right)t\right)}\prod_{t=\kappa}^{\infty}\frac{1}{1-\exp\left(-\frac{1}{\kappa}t\right)}\\&\leq\prod_{t=1}^{\kappa-1}\frac{1}{1-\exp\left(-\left(\frac{1-\epsilon}{i^\epsilon}\right)t\right)}\prod_{j=1}^{\infty}\prod_{t=j\kappa}^{(j+1)\kappa-1}\frac{1}{1-\exp\left(-\frac{1}{\kappa}t\right)}\\&\leq\prod_{t=1}^{\kappa-1}\frac{1}{1-\exp\left(-\left(\frac{1-\epsilon}{i^\epsilon}\right)t\right)}\prod_{j=1}^{\infty}\left(\frac{1}{1-\exp(-j)}\right)^\kappa\\&\leq(c_u)^\kappa\prod_{t=1}^{\kappa-1}\frac{1}{1-\exp\left(-\left(\frac{1-\epsilon}{i^\epsilon}\right)t\right)}.\end{aligned} \tag{A.5}$$

The last inequality is true as follows. For $a_j = (\exp(j) - 1)^{-1}$, using the fact that $1 + x \leq \exp(x)$, we have

$$\prod_{j=1}^{\infty} \frac{1}{1 - \exp(-j)} = \prod_{j=1}^{\infty} (1 + a_j) \leq \exp\Big(\sum_{j=1}^{\infty} a_j \Big) = c_u,$$

For $1 \leq t \leq \kappa - 1$, we have

$$\frac{1-\epsilon}{i^{\epsilon}} t \leq \frac{1-\epsilon}{i^{\epsilon}} (\kappa - 1) < 1,$$

and $1 - \exp(-x) \geq x/2$ for $x \leq 1$. Therefore, we can write

$$1 - \exp\big(- (\frac{1-\epsilon}{i^{\epsilon}}) t \big) \geq \frac{1}{2} \frac{1-\epsilon}{i^{\epsilon}} t.$$

As a result, we can further simplify the second product term in (A.4) as follows,

$$\prod_{j=1}^{b} \frac{1}{1 - \exp\big(-\frac{1-\epsilon}{i^{\epsilon}} (i-j)\big)} \leq (c_u)^{\kappa} \prod_{t=1}^{\kappa-1} 2 \frac{i^{\epsilon}}{(1-\epsilon)\, t} \leq (c_u)^{\kappa} 2^{\kappa-1} \frac{1}{(\kappa-1)!} \left(\frac{i^{\epsilon}}{1-\epsilon} \right)^{\kappa-1}. \quad \text{(A.6)}$$

For $x > 0$ and $k \in \mathbb{N}$, $x^k / k! \leq \exp(x)$. Thus,

$$\frac{e c_u^{\kappa} 2^{\kappa-1}}{(\kappa-1)! \, (1-\epsilon)^{\kappa-1}} \leq e c_u \exp(\frac{2c_u}{1-\epsilon}) =: c_e,$$

which is an $\epsilon-$dependent constant. Next we upper bound the term $\prod_{j=b+1}^{i} p_j$ using integral lower bound as below:

$$(b+1)^{1-\epsilon} + \ldots + i^{1-\epsilon} \geq \frac{1}{2-\epsilon} \left(i^{2-\epsilon} - b^{2-\epsilon} \right). \quad \text{(A.7)}$$

Thus, using the above discussion, we simplify (A.4) to get

$$\mathbb{P}\left(y_1 + \cdots + y_i \leq n \right) \leq c_e \exp\Big(- \frac{1}{2-\epsilon} \left(i^{2-\epsilon} - b^{2-\epsilon} \right) \Big) n^{i-b} i^{\epsilon(\kappa-1)}. \quad \text{(A.8)}$$

We upper bound the summation given in the statement of Lemma 5. From (A.8) and using the fact that $d \geq b$,

$$\begin{aligned}
&\sum_{i=d}^{n} i\,\mathbb{P}\left(y_1 + ... + y_i \leq n\right) \\
&\leq c_e \sum_{i=d}^{n} i \exp\left(-\frac{1}{2-\epsilon}\left(i^{2-\epsilon} - b^{2-\epsilon}\right)\right)(n+1)^{i-b} i^{\epsilon(\kappa-1)} \\
&\leq c_e (n+1)^{-b} \exp\left(\frac{b^{2-\epsilon}}{2-\epsilon}\right) \sum_{i=d}^{\infty} i \exp\left(-\frac{i^{2-\epsilon}}{2-\epsilon} + i\log(n+1) + \frac{\epsilon}{1-\epsilon}\log\left(i\right) i^{\epsilon}\right) \\
&\leq \tilde{c}_e \exp\left(-b\log(n+1) + \frac{b^{2-\epsilon}}{2-\epsilon}\right) \\
&\leq \tilde{c}_e \exp\Big(-b(b-1)^{1-\epsilon} + \frac{b^{2-\epsilon}}{2-\epsilon}\Big) \\
&= \tilde{c}_e \exp\left(-b^{2-\epsilon}\left(\left(1-\frac{1}{b}\right)^{1-\epsilon} - \frac{1}{2-\epsilon}\right)\right),
\end{aligned}$$

where we have used $b = \lceil (\log^{\frac{1}{1-\epsilon}}(n+1)) \rceil$ in the last line. The third inequality holds as for $i \geq d$, the negative term inside the second exponential function is dominating. Further, as n grows, b converges to infinity; hence, in the final term, the exponential term converges to zero. Thus, we can bound the sum with a constant. □

A.1.6 Proof of Corollary 1

Proof. We follow the same arguments as in Theorem 7 to show a $O(\log(n))$ regret. As a parallel to Lemma 5, we bound $\sum_{i=\tilde{d}}^{n-1} i\,\mathbb{P}\left(\sum_{j=1}^{i} y_j < n, \sum_{j=1}^{i+1} y_j \geq n\right)$ for independent geometric random variables $\{y_i\}_{i=1}^{n}$ with success probability $\{f(i)^{-1}\}_{i=1}^{n}$ following similar arguments to Lemma 5. Denote the smallest i that satisfies $i^{1-\epsilon_i} \geq \log(n+1)$ as b and let $\tilde{d}$ be the smallest integer i such that $\log(n+1) \leq \frac{1}{3} i^{1-\epsilon_{b+1}}$. We note that $i^{1-\epsilon_i}$ is increasing for $i \geq 1$ as ϵ_i is a decreasing sequence. Take $p_i = \exp\left(-i^{1-\epsilon_i}\right)$ and $t \geq 0$ such that $e^t = \frac{n+1}{n}(1-p_i)$, which exists for $i > b$. From (A.4), for $i > b$,

$$\mathbb{P}\left(y_1 + \cdots + y_i \leq n\right) \leq e n^{i-b} \prod_{j=b+1}^{i} p_j \prod_{j=1}^{b} \frac{1}{1 - \exp\left(-\left(i^{1-\epsilon_i} - j^{1-\epsilon_j}\right)\right)}. \tag{A.9}$$

Moreover, for $1 \leq j \leq i$, by concavity and gradient inequality, we have $\epsilon_j \geq \epsilon_i$ and

$$i^{1-\epsilon_i} - j^{1-\epsilon_j} \geq i^{1-\epsilon_i} - j^{1-\epsilon_i} \geq \frac{1-\epsilon_i}{i^{\epsilon_i}}\left(i-j\right). \tag{A.10}$$

We define $\kappa = \lceil i^{\epsilon_i}/(1-\epsilon_i) \rceil$ and using (A.6), simplify the second product term in the RHS of (A.9) to get

$$\prod_{j=1}^{b} \frac{1}{1-\exp\left(-\left(i^{1-\epsilon_i} - j^{1-\epsilon_j}\right)\right)} \leq \prod_{j=1}^{b} \frac{1}{1-\exp\left(-\frac{1-\epsilon_i}{i^{\epsilon_i}}\left(i-j\right)\right)} \leq c_u^{\kappa} 2^{\kappa-1} \frac{1}{(\kappa-1)!} \left(\frac{i^{\epsilon_i}}{1-\epsilon_i}\right)^{\kappa-1}. \tag{A.11}$$

Furthermore, using an integral lower bound, we find an upper bound for the term $\prod_{j=b+1}^{i} p_j$:

$$(b+1)^{1-\epsilon_{b+1}} + \ldots + i^{1-\epsilon_i} \geq (b+1)^{1-\epsilon_{b+1}} + \ldots + i^{1-\epsilon_{b+1}} \geq \frac{1}{2-\epsilon_{b+1}} \left(i^{2-\epsilon_{b+1}} - b^{2-\epsilon_{b+1}}\right). \tag{A.12}$$

Using (A.11), (A.12), and the fact that $\frac{e c_u^{\kappa} 2^{\kappa-1}}{(\kappa-1)!(1-\epsilon_i)^{\kappa-1}} \leq e c_u \exp(\frac{2c_u}{1-\varepsilon}) =: c_e$, we simplify (A.9) to get

$$\mathbb{P}\left(y_1 + \cdots + y_i \leq n\right) \leq c_e \exp\Big(- \frac{1}{2-\epsilon_{b+1}} \left(i^{2-\epsilon_{b+1}} - b^{2-\epsilon_{b+1}}\right)\Big) n^{i-b} i^{\epsilon_i(\kappa-1)}. \tag{A.13}$$

Finally, we can bound $\sum_{i=d}^{n-1} i\, \mathbb{P}\left(y_1 + \cdots + y_i < n, y_1 + \cdots + y_{i+1} \geq n\right)$ using (A.13) as follows

$$\begin{aligned}
&\sum_{i=\tilde{d}}^{n} i\, \mathbb{P}(y_1 + ... + y_i \leq n) \\
&\leq c_e (n+1)^{-b} \exp\Big(\frac{b^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}}\Big) \sum_{i=\tilde{d}}^{\infty} i \exp\Big(\frac{-i^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}} + i\log(n+1) + \frac{\epsilon_i}{1-\epsilon_i}\log(i) i^{\epsilon_i}\Big) \\
&\leq c_e (n+1)^{-b} \exp\Big(\frac{b^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}}\Big) \sum_{i=\tilde{d}}^{\infty} i \exp\Big(\frac{-i^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}} + \frac{i^{2-\epsilon_{b+1}}}{3} + \frac{\epsilon_i}{1-\epsilon_i}\log(i) i^{\epsilon_i}\Big) \\
&\leq c_e (n+1)^{-b} \exp\Big(\frac{b^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}}\Big),
\end{aligned} \tag{A.14}$$

where the second line follows from $\log(n+1) \leq \frac{1}{3}(\tilde{d})^{1-\epsilon_{b+1}} \leq \frac{1}{3} i^{1-\epsilon_{b+1}}$ for $i \geq \tilde{d}$. As the negative term inside the second exponential function is the dominating term, we can bound the summation with a constant independent of n. From the definition of b, we have

$$(b-1)^{1-\epsilon_{b-1}} < \log(n+1) \leq b^{1-\epsilon_b}.$$

Thus

$$(n+1)^{-b}\exp\Big(\frac{b^{2-\epsilon_{b+1}}}{2-\epsilon_{b+1}}\Big) = \exp\Big(b\Big(\frac{b^{1-\epsilon_{b+1}}}{2-\epsilon_{b+1}} - \log(n+1)\Big)\Big)$$
$$\leq \exp\Big(b\Big(\frac{b^{1-\epsilon_{b+1}}}{2-\epsilon_{b+1}} - (b-1)^{1-\epsilon_{b-1}}\Big)\Big)$$
$$= \exp\Big(-b^{2-\epsilon_{b+1}}\Big(b^{\epsilon_{b+1}-\epsilon_{b-1}}\Big(1-\frac{1}{b}\Big)^{1-\epsilon_{b-1}} - \frac{1}{2-\epsilon_{b+1}}\Big)\Big). \quad \text{(A.15)}$$

We note that as b grows to infinity, the term $\big(1-\frac{1}{b}\big)^{1-\epsilon_{b-1}}$ converges to 1, and the term $b^{2-\epsilon_{b+1}}$ converges to ∞. Since $\epsilon_{b+1} < \epsilon_{b-1}$, the term $b^{\epsilon_{b+1}-\epsilon_{b-1}}$ is less than 1. However, we also note that for large enough b,

$$\begin{aligned}
1 &> b^{\epsilon_{b+1}-\epsilon_{b-1}} \\
&= b^{\frac{\varepsilon}{\sqrt{1+\log(b+2)}} - \frac{\varepsilon}{\sqrt{1+\log(b)}}} \\
&= \exp\left(\frac{\varepsilon\log(b)}{\sqrt{1+\log(b+2)}} - \frac{\varepsilon\log(b)}{\sqrt{1+\log(b)}}\right) \\
&> \exp\big(\sqrt{\log(b+2)-1} - \sqrt{\log(b)+1}\big),
\end{aligned}$$

which follows from $\varepsilon < 1$ and $(\log(b))^2 > (\log(b+2))^2 - 1$ for sufficiently large b (since $(\log(b+2))^2 - (\log(b))^2$ converges to 0 as b grows). Thus, $b^{\epsilon_{b+1}-\epsilon_{b-1}}$ converges to 1 as b increases without bound. Using all of these, we can assert that the RHS of (A.15) goes to 0 as b increases to infinity, and so we can bound it by a constant independent of n. Finally, by repeating the arguments of Theorem 3, the expected regret is upper bounded by a linear function of $\tilde{d}$ and we conclude that the expected regret is of the order $O(\log(n))$. □

A.2 Analysis of the Multi-server Erlang-B Queueing System

A.2.1 Lemma 18

Lemma 18. *In a multi-server Erlang-B queueing system following policy Π_{Alg1}, the number of accepted arrivals that find the system empty is almost surely infinite.*

Proof. By observing Markov process $\{\tilde{X}_n\}_{n=0}^{\infty}$, we first argue that the system becomes empty

infinitely often following our proposed policy. By coupling the two systems, we get

$$\begin{aligned}
&\mathbb{P}\left(\text{returns to state 0 at a finite time} \,\middle|\, N_n = 0, X_n = x, \alpha_n = \alpha\right) \\
&\geq \mathbb{P}\left(\text{returns to state 0 at a finite time in a system that accepts all arrivals} \,\middle|\, N_n = 0\right) \\
&= 1.
\end{aligned}$$

Thus, state 0 is visited infinitely often. Let A be the event that the system admits a finite number of arrivals at instances when the server is empty, A_1 be the event that the system admits a finite number of arrivals, and A_2 be the event that the system gets empty a finite number of times. We have $\mathbb{P}(A) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) = 0$, wherein $\mathbb{P}(A_1) = 0$ follows from the same arguments as Lemma 17. □

A.2.2 Lemma 10

We first present the following lemma, which is used in the proof of Lemma 10.

Lemma 19. *[102, Theorem 2.19] let $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a martingale difference sequence such that for $\nu_i, \alpha_i > 0$, we have $\mathbb{E}\big[\exp(\tilde{\lambda} D_i) \,\big|\, \mathcal{F}_{i-1}\big] \leq \exp\left(\frac{\tilde{\lambda}^2 \nu_i^2}{2}\right)$* a.s. *for any $|\tilde{\lambda}| < 1/\alpha_i$. Then the sum $\sum_{i=1}^{n} D_i$ satisfies the concentration inequality*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{n} D_i\Big| \geq t\right) \leq 2\exp\Big(-\min\Big(\frac{t^2}{2\sum_{i=1}^{n}\nu_i^2}, \frac{t}{2\max\limits_{i=1,\ldots,n}\alpha_i}\Big)\Big).$$

Proof of Lemma 10. Without loss of generality, we assume $\mu > c/R$. Note that $\tilde{\delta}_1$ and $\tilde{\delta}_2$ are as defined in Lemma 8. We define the martingale difference sequence $\{Y_n^D\}_{n=0}^{\infty}$ as $Y_n^D = Y_{n+1}^M - Y_n^M$. To verify the conditions of Lemma 19, we argue that $\mathbb{E}\big[\exp(\tilde{\lambda}\,|Y_i^D|) \,\big|\, \mathcal{F}_{i-1}\big]$ is bounded for some positive $\tilde{\lambda}$. We show this by proving $\mathbb{E}\big[\exp(\tilde{\lambda} Y_i^D) \,\big|\, \mathcal{F}_{i-1}\big]$ and $\mathbb{E}\big[\exp(-\tilde{\lambda} Y_i^D) \,\big|\, \mathcal{F}_{i-1}\big]$ are bounded for some positive $\tilde{\lambda}$. From (2.43),

$$\mathbb{E}\big[\exp(\tilde{\lambda} Y_i^D) \,\big|\, \mathcal{F}_{i-1}\big] \leq \mathbb{E}\big[\exp\big(\tilde{\lambda} k \frac{R}{c}\tau_i\big) \,\big|\, \mathcal{F}_{i-1}\big] \leq \mathbb{E}\big[\exp\big(\tilde{\lambda} k \frac{R}{c}\zeta_i\big)\big], \tag{A.16}$$

where ζ_i is the first passage time of state zero starting from zero in a finite-state irreducible Markov chain, and thus, sub-exponential. From [101, Theorem 2.8.2], the moment generating function of ζ_i is bounded at some $\tilde{\lambda}_1$ independent of i, which leads to a finite bound. For $\mathbb{E}\big[\exp(-\tilde{\lambda} Y_i^D) \,\big|\, \mathcal{F}_{i-1}\big]$,

using (2.43),

$$\mathbb{E}\big[\exp(-\tilde{\lambda}Y_i^D)\,\Big|\,\mathcal{F}_{i-1}\big] \le \mathbb{E}\Big[\exp\Big(\tilde{\lambda}\Big(k\sum_{j=1}^{\tau_i} T_{\beta_i+j} + c_{\tilde{\delta}}\Big)\Big)\,\Big|\,\mathcal{F}_{i-1}\Big]$$
$$\le \mathbb{E}\Big[\exp\Big(\tilde{\lambda}\Big(k\sum_{j=1}^{\zeta_i} T_{\beta_i+j} + c_{\tilde{\delta}}\Big)\Big)\Big].$$

From the above inequality, it suffices to show $\sum_{j=1}^{\zeta_i} T_{\beta_i+j}$ is sub-exponential. From [101, Theorem 2.8.2], we need to argue that for some positive $\tilde{\lambda}$, $\mathbb{E}\Big[\exp\Big(\tilde{\lambda}\sum_{j=1}^{\zeta_i} T_{\beta_i+j}\Big)\Big] \le 2$. For $\tilde{\lambda} < \lambda$, we define the martingale sequence $\{M_{i,m}\}_{m=0}^{\infty}$ with respect to filtration $\{\mathcal{G}_{i,m}\}_{m=0}^{\infty}$ as

$$M_{i,m} = \frac{\exp\Big(\tilde{\lambda}\sum_{j=1}^{m} T_{\beta_i+j}\Big)}{\mathbb{E}\left[\exp\Big(\tilde{\lambda}\sum_{j=1}^{m} T_{\beta_i+j}\Big)\right]} = \frac{\exp\Big(\tilde{\lambda}\sum_{j=1}^{m} T_{\beta_i+j}\Big)}{\Big(\frac{\lambda}{\lambda-\tilde{\lambda}}\Big)^m}.$$

The passage time ζ_i is a finite-mean stopping time for the martingale sequence $\{M_{i,m}\}_{m=0}^{\infty}$. Therefore, using the optional stopping theorem for non-negative supermartingale sequences, we have

$$\mathbb{E}\left[M_{i,\zeta_i}\right] \le \mathbb{E}\left[M_{i,0}\right],$$

or

$$\mathbb{E}\big[\exp\big(\tilde{\lambda}\sum_{j=1}^{\zeta_i} T_{\beta_i+j}\big)\Big(\frac{\lambda}{\lambda-\tilde{\lambda}}\Big)^{-\zeta_i}\big] \le 1.$$

Using Cauchy-Schwarz inequality, we have

$$\mathbb{E}\big[\exp\big(\frac{\tilde{\lambda}}{2}\sum_{j=1}^{\zeta_i} T_{\beta_i+j}\big)\big] \le \sqrt{\mathbb{E}\big[\big(\frac{\lambda}{\lambda-\tilde{\lambda}}\big)^{\zeta_i}\big]} = \sqrt{\mathbb{E}\big[\exp\big(\log\big(\frac{\lambda}{\lambda-\tilde{\lambda}}\big)\zeta_i\big)\big]}. \tag{A.17}$$

As ζ_i is a sub-exponential random variable, we can choose $\tilde{\lambda}$ such that the RHS of (A.17) is less than or equal to 2 and the conditions of Lemma 19 are verified. Consequently, we apply Lemma 19 to conclude that

$$\mathbb{P}\Big(Y_n^M \le -\tilde{\delta}_1 n\Big) = \mathbb{P}\Big(\sum_{i=0}^{n-1}\big(Y_{i+1}^M - Y_i^M\big) \le -\tilde{\delta}_1 n\Big)$$
$$\le \exp\Big(-\min\Big(\frac{\tilde{\delta}_1^2 n^2}{2nv^2}, \frac{\tilde{\delta}_1 n}{2\alpha}\Big)\Big)$$
$$= \exp\left(-c_3 n\right),$$

where ν and α are positive constants independent of n. $\square$

APPENDIX B

Appendix of Chapter 3

B.1 Proofs related to problem formulation

B.1.1 Ergodicity definitions

Suppose that Markov process $\boldsymbol{X}$ on $\mathcal{X}$ with transition kernel P is irreducible, aperiodic and positive recurrent with stationary distribution μ and let $f : \mathcal{X} \mapsto [1, \infty)$ be a measurable function such that $\mu(f) := \mathbb{E}_\mu[f(Y)] < +\infty$ with $Y \sim \mu$. We are interested in conditions under which for a sequence of positive numbers $\rho := (\rho(n))_{n\geq 0}$,

$$\lim_{n\to\infty} \rho(n)\|P^n(\boldsymbol{x}, \cdot) - \mu(\cdot)\|_f = 0, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \tag{B.1}$$

where for a signed measure $\tilde{\mu}$ on $\mathcal{X}$, $\|\tilde{\mu}\|_f := \sup_{|g|\leq f} |\tilde{\mu}(g)|$. The sequence ρ is interpreted as the rate function, and three different notions of ergodicity are distinguished based on the following rate functions: $\rho(n) \equiv 1$, $\rho(n) = \zeta^n$ for $\zeta > 1$, and $\rho(n) = n^{\zeta-1}$ for $\zeta \geq 1$. Further, for each rate function ρ, we state the Foster-Lyapunov characterization of ergodicity of the Markov process $\boldsymbol{X}$, which provides sufficient conditions for (B.1) to hold.

1. If $\rho(n) \equiv 1$ for all $n \geq 0$, the Markov process $\boldsymbol{X}$ satisfying (B.1) is said to be f-**ergodic**. From [71], for an irreducible and aperiodic chain, f-ergodicity is *equivalent* to the existence of a function $V : \mathcal{X} \mapsto [0, \infty)$, a finite set C, and positive constant b such that

$$\Delta V \leq -f + b\mathbb{I}_C, \tag{B.2}$$

   where $\Delta V := PV - V$ with $PV(\boldsymbol{x}) := \sum_{\boldsymbol{x}'\in\mathcal{X}} P(\boldsymbol{x}, \boldsymbol{x}')V(\boldsymbol{x}')$. The drift condition (B.2) implies positive recurrence of the Markov process, existence of a unique stationary distribution μ, and $\mu(f) \leq b < +\infty$ ([71], Theorem 14.3.7).

2. If $\rho(n) = \zeta^n$ for some $\zeta > 1$, the Markov process $\boldsymbol{X}$ satisfying (B.1) is said to be f-**geometrically ergodic**. From [71], for an irreducible and aperiodic chain, f-geometric

ergodicity is *equivalent* to the existence of a function $V : \mathcal{X} \mapsto [1, \infty)$, a finite set C, a constant $\gamma \in (0, 1)$ and positive constant b such that

$$\Delta V \leq -(1-\gamma)V + b\mathbb{I}_C. \tag{B.3}$$

The drift condition (B.3) implies positive recurrence of the Markov process, existence of a unique stationary distribution μ, and $\mu(V) \leq \frac{b}{1-\gamma} < +\infty$ ([71], Theorem 14.3.7). Moreover, if $f(\cdot) \equiv 1$ in (B.1), then the Markov process $\boldsymbol{X}$ is called **geometrically ergodic**.

3. If $\rho(n) = n^{\zeta-1}$ for some $\zeta \geq 1$, the Markov process $\boldsymbol{X}$ satisfying (B.1) is said to be f-**polynomially ergodic**. From [71, 44], for an irreducible and aperiodic chain, the existence of a function $V : \mathcal{X} \mapsto [1, \infty)$, a finite set C, a constant $\alpha \in [0, 1)$, and positive constants c and b such that

$$\Delta V \leq -cV^{\alpha} + b\mathbb{I}_C \tag{B.4}$$

implies V_ζ-polynomial ergodicity of $\boldsymbol{X}$ at rate $\rho(n) = n^{\zeta-1}$ for all $\zeta \in [1, 1/(1-\alpha)]$ with $V_\zeta = V^{1-\zeta(1-\alpha)}$. The drift condition (B.4) implies positive recurrence of the Markov process, existence of a unique stationary distribution μ, and $\mu(V^\alpha) \leq \frac{b}{c} < +\infty$.

B.1.2 Lemma 20

Lemma 20. *For any state $\boldsymbol{x} \neq 0^d$, there exists constants $\kappa > 1$ and c_1 such that the following holds for the hitting time of state 0^d, τ_{0^d},*

$$\mathbb{E}_{\boldsymbol{x}}[\kappa^{\tau_{0^d}}] \leq c_1 V^g(\boldsymbol{x}).$$

Proof. We define $\tilde{V} := \sum_{n=0}^{\infty} {}_{0^d}P^n V^g$ where ${}_{0^d}P^n$ is the n-step taboo probability [71] defined as

$${}_{A}P^n_{\boldsymbol{x}B} = \mathbb{P}_{\boldsymbol{x}}\left(\boldsymbol{X}_n \in B, \tau_A > n\right),$$

for $A, B \subseteq \mathcal{X}$, and τ_A is the first hitting time of set A. We also let ${}_A P^0_{\boldsymbol{x}B} = \mathbb{I}_B(\boldsymbol{x})$. We have

$$\begin{aligned}
{}_{0^d}P\tilde{V}(\boldsymbol{x}) &= \sum_{\boldsymbol{y}\neq 0^d} P_{\boldsymbol{xy}}\tilde{V}(\boldsymbol{y}) \\
&= \sum_{n=0}^{\infty} \sum_{\boldsymbol{y},\boldsymbol{z}\neq 0^d} P_{\boldsymbol{xy}}\, {}_{0^d}P^n_{\boldsymbol{yz}} V^g(\boldsymbol{z}) \\
&= \sum_{n=0}^{\infty} \sum_{\boldsymbol{z}\neq 0^d} {}_{0^d}P^{n+1}_{\boldsymbol{xz}} V^g(\boldsymbol{z}) \\
&= \tilde{V}(\boldsymbol{x}) - V^g(\boldsymbol{x}).
\end{aligned}$$

In Appendix B.4.3, we argue that there exists $\tilde{b}^g > 1$ such that $\tilde{V}(\boldsymbol{y}) \leq \tilde{b}^g V^g(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{X}$, which leads to

$$ {}_{0^d}P\tilde{V} = \tilde{V} - V^g \leq \tilde{V} - \frac{1}{\tilde{b}^g}\tilde{V} = \left(1 - \frac{1}{\tilde{b}^g}\right)\tilde{V}. \tag{B.5}$$

Define Lyapunov function

$$\tilde{V}^g(\boldsymbol{x}) = \begin{cases} (1 + 2\tilde{b}^g)\tilde{V}(\boldsymbol{x}), & \text{if } \boldsymbol{x} \neq 0^d, \\ 1 + \left(2\tilde{b}^g\right)^{-1}, & \text{if } \boldsymbol{x} = 0^d. \end{cases}$$

From the above equation and (B.5), we get

$$\begin{aligned}
P\tilde{V}^g(\boldsymbol{x}) &= \sum_{\boldsymbol{y}\neq 0^d} P_{\boldsymbol{xy}}\tilde{V}^g(\boldsymbol{y}) + P_{\boldsymbol{x}0^d}\tilde{V}^g(0^d) \\
&= \sum_{\boldsymbol{y}\neq 0^d} P_{\boldsymbol{xy}}(1 + 2\tilde{b}^g)\tilde{V}(\boldsymbol{y}) + P_{\boldsymbol{x}0^d}\left(1 + \frac{1}{2\tilde{b}^g}\right) \\
&\leq \left(1 - \frac{1}{\tilde{b}^g}\right)(1 + 2\tilde{b}^g)\tilde{V}(\boldsymbol{x}) + 1 + \frac{1}{2\tilde{b}^g} \\
&\leq \left(1 - \frac{1}{\tilde{b}^g}\right)(1 + 2\tilde{b}^g)\tilde{V}(\boldsymbol{x}) + \left(1 + \frac{1}{2\tilde{b}^g}\right)\tilde{V}(\boldsymbol{x}) \\
&= \left(1 - \frac{1}{2\tilde{b}^g}\right)(1 + 2\tilde{b}^g)\tilde{V}(\boldsymbol{x}).
\end{aligned}$$

Thus,

$$P\tilde{V}^g(\boldsymbol{x}) \leq \left(1 - \frac{1}{2\tilde{b}^g}\right)\tilde{V}^g(\boldsymbol{x}) + \left(1 - \frac{1}{2\tilde{b}^g}\right)(1 + 2\tilde{b}^g)\tilde{V}(0^d)\mathbb{I}_{0^d}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}\,.$$

To find an upper bound for $\mathbb{E}_{\boldsymbol{x}}[\kappa^{\tau_{0^d}}]$, we apply [71, Theorem 15.2.5], which is a generalization of

Lemma 25. For any $1 \leq \kappa \leq \frac{2\tilde{b}^g}{2\tilde{b}^g-1}$, there exists $\epsilon > 0$ such that

$$\mathbb{E}_{\boldsymbol{x}}\Big[\sum_{i=0}^{\tau_{0^d}-1} \tilde{V}^g(\boldsymbol{X}_i)\kappa^i\Big] \leq \epsilon^{-1}\kappa^{-1}\tilde{V}^g(\boldsymbol{x}).$$

As $\tilde{V}^g(\boldsymbol{y}) \geq 1$ for all $\boldsymbol{y} \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}[\kappa^{\tau_{0^d}}] &\leq \kappa\mathbb{E}_{\boldsymbol{x}}\Big[\sum_{i=0}^{\tau_{0^d}-1} \tilde{V}^g(\boldsymbol{X}_i)\kappa^i\Big] \\
&\leq \epsilon^{-1}\tilde{V}^g(\boldsymbol{x}) \\
&= \epsilon^{-1}\left(1+2\tilde{b}^g\right)\tilde{V}(\boldsymbol{x}) \\
&\leq \tilde{b}^g\epsilon^{-1}\left(1+2\tilde{b}^g\right)V^g(\boldsymbol{x}),
\end{aligned}$$

and the claim holds for any $\kappa \in [1, \frac{2\tilde{b}^g}{2\tilde{b}^g-1}]$ and $c_1 = \tilde{b}^g\epsilon^{-1}\left(1+2\tilde{b}^g\right)$. □

B.1.3 Poisson equation

For an irreducible Markov process on the countably-infinite space $\mathcal{X}$ with time-homogeneous transition kernel P and cost function $\bar{c}(\cdot)$, a solution pair to the Poisson equation [65] is a scalar J and function $v(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ such that $J + v = \bar{c} + Pv$, where $v(\boldsymbol{z}) = 0$ for some $\boldsymbol{z} \in \mathcal{X}$. If the Markov process is also positive recurrent and $\mathbb{E}_{\boldsymbol{x}}\left[\sum_{i=0}^{\tau_{\boldsymbol{y}}-1} |\bar{c}(\boldsymbol{X}(i))|\right] < \infty$, where $\tau_{\boldsymbol{y}}$ is the first hitting time of some state $\boldsymbol{y} \in \mathcal{X}$, then solution pair (J, v) given as

$$J = \frac{\mathbb{E}_{\boldsymbol{y}}\left[\sum_{i=0}^{\tau_{\boldsymbol{y}}-1} |\bar{c}(\boldsymbol{X}(i))|\right]}{\mathbb{E}_{\boldsymbol{y}}[\tau_{\boldsymbol{y}}]} \text{ and } v(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}}\Big[\sum_{i=0}^{\tau_{\boldsymbol{x}}-1} |\bar{c}(\boldsymbol{X}(i))|\Big] - J\mathbb{E}_{\boldsymbol{x}}[\tau_{\boldsymbol{y}}], \quad \forall \boldsymbol{x} \in \mathcal{X},$$

is a solution to the Poisson equation $J + v = \bar{c} + Pv$ with $v(\boldsymbol{z}) = 0$ [65, Theorem 9.5].

Lemma 21. *Consider Markov Decision Processes $(\mathcal{X}, \mathcal{A}, c, P_\theta)$ governed by parameter $\theta \in \Theta$ following the best-in-class policy π^*_θ. Then the pair $\left(J(\theta), v^{\pi^*_\theta}\right)$ given as*

$$J(\theta) := \frac{\bar{C}^{\pi^*_\theta}(0^d)}{\mathbb{E}^{\pi^*_\theta}_{0^d}[\tau_{0^d}]} \text{ and } v^{\pi^*_\theta}(\boldsymbol{x}) = \bar{C}^{\pi^*_\theta}(\boldsymbol{x}) - J(\theta)\mathbb{E}^{\pi^*_\theta}_x[\tau_{0^d}], \quad \forall \boldsymbol{x} \in \mathcal{X},$$

is a solution to the Poisson equation $v + J = c + P^{\pi^*_\theta}_\theta v$, where $v^{\pi^*_\theta}(0^d) = 0$ and $\bar{C}^{\pi^*_\theta}(\boldsymbol{x}) = \mathbb{E}^{\pi^*_\theta}_{\boldsymbol{x}}\left[\sum_{i=0}^{\tau_{0^d}-1} c(\boldsymbol{X}(i), \pi^*_\theta(\boldsymbol{X}(i)))\right]$.

Proof. From [65, Theorem 9.5], a solution pair to the Poisson equation exists if $\mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}[\tau_{0^d}]$ and $\bar{C}^{\pi_\theta^*}(\boldsymbol{x})$ are finite for all $\boldsymbol{x} \in \mathcal{X}$. The former follows from positive recurrence assumed in Assumption 3 and for the latter, from Assumptions 1 and 2,

$$\begin{aligned}
\bar{C}^{\pi_\theta^*}(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}\Big[\sum_{i=0}^{\tau_{0^d}-1} c(\boldsymbol{X}(i), \pi_\theta^*(\boldsymbol{X}(i)))\Big] \\
&\leq \mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}\Big[\sum_{i=0}^{\tau_{0^d}-1}\sum_{j=1}^{d} K\left(X_j(i)\right)^r\Big] \\
&\leq \mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}\Big[\sum_{i=0}^{\tau_{0^d}-1} Kd\left(\|\boldsymbol{x}\|_\infty + hi\right)^r\Big] \\
&\leq \mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}\left[Kd\left(\|\boldsymbol{x}\|_\infty + h\tau_{0^d}\right)^r \tau_{0^d}\right],
\end{aligned}$$

which is finite from geometric ergodicity (Assumption 3) and the discussion following that. □

B.2 Proofs of regret analysis

In the subsequent sections, several equalities and inequalities in the proofs are between random variables and hold almost surely (*a.s.*). Throughout the remainder, we will omit the explicit mention of *a.s.*, but any such statement should be interpreted in this context.

B.2.1 Proof of Lemma 11

Proof. Let $\{\alpha_i\}_{i\geq 0}$ be the sequence of hitting times of state 0^d starting from 0^d (set $\alpha_0 = 0$). Define $\tau_{0^d}^{(i)}$ as the length of the i-th recurrence time of state 0^d for $i \in \mathbb{N}$, i.e., $\tau_{0^d}^{(i)} = \alpha_i - \alpha_{i-1}$. For simplicity, we take $\tau_{0^d} = \tau_{0^d}^{(1)}$. Each such recurrence time is generated using policy $\pi_{\theta_i}^*$ that is determined using the algorithm in operation in an MDP governed by parameter $\boldsymbol{\theta}^*$. Furthermore, $\{\tau_{0^d}^{(i)}\}_{i\in\mathbb{N}}$ are independent with length at least 1, but they need not be identically distributed. The time T can be in the middle of one of these recurrence times, hence the current recurrence interval count is $N(T) = \inf\{n : \sum_{i=1}^{n} \tau_{0^d}^{(i)} \geq T\}$. Note that the lower bound of 1 on every $\tau_{0^d}^{(i)}$ says that $N(T) \leq T$ a.s. Further, from the skip-free to the right property, the most any component of state can increase in during recurrence time $\tau_{0^d}^{(i)}$ is $h\tau_{0^d}^{(i)}$. Hence, the most any component of the state (and also the $\|\cdot\|_\infty$ norm of the state) can increase is given by $h\max_{i=1,\ldots,T} \tau_{0^d}^{(i)}$ where the random variables are independent with geometrically decaying tails with a worst case rate of

$$\sup_{\theta_1,\theta_2\in\Theta} \tilde{\gamma}_{\theta_1,\theta_2}^g = 1 - \Big(\sup_{\theta_1,\theta_2\in\Theta} \tilde{b}_{\theta_1,\theta_2}^g\Big)^{-1};$$

see Lemma 24. From Lemma 23, we have

$$
\begin{aligned}
\tilde{b}^g_{\theta_1,\theta_2} &= \frac{3b^g_{\theta_1,\theta_2}+1}{1-\gamma^g_{\theta_1,\theta_2}}\Big(|C^g_{\theta_1,\theta_2}|^2 \max\Big(1, \max_{\boldsymbol{u}\in C^g_{\theta_1,\theta_2}\setminus\{0^d\}} \mathbb{E}^{\pi^*_{\theta_2}}_{\boldsymbol{u}}[\tau_{0^d}]\Big)\Big) \\
&\le \frac{3b^g_*+1}{1-\gamma^g_*}\left(|C^g_*|^2 \max\Big(1, \sup_{\substack{\boldsymbol{u}\in C^g_*\setminus\{0^d\}\\ \theta_1,\theta_2\in\Theta}} \phi^p_{\theta_1,\theta_2}(1)\Big(V^p_{\theta_1,\theta_2}(\boldsymbol{u}) + b^p_{\theta_1,\theta_2}\alpha_{C^p_{\theta_1,\theta_2}}\Big)\Big)\right) \\
&\le \frac{3b^g_*+1}{1-\gamma^g_*}\left(|C^g_*|^2 \max\Big(1, \sup_{\substack{\boldsymbol{u}\in C^g_*\setminus\{0^d\}\\ \theta_1,\theta_2\in\Theta}} \frac{1}{\beta^p_{\theta_1,\theta_2}}\Big(s^p_{\theta_1,\theta_2}\|\boldsymbol{u}\|^{r^p_{\theta_1,\theta_2}}_\infty + \frac{b^p_{\theta_1,\theta_2}}{\min_{\boldsymbol{y}\in C^p_{\theta_1,\theta_2}} K_{\theta_1,\theta_2}(\boldsymbol{y})}\Big)\Big)\right) \\
&\le \frac{3b^g_*+1}{1-\gamma^g_*}\left(|C^g_*|^2 \max\Big(1, \sup_{\boldsymbol{u}\in C^g_*\setminus\{0^d\}} \frac{1}{\beta^p_*}\Big(s^p_*\|\boldsymbol{u}\|^{r^p_*}_\infty + \frac{b^p_*}{K_*}\Big)\Big)\right) \qquad \text{(B.6)} \\
&:= \tilde{b}^g_*,
\end{aligned}
$$

and we define $\tilde{\gamma}^g_* := 1-(\tilde{b}^g_*)^{-1}$. From the definition of $b^g_{\theta_1,\theta_2}$ in Assumption 3, $b^g_{\theta_1,\theta_2}$ is greater than or equal to 2. Thus, $\tilde{b}^g_{\theta_1,\theta_2} \ge 7$ and we have

$$
\sup_{\theta_1,\theta_2\in\Theta} c^g_{\theta_1,\theta_2} = \sup_{\theta_1,\theta_2\in\Theta} \frac{b^g_{\theta_1,\theta_2}\left(\tilde{b}^g_{\theta_1,\theta_2}\right)^2}{\tilde{b}^g_{\theta_1,\theta_2}-1} \le \frac{b^g_*\left(\tilde{b}^g_*\right)^2}{6} := c^g_*,
$$

and as a result of Lemma 24,

$$
\mathbb{P}_{0^d}(\tau^{(i)}_{0^d} > n) \le c^g_*\,(\gamma^g_*)^n\,, \qquad 1\le i\le T. \qquad \text{(B.7)}
$$

We upper bound $\mathbb{E}\left[M_{\boldsymbol{\theta}^*}^T\right]$ using the independence of $\{\tau_{0^d}^{(i)}\}_{i\in\mathbb{N}}$ and the above equation,

$$\begin{aligned}
\mathbb{E}\left[M_{\boldsymbol{\theta}^*}^T\right] &\leq h\,\mathbb{E}[\max_{1\leq i\leq T}\tau_{0^d}^{(i)}] \\
&= h\sum_{n=0}^{\infty}\mathbb{P}(\max_{1\leq i\leq T}\tau_{0^d}^{(i)} > n) \\
&= h\sum_{n=0}^{\infty}\left(1-\mathbb{P}(\max_{1\leq i\leq T}\tau_{0^d}^{(i)} \leq n)\right) \\
&= h\sum_{n=0}^{\infty}\left(1-\prod_{i=1}^{T}\mathbb{P}\left(\tau_{0^d}^{(i)} \leq n\right)\right) \\
&\leq hn_0 + h\sum_{n=n_0}^{\infty} 1-\left(1-c_*^g\left(\gamma_*^g\right)^{n_0}\left(\gamma_*^g\right)^{n-n_0}\right)^T \\
&\leq h(n_0+1) + h\sum_{n=n_0+1}^{\infty} 1-\left(1-\left(\gamma_*^g\right)^{n-n_0}\right)^T,
\end{aligned}$$

where n_0 is the smallest $n \geq 0$ such that $c_*^g\left(\gamma_*^g\right)^n < 1$. By Reimann sum approximation, we get

$$\begin{aligned}
\mathbb{E}\left[M_{\boldsymbol{\theta}^*}^T\right] &\leq h(n_0+1) + h\sum_{n=1}^{\infty} 1-\left(1-\left(\gamma_*^g\right)^n\right)^T \\
&< h(n_0+1) + h\int_0^{\infty} 1-\left(1-\left(\gamma_*^g\right)^u\right)^T\,du \\
&= h(n_0+1) + \frac{h}{\log\gamma_*^g}\int_0^1 \frac{1-u^T}{1-u}\,du \\
&\leq h(n_0+1) + \frac{h}{\log\gamma_*^g}\left(\log T+1\right),
\end{aligned}$$

where the last inequality follows from $\sum_{n=1}^{T} n^{-1} \leq \log T + 1$ and thus $\mathbb{E}\left[M_{\boldsymbol{\theta}^*}^T\right]$ is $O(h\log T)$. We now extend the result to moments of order greater than one. From (B.7), for $1 \leq i \leq T$,

$$\mathbb{P}_{0^d}(\tau_{0^d}^{(i)} > n) \leq c_*^g\left(\gamma_*^g\right)^n = c_*^g\left(\gamma_*^g\right)^{n_0}\left(\gamma_*^g\right)^{n-n_0} < \left(\gamma_*^g\right)^{n-n_0}.$$

For $n \geq n_0$, let $t = n - n_0 \geq 0$ and $Y_i = \max(\tau_{0^d}^{(i)} - n_0, 0)$ to get

$$\mathbb{P}_{0^d}(Y_i > t) = \mathbb{P}_{0^d}(\tau_{0^d}^{(i)} - n_0 > t) < \left(\gamma_*^g\right)^t,$$

which means random variables $\{Y_i\}_{i=1}^T$ are stochastically dominated by independent and identically distributed geometric random variables with parameter $1-\gamma_*^g$. Furthermore, [93] argues that the p-th moment of the maximum of T independent and identically distributed geometric random

variables is $O(\log^p T)$. Thus, the p-th moment of $\max_{1\le i\le T} Y_i$ is $O(\log^p T)$ and

$$\begin{aligned}\max_{1\le i\le T} Y_i &= \max(\tau_{0^d}^{(1)} - n_0, \dots, \tau_{0^d}^{(T)} - n_0, 0)\\ &= \max(\tau_{0^d}^{(1)}, \dots, \tau_{0^d}^{(T)}, n_0) - n_0\\ &\ge \max(\tau_{0^d}^{(1)}, \dots, \tau_{0^d}^{(T)}) - n_0\\ &\ge h^{-1} M_{\boldsymbol{\theta}^*}^T - n_0,\end{aligned}$$

which gives

$$\mathbb{E}\left[\left(M_{\boldsymbol{\theta}^*}^T\right)^p\right] \le h^p\, \mathbb{E}\left[\left(\max_{1\le i\le T} \tau_{0^d}^{(i)}\right)^p\right] \le h^p\, \mathbb{E}\left[\left(\max_{1\le i\le T} Y_i + n_0\right)^p\right].$$

Since the right-hand side of the above equation is $O(h^p \log^p T)$, the claim is proved. □

B.2.2 Proof of Lemma 12

Proof. Let $K_M(\boldsymbol{x}, a)$ be the number of episodes k such that $1 \le k \le K_T$ and in which the number of visits to the state-action pair $(\boldsymbol{x}, a)$ is increased more than twice at episode k, or

$$K_M(\boldsymbol{x}, a) = |\{k \le K_T : N_{\tilde{t}_{k+1}}(\boldsymbol{x}, a) > 2N_{t_k}(\boldsymbol{x}, a)\}|.$$

As for every episode in the above set the number of visits to (x, a) doubles,

$$K_M(\boldsymbol{x}, a) \le \log_2(N_{T+1}(\boldsymbol{x}, a)) + 1,$$

and we can upper bound K_M as follows

$$\begin{aligned}K_M &= \sum_{\boldsymbol{x}\in\mathcal{X}, a\in\mathcal{A}} K_M(\boldsymbol{x}, a)\\ &= \sum_{\substack{\|\boldsymbol{x}\|_\infty \le M_{\boldsymbol{\theta}^*}^T\\ a\in\mathcal{A}}} K_M(\boldsymbol{x}, a)\\ &\le \sum_{\substack{\|\boldsymbol{x}\|_\infty \le M_{\boldsymbol{\theta}^*}^T\\ a\in\mathcal{A}}} (1 + \log_2 N_{T+1}(\boldsymbol{x}, a))\\ &\le |\,\mathcal{A}\,| \left(M_{\boldsymbol{\theta}^*}^T + 1\right)^d (1 + \log_2 T).\end{aligned}$$

This completes the proof. □

B.2.3 Proof of Lemma 13

Proof. We define macro episodes with start times t_{n_k}, $k = 1, 2, \ldots, K_M + 1$ where $t_{n_1} = t_1$, $t_{n_{K_M+1}} = T + 1$ (which is equivalent to $n_{K_M+1} = K_T + 1$), and for $1 < k < K_M + 1$

$$t_{n_{k+1}} = \min\{t_j > t_{n_k} : \quad N_{t_j}(\boldsymbol{x}, a) > 2N_{t_{j-1}}(\boldsymbol{x}, a) \text{ for some } (\boldsymbol{x}, a)\},$$

which are episodes wherein the second stopping criterion is triggered. Any episode (except for the last episode) in a macro episode must be triggered by the first stopping criterion; equivalently, $\tilde{T}_j = \tilde{T}_{j-1} + 1$ for all $j = n_k, n_k + 1, \ldots, n_{k+1} - 2$. For $1 \leq k \leq K_M$, let $T_k^M = \sum_{j=n_k}^{n_{k+1}-1} T_j$ be the length of the k-th macro episode. We have

$$T_k^M = \sum_{j=n_k}^{n_{k+1}-1} T_j \geq \sum_{j=n_k}^{n_{k+1}-1} \tilde{T}_j \geq 1 + \sum_{j=n_k}^{n_{k+1}-2} (j - n_k + 2) = 0.5(n_{k+1} - n_k)(n_{k+1} - n_k + 1).$$

Consequently, $n_{k+1} - n_k \leq \sqrt{2T_k^M}$ for all $1 \leq k \leq K_M$. From this, we obtain

$$K_T = n_{K_M+1} - 1 = \sum_{k=1}^{K_M} (n_{k+1} - n_k) \leq \sum_{k=1}^{K_M} \sqrt{2T_k^M}.$$

Using the above equation and the fact that $\sum_{k=1}^{K_M} T_k^M = T$ we get

$$K_T \leq \sum_{k=1}^{K_M} \sqrt{2T_k^M} \leq \sqrt{K_M \sum_{k=1}^{K_M} 2T_k^M} = \sqrt{2K_M T}.$$

Finally, from Lemma 12 we get

$$K_T \leq \sqrt{2K_M T} \leq 2\sqrt{|\mathcal{A}| \left(M_{\boldsymbol{\theta}^*}^T + 1\right)^d T \log_2 T}.$$

This completes the proof. □

B.2.4 Proof of Lemma 14

Proof. Let $E_k = T_k - \tilde{T}_k \geq 0$ be the settling time needed to return to state 0^d after a stopping criterion is realized in episode k. We have

$$\begin{aligned} R_0 &= \mathbb{E}\Big[\sum_{k=1}^{K_T} T_k J(\theta_k)\Big] - T\,\mathbb{E}\Big[J(\boldsymbol{\theta}^*)\Big] \\ &= \mathbb{E}\Big[\sum_{k=1}^{K_T} \tilde{T}_k J(\theta_k)\Big] + \mathbb{E}\Big[\sum_{k=1}^{K_T} E_k J(\theta_k)\Big] - T\,\mathbb{E}\Big[J(\boldsymbol{\theta}^*)\Big]. \end{aligned} \tag{B.8}$$

We first simplify the first term in the above summation. From the monotone convergence theorem,

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} \tilde{T}_k J(\theta_k)\Big] = \sum_{k=1}^{\infty} \mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} \tilde{T}_k J(\theta_k)\Big].$$

Note that the first stopping criterion of Algorithm 2 ensures that $\tilde{T}_k \leq \tilde{T}_{k-1} + 1$ at all episodes $k \geq 1$. Hence

$$\mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} \tilde{T}_k J(\theta_k)\Big] \leq \mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} (\tilde{T}_{k-1} + 1) J(\theta_k)\Big].$$

Since $\mathbb{I}_{\{t_k \leq T\}}(\tilde{T}_{k-1} + 1)$ is measurable with respect to $\mathcal{H}_{t_k}$, by (3.15) we get

$$\mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} (\tilde{T}_{k-1} + 1) J(\theta_k)\Big] = \mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} (\tilde{T}_{k-1} + 1) J(\boldsymbol{\theta}^*)\Big].$$

Therefore,

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} \tilde{T}_k J(\theta_k)\Big] \leq \sum_{k=1}^{\infty} \mathbb{E}\Big[\mathbb{I}_{\{t_k \leq T\}} (\tilde{T}_{k-1} + 1) J(\boldsymbol{\theta}^*)\Big] = \mathbb{E}\Big[\sum_{k=1}^{K_T} (\tilde{T}_{k-1} + 1) J(\boldsymbol{\theta}^*)\Big].$$

Thus,

$$\begin{aligned} \mathbb{E}\Big[\sum_{k=1}^{K_T} \tilde{T}_k J(\theta_k)\Big] - T\,\mathbb{E}\Big[J(\boldsymbol{\theta}^*)\Big] &\leq \mathbb{E}\Big[J(\boldsymbol{\theta}^*) \sum_{k=1}^{K_T} (\tilde{T}_{k-1} + 1)\Big] - \mathbb{E}\Big[J(\boldsymbol{\theta}^*) \sum_{k=1}^{K_T} T_k\Big] \\ &= \mathbb{E}\Big[J(\boldsymbol{\theta}^*)\Big(K_T + 1 - T_{K_T} - \sum_{k=1}^{K_T - 1} E_k\Big)\Big] \\ &\leq \mathbb{E}\Big[J(\boldsymbol{\theta}^*) K_T\Big]. \end{aligned} \tag{B.9}$$

For the second term in (B.8), from Assumption 5

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} E_k J(\theta_k)\Big] \leq J^* \,\mathbb{E}\Big[\sum_{k=1}^{K_T} E_k\Big] \leq J^* \,\mathbb{E}[K_T \max_{1\leq i\leq T} \tau_{0^d}^{(i)}]. \tag{B.10}$$

Substitutinh (B.9) and (B.10) in (B.8), we get

$$\begin{aligned} R_0 &\leq \mathbb{E}\left[K_T J(\boldsymbol{\theta}^*)\right] + J^* \,\mathbb{E}[K_T \max_{1\leq i\leq T} \tau_{0^d}^{(i)}] \\ &\leq J^* \,\mathbb{E}\left[K_T\right] + J^* \,\mathbb{E}[K_T \max_{1\leq i\leq T} \tau_{0^d}^{(i)}] \\ &= J^* \,\mathbb{E}\left[K_T\Big(\max_{1\leq i\leq T} \tau_{0^d}^{(i)} + 1\Big)\right]. \end{aligned}$$

□

B.2.5 Proof of Lemma 15

Proof. We note that the state of the MDP is equal to 0^d at the beginning of all episodes and the relative value function $v(\boldsymbol{x};\theta)$ is equal to 0 at $\boldsymbol{x} = 0^d$ for all θ. Thus,

$$\begin{aligned} R_1 &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big[v\left(\boldsymbol{X}\left(t\right);\theta_k\right) - v\left(\boldsymbol{X}\left(t+1\right);\theta_k\right)\Big]\Big] \\ &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\Big[v\left(\boldsymbol{X}\left(t_k\right);\theta_k\right) - v\left(\boldsymbol{X}\left(t_{k+1}\right);\theta_k\right)\Big]\Big] \\ &= \mathbb{E}\Big[\sum_{k=1}^{K_T-1}\Big[v\left(0^d;\theta_k\right) - v\left(0^d;\theta_k\right)\Big] + v\left(0^d;\theta_{K_T}\right) - v\left(\boldsymbol{X}(T+1);\theta_{K_T}\right)\Big] \\ &= -\,\mathbb{E}[v\left(\boldsymbol{X}(T+1);\theta_{K_T}\right)]. \end{aligned}$$

From the lower bound derived for the relative value function in (3.17),

$$-v(\boldsymbol{x};\theta) \leq J^* \mathbb{E}_{\boldsymbol{x}}^{\pi_\theta^*}[\tau_{0^d}] \leq \frac{J^*}{\beta_*^p}\Big(s_*^p \|\boldsymbol{x}\|_\infty^{r_*^p} + \frac{b_*^p}{K_*}\Big),$$

where the second inequality follows from (B.6) in the proof of Lemma 11. We also note that $\|\boldsymbol{X}(T+1)\|_\infty \leq M_{\boldsymbol{\theta}^*}^T + h$. Thus,

$$R_1 = -\,\mathbb{E}[v\left(\boldsymbol{X}(T+1);\theta_{K_T}\right)] \leq \mathbb{E}\Big[\frac{J^*}{\beta_*^p}\Big(s_*^p (M_{\boldsymbol{\theta}^*}^T + h)^{r_*^p} + \frac{b_*^p}{K_*}\Big)\Big].$$

From the inequality $(a+b)^r \leq 2^r(a^r+b^r)$, we have

$$R_1 \leq \frac{J^* 2^{r_*^p} s_*^p}{\beta_*^p} \mathbb{E}\left[\left(M_{\boldsymbol{\theta}^*}^T\right)^{r_*^p} \right] + \frac{J^*}{\beta_*^p} \left(s_*^p (2h)^{r_*^p} + \frac{b_*^p}{K_*} \right).$$

□

B.2.6 Proof of Lemma 16

Proof. Let $\boldsymbol{Z}(t) = \left(\boldsymbol{X}(t), \pi_{\theta_k}^*(\boldsymbol{X}(t))\right)$ be the state-action pair at $t_k \leq t < t_{k+1}$. R_2 can be upper bounded as

$$\begin{aligned}
R_2 &= \mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \Big[v(\boldsymbol{X}(t+1);\theta_k) - \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}\Big(\boldsymbol{y} \,\Big|\, \boldsymbol{X}(t), \pi_{\theta_k}^*(\boldsymbol{X}(t))\Big) v(\boldsymbol{y};\theta_k) \Big]\Big] \\
&\leq \mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \Big[\sum_{\boldsymbol{y}\in\mathcal{X}} |P_{\boldsymbol{\theta}^*}(\boldsymbol{y}|\boldsymbol{Z}(t)) - P_{\theta_k}(\boldsymbol{y}|\boldsymbol{Z}(t))|\, |v(\boldsymbol{y};\theta_k)| \Big]\Big] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\Big[\Big(\max_{\substack{1\leq k\leq K_T \\ \|\boldsymbol{x}\|_\infty \leq M_{\boldsymbol{\theta}^*}^T}} |v(\boldsymbol{x};\theta_k)| \Big) \|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{Z}(t)) - P_{\theta_k}(\cdot|\boldsymbol{Z}(t))\|_1 \Big]. \qquad \text{(B.11)}
\end{aligned}$$

We have

$$\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{Z}(t)) - P_{\theta_k}(\cdot|\boldsymbol{Z}(t))\|_1 \leq \|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 + \|P_{\theta_k}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1,$$

where $P_{\hat{\theta}_k}(\boldsymbol{y}|\boldsymbol{Z}(t))$ is the empirical transition probability defined as

$$P_{\hat{\theta}_k}(\boldsymbol{y}|\boldsymbol{Z}(t)) = \frac{N_{t_k}(\boldsymbol{Z}(t), \boldsymbol{y})}{\max(1, N_{t_k}(\boldsymbol{Z}(t)))},$$

and for any tuple $(\boldsymbol{x}, a, \boldsymbol{y})$, we define $N_1(\boldsymbol{x}, a, \boldsymbol{y}) = 0$ and for $t > 1$,

$$N_t(\boldsymbol{x}, a, \boldsymbol{y}) = |\{t_k \leq i < \tilde{t}_{k+1} \leq t \text{ for some } k \geq 1 : (\boldsymbol{X}(i), A(i), \boldsymbol{X}(i+1)) = (\boldsymbol{x}, a, \boldsymbol{y})\}|.$$

Thus, from (B.11) and defining random variable $v_M = \max_{\substack{1\leq k\leq K_T \\ \|\boldsymbol{x}\|_\infty \leq M_{\boldsymbol{\theta}^*}^T}} |v(\boldsymbol{x};\theta_k)|$,

$$R_2 \leq \sum_{t=1}^{T} \mathbb{E}\Big[v_M \|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 \Big] + \sum_{t=1}^{T} \mathbb{E}\Big[v_M \|P_{\theta_k}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 \Big]. \qquad \text{(B.12)}$$

We define set B_k as the set of parameters θ for which the transition kernel $P_\theta(\cdot|\boldsymbol{z})$ is close to the empirical transition kernel $P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})$ at episode k for every state-action pair $\boldsymbol{z} = (\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A}$, or

$$B_k = \left\{\theta : \|P_\theta(\cdot|\boldsymbol{z}) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1 \leq \beta_k(\boldsymbol{z}),\ \boldsymbol{z} = (\boldsymbol{x}, a) \in \{0, 1, \cdots, hT\}^d \times \mathcal{A}\right\},$$

where $\beta_k(\boldsymbol{z}) = \sqrt{\frac{14\prod_{i=1}^d (x_i+h)}{\max(1, N_{t_k}(\boldsymbol{z}))} \log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}$ for $\boldsymbol{x} = (x_1, \ldots, x_d)$ and some $0 < \tilde{\delta} < 1$, which will be determined later. We simplify the ℓ_1-difference of the real and empirical transition kernels as follows

$$\begin{aligned}
&\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 \\
&= \mathbb{I}_{\{\boldsymbol{\theta}^* \notin B_k\}} \|P_{\theta_*}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 + \mathbb{I}_{\{\boldsymbol{\theta}^* \in B_k\}} \|P_{\theta_*}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 \\
&\leq 2\mathbb{I}_{\{\boldsymbol{\theta}^* \notin B_k\}} + \beta_k(\boldsymbol{Z}(t)).
\end{aligned}$$

Similarly, we have

$$\|P_{\theta_k}(\cdot|\boldsymbol{Z}(t)) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{Z}(t))\|_1 \leq 2\mathbb{I}_{\{\theta_k \notin B_k\}} + \beta_k(\boldsymbol{Z}(t)).$$

Substituting in (B.12), we get

$$R_2 \leq \mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2v_M \left[\mathbb{I}_{\{\boldsymbol{\theta}^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right]\Big] + \mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2v_M \beta_k(\boldsymbol{Z}(t))\Big]. \tag{B.13}$$

We first find an upper bound for $v_M = \max_{\substack{1 \leq k \leq K_T \\ \|\boldsymbol{x}\|_\infty \leq M^T_{\boldsymbol{\theta}^*}}} |v(\boldsymbol{x}; \theta_k)|$ using the bounds derived in (3.16) and (3.17). From (3.16),

$$\begin{aligned}
v(\boldsymbol{x}; \theta_k) &\leq \mathbb{E}_{\boldsymbol{x}}^{\pi^*_{\theta_k}} \left[Kd\left(\|\boldsymbol{x}\|_\infty + h\tau_{0^d}\right)^r \tau_{0^d}\right] \\
&\leq \mathbb{E}_{\boldsymbol{x}}^{\pi^*_{\theta_k}} \left[2^r Kd\left(\|\boldsymbol{x}\|_\infty^r + h^r(\tau_{0^d})^r\right)\tau_{0^d}\right] \\
&= Kd(2\|\boldsymbol{x}\|_\infty)^r \mathbb{E}_{\boldsymbol{x}}^{\pi^*_{\theta_k}} \left[\tau_{0^d}\right] + Kd(2h)^r \mathbb{E}_{\boldsymbol{x}}^{\pi^*_{\theta_k}} \left[(\tau_{0^d})^{r+1}\right] \\
&\leq Kd2^r \left(\|\boldsymbol{x}\|_\infty^r + h^r\right) \mathbb{E}_{\boldsymbol{x}}^{\pi^*_{\theta_k}} \left[(\tau_{0^d})^{r+1}\right] \\
&\leq Kd(r+1)2^r \left(\|\boldsymbol{x}\|_\infty^r + h^r\right) \phi^p_{\theta_k}(r+1) \left(V^p_{\theta_k}(\boldsymbol{x}) + b^p_{\theta_k} \alpha_{C^p_{\theta_k}}\right) \\
&\leq Kd(r+1)2^r \left(\|\boldsymbol{x}\|_\infty^r + h^r\right) \phi^p_{\theta_k}(r+1) \left(s^p_* \|\boldsymbol{x}\|^{r^p_*}_\infty + b^p_*(K_*)^{-1}\right),
\end{aligned} \tag{B.14}$$

where the second line follows from the inequality $(a+b)^r \leq 2^r(a^r+b^r)$, the fifth line from Lemma 23, and the last line from Assumption 4 and (B.6). We further have

$$\begin{aligned}\phi^p_{\theta_1,\theta_2}(r+1) &= \prod_{j=1}^{r+1} \frac{1}{\beta^{\eta_j}_{\theta_1,\theta_2}} \left(2^{j-1} + (j-1)\,\alpha_{C^p_{\theta_1,\theta_2}} b^{\eta_j}_{\theta_1,\theta_2}\right) \\ &\leq \prod_{j=1}^{r+1} \frac{r+1}{\min(1,\beta^p_*)} \left(2^{j-1} + (j-1)\,(K_*)^{-1} b^{\eta_j}_{\theta_1,\theta_2}\right),\end{aligned}$$

where using the definition of $b^{\eta_j}_{\theta_1,\theta_2}$ in (B.23),

$$b^{\eta_j}_{\theta_1,\theta_2} = \left(b^p_{\theta_1,\theta_2}\right)^{\eta_j} + \eta_j \tilde{\beta}^p_{\theta_1,\theta_2} \max\left(1, \left(\tilde{\beta}^p_{\theta_1,\theta_2}\right)^{(\alpha^p_{\theta_1,\theta_2}+\eta_j-1)/(1-\alpha^p_{\theta_1,\theta_2})}\right) \leq 1 + b^p_* + \beta^p_*.$$

We also define

$$\phi^p_*(r+1) := \prod_{j=1}^{r+1} \frac{r+1}{\min(1,\beta^p_*)} \left(2^{j-1} + (j-1)\,(K_*)^{-1}(1+b^p_*+\beta^p_*)\right).$$

We next find a lower bound for $v(\boldsymbol{x};\theta_k)$ using (3.17) as follows:

$$v(\boldsymbol{x};\theta_k) \geq -J^* \mathbb{E}^{\pi^*_{\theta_k}}_{\boldsymbol{x}}\left[\tau_{0^d}\right] \geq -\frac{J^*}{\beta^p_*}\left(s^p_* \|\boldsymbol{x}\|^{r^p_*}_\infty + \frac{b^p_*}{K_*}\right).$$

Combining (B.14) and the above equation, we get a uniform upper bound for $|v(\boldsymbol{x};\theta_k)|$ over Θ, which we use to upper bound $v_M = \max_{\substack{1\leq k\leq K_T \\ \|\boldsymbol{x}\|_\infty \leq M^T_{\boldsymbol{\theta}^*}}} |v(\boldsymbol{x};\theta_k)|$ as below

$$\begin{aligned} v_M &\leq (J^* + Kd(r+1)2^r)\,\phi^p_*(r+1)\left(\left(M^T_{\boldsymbol{\theta}^*}\right)^r + h^r\right)\left(s^p_*\left(M^T_{\boldsymbol{\theta}^*}\right)^{r^p_*} + b^p_*(K_*)^{-1}\right) \\ &= c_{p_1}\left(\left(M^T_{\boldsymbol{\theta}^*}\right)^r + h^r\right)\left(s^p_*\left(M^T_{\boldsymbol{\theta}^*}\right)^{r^p_*} + b^p_*(K_*)^{-1}\right) \\ &\leq c_{p_2}\left(M^T_{\boldsymbol{\theta}^*}\right)^{r+r^p_*}, \end{aligned} \tag{B.15}$$

where the constant terms are defined as

$$c_{p_1} := (J^* + Kd(r+1)2^r)\,\phi^p_*(r+1), \quad c_{p_2} := \max\left(1, c_{p_1}(h^r+1)(s^p_* + b^p_*(K_*)^{-1})\right).$$

A deterministic upper bound on v_M can also be found from the above equation. Noting that from Assumption 2, until time T only states with each component less than or equal to hT are visited, we have

$$v_M \leq c_{p_2}\left(M^T_{\boldsymbol{\theta}^*}\right)^{r+r^p_*} \leq c_{p_2}(Th)^{r+r^p_*} := Q(T),$$

where $Q(T)$ is a polynomial defined as above. Using the bounds derived for v_M, we bound R_2 starting with the first term on the right-hand side of (B.13). We have

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} 2v_M\left[\mathbb{I}_{\{\boldsymbol{\theta}^*\notin B_k\}}+\mathbb{I}_{\{\theta_k\notin B_k\}}\right]\Big] &\leq 2Q(T)\,\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{I}_{\{\boldsymbol{\theta}^*\notin B_k\}}+\mathbb{I}_{\{\theta_k\notin B_k\}}\Big]\\
&\leq 2TQ(T)\,\mathbb{E}\Big[\sum_{k=1}^{K_T}\mathbb{I}_{\{\boldsymbol{\theta}^*\notin B_k\}}+\mathbb{I}_{\{\theta_k\notin B_k\}}\Big]\\
&\leq 2TQ(T)\sum_{k=1}^{T}\mathbb{E}\Big[\mathbb{I}_{\{\boldsymbol{\theta}^*\notin B_k\}}+\mathbb{I}_{\{\theta_k\notin B_k\}}\Big]\\
&\leq 4TQ(T)\sum_{k=1}^{T}\mathbb{P}\{\boldsymbol{\theta}^*\notin B_k\}, \qquad \text{(B.16)}
\end{aligned}$$

where the last inequality follows from (3.15) and the fact that set B_k is $\mathcal{H}_{t_k}-$measurable. To further simplify the first term in (B.13), we find an upper bound for $\mathbb{P}\{\boldsymbol{\theta}^*\notin B_k\}$ using [105]. For a fixed $\boldsymbol{z}=(\boldsymbol{x},a)$ and n independent samples of the distribution $P_{\boldsymbol{\theta}^*}(.|\boldsymbol{z})$, the L^1-deviation of the true distribution $P_{\boldsymbol{\theta}^*}(.|\boldsymbol{z})$ and empirical distribution at the end of episode k, $P_{\hat{\theta}_k}(.|\boldsymbol{z})$, is bounded in [12] as

$$\mathbb{P}\left\{\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{z})-P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1\geq\sqrt{\frac{14\prod_{i=1}^{d}(x_i+h)}{n}\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\right\}\leq\frac{\tilde{\delta}}{20|\mathcal{A}|T^7\prod_{i=1}^{d}(x_i+h)}.$$

Therefore,

$$\mathbb{P}\left\{\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{z})-P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1\geq\beta_k(\boldsymbol{z})\,\Big|\,N_{t_k}(\boldsymbol{z})=n\right\}\leq\frac{\tilde{\delta}}{20|\mathcal{A}|T^7\prod_{i=1}^{d}(x_i+h)},$$

and

$$\begin{aligned}
&\mathbb{P}\left\{\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{z})-P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1\geq\beta_k(\boldsymbol{z})\right\}\\
&=\sum_{n=1}^{T}\mathbb{P}\left\{\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{z})-P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1\geq\beta_k(\boldsymbol{z})\,\Big|\,N_{t_k}(\boldsymbol{z})=n\right\}\mathbb{P}\left\{N_{t_k}(\boldsymbol{z})=n\right\}\\
&\leq\frac{\tilde{\delta}}{20|\mathcal{A}|T^6\prod_{i=1}^{d}(x_i+h)}.
\end{aligned}$$

The probability that at episode $k \leq T$, the true parameter $\boldsymbol{\theta}^*$ does not belong to the confidence set B_k can be bounded using the above and union bound as

$$\begin{aligned}
\mathbb{P}\{\boldsymbol{\theta}^* \notin B_k\} &\leq \sum_{\boldsymbol{z} \in \{0,1,\cdots,hT\}^d \times \mathcal{A}} \mathbb{P}\left\{\|P_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{z}) - P_{\hat{\theta}_k}(\cdot|\boldsymbol{z})\|_1 \geq \beta_k(\boldsymbol{z})\right\} \\
&\leq \sum_{\boldsymbol{z} \in \{0,1,\cdots,hT\}^d \times \mathcal{A}} \frac{\tilde{\delta}}{20|\mathcal{A}|T^6 \prod_{i=1}^d (x_i + h)} \\
&= \sum_{\boldsymbol{x} \in \{0,1,\cdots,hT\}^d} \frac{\tilde{\delta}}{20T^6 \prod_{i=1}^d (x_i + h)} \\
&\leq \frac{\tilde{\delta}}{20T^6} \left(\log\left(h(T+1)\right) + 1\right)^d \\
&\leq \frac{\tilde{\delta}}{20k^6} \left(\log\left(h(T+1)\right) + 1\right)^d.
\end{aligned}$$

In the summation in the above equation, we have simplified the expression by summing over $x_i \leq hT$ instead of considering the more detailed summation over $x_i \leq M_{\boldsymbol{\theta}^*}^T$. However, this simplification does not affect the final evaluation of regret, as this term is not dominant and only contributes to a logarithmic term in the regret bound. Substituting in (B.16),

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2v_M \left[\mathbb{I}_{\{\boldsymbol{\theta}^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right]\Big] &\leq 4TQ(T) \sum_{k=1}^{T} \mathbb{P}\{\boldsymbol{\theta}^* \notin B_k\} \\
&\leq \frac{\tilde{\delta}\left(\log\left(h(T+1)\right) + 1\right)^d TQ(T)}{5} \sum_{k=1}^{\infty} \frac{1}{k^6} \\
&< \tilde{\delta}\left(\log\left(h(T+1)\right) + 1\right)^d TQ(T). \qquad \text{(B.17)}
\end{aligned}$$

We now upper bound the second term in (B.13). From (B.15),

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2v_M \beta_k\left(\boldsymbol{Z}\left(t\right)\right)\Big] \leq 2c_{p_2}\, \mathbb{E}\Big[\left(M_{\boldsymbol{\theta}^*}^T\right)^{r+r_*^p} \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \beta_k\left(\boldsymbol{Z}\left(t\right)\right)\Big]. \qquad \text{(B.18)}$$

To bound the regret term resulting from the summation of $\beta_k(\boldsymbol{Z}(t))$, we note that from the second stopping criterion, $N_t(\boldsymbol{Z}(t)) \leq 2N_{t_k}(\boldsymbol{Z}(t))$ for all $t_k \leq t < t_{k+1}$ and

$$\begin{aligned}
&\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\beta_k(\boldsymbol{Z}(t)) \\
&= \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{14\prod_{i=1}^{d}(\boldsymbol{X}_i(t)+h)}{\max(1,N_{t_k}(\boldsymbol{Z}(t)))}\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)} \\
&\leq \sqrt{14\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{\tilde{t}_{k+1}-1}\sqrt{\frac{2\prod_{i=1}^{d}(\boldsymbol{X}_i(t)+h)}{\max(1,N_t(\boldsymbol{Z}(t)))}}+\sum_{k=1}^{K_T}\sum_{t=\tilde{t}_{k+1}}^{t_{k+1}-1}\sqrt{\prod_{i=1}^{d}(\boldsymbol{X}_i(t)+h)}\right].
\end{aligned} \tag{B.19}$$

The first summation can be simplified as

$$\begin{aligned}
\sum_{k=1}^{K_T}\sum_{t=t_k}^{\tilde{t}_{k+1}-1}\sqrt{\frac{2\prod_{i=1}^{d}(\boldsymbol{X}_i(t)+h)}{\max(1,N_t(\boldsymbol{Z}(t)))}} &\leq \sqrt{2(M_{\boldsymbol{\theta}^*}^T+h)^d}\sum_{k=1}^{K_T}\sum_{t=t_k}^{\tilde{t}_{k+1}-1}\frac{1}{\sqrt{\max(1,N_t(\boldsymbol{Z}(t)))}} \\
&\leq 3\sqrt{2(M_{\boldsymbol{\theta}^*}^T+h)^d}\sum_{\boldsymbol{z}\in\{0,1,\cdots,M_{\boldsymbol{\theta}^*}^T\}^d\times\mathcal{A}}\sqrt{N_{T+1}(\boldsymbol{z})} \\
&\leq 3\sqrt{2|\mathcal{A}|}(M_{\boldsymbol{\theta}^*}^T+h)^d\sqrt{\sum_{\boldsymbol{z}\in\{0,1,\cdots,M_{\boldsymbol{\theta}^*}^T\}^d\times\mathcal{A}}N_{T+1}(\boldsymbol{z})}\Big] \\
&\leq 3\sqrt{2|\mathcal{A}|T}(M_{\boldsymbol{\theta}^*}^T+h)^d,
\end{aligned}$$

where the second inequality is due to the following arguments,

$$\begin{aligned}
\sum_{k=1}^{K_T}\sum_{t=t_k}^{\tilde{t}_{k+1}-1}\frac{1}{\sqrt{\max(1,N_t(\boldsymbol{Z}(t)))}} &= \sum_{\boldsymbol{z}\in\{0,1,\cdots,M_{\boldsymbol{\theta}^*}^T\}^d\times\mathcal{A}}\left(\mathbb{I}_{\{N_{T+1}(\boldsymbol{z})>0\}}+\sum_{i=1}^{N_{T+1}(\boldsymbol{z})-1}\frac{1}{\sqrt{i}}\right) \\
&\leq 3\sum_{\boldsymbol{z}\in\{0,1,\cdots,M_{\boldsymbol{\theta}^*}^T\}^d\times\mathcal{A}}\sqrt{N_{T+1}(\boldsymbol{z})}.
\end{aligned}$$

For the second term in (B.19), we get

$$\sum_{k=1}^{K_T}\sum_{t=\tilde{t}_{k+1}}^{t_{k+1}-1}\sqrt{\prod_{i=1}^{d}(\boldsymbol{X}_i(t)+h)} = \sqrt{(M_{\boldsymbol{\theta}^*}^T+h)^d}\sum_{k=1}^{K_T}E_k$$
$$\leq K_T\left(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\right)\sqrt{(M_{\boldsymbol{\theta}^*}^T+h)^d}$$
$$\leq 2\sqrt{|\mathcal{A}|T\log_2 T}\left(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\right)(M_{\boldsymbol{\theta}^*}^T+h)^d,$$

where $E_k = T_k - \tilde{T}_k$, and K_T is bounded from Lemma 13. Thus $\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\beta_k(\boldsymbol{Z}(t))$ is bounded as

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\beta_k(\boldsymbol{Z}(t)) \leq 24\sqrt{|\mathcal{A}|T\log_2 T\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\left(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\right)(M_{\boldsymbol{\theta}^*}^T+h)^d.$$

Substituting the above bound in (B.18),

$$\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}2v_M\beta_k(\boldsymbol{Z}(t))\Big]$$
$$\leq 48c_{p_2}\sqrt{|\mathcal{A}|T\log_2 T\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\,\mathbb{E}\Big[\left(M_{\boldsymbol{\theta}^*}^T\right)^{r+r_*^p}(M_{\boldsymbol{\theta}^*}^T+h)^d\left(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\right)\Big]$$
$$\leq c_{p_3}\sqrt{|\mathcal{A}|T\log_2 T\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\,\mathbb{E}\Big[(M_{\boldsymbol{\theta}^*}^T+h)^{d+r+r_*^p}\left(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\right)\Big],$$

where $c_{p_3} := 48c_{p_2}$. Finally, from the above equation, (B.17), and (B.13),

$$R_2 \leq \tilde{\delta}\left(\log\left(h(T+1)\right)+1\right)^d TQ(T)$$
$$+ c_{p_3}\sqrt{|\mathcal{A}|T\log_2 T\log\left(\frac{2|\mathcal{A}|T}{\tilde{\delta}}\right)}\,\mathbb{E}\Big[(M_{\boldsymbol{\theta}^*}^T+h)^{d+r+r_*^p}\Big(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\Big)\Big].$$

By choosing $\tilde{\delta} = \frac{1}{TQ(T)}$, we get

$$R_2$$
$$\leq \left(\log(h(T+1))+1\right)^d + c_{p_3}\sqrt{|\mathcal{A}|T\log_2 T\log(2|\mathcal{A}|T^2Q(T))}\,\mathbb{E}\Big[(M_{\boldsymbol{\theta}^*}^T+h)^{d+r+r_*^p}\Big(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\Big)\Big],$$
$$\leq \left(\log(h(T+1))+1\right)^d + c_{p_3}\sqrt{|\mathcal{A}|T}\log_2\left(2|\mathcal{A}|T^2Q(T)\right)\mathbb{E}\Big[(M_{\boldsymbol{\theta}^*}^T+h)^{d+r+r_*^p}\Big(\max_{1\leq i\leq T}\tau_{0^d}^{(i)}\Big)\Big],$$

where $Q(T) = c_{p_2}(Th)^{r+r_*^p}$. □

B.2.7 Proof of Theorem 8

Proof. Lemmas 14, 15, and 16 along with Cauchy-Schwarz inequality showed that the regret terms R_0 and R_2 are of the order $\tilde{O}(KrdJ^*h^{d+2r+r_*^p}\sqrt{|\mathcal{A}|T})$ and the term R_1 is $\tilde{O}(J^*(h)^{r_*^p})$. Therefore, from $R(T, \pi_{TSDE}) = R_0 + R_1 + R_2$, the regret of Algorithm 2, $R(T, \pi_{TSDE})$, is $\tilde{O}(KrdJ^*h^{d+2r+r_*^p}\sqrt{|\mathcal{A}|T})$. □

B.2.8 Requirement of an optimal policy oracle.

To implement our algorithm, we need to find the optimal policy for each model sampled by the algorithm—optimal policy for Theorem 8 and optimal policy within policy class Π for Corollary 3; this has also been used in past work [36, 37, 56]. In the finite state-space setting, [80] provides a schedule of ϵ values and selects ϵ-optimal policies to obtain $\tilde{O}(\sqrt{T})$ regret guarantees. The issue with extending the analysis of [80] to the countable state-space setting is that we need to ensure (uniform) ergodicity for the chosen ϵ-optimal policies; the $\limsup$ or $\liminf$ of the time-average expected reward (used to define the average cost problem) being finite doesn't imply ergodicity. In other words, we must formulate (and verify) ergodicity assumptions for a potentially large set of close-to-optimal algorithms whose structure is undetermined. Another issue is that, to the best of our knowledge, there isn't a general structural characterization of all ϵ-optimal stationary policies for countable state-space MDPs or even a characterization of the policy within this set that is selected by any computational procedure in the literature; current results only discuss existence and characterization of the stationary optimal policy. In the absence of such results, stability assumptions with the same uniformity across models as in our submission will be needed, which are likely too strong to be useful.

If we could verify the stability requirements of Assumptions 3 and 4 for a subset of policies, the optimal oracle is not needed, and instead, by choosing approximately optimal policies within this subset, we can follow the same proof steps as [80] to guarantee regret performance similar to Corollary 3 (without knowledge of model parameters). To theoretically analyze the performance of the algorithm that follows an approximately optimal policy rather than the optimal one, we assume that for a specific sequence of $\{\epsilon_k\}_{k=1}^{\infty}$, an ϵ_k-optimal policy is given, which is defined below.

Definition 2. *Policy $\pi \in \Pi$ is called an ϵ-optimal policy if for every $\theta \in \Theta$,*

$$c(\boldsymbol{x}, \pi(\boldsymbol{x})) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_\theta(\boldsymbol{y}|\boldsymbol{x}, \pi(\boldsymbol{x}))v(\boldsymbol{y};\theta) \leq c(\boldsymbol{x}, \pi_\theta^*(\boldsymbol{x})) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_\theta(\boldsymbol{y}|\boldsymbol{x}, \pi_\theta^*(\boldsymbol{x}))v(\boldsymbol{y};\theta) + \epsilon,$$

where π^*_θ is the optimal policy in the policy class Π corresponding to parameter θ and $v(.;\theta)$ is the solution to Poisson equation (3.6)*.*

Given ϵ-optimal policies that satisfy Assumptions 3 and 4, in Theorem 9 we extend the regret guarantees of Corollary 3 to the algorithm employing ϵ-optimal policy, instead of the best-in-class policy, and show that the same regret upper bounds continue to apply.

Theorem 10. *Consider a non-negative sequence $\{\epsilon_k\}_{k=1}^{\infty}$ such that for every $k \in \mathbb{N}$, ϵ_k is bounded above by $\frac{1}{k+1}$ and an ϵ_k-optimal policy satisfying Assumptions 3 and 4 is given. The regret incurred by Algorithm 2 while using the ϵ_k-optimal policy during any episode k is $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$.*

Proof. For the ϵ_k-optimal policy used in episode k, shown by π^{ϵ_k}, we have

$$\begin{aligned}
&c(\boldsymbol{x}, \pi^{\epsilon_k}(\boldsymbol{x})) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}(\boldsymbol{y}|\boldsymbol{x}, \pi^{\epsilon_k}(\boldsymbol{x}))v(\boldsymbol{y};\theta_k)\\
&\leq c(\boldsymbol{x}, \pi^*_{\theta_k}(\boldsymbol{x})) + \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}(\boldsymbol{y}|\boldsymbol{x}, \pi^*_{\theta_k}(\boldsymbol{x}))v(\boldsymbol{y};\theta_k) + \epsilon_k\\
&= J(\theta_k) + v(\boldsymbol{x};\theta_k) + \epsilon_k.
\end{aligned}$$

Thus,

$$\begin{aligned}
R(T, \pi_{TSDE}) &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} c(\boldsymbol{X}(t), \pi^{\epsilon_k}(\boldsymbol{X}(t)))\Big] - T\,\mathbb{E}\left[J\left(\boldsymbol{\theta}^*\right)\right]\\
&= R_0 + R_1 + R_2 + \mathbb{E}\Big[\sum_{k=1}^{K_T} T_k\epsilon_k\Big]\\
\text{with } R_0 &= \mathbb{E}\Big[\sum_{k=1}^{K_T} T_k J(\theta_k)\Big] - T\,\mathbb{E}\Big[J(\boldsymbol{\theta}^*)\Big],\\
R_1 &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big[v(\boldsymbol{X}(t);\theta_k) - v(\boldsymbol{X}(t+1);\theta_k)\Big]\Big],\\
R_2 &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big[v(\boldsymbol{X}(t+1);\theta_k) - \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\theta_k}(\boldsymbol{y}|\boldsymbol{X}(t), \pi^{\epsilon_k}(\boldsymbol{X}(t)))v(\boldsymbol{y};\theta_k)\Big]\Big].
\end{aligned}$$

We assumed that given ϵ-optimal policies satisfy Assumptions 3 and 4. As a result, we can utilize the proof of Theorem 8 to deduce that the term $R_0 + R_1 + R_2$ is of the order $\tilde{O}(dh^d\sqrt{|\mathcal{A}|T})$. Moreover, we can simplify the term $\mathbb{E}\Big[\sum_{k=1}^{K_T} T_k\epsilon_k\Big]$ as below:

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} T_k\epsilon_k\Big] = \mathbb{E}\Big[\sum_{k=1}^{K_T} \tilde{T}_k\epsilon_k\Big] + \mathbb{E}\Big[\sum_{k=1}^{K_T} E_k\epsilon_k\Big]. \tag{B.20}$$

From the second stopping condition of Algorithm 2, we have $\tilde{T}_k \leq \tilde{T}_{k-1} + 1 \leq \ldots \leq k+1$ and

$$\mathbb{E}\Big[\sum_{k=1}^{K_T} T_k \epsilon_k\Big] \leq \mathbb{E}[K_T],$$

where we have used the assumption that $\epsilon_k \leq \frac{1}{k+1}$. For the second term of (B.20), from (B.10)

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{K_T} E_k \epsilon_k\Big] &\leq \mathbb{E}\Big[\sum_{k=1}^{K_T} \frac{E_k}{k+1}\Big] \\
&\leq \mathbb{E}\Big[\max_{1\leq i\leq T} \tau_{0^d}^{(i)} \sum_{k=1}^{K_T} \frac{1}{k+1}\Big] \\
&\leq \mathbb{E}\Big[\max_{1\leq i\leq T} \tau_{0^d}^{(i)} \log(K_T+1)\Big], \qquad \text{(B.21)}
\end{aligned}$$

where in the last inequality we have used $\sum_{i=1}^{n} \frac{1}{n} \leq 1 + \log(n)$. Finally, as a result of Lemma 11 and Lemma 13, the result follows. □

B.3 Bounds on hitting times under polynomial and geometric ergodicity

B.3.1 Polynomial upper bounds for the moments of hitting time of state 0^d

For any $\theta_1, \theta_2 \in \Theta$, consider the Markov process with transition kernel $P_{\theta_1}^{\pi^*_{\theta_2}}$ obtained from the MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy $\pi^*_{\theta_2}$. [44, Lemma 3.5] establishes that if the process is polynomially ergodic, equivalently satisfies (3.4), then for every $0 < \eta \leq 1$, there exists constants $\beta^{\eta}_{\theta_1,\theta_2}, b^{\eta}_{\theta_1,\theta_2} > 0$ such that the following holds:

$$\Delta\left(V^p_{\theta_1,\theta_2}\right)^{\eta}(\boldsymbol{x}) \leq -\beta^{\eta}_{\theta_1,\theta_2}\left(V^p_{\theta_1,\theta_2}(\boldsymbol{x})\right)^{\alpha^p_{\theta_1,\theta_2}+\eta-1} + b^{\eta}_{\theta_1,\theta_2}\mathbb{I}_{C^p_{\theta_1,\theta_2}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \qquad \text{(B.22)}$$

where for $\eta \in (0,1)$, $\tilde{\beta}^p_{\theta_1,\theta_2} := \min(\beta^p_{\theta_1,\theta_2}, 1)$ and

$$\beta^{\eta}_{\theta_1,\theta_2} = \eta\tilde{\beta}^p_{\theta_1,\theta_2},\ b^{\eta}_{\theta_1,\theta_2} = \left(b^p_{\theta_1,\theta_2}\right)^{\eta} + \eta\tilde{\beta}^p_{\theta_1,\theta_2}\max\left(1, \left(\tilde{\beta}^p_{\theta_1,\theta_2}\right)^{(\alpha^p_{\theta_1,\theta_2}+\eta-1)/(1-\alpha^p_{\theta_1,\theta_2})}\right), \qquad \text{(B.23)}$$

and for $\eta = 1$, $\beta^{\eta}_{\theta_1,\theta_2} = \beta^p_{\theta_1,\theta_2}$ and $b^{\eta}_{\theta_1,\theta_2} = b^p_{\theta_1,\theta_2}$. Consequently, the following result is immediate from the proof of [44, Theorem 3.6]; for completeness, we provide the proof in Appendix B.4.1.

Lemma 22. *Suppose a finite set $C^p_{\theta_1,\theta_2}$, constants $\beta^p_{\theta_1,\theta_2}, b^p_{\theta_1,\theta_2} > 0$, $r/(r+1) \leq \alpha^p_{\theta_1,\theta_2} < 1$, and*

a function $V^p_{\theta_1,\theta_2} : \mathcal{X} \to [1, +\infty)$ *exist such that* (3.4) *holds. Then, there exist a sequence of non-negative functions* $V^i_{\theta_1,\theta_2} : \mathcal{X} \to [1, +\infty)$ *for* $i = 0, \ldots, r+1$ *that satisfy the following system of drift equations for finite sets* $C^i_{\theta_1,\theta_2}$*, constants* $b^i_{\theta_1,\theta_2} \geq 0$ *and* $\beta^i_{\theta_1,\theta_2} > 0$*:*

$$\Delta V^{i-1}_{\theta_1,\theta_2}(\boldsymbol{x}) \leq -\beta^i_{\theta_1,\theta_2} V^i_{\theta_1,\theta_2}(\boldsymbol{x}) + b^i_{\theta_1,\theta_2} \mathbb{I}_{C^i_{\theta_1,\theta_2}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X},\ i = 1, \ldots, r+1. \tag{B.24}$$

Notice that r is the maximum degree of the cost function c defined in Assumption 1. Following the proof and approach of [44] and using the set of equations (B.24), we can find an upper-bound for $\mathbb{E}_{\boldsymbol{x}}[\tau^i_{0^d}]$ for $i = 1, \ldots, r+1$ in Lemma 23. In order to establish upper bounds for the first $r+1$ moments of τ_{0^d}, it is crucial to choose the value of $\alpha^p_{\theta_1,\theta_2}$ greater than or equal to $\frac{r}{r+1}$, as demonstrated in the proof of Lemma 23 in Appendix B.4.2

Lemma 23. *For* $i = 1, \ldots, r+1$*, and for all* $\boldsymbol{x} \in \mathcal{X}$

$$\mathbb{E}^{\pi^*_{\theta_2}}_{\boldsymbol{x}}[(\tau_{0^d})^i] \leq i \phi^p_{\theta_1,\theta_2}(i) \left(V^p_{\theta_1,\theta_2}(\boldsymbol{x}) + b^p_{\theta_1,\theta_2} \alpha_{C^p_{\theta_1,\theta_2}} \right),$$

where $\phi^p_{\theta_1,\theta_2}(i) := \prod_{j=1}^{i} \frac{1}{\beta^{\eta_j}_{\theta_1,\theta_2}} \left(2^{j-1} + (j-1)\, \alpha_{C^p_{\theta_1,\theta_2}} b^{\eta_j}_{\theta_1,\theta_2} \right)$*,* $\eta_i = 1 - (i-1)(1 - \alpha^p_{\theta_1,\theta_2})$ *,* $b^{\eta_i}_{\theta_1,\theta_2}$ *and* $\beta^{\eta_i}_{\theta_1,\theta_2}$ *defined in* (B.23)*, and* $\alpha_{C^p_{\theta_1,\theta_2}} = \left(\min_{\boldsymbol{y} \in C^p_{\theta_1,\theta_2}} K_{\theta_1,\theta_2}(\boldsymbol{y}) \right)^{-1}$*.*

Based on Lemma 23, we impose the conditions of Assumption 4 to obtain uniform (over model class) and polynomial (in norm of the state) upper-bounds on the moments of hitting times to 0^d. Moreover, these conditions lead to a uniform characterization of parameters of Lemma 23 over all models in our class.

B.3.2 Distribution of return times to state 0^d

For any $\theta_1, \theta_2 \in \Theta$, consider the Markov process with transition kernel $P^{\pi^*_{\theta_2}}_{\theta_1}$ obtained from the MDP $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ by following policy $\pi^*_{\theta_2}$. In the following lemma, we show that the tail probabilities of the return times to the common state 0^d, again τ_{0^d}, converge geometrically fast to 0, and characterize the convergence parameters in terms of the constants given in Assumption 3. Explicitly, we show

$$\mathbb{P}_{0^d}(\tau_{0^d} > n) \leq c^g_{\theta_1,\theta_2} \left(\tilde{\gamma}^g_{\theta_1,\theta_2} \right)^n,$$

for problem and policy dependent constants $c^g_{\theta_1,\theta_2}$ and $\tilde{\gamma}^g_{\theta_1,\theta_2}$. We will follow the method outlined in [42] with the goal to identify problem dependent parameters that will be relevant to our results. Proof of the following lemma is given in Appendix B.4.3 and follows the methodology of [42].

Lemma 24. *For every* $\theta_1, \theta_2 \in \Theta$ *in the Markov process obtained from the Markov decision process* $(\mathcal{X}, \mathcal{A}, c, P_{\theta_1})$ *following policy* $\pi^*_{\theta_2}$*, the return time to state* 0 *starting from state* 0 *satisfies the*

following:

$$\mathbb{P}_{0^d}(\tau_{0^d} > n) \leq c^g_{\theta_1,\theta_2} \left(\tilde{\gamma}^g_{\theta_1,\theta_2}\right)^n,$$

where

$$c^g_{\theta_1,\theta_2} = \frac{b^g_{\theta_1,\theta_2}\left(\tilde{b}^g_{\theta_1,\theta_2}\right)^2}{\tilde{b}^g_{\theta_1,\theta_2} - 1} \quad \textit{and} \quad \tilde{\gamma}^g_{\theta_1,\theta_2} = 1 - \frac{1}{\tilde{b}^g_{\theta_1,\theta_2}},$$

with

$$\tilde{b}^g_{\theta_1,\theta_2} = \frac{3b^g_{\theta_1,\theta_2}+1}{1-\gamma^g_{\theta_1,\theta_2}}\left(|C^g_{\theta_1,\theta_2}|^2 \max\left(1, \max_{\boldsymbol{u}\in C^g_{\theta_1,\theta_2}\setminus\{0^d\}} \mathbb{E}^{\pi^*_{\theta_2}}_{\boldsymbol{u}}[\tau_{0^d}]\right)\right).$$

Based on Lemma 24, it is necessary to impose the conditions in Assumption 3 to obtain uniform tail probability bounds on τ_{0^d} for all model parameters and policy choices in Θ. Moreover, these conditions lead to a uniform characterization of $c^g_{\theta_1,\theta_2}$ and $\tilde{\gamma}^g_{\theta_1,\theta_2}$ over Θ. Furthermore, as a result of Lemma 23 and uniformity conditions of Assumption 4, $\mathbb{E}^{\pi^*_{\theta_2}}_{\boldsymbol{u}}[\tau_{0^d}]$ has a uniform bound over Θ and $C^g_{\theta_1,\theta_2} \setminus \{0^d\}$, which can be characterized in terms of the polynomial Lyapunov function.

B.4 Proofs of hitting time bounds

B.4.1 Proof of Lemma 22

Proof. In the proof, to avoid cumbersome notation we will drop the indices θ_1, θ_2. Following the proof of Theorem 3.6 in [44], we choose $\eta_i = 1 - (i-1)(1-\alpha^p)$ for $i = 1, \ldots, r+1$ and note that as $\alpha^p \in [\frac{r}{r+1}, 1)$, we have $\eta_i \in [\frac{1}{r+1}, 1]$. As a result, we can apply (B.22) to each η_i to get

$$\Delta\left(V^p\right)^{\eta_i}(\boldsymbol{x}) \leq -\beta^{\eta_i}\left(V^p(\boldsymbol{x})\right)^{i\alpha^p - i + 1} + b^{\eta_i}\mathbb{I}_{C^p}(\boldsymbol{x}), \quad i = 1, \ldots, r+1.$$

Thus, the system of drift equations (B.24) hold for

$$\begin{aligned}
V_i &= \left(V^p\right)^{1-i(1-\alpha^p)}, && i = 0, \ldots, r+1,\\
\beta_i &= \beta^{\eta_i}, && i = 1, \ldots, r+1,\\
b_i &= b^{\eta_i}, && i = 1, \ldots, r+1,\\
C_i &= C^p, && i = 1, \ldots, r+1,
\end{aligned}$$

where β^{η_i} and b^{η_i} are defined in (B.23). □

B.4.2 Proof of Lemma 23

The proof of Lemma 23 uses the following lemma.

Lemma 25 (Proposition 11.3.2, [71])**.** *Suppose for nonnegative functions f, g, and V on the state space $\mathcal{X}$ and every $k \in \mathbb{Z}_+$, the following holds:*

$$\mathbb{E}[V(X_{k+1})|\mathcal{F}_k] \leq V(X_k) - f(X_k) + g(X_k).$$

Then, for any initial condition x and stopping time τ

$$\mathbb{E}_x\left[\sum_{k=0}^{\tau-1} f(X_k)\right] \leq V(x) + \mathbb{E}_x\left[\sum_{k=0}^{\tau-1} g(X_k)\right].$$

Proof of Lemma 23. Following [44], the proof uses an induction argument. We will use the notation of Lemma 22 for simplicity. Similarly, in this proof we will also denote $\phi^p_{\theta_1,\theta_2}(i)$ as $\phi(i)$, $K_{\theta_1,\theta_2}(\cdot)$ as $K(\cdot)$, and $V^i_{\theta_1,\theta_2}$, $b^i_{\theta_1,\theta_2}$, $\beta^i_{\theta_1,\theta_2}$, $C^i_{\theta_1,\theta_2}$ as V_i, b_i, β_i, C_i.

From irreducibility, for all $\boldsymbol{x} \in \mathcal{X}$, $K(\boldsymbol{x})$ is positive and finite. Considering the system of drift equations found in Lemma 22, $C_i = C^p$ is a finite set for all $i = 1, \ldots, r+1$. Thus, $\min_{\boldsymbol{y}\in C_i} K(\boldsymbol{y})$ is strictly positive. For all $\boldsymbol{x} \in \mathcal{X}$ and $i = 1, \ldots, r+1$, we have

$$\mathbb{I}_{C_i}(\boldsymbol{x}) \leq \left(\min_{\boldsymbol{y}\in C_i} K(\boldsymbol{y})\right)^{-1} K(\boldsymbol{x}). \tag{B.25}$$

We set $\alpha_{C^p} := \left(\min_{\boldsymbol{y}\in C_i} K(\boldsymbol{y})\right)^{-1} = \left(\min_{\boldsymbol{y}\in C^p} K(\boldsymbol{y})\right)^{-1}$. From Lemma 22, for $j = 1$ and $\boldsymbol{x} \in \mathcal{X}$

$$\Delta V_0(\boldsymbol{x}) \leq -\beta_1 V_1(\boldsymbol{x}) + b_1 \mathbb{I}_{C_1}(\boldsymbol{x}).$$

By applying Lemma 25, for all $\boldsymbol{x} \in \mathcal{X}$ we get

$$\beta_1 \mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1} V_1\left(\boldsymbol{X}_k\right)\right] \leq V_0(\boldsymbol{x}) + b_1 \mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1} \mathbb{I}_{C_1}\left(\boldsymbol{X}_k\right)\right]. \tag{B.26}$$

Using (B.25) and (B.26), followed by noting that

$$K(\boldsymbol{x}) = \sum_{n=0}^{\infty} 2^{-n-2} P^n(\boldsymbol{x}, 0^d) = \sum_{n=0}^{\infty} 2^{-n-2} \mathbb{E}_{\boldsymbol{x}}[\mathbb{I}_{0^d}\left(\boldsymbol{X}_n\right)],$$

we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1} V_1\left(\boldsymbol{X}_k\right)\right] &\leq \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{n=0}^{\infty}2^{-n-2}\sum_{k=0}^{\tau_{0^d}-1}\mathbb{I}_{0^d}\left(\boldsymbol{X}_{k+n}\right)\right] \\
&= \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{n=0}^{\infty}2^{-n-2}\sum_{k=n}^{\tau_{0^d}-1+n}\mathbb{I}_{0^d}\left(\boldsymbol{X}_k\right)\right] \\
&\leq \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{n=0}^{\infty}2^{-n-2}\sum_{k=n\vee\tau_{0^d}}^{\tau_{0^d}-1+n}\mathbb{I}_{0^d}\left(\boldsymbol{X}_k\right)\right] \\
&\leq \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}\sum_{n=0}^{\infty}2^{-n-2}(n+1) \\
&= \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}.
\end{aligned}$$

As $V_1(\boldsymbol{x}) \geq 1$, this gives us a bound on $\mathbb{E}_{\boldsymbol{x}}[\tau_{0^d}]$ as follows:

$$\mathbb{E}_{\boldsymbol{x}}[\tau_{0^d}] \leq \frac{1}{\beta_1}V_0(\boldsymbol{x}) + \frac{b_1\alpha_{C^p}}{\beta_1}.$$

Assume for $i \geq 1$, by the induction assumption we have

$$\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^{i-1}V_i\left(\boldsymbol{X}_k\right)\right] \leq \phi(i)\left(V_0(\boldsymbol{x}) + b_1\alpha_{C^p}\right). \tag{B.27}$$

Set $j = i+1$ in (B.24), which yields

$$\Delta V_i(\boldsymbol{x}) \leq -\beta_{i+1}V_{i+1}(\boldsymbol{x}) + b_{i+1}\mathbb{I}_{C^p}(\boldsymbol{x}).$$

Define $Z_k = k^i V_i(\boldsymbol{X}_k)$. From the above equation, we have

$$\begin{aligned}
\mathbb{E}[Z_{k+1}|\boldsymbol{X}_k] &\leq (k+1)^i\left(V_i\left(\boldsymbol{X}_k\right) - \beta_{i+1}V_{i+1}(\boldsymbol{X}_k) + b_{i+1}\mathbb{I}_{C^p}(\boldsymbol{X}_k)\right) \\
&\leq Z_k + 2^i(k+1)^{i-1}V_i\left(\boldsymbol{X}_k\right) + (k+1)^i b_{i+1}\mathbb{I}_{C^p}(\boldsymbol{X}_k) - (k+1)^i\beta_{i+1}V_{i+1}(\boldsymbol{X}_k).
\end{aligned}$$

By applying Lemma 25 to the above equation, we get

$$\begin{aligned}
&\beta_{i+1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^i V_{i+1}\left(\boldsymbol{X}_k\right)\right] \\
&\leq 2^i\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^{i-1}V_i\left(\boldsymbol{X}_k\right)\right]+b_{i+1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^i\mathbb{I}_{C^p}\left(\boldsymbol{X}_k\right)\right] \\
&\leq 2^i\phi(i)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right)+\alpha_{C^p}b_{i+1}\mathbb{E}_{\boldsymbol{x}}[(\tau_{0^d})^i], \qquad \text{(B.28)}
\end{aligned}$$

where the second inequality follows from (B.25) and the induction hypothesis (B.27). Thereafter, from (B.27) (by using integral lower bound after using $V_i \geq 1$), we have

$$\frac{1}{i}\mathbb{E}_{\boldsymbol{x}}[(\tau_{0^d})^i] \leq \mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^{i-1}V_i\left(\boldsymbol{X}_k\right)\right] \leq \phi(i)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right).$$

Substituting in (B.28), we get

$$\begin{aligned}
\beta_{i+1}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=0}^{\tau_{0^d}-1}(k+1)^i V_{i+1}\left(\boldsymbol{X}_k\right)\right] &\leq 2^i\phi(i)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right)+ib_{i+1}\alpha_{C^p}\phi(i)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right) \\
&= \left(2^i+ib_{i+1}\alpha_{C^p}\right)\phi(i)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right) \\
&= \beta_{i+1}\phi(i+1)\left(V_0(\boldsymbol{x})+b_1\alpha_{C^p}\right).
\end{aligned}$$

This completes the proof. □

B.4.3 Proof of Lemma 24

Proof. In the proof, to avoid cumbersome notation we will drop the indices θ_1, θ_2. Based on Assumption 3, there exists a finite set C^g, constants b^g, $\gamma^g \in (0,1)$, and a function $V^g : \mathcal{X} \to [1, +\infty)$ satisfying

$$\Delta V^g(\boldsymbol{x}) \leq -\left(1-\gamma^g\right)V^g(\boldsymbol{x})+b^g\mathbb{I}_{C^g}(\boldsymbol{x}), \quad \boldsymbol{x}\in\mathcal{X}. \qquad \text{(B.29)}$$

For $n \geq 1$, define the n-step taboo probabilities [71] as

$${}_A P^n_{\boldsymbol{x}B} = \mathbb{P}_{\boldsymbol{x}}\left(\boldsymbol{X}_n \in B, \tau_A > n\right),$$

where $A, B \subseteq \mathcal{X}$, and τ_A is the first hitting time of set A. We also let ${}_A P^0_{\boldsymbol{x}B} = \mathbb{I}_B(\boldsymbol{x})$ and $\tilde{V}^g = \sum_{n=0}^{\infty} {}_{0^d} P^n V^g$. Applying the last exit decomposition on $C^g \setminus \{0^d\}$ for all $x \in \mathcal{X}$, we obtain

$$
\begin{aligned}
&\tilde{V}^g(\boldsymbol{x}) \\
&= \sum_{n=0}^{\infty} \sum_{\boldsymbol{y}\in\mathcal{X}} {}_{0^d} P^n_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \\
&= V^g(\boldsymbol{x}) + \sum_{n=1}^{\infty} \sum_{\boldsymbol{y}\in\mathcal{X}} {}_{C^g} P^n_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \\
&+ \sum_{n=1}^{\infty} \sum_{\boldsymbol{y}\in\mathcal{X}} \sum_{m=1}^{n-1} \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} {}_{0^d} P^m_{\boldsymbol{xz}} \, {}_{C^g} P^{n-m}_{\boldsymbol{zy}} V^g(\boldsymbol{y}) + \sum_{n=1}^{\infty} \sum_{\boldsymbol{y}\in\mathcal{X}} \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} {}_{0^d} P^n_{\boldsymbol{xz}} \, {}_{C^g} P^0_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \\
&= V^g(\boldsymbol{x}) + \sum_{n=1}^{\infty} \sum_{\boldsymbol{y}\in\mathcal{X}} {}_{C^g} P^n_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \qquad \text{(B.30)} \\
&+ \underbrace{\sum_{\boldsymbol{y}\in\mathcal{X}} \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} \left(\sum_{m=1}^{\infty} {}_{0^d} P^m_{\boldsymbol{xz}} \right) \left(\sum_{n=1}^{\infty} {}_{C^g} P^n_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \right)}_{\text{Term 1}} + \underbrace{\sum_{n=1}^{\infty} \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} {}_{0^d} P^n_{\boldsymbol{xz}} V^g(\boldsymbol{z})}_{\text{Term 2}}, \qquad \text{(B.31)}
\end{aligned}
$$

where we break up the trajectories starting at state $\boldsymbol{x}$ and reaching state $\boldsymbol{y}$ while avoiding state 0^d into two: ones that never visit the set C^g, and the others that visit $C^g \setminus \{0^d\}$ up until time m but not afterwards and exit $C^g \setminus \{0^d\}$ at time m.

We first bound Term 1 in (B.31) by finding an upper bound for the probability term $\sum_{m=1}^{\infty} {}_{0^d} P^m_{\boldsymbol{xz}}$ using the first entrance decomposition on $C^g \setminus \{0^d\}$ while noting that $\boldsymbol{z} \in C^g \setminus \{0^d\}$:

$$
\begin{aligned}
\sum_{m=1}^{\infty} {}_{0^d} P^m_{\boldsymbol{xz}} &= \sum_{m=1}^{\infty} \sum_{l=1}^{m} \sum_{\substack{\boldsymbol{u}\in C^g\setminus\{0^d\} \\ \boldsymbol{v}\notin C^g}} {}_{C^g} P^{l-1}_{\boldsymbol{xv}} P_{\boldsymbol{vu}} \, {}_{0^d} P^{m-l}_{\boldsymbol{uz}} \\
&= \sum_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \left(\sum_{l=0}^{\infty} \sum_{\boldsymbol{v}\notin C^g} {}_{C^g} P^l_{\boldsymbol{xv}} P_{\boldsymbol{vu}} \right) \left(\sum_{m=0}^{\infty} {}_{0^d} P^m_{\boldsymbol{uz}} \right) \\
&\le \sum_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \sum_{m=0}^{\infty} {}_{0^d} P^m_{\boldsymbol{uz}} \\
&\le \sum_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \sum_{m=0}^{\infty} \mathbb{P}_{\boldsymbol{u}}(\tau_{0^d} > m) \\
&\le |C^g| \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}], \qquad \text{(B.32)}
\end{aligned}
$$

where the third line follows from the fact that $\sum_{l=0}^{\infty} \sum_{\boldsymbol{v} \notin C^g} {}_{C^g} P^l_{\boldsymbol{xv}} P_{\boldsymbol{vu}}$ is the probability of entrance to C^g through $\boldsymbol{u} \in C^g \setminus \{0\}$, so it is less than 1. Irreducibility and positive recurrence combined with $|C^g| < \infty$ imply that $\max_{\boldsymbol{u} \in C^g \setminus \{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}] < \infty$, which shows $\sum_{m=0}^{\infty} {}_{0^d} P^m_{\boldsymbol{xz}}$ is finite. Next, by induction we prove that for $n \geq 1$ and $\boldsymbol{z} \in C^g \setminus \{0^d\}$ we have

$$\sum_{\boldsymbol{y} \in \mathcal{X}} {}_{C^g} P^n_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \leq (\gamma^g)^{n-1} b^g. \tag{B.33}$$

For $n = 1$, we have using Assumption 3 that

$$\sum_{\boldsymbol{y} \in \mathcal{X}} {}_{C^g} P_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \leq \sum_{\boldsymbol{y} \in \mathcal{X}} P_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \leq b^g.$$

Assuming that (B.33) holds for n, for $n+1$ we have

$$\begin{aligned}
\sum_{\boldsymbol{y} \in \mathcal{X}} {}_{C^g} P^{n+1}_{\boldsymbol{zy}} V^g(\boldsymbol{y}) &\leq \sum_{\substack{\boldsymbol{y} \in \mathcal{X} \\ \boldsymbol{v} \notin C^g}} {}_{C^g} P^n_{\boldsymbol{zv}} P_{\boldsymbol{vy}} V^g(\boldsymbol{y}) \\
&\leq \gamma^g \sum_{\boldsymbol{v} \notin C^g} {}_{C^g} P^n_{\boldsymbol{zv}} V^g(\boldsymbol{v}) && \text{(Using (B.29))} \\
&\leq \gamma \sum_{\boldsymbol{v} \in \mathcal{X}} {}_{C^g} P^n_{\boldsymbol{zv}} V^g(\boldsymbol{v}) \\
&\leq (\gamma^g)^n b^g, && \text{(By induction step)}
\end{aligned}$$

so (B.33) is shown. We collect these bounds later on for our result on Term 2.

We now simplify the summation in (B.30). Similar to previous arguments, we will use induction for $n \geq 1$ and show for all $\boldsymbol{x} \in \mathcal{X}$

$$\sum_{\boldsymbol{y} \in \mathcal{X}} {}_{C^g} P^n_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \leq (\gamma^g)^{n-1} \left(\gamma^g V^g(\boldsymbol{x}) + b^g\right). \tag{B.34}$$

For $n = 1$, we have

$$\sum_{\boldsymbol{y} \in \mathcal{X}} {}_{C^g} P_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \leq \sum_{\boldsymbol{y} \in \mathcal{X}} P_{\boldsymbol{xy}} V^g(\boldsymbol{y}) \leq \gamma^g V^g(\boldsymbol{x}) + b^g.$$

Assuming that (B.34) holds for n, for $n+1$ we have

$$\begin{aligned}\sum_{\boldsymbol{y}\in\mathcal{X}} {}_{C^g}P^{n+1}_{\boldsymbol{xy}} V^g(\boldsymbol{y}) &\le \sum_{\boldsymbol{z}\notin C^g} {}_{C^g}P^{n}_{\boldsymbol{xz}} \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\boldsymbol{zy}} V^g(\boldsymbol{y}) \\ &\le \gamma^g \sum_{\boldsymbol{z}\notin C^g} {}_{C^g}P^{n}_{\boldsymbol{xz}} V^g(\boldsymbol{z}) \\ &\le \gamma^g \sum_{\boldsymbol{z}\in\mathcal{X}} {}_{C^g}P^{n}_{\boldsymbol{xz}} V^g(\boldsymbol{z}) \\ &\le (\gamma^g)^n \left(\gamma^g V^g(\boldsymbol{x}) + b^g\right),\end{aligned}$$

where the first and second inequalities follow from the definition of taboo probabilities and (B.29). Thus, (B.34) is proved. Lastly, for Term 2 in (B.31), we note

$$\begin{aligned}\sum_{n=1}^{\infty} \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} {}_{0^d}P^{n}_{\boldsymbol{xz}} V^g(\boldsymbol{z}) &\le \max_{\boldsymbol{y}\in C^g\setminus\{0^d\}} V^g(\boldsymbol{y}) \sum_{\boldsymbol{z}\in C^g\setminus\{0^d\}} \sum_{n=1}^{\infty} {}_{0^d}P^{n}_{\boldsymbol{xz}} \\ &\le b^g |C^g|^2 \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}] \qquad \text{(From (B.32)).}\end{aligned}$$

From the above equation, (B.32), (B.33), and (B.34), we bound $\tilde{V}^g(\boldsymbol{x})$ as follows:

$$\begin{aligned}&\tilde{V}^g(\boldsymbol{x}) \\ &\le V^g(\boldsymbol{x}) + \left(\gamma^g V^g(\boldsymbol{x}) + b^g\right) \sum_{n=1}^{\infty} (\gamma^g)^{n-1} + |C^g|^2 b^g \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}] \left(1 + \sum_{n=1}^{\infty} (\gamma^g)^{n-1}\right) \\ &\le \frac{V^g(\boldsymbol{x})}{1-\gamma^g} + \frac{3|C^g|^2 b^g}{1-\gamma^g} \max\left(1, \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}]\right) \\ &\le V^g(\boldsymbol{x}) \left(\frac{3b^g+1}{1-\gamma^g} \left(|C^g|^2 \max\left(1, \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}]\right)\right)\right),\end{aligned}$$

where the last line is due to $V^g(\boldsymbol{x}) \ge 1$. Taking

$$\tilde{b}^g := \frac{3b^g+1}{1-\gamma^g} \left(|C^g|^2 \max\left(1, \max_{\boldsymbol{u}\in C^g\setminus\{0^d\}} \mathbb{E}_{\boldsymbol{u}}[\tau_{0^d}]\right)\right) > 1,$$

we have shown that

$$\tilde{V}^g(\boldsymbol{x}) \le \tilde{b}^g V^g(\boldsymbol{x}), \quad \boldsymbol{x}\in\mathcal{X}. \tag{B.35}$$

We now upper-bound $\mathbb{P}_{0^d}(\tau_{0^d} > n)$ for all $n \ge 1$ in an inductive manner, starting with $\mathbb{P}_{0^d}(\tau_{0^d} > 1)$.

As a part of showing this, for every $\boldsymbol{x} \neq 0^d$ we argue that for all $n \geq 1$

$$\mathbb{P}_{\boldsymbol{x}}(\tau_{0^d} > n) \leq \tilde{V}^g(\boldsymbol{x}) \left(1 - \frac{1}{\tilde{b}^g}\right)^n. \tag{B.36}$$

First note that

$$\tilde{V}^g(\boldsymbol{x}) \geq V^g(\boldsymbol{x}) \geq 1. \tag{B.37}$$

Thus,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{x}}(\tau_{0^d} > 1) &= \sum_{\boldsymbol{y} \in \mathcal{X}_{0^d}} P_{\boldsymbol{xy}} \\
&\leq \sum_{\boldsymbol{y} \in \mathcal{X}_{0^d}} P_{\boldsymbol{xy}} \tilde{V}^g(\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in \mathcal{X}_{0^d}} P_{\boldsymbol{xy}} \sum_{n=0}^{\infty} \sum_{\boldsymbol{z} \in \mathcal{X}_{0^d}} P^n_{\boldsymbol{yz}} V^g(\boldsymbol{z}) \\
&= \sum_{\boldsymbol{z} \in \mathcal{X}} \sum_{n=1}^{\infty}{}_{0^d} P^n_{\boldsymbol{xz}} V^g(\boldsymbol{z}).
\end{aligned} \tag{B.38}$$

We now apply the bound in (B.35) to get

$$\mathbb{P}_{\boldsymbol{x}}(\tau_{0^d} > 1) \leq \sum_{\boldsymbol{z} \in \mathcal{X}} \sum_{n=1}^{\infty}{}_{0^d} P^n_{\boldsymbol{xz}} V^g(\boldsymbol{z}) = \tilde{V}^g(\boldsymbol{x}) - V^g(\boldsymbol{x}) \leq \tilde{V}^g(\boldsymbol{x}) \left(1 - \frac{1}{\tilde{b}^g}\right). \tag{B.39}$$

With the base of induction established, we assume the statement in (B.36) is true for n, and show that it continues to hold for $n+1$ as follows:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{x}}(\tau_{0^d} > n+1) &= \sum_{\boldsymbol{y} \neq 0^d} P_{\boldsymbol{xy}} \mathbb{P}_{\boldsymbol{y}}(\tau_{0^d} > n) \\
&\leq \left(1 - \frac{1}{\tilde{b}^g}\right)^n \sum_{\boldsymbol{y} \neq 0^d} P_{\boldsymbol{xy}} \tilde{V}^g(\boldsymbol{y}) \\
&\leq \tilde{V}^g(\boldsymbol{x}) \left(1 - \frac{1}{\tilde{b}^g}\right)^{n+1},
\end{aligned}$$

where the final inequality uses the same arguments as in (B.38) and (B.39).

Finally, using the tail probabilities of hitting time of state 0^d from any state $\boldsymbol{x} \neq 0^d$, we bound

the tail probability of the return time to state 0^d (starting from 0^d) as follows

$$\begin{aligned}\mathbb{P}_{0^d}(\tau_{0^d} > n+1) &= \sum_{\boldsymbol{x}\neq 0^d} P_{0\boldsymbol{x}}\mathbb{P}_{\boldsymbol{x}}(\tau_{0^d} > n) \\ &\leq \left(1-\frac{1}{\tilde{b}^g}\right)^n \sum_{\boldsymbol{x}\neq 0^d} P_{0\boldsymbol{x}}\tilde{V}^g(\boldsymbol{x}) \\ &\leq \tilde{b}^g\left(1-\frac{1}{\tilde{b}^g}\right)^n \sum_{\boldsymbol{x}\neq 0^d} P_{0\boldsymbol{x}}V^g(\boldsymbol{x}) \\ &\leq b^g\tilde{b}^g\left(1-\frac{1}{\tilde{b}^g}\right)^n,\end{aligned}$$

where the final inequality follows from the definition of b^g, and we have

$$\tilde{\gamma}^g = 1-\frac{1}{\tilde{b}^g}, \text{ and } c^g = \frac{b^g\left(\tilde{b}^g\right)^2}{\tilde{b}^g-1},$$

and the proof is complete. □

B.5 Queueing model examples

B.5.1 Model 1: Two-server queueing system with a common buffer

We consider a continuous-time queueing system with two heterogeneous servers with unknown service rate vector $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$ and a common infinite buffer, shown in Figure 3.2a. Arrivals to the system are according to a Poisson process with rate λ and service times are exponentially distributed with parameter θ_i^*, depending on the assigned server. The service rate vector $\boldsymbol{\theta}^*$ is sampled from the prior distribution ν_0 defined on the space Θ given as

$$\Theta = \left\{(\theta_1,\theta_2)\in\mathbb{R}_+^2 : \frac{\lambda}{\theta_1+\theta_2} \leq \frac{1-\delta}{1+\delta}, 1 \leq \frac{\theta_1}{\theta_2} \leq R\right\}, \tag{B.40}$$

for fixed $\delta \in (0, 0.5)$ and $R \geq 1$. Note that for any $(\theta_1, \theta_2) \in \Theta$, we have $\theta_1 \geq \theta_2$ and the stability requirement $\lambda < \theta_1 + \theta_2$ holds. The countable state space $\mathcal{X}$ is defined as $\mathcal{X} = \{\boldsymbol{x} = (x_0, x_1, x_2) : x_0 \in \mathbb{N}\cup\{0\}, x_1, x_2 \in \{0, 1\}\}$, in which x_0 is the length of the queue, and $x_i, i = 1, 2$ is equal to 1 if server i is busy serving a job. At each time instance $r \in \mathbb{R}_+$, the dispatcher can assign jobs from the (non-empty) buffer to an available server. Thus, the action space $\mathcal{A}$ is equal to

$$\mathcal{A} = \{h, b, 1, 2\},$$

where h indicates no action, b sends a job to both of the servers, and $i = 1, 2$ assigns a job to server i. The goal of the dispatcher is to minimize the expected sojourn time of customers, which by Little's law [87] is equivalent to minimizing the average number of customers in the system, or

$$\inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \|\boldsymbol{X}(r)\|_1 \, dr, \tag{B.41}$$

where $\boldsymbol{X}(r)$ is the state of the system at time $r \in \mathbb{R}_+$, immediately after the arrival/departure and just before the action is taken. In [59], it is argued that from uniformization [60] and sampling the continuous-time Markov process at a rate of $\lambda + \theta_1^* + \theta_2^*$, a discrete-time Markov chain is obtained, which converts the original continuous-time problem shown in (B.41) to an equivalent discrete-time problem as below:

$$\inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \|\boldsymbol{X}(r)\|_1 \, dr = \inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \sum_{i=0}^{T-1} \|\boldsymbol{X}(i)\|_1. \tag{B.42}$$

To obtain a uniform sampling rate of $\lambda + \theta_1^* + \theta_2^*$, the continuous-time system is sampled at arrivals, real and dummy customer departures. In [59], it is further shown that the optimal policy that achieves the infimum in (B.42) is a threshold policy π_t with the optimal finite threshold $t(\theta) \in \mathbb{N}$, with the policy defined as below:

$$\pi_t(\boldsymbol{x}) = \begin{cases} \text{h} & \text{if } \{x_0 = 0\} \text{ or } \{\|\boldsymbol{x}\|_1 \leq t, x_1 = 1\} \text{ or } \{x_1 = x_2 = 1\} \\ 1 & \text{if } \{x_0 \geq 1, x_1 = 0\} \\ 2 & \text{if } \{x_0 \geq 1, \|\boldsymbol{x}\|_1 \geq t + 1, x_1 = 1, x_2 = 0\}; \end{cases}$$

note that action b is not used. Policy π_t assigns a job to the faster (first) server whenever there is a job waiting in the queue and the first server is available. In contrast, π_t dispatches a job to the second server only if the number of jobs in the system are greater than threshold t and the second server is available. If neither of these conditions hold, no action or h is taken. Consequently, we can restrict the set of all policies Π in (B.42) to the set Π_t, which is the set of all possible threshold policies corresponding to some $t \in \mathbb{N}$.

In the rest of this subsection, our aim is to show that Assumptions 1-5 are satisfied for the discrete-time Markov process obtained by uniformization of the described queueing system and hence, conclude that Algorithm 2 can be used to learn the unknown service rate vector $\boldsymbol{\theta}^*$ with the expected regret of order $\tilde{O}(\sqrt{T})$.

Assumption 1. Cost function is given as $c(\boldsymbol{x}, a) = \|\boldsymbol{x}\|_1$, which satisfies Assumption 1 with $f_c(\boldsymbol{x}) = x_0 + x_1 + x_2$ and $K = r = 1$.

Assumption 2. For any state-action pair $(\boldsymbol{x}, a)$ and $\theta \in \Theta$, we have $P_\theta(A(\boldsymbol{x}); \boldsymbol{x}, a) = 0$ where

$A(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathcal{X} : |\|\boldsymbol{y}\|_1 - \|\boldsymbol{x}\|_1| > 1\}$; thus, Assumption 2 holds with $h = 1$.

Assumption 3. Consider a queueing system with parameter θ following threshold policy π_t for some $t \in \mathbb{N}$. The uniformized discrete-time Markov chain is irreducible and aperiodic on a subset of state space given as $\mathcal{X}_t = \mathcal{X} \setminus (\{(i,0,0) : i \geq \min(t,2)\} \cup \{(0,1,1)\})$. In [59], it is proved that for every t, the chain consists of a single positive recurrent class and the corresponding average number of customers, depicted by $J^t(\theta)$, is calculated. Moreover, it is shown that for every $\theta \in \Theta$ the optimal threshold $t(\theta)$ can be numerically found as the smallest $i \in \mathbb{N}$ for which $J^i(\theta) < J^{i+1}(\theta)$. Define the set T^* as the set of all optimal thresholds corresponding to at least one $\theta \in \Theta$, or

$$T^* = \{t : t = t(\theta) \text{ for } \theta \in \Theta\}.$$

Remark 12. *There is a discrepancy between the class of MDPs defined in this section and in Section 3.2, as in the former the MDPs are not irreducible in the whole state space $\mathcal{X}$. Specifically, for every Markov process generated by a queueing system with parameter θ following threshold policy π_t, irreducibility holds on $\mathcal{X}_t \subset \mathcal{X}$. Nevertheless, the results of Section 3.4 are valid as starting from state $(0,0)$, the visited states are positive recurrent; see Remark 10.*

In the following proposition, we verify the geometric ergodicity of the discrete-time chain governed by any parameter $\theta \in \Theta$ and obtained by following any threshold policy π_t for $t \in T^*$; proof is given in Appendix B.6.1.

Proposition 2. *The discrete-time Markov process obtained from the queueing system governed by parameter $\theta = (\theta_1, \theta_2) \in \Theta$ and following threshold policy π_t for some $t \in T^*$ is geometrically ergodic. Equivalently, the following holds*

$$\Delta V^g_{\theta,t}(\boldsymbol{x}) \leq -\left(1 - \gamma^g_{\theta,t}\right) V^g_{\theta,t}(\boldsymbol{x}) + b^g_{\theta,t} \mathbb{I}_{C^g_{\theta,t}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}_t,$$

for

$$V^g_{\theta,t}(\boldsymbol{x}) = \exp(-\log(1-\delta)\|\boldsymbol{x}\|_1),$$

$$C^g_{\theta,t} = \{(x_0, x_1, 0) : x_0 < t\} \cup \{(0,0,1)\}, \tag{B.43}$$

$$b^g_{\theta,t} = \max_{\boldsymbol{x} \in C^g_{\theta,t}} \exp\left(-\log(1-\delta)\left(\|\boldsymbol{x}\|_1 + 1\right)\right), \tag{B.44}$$

$$\gamma^g_{\theta,t} = \frac{1}{2} - \frac{1}{2(\theta_1 + \theta_2 + \lambda)}\left((\theta_1 + \theta_2)(1-\delta) + \lambda\,(1-\delta)^{-1}\right). \tag{B.45}$$

Having described all the terms explicitly, we verify the rest of the conditions of Assumption 3, which lead to uniform (over model class) upper-bounds on the moments of hitting time to 0^d as follows:

1. From (B.45), $\sup_{\theta\in\Theta,t\in T^*}\gamma^g_{\theta,t}\le 1/2<1$.

2. From (B.43), we can see that state $(0,0)$ belongs to $C^g_{\theta,t}$ for all $\theta\in\Theta$ and $t\in T^*$. In order for $C^g_*=\cup_{\theta\in\Theta,t\in T^*}C^g_{\theta,t}$ to be a finite set, the supremum of the optimal threshold $t(\theta)$ over Θ should be finite. In [57] with service rate vector (θ_1,θ_2), it is shown that the optimal threshold is bounded above by $\sqrt{2}\theta_1/\theta_2$, which further gives

$$t(\theta)\le\sqrt{2}\frac{\theta_1}{\theta_2}\le\sqrt{2}R. \tag{B.46}$$

Thus, $\sup_{\theta\in\Theta}t(\theta)\le\sqrt{2}R$, which is finite. To confirm a uniform upper bound for $b^g_{\theta,t}$, we note that from (B.44),

$$\sup_{\theta\in\Theta,t\in T^*}b^g_{\theta,t}=\frac{2-\delta}{1-\delta}\max_{x\in C^g_*}\exp(-\log(1-\delta)\|\boldsymbol{x}\|_1),$$

which is finite as $|C^g_*|<\infty$.

Assumption 4. To find an upper bound on the second moment of hitting times, we verify Assumption 4 and show that there exists a finite set $C^p_{\theta,t}$, constants $\beta^p_{\theta,t}$, $b^p_{\theta,t}>0$, $r/(r+1)\le\alpha^p_{\theta,t}<1$, and a function $V^p_{\theta,t}:\mathcal{X}_t\to[1,+\infty)$ satisfying

$$\Delta V^p_{\theta,t}(\boldsymbol{x})\le-\beta^p_{\theta,t}\left(V^p_{\theta,t}(\boldsymbol{x})\right)^{\alpha^p_{\theta,t}}+b^p_{\theta,t}\mathbb{I}_{C^p_{\theta,t}}(\boldsymbol{x}),\quad \boldsymbol{x}\in\mathcal{X}_t. \tag{B.47}$$

Proposition 3. *The discrete-time Markov process obtained from the queueing system governed by parameter $\theta=(\theta_1,\theta_2)\in\Theta$ and following threshold policy π_t for some $t\in T^*$ is polynomially ergodic. This is true because* (B.47) *holds for*

$$V^p_{\theta,t}(\boldsymbol{x})=\|\boldsymbol{x}\|_1^2, \tag{B.48}$$

$$C^p_{\theta,t}=\{(x_0,x_1,0):x_0<t\}\cup\left\{(x_0,x_1,x_2):x_0<\frac{2\lambda}{\theta_1+\theta_2-\lambda},x_1+x_2\ge1\right\}, \tag{B.49}$$

$$b^p_{\theta,t}=\max_{\boldsymbol{x}\in C^p_{\theta,t}}\left(\|\boldsymbol{x}\|_1+1\right)^2), \tag{B.50}$$

$$\beta^p_{\theta,t}=1-\frac{2\lambda}{\theta_1+\theta_2+\lambda}, \tag{B.51}$$

$$\alpha^p_{\theta,t}=\frac{1}{2}. \tag{B.52}$$

Proof of Proposition 3 is given in Appendix B.6.2. We define the normalized rates as $\tilde{\lambda}=\frac{\lambda}{\lambda+\theta_1+\theta_2}$ and $\tilde{\theta}_i=\frac{\theta_i}{\lambda+\theta_1+\theta_2}$, for $i=1,2$. From the choice of parameter space Θ, we have $\tilde{\lambda}\le0.5-0.5\delta$, $\tilde{\theta}_1+\tilde{\theta}_2\ge0.5+0.5\delta$, and $\tilde{\theta}_1\ge0.25+0.25\delta$. We verify the remaining conditions of Assumption 4 as follows:

1. From (B.48), the first condition holds with $r_*^p = 2$ and $s_*^p = 2$.

2. From (B.49), we can see that state $(0, 0)$ belongs to $C_{\theta,t}^p$ for all $\theta \in \Theta$ and $t \in T^*$. Furthermore,

$$\sup_{\theta\in\Theta, t\in T^*} \frac{2\lambda}{\theta_1 + \theta_2 - \lambda} \leq \frac{1-\delta}{\delta},$$

   which follows from the stability condition $\tilde{\lambda} \leq 0.5 - 0.5\delta$. Thus, from the definition of $C_{\theta,t}^p$ in (B.49), and the fact that $\sup_{\theta\in\Theta} t(\theta) \leq \sqrt{2}R$ as argued in in (B.46), $C_*^p = \cup_{\theta\in\Theta, t\in T^*} C_{\theta,t}^p$ is a finite set. We also note that $\sup_{\theta\in\Theta, t\in T^*} b_{\theta,t}^p$ is finite as $|C_*^p| < \infty$. It remains to show that $\inf_{\theta\in\Theta, t\in T^*} \beta_{\theta,t}^p$ is positive, which is equivalent to verifying that $\sup_{\theta\in\Theta, t\in T^*} \tilde{\lambda} < 1/2$, which follows from the stability condition $\tilde{\lambda} \leq 0.5 - 0.5\delta$.

3. We need to show that $K_{\theta,t}(\boldsymbol{x}) := \sum_{n=0}^{\infty} 2^{-n-2} \left(P_\theta^t\right)^n (\boldsymbol{x}, 0^d)$ is strictly bounded away from zero. We notice that from any non-zero state $\boldsymbol{x}$, the queueing system hits 0^d in $\|\boldsymbol{x}\|_1$ transitions only if all transitions are real departures. Hence,

$$\begin{aligned}
K_{\theta,t}(\boldsymbol{x}) &\geq 2^{-\|\boldsymbol{x}\|_1-2} \left(P_\theta^t\right)^{\|\boldsymbol{x}\|_1} (\boldsymbol{x}, 0^d) \\
&\geq 2^{-\|\boldsymbol{x}\|_1-2} \left(\tilde{\theta}_1\right)^{\|\boldsymbol{x}\|_1} \left(\tilde{\theta}_2\right)^{\|\boldsymbol{x}\|_1} \\
&\geq 2^{-\|\boldsymbol{x}\|_1-2} R^{-\|\boldsymbol{x}\|_1} \left(\tilde{\theta}_1\right)^{2\|\boldsymbol{x}\|_1} \\
&\geq 2^{-\|\boldsymbol{x}\|_1-2} R^{-\|\boldsymbol{x}\|_1} \left(\frac{1}{4} + \frac{\delta}{4}\right)^{2\|\boldsymbol{x}\|_1},
\end{aligned}$$

   where the third and fourth inequalities follow from the definition of Θ in (B.40). Thus, the infimum of $K_{\theta,t}(\boldsymbol{x})$ over the finite set C_*^p and sets Θ and T^* is strictly greater than zero.

Assumption 5. We finally verify Assumption 5, which asserts that $\sup_{\theta\in\Theta} J(\theta)$ is finite. We have

$$J(\theta) = \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,t(\theta)}} [c(\boldsymbol{X})] = \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,t(\theta)}} [\|\boldsymbol{X}\|_1] = \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,t(\theta)}} \left[\sqrt{V_{\theta,t(\theta)}^p (\boldsymbol{X})}\right],$$

where $\mu_{\theta,t(\theta)}$ is the stationary distribution of the discrete-time process governed by parameter θ and following the optimal policy according to θ. From (B.47) and [71, Theorem 14.3.7],

$$\mu_{\theta,t(\theta)} \left(\sqrt{V_{\theta,t(\theta)}^p (\boldsymbol{X})}\right) \leq \frac{b_*^p}{\beta_*^p},$$

which is finite from the the previously verified assumption. Consequently,

$$\sup_{\theta \in \Theta} J(\theta) \leq \frac{b_*^p}{\beta_*^p} < \infty.$$

B.5.2 Model 2: Two heterogeneous parallel queues

We consider two parallel queues with infinite buffers, each with its own single server, and unknown service rate vector $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$, shown in Figure 3.2b. The service rate vector $\boldsymbol{\theta}^*$ is sampled from the prior distribution ν_0 defined on the space Θ given as

$$\Theta = \left\{ (\theta_1, \theta_2) \in \mathbb{R}_+^2 : \frac{\lambda}{\theta_1 + \theta_2} \leq \frac{1-\delta}{1+\delta}, 1 \leq \frac{\theta_1}{\theta_2} \leq R \right\}, \tag{B.53}$$

for fixed $\delta \in (0, 0.5)$ and $R \geq 1$, which ensures the stability of the queueing system. Consider the discrete-time MDP $(\mathcal{X}, \mathcal{A}, P_{\boldsymbol{\theta}^*}, c)$ obtained by sampling the queueing system at the Poisson arrival sequence. The countably infinite state space $\mathcal{X}$ is defined as below

$$\mathcal{X} = \{ \boldsymbol{x} = (x_1, x_2) : x_i \in \mathbb{N} \cup \{0\} \},$$

where the state of the system is the number of jobs in the server-queue pair i just before an arrival. Furthermore, the action space $\mathcal{A}$ is equal to

$$\mathcal{A} = \{1, 2\},$$

where action $i \in \mathcal{A}$ indicates the arrival dispatched to queue i. The unbounded cost function $c : \mathcal{X} \times \mathcal{A} \to \mathbb{N} \cup \{0\}$ is defined as the total number of jobs in the queueing system, i.e., $c(\boldsymbol{x}, a) = \|\boldsymbol{x}\|_1$. For every $\omega \in \mathbb{R}_+$, we define policy $\pi_\omega : \mathcal{X} \to \mathcal{A}$, which routes the arrival according to the weighted queue lengths, as

$$\pi_\omega(\boldsymbol{x}) = \arg\min \left(1 + x_1, \omega \left(1 + x_2\right)\right),$$

where the tie is broken in favor of the first server. We also define policy class $\tilde{\Pi}$ as the set of policies π_ω such that ω belongs to a compact interval; in other words,

$$\tilde{\Pi} = \left\{ \pi_\omega;\, \omega \in \left[\frac{1}{c_R R}, c_R R \right] \right\},$$

where R is defined in (B.53) and $c_R \geq 1$. We aim to minimize the infinite-horizon average cost in the policy class $\tilde{\Pi}$, that is,

$$J(\theta) = \inf_{\pi \in \tilde{\Pi}} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c\left(\boldsymbol{X}\left(t\right), A\left(t\right)\right)\right], \tag{B.54}$$

where $\boldsymbol{X}(t) = (X_1(t), X_2(t))$ is the occupancy vector of the queueing system just before arrival t. Even with the controlled Markov process transition kernel fully-specified (by the values of the arrival rate and the two service rates), the optimal policy[1] that satisfies (B.54) in policy class $\tilde{\Pi}$ is not known except when $\theta_1 = \theta_2$ where the optimal value is $\omega = 1$, and so, to learn it, we will use Proximal Policy Optimization for countable state-space controlled Markov processes as developed in [27]. Note that [27] requires full knowledge of the controlled Markov process, which holds in our learning scheme since we use the parameters sampled from the posterior for determining the policy at the beginning of each episode. Furthermore, for each policy in the set of applicable policies $\tilde{\Pi}$, [27] also requires that the resulting Markov process be geometrically ergodic, which we will establish below.

Proposition 4. *The discrete-time Markov process obtained from the queueing system governed by parameter $\theta = (\theta_1, \theta_2) \in \Theta$ and following policy $\pi_\omega \in \tilde{\Pi}$ is geometrically ergodic. Equivalently, the following holds*

$$\Delta V^g_{\theta,\omega}(\boldsymbol{x}) \leq -\left(1 - \gamma^g_{\theta,\omega}\right) V^g_{\theta,\omega}(\boldsymbol{x}) + b^g_{\theta,\omega} \mathbb{I}_{C^g_{\theta,\omega}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \tag{B.55}$$

for

$$V^g_{\theta,\omega}(\boldsymbol{x}) = \frac{\omega}{\omega+1} \exp\left(a^g_{\theta,\omega} \frac{x_1+1}{\omega}\right) + \frac{1}{\omega+1} \exp\left(a^g_{\theta,\omega}\left(x_2+1\right)\right),$$

$$a^g_{\theta,\omega} = \min\left(\omega \log(1+\delta), \log(1+\delta), \omega \log \frac{1-0.5\delta}{1-\delta}, \log \frac{1-0.5\delta}{1-\delta}, \frac{\delta(1-\delta^2)}{4c_R R(1-0.5\delta)}\right), \tag{B.56}$$

$$C^g_{\theta,\omega} = \left\{(x_1, x_2) \in \mathcal{X} : x_i \leq \max\left(x^{g_j}_{i,\theta,\omega}, 0\right), i, j = 1, 2\right\}, \tag{B.57}$$

$$b^g_{\theta,\omega} = \max_{\boldsymbol{x} \in C^g_{\theta,\omega}} \left(\frac{2\omega}{\omega+1} \exp\left(a^g_{\theta,\omega} \frac{x_1+2}{\omega}\right) + \frac{2}{\omega+1} \exp\left(a^g_{\theta,\omega}\left(x_2+2\right)\right)\right), \tag{B.58}$$

$$\gamma^g_{\theta,\omega} = \frac{1}{2} + \frac{1}{2} \max\left(\zeta_{1,\theta,\omega}, \zeta_{2,\theta,\omega}, \frac{\zeta_{1,\theta,\omega}\omega}{1+\omega} \exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) + \frac{\zeta_{2,\theta,\omega}}{1+\omega}, \frac{\zeta_{1,\theta,\omega}\omega}{1+\omega} + \frac{\zeta_{2,\theta,\omega}}{1+\omega} \exp\left(a^g_{\theta,\omega}\right)\right), \tag{B.59}$$

and problem-dependent constants $x^{g_j}_{i,\theta,\omega}$ and $\zeta_{i,\theta,\omega}$ for $i, j = 1, 2$.

[1] When $\theta_1 = \theta_2$, then the policy with $\omega = 1$ (Join-the-Shortest-Queue) is the optimal policy [29] for the underling MDP.

Proof of Proposition 4 is given in Appendix B.6.3. In the rest of this subsection, our aim is to show that Assumptions 1-5 are satisfied for the discrete-time MDP and conclude that Algorithm 2 can be used to learn the unknown service rate vector $\boldsymbol{\theta}^*$ with expected regret of order $\tilde{O}(\sqrt{T})$.

Assumption 1. Cost function is given as $c(\boldsymbol{x}, a) = \|\boldsymbol{x}\|_1$, which satisfies Assumption 1 with $f_c(\boldsymbol{x}) = x_0 + x_1 + x_2$ and $K = r = 1$.

Assumption 2. For any state-action pair $(\boldsymbol{x}, a)$ and $\theta \in \Theta$, we have $P_\theta(A(\boldsymbol{x}); \boldsymbol{x}, a) = 0$ where $A(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathcal{X} : \|\boldsymbol{y}\|_1 - \|\boldsymbol{x}\|_1 > 1\}$; thus, the MDP is skip-free to the right with $h = 1$. Moreover, from any $(\boldsymbol{x}, a)$, the finite set $\{\boldsymbol{y} \in \mathcal{X} : \|\boldsymbol{y}\|_1 \leq \|\boldsymbol{x}\|_1 + 1\}$ is only accessible in one step; thus, Assumption 2 holds.

Assumption 3. In Proposition 4, we verified the geometric ergodicity of the discrete-time chain governed by parameter $\theta = (\theta_1, \theta_2) \in \Theta$ and following policy $\pi_\omega \in \tilde{\Pi}$ and thus, it only remains to verify the uniform model conditions. We define the normalized rates as $\tilde{\lambda} = \frac{\lambda}{\lambda+\theta_1+\theta_2}$ and $\tilde{\theta}_i = \frac{\theta_i}{\lambda+\theta_1+\theta_2}$, for $i = 1, 2$. From the choice of parameter space Θ, we have $\tilde{\lambda} \leq 0.5 - 0.5\delta$, $\tilde{\theta}_1 + \tilde{\theta}_2 \geq 0.5 + 0.5\delta$, and $\tilde{\theta}_1 \geq 0.25 + 0.25\delta$.

1. We first argue that $\zeta_{1,\theta,\omega}$ is bounded away from 1 as follows

$$
\begin{aligned}
1 - \zeta_{1,\theta,\omega} &= 1 - \frac{\frac{\lambda}{\theta_1+\lambda}}{1 - \exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}} \\
&= \frac{\frac{\theta_1}{\theta_1+\lambda}\left(1 - \exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\right)}{1 - \exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}} \\
&\geq \frac{\theta_1}{\theta_1+\lambda}\left(1 - \exp\left(-\frac{a^g_{\theta,(c_R R)^{-1}}}{c_R R}\right)\right) \\
&> \tilde{\theta}_1\left(1 - \exp\left(-\frac{a^g_{\theta,(c_R R)^{-1}}}{c_R R}\right)\right) \\
&> (0.25 + 0.25\delta)\left(1 - \exp\left(-\frac{a^g_{\theta,(c_R R)^{-1}}}{c_R R}\right)\right),
\end{aligned}
$$

where the first line follows from the definition of $\zeta_{1,\theta,\omega}$ in Appendix B.6.3, the second line from (B.56) and the definition of policy class $\tilde{\Pi}$. As $a^g_{\theta,\omega}$ does not depend on θ, $\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} \zeta_{1,\theta,\omega} < 1$. Furthermore, by similar arguments it can be shown that $\zeta_{2,\theta,\omega}$ is bounded away from 1. We next argue that $\frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) + \frac{\zeta_{2,\theta,\omega}}{1+\omega}$ is bounded away from 1

using an upper bound found in Appendix B.6.3 as below,

$$
\begin{aligned}
&1-\frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)-\frac{\zeta_{2,\theta,\omega}}{1+\omega}\\
&\geq 1-\frac{\frac{\lambda}{1+\omega}\left(\omega+a^g_{\theta,\omega}\zeta_4\right)}{\lambda+\frac{\theta_1 a^g_{\theta,\omega}\zeta_3}{\omega}}-\frac{\frac{\lambda}{1+\omega}}{\lambda+\theta_2 a^g_{\theta,\omega}\zeta_3}\\
&=\frac{a^g_{\theta,\omega}\left(-a^g_{\theta,\omega}\zeta_3\theta_2\left(\lambda\zeta_4-\frac{\zeta_3\theta_1(1+\omega)}{\omega}\right)+\lambda\zeta_3\left(\theta_1+\theta_2\right)-\lambda^2\zeta_4\right)}{(1+\omega)(\lambda+\theta_1 a^g_{\theta,\omega}\zeta_3\omega^{-1})(\lambda+\theta_2 a^g_{\theta,\omega}\zeta_3)}\\
&>\frac{\left(\zeta_3 a^g_{\theta,\omega}\right)^2\tilde{\theta}_1\tilde{\theta}_2}{\omega(\tilde{\lambda}+\tilde{\theta}_1 a^g_{\theta,\omega}\zeta_3\omega^{-1})(\tilde{\lambda}+\tilde{\theta}_2 a^g_{\theta,\omega}\zeta_3)}\\
&>\frac{(\zeta_3 a^g_{\theta,(c_R R)^{-1}})^2(0.25+0.25\delta)^2}{c_R R^2(1+c_R R\zeta_3 a^g_{\theta,c_R R})^2},
\end{aligned}
\tag{B.60}
$$

where $\zeta_3=(1+\delta)^{-1}$, $\zeta_4=\frac{1-0.5\delta}{1-\delta}$, and we have used the arguments of Appendix B.6.3 and the definition of Θ. Using a similar argument, we can show that $\frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}+\frac{\zeta_{2,\theta,\omega}}{1+\omega}\exp\left(a^g_{\theta,\omega}\right)$ is bounded away from one, and finally, we conclude that $\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]}\gamma^g_{\theta,\omega}<1$.

2. From (B.57), we can see that state $(0,0)$ belongs to $C^g_{\theta,\omega}$ for all $\theta\in\Theta$ and $\omega\in[\frac{1}{c_R R},c_R R]$. In order for C^g_* to be a finite set, the supremum of $x^{g_j}_{i,\theta,\omega}$ over Θ and $\tilde{\Pi}$ should be finite. From the definition of $x^{g_1}_{1,\theta,\omega}$ in Appendix B.6.3,

$$
\begin{aligned}
x^{g_1}_{1,\theta,\omega}&=\frac{\omega}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega})}{(\omega+1)\gamma^g_{\theta,\omega}-\omega\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)-\zeta_{2,\theta,\omega}}\\
&\leq\frac{c_R R}{a^g_{\theta,(c_R R)^{-1}}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,c_R R})}{(\omega+1)\gamma^g_{\theta,\omega}-\omega\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)-\zeta_{2,\theta,\omega}},
\end{aligned}
$$

and we can derive a lower bound for the denominator from (B.60). Similarly, we can show that $\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]}x^{g_2}_{2,\theta,\omega}$ is finite. We next find a uniform upper bound for $x^{g_1}_{2,\theta,\omega}$ from Appendix B.6.3,

$$
\begin{aligned}
&x^{g_1}_{2,\theta,\omega}\\
&=\frac{1}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega})+\omega\exp\left(a^g_{\theta,\omega}\frac{x^{g_1}_{1,\theta,\omega}+1}{\omega}\right)\left(\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)-\gamma^g_{\theta,\omega}\right)}{\gamma^g_{\theta,\omega}-\zeta_{2,\theta,\omega}}\\
&\leq\frac{1}{a^g_{\theta,(c_R R)^{-1}}}\log\frac{(2c_R R+1)\exp\left(c_R R a^g_{\theta,c_R R}\left(x^{g_1}_{1,\theta,\omega}+2\right)\right)}{1-\gamma^g_{\theta,\omega}},
\end{aligned}
$$

which is uniformly bounded as $\gamma^g_{\theta,\omega}$ is unformly bounded away from 1 and the second line follows from (B.59) and the fact that $\gamma^g_{\theta,\omega} - \zeta_{2,\theta,\omega} \geq 1 - \gamma^g_{\theta,\omega}$. Arguments verifying the finiteness of the supremum of $x^{g_2}_{1,\theta,\omega}$ follow similarly, and we conclude that $|C^g_*| < \infty$. To confirm a uniform upper bound for $b^g_{\theta,\omega}$, we note that from (B.58),

$$\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]} b^g_{\theta,\omega} \leq \max_{x\in C^g_*} \left(2\exp\left(c_R R a^g_{\theta,c_R R}(x_1+2)\right) + 2\exp\left(a^g_{\theta,c_R R}\left(x_2+2\right)\right)\right),$$

which is finite as $a^g_{\theta,c_R R}$ is independent of the choice of θ and $|C^g_*| < \infty$.

Assumption 4. We next verify Assumption 4 and show that there exists a finite set $C^p_{\theta,\omega}$, constants $\beta^p_{\theta,\omega}$, $b^p_{\theta,\omega} > 0$, $r/(r+1) \leq \alpha^p_{\theta,\omega} < 1$, and a function $V^p_{\theta,\omega} : \mathcal{X} \to [1, +\infty)$ satisfying

$$\Delta V^p_{\theta,\omega}(\boldsymbol{x}) \leq -\beta^p_{\theta,\omega}\left(V^p_{\theta,\omega}(\boldsymbol{x})\right)^{\alpha^p_{\theta,\omega}} + b^p_{\theta,t}\mathbb{I}_{C^p_{\theta,\omega}}(\boldsymbol{x}), \quad \boldsymbol{x}\in\mathcal{X}. \tag{B.61}$$

Proposition 5. *The discrete-time Markov process obtained from the queueing system governed by parameter $\theta = (\theta_1, \theta_2) \in \Theta$ and following policy $\pi_\omega \in \tilde{\Pi}$ is polynomially ergodic. This follow because* (B.61) *holds for*

$$V^p_{\theta,\omega}(\boldsymbol{x}) = \frac{x_1^2}{\omega} + x_2^2, \tag{B.62}$$

$$C^p_{\theta,\omega} = \left\{(x_1,x_2)\in\mathcal{X} : x_i \leq \left(16c_R^2R^{3-i} + 101c_R R\right)\frac{\lambda+\theta_i}{\theta_i}, i = 1,2\right\}, \tag{B.63}$$

$$\beta^p_{\theta,\omega} = \min\left(\frac{\theta_2}{2(\theta_2+\lambda)\sqrt{\omega+1}}, \frac{\theta_1+\theta_2-\lambda}{(\theta_1+\theta_2+\lambda)\sqrt{\omega+1}}, \frac{\theta_2}{2(\theta_2+\lambda)}, \frac{\theta_1}{2(\theta_1+\lambda)\sqrt{\omega}}\right), \tag{B.64}$$

$$b^p_{\theta,\omega} = (\beta^p_{\theta,\omega}+1)\max_{\boldsymbol{x}\in C^p_{\theta,\omega}}\left(\frac{(x_1+1)^2}{\omega} + (x_2+1)^2\right), \tag{B.65}$$

$$\alpha^p_{\theta,\omega} = \frac{1}{2}. \tag{B.66}$$

Proof of Proposition 5 is given in Appendix B.6.4. Next, we verify the remaining conditions of Assumption 4.

1. From (B.62) and the fact that $\omega \in [\frac{1}{c_R R}, c_R R]$, the first condition holds with $r^p_* = 2$ and $s^p_* = \sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]} s_{\theta,\omega} = c_R R + 1$.

2. From (B.63), state $(0,0)$ belongs to $C^p_{\theta,\omega}$ for all $\theta\in\Theta$ and $\omega\in[\frac{1}{c_R R}, c_R R]$. Furthermore, for $i = 1, 2$,

$$\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]}\frac{\lambda+\theta_i}{\theta_i} \leq \sup_{\theta\in\Theta}\frac{1}{\tilde{\theta}_2} \leq \sup_{\theta\in\Theta}\frac{R}{\tilde{\theta}_1} \leq \frac{4R}{1+\delta}, \tag{B.67}$$

which follows from the fact that $\theta_1 \leq R\theta_2$ and $\tilde{\theta}_1 \geq 0.25 + 0.25\delta$. Thus, from the definition of $C^p_{\theta,\omega}$ in (B.63), $C^p_* = \cup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} C^p_{\theta,\omega}$ is a finite set. We next verify that the infimum of $\beta^p_{\theta,\omega}$, found in (B.64), is positive. In (B.67), we showed that infimum of $\frac{\lambda+\theta_i}{\theta_i}$ over Θ is lower bounded by $\frac{1+\delta}{4}$. From this, the fact that ω belongs to a compact set, and $\theta_1 + \theta_2 + \lambda \geq \delta$, it follows that $\inf_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} \beta^p_{\theta,\omega} > 0$. Furthermore, it is easy to see that $\beta^p_{\theta,\omega} \leq \sqrt{c_R R}$. Hence, from (B.65),

$$\begin{aligned}\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} b^p_{\theta,\omega} &= \sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} (\beta^p_{\theta,\omega} + 1) \max_{\boldsymbol{x}\in C^p_{\theta,\omega}} \left(\frac{(x_1+1)^2}{\omega} + (x_2+1)^2 \right) \\ &\leq (\sqrt{c_R R} + 1) \max_{\boldsymbol{x}\in C^p_*} \left(C_R R (x_1+1)^2 + (x_2+1)^2 \right),\end{aligned}$$

which is finite as $|C^p_*| < \infty$.

3. We need to show that $K_{\theta,\omega}(\boldsymbol{x}) := \sum_{n=0}^{\infty} 2^{-n-2} \left(P^{\pi_\omega}_\theta\right)^n (\boldsymbol{x}, 0^d)$ is strictly bounded away from zero. We show this using the fact that from any state $\boldsymbol{x}$, the queueing system hits $(0,0)$ in one step with positive probability. Take $x_{i,\theta,\omega} = \max_{\boldsymbol{x}\in C_{\theta,\omega}} x_i$ for $i = 1, 2$. We have

$$\begin{aligned}\inf_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} \min_{\boldsymbol{x}\in C_{\theta,\omega}} K(\boldsymbol{x}) &\geq \inf_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} \min_{\boldsymbol{x}\in C_{\theta,\omega}} P(\boldsymbol{x}, 0^d) \\ &\geq \inf_{\theta\in\Theta,\omega\in[\frac{1}{c_R R}, c_R R]} P\left((x_{1,\theta,\omega}, x_{2,\theta,\omega}), 0^d \right).\end{aligned}$$

The infimum in the right-hand side of the above equation is attained for the minimum normalized service rates possible for each server, or $\tilde{\theta}_1 = \frac{1+\delta}{4}$ and $\tilde{\theta}_2 = \frac{1+\delta}{4R}$. Therefore, the infimum of $K_{\theta,\omega}(\boldsymbol{x})$ over the finite set C^p_*, Θ, and interval $[\frac{1}{c_R R}, c_R R]$ is strictly greater than zero.

Assumption 5. We finally verify that $\sup_{\theta\in\Theta} J(\theta)$ is finite. We first note that for $\boldsymbol{x} = (x_1, x_2)$,

$$(x_1 + x_2)^2 \leq 2\max(\omega^*(\theta), 1) \left(\frac{x_1^2}{\omega^*(\theta)} + x_2^2 \right) = 2\max(\omega^*(\theta), 1) V^p_{\theta,\omega^*(\theta)}(\boldsymbol{x}).$$

From the above equation,

$$\begin{aligned}J(\theta) &= \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,\omega^*(\theta)}} [c(\boldsymbol{X})] \\ &= \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,\omega^*(\theta)}} [\|\boldsymbol{X}\|_1] \\ &\leq \sqrt{2\max(\omega^*(\theta), 1)} \mathbb{E}_{\boldsymbol{X}\sim\mu_{\theta,\omega^*(\theta)}} \left[\sqrt{V^p_{\theta,\omega^*(\theta)}(\boldsymbol{X})} \right],\end{aligned}$$

where $\mu_{\theta,\omega^*(\theta)}$ is the stationary distribution of the discrete-time process governed by parameter θ

and following the best in-class policy according to θ, shown by $\pi_{\omega^*(\theta)}$. From [71], Theorem 14.3.7,

$$\mu_{\theta,\omega^*(\theta)}\left(\sqrt{V^p_{\theta,\omega^*(\theta)}(\boldsymbol{X})}\right) \leq \frac{\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]} b^p_{\theta,\omega}}{\beta^p_*},$$

which is finite from the the previous verified assumption. Thus,

$$\sup_{\theta\in\Theta} J(\theta) \leq \frac{\sqrt{2c_R R}\left(\sup_{\theta\in\Theta,\omega\in[\frac{1}{c_R R},c_R R]} b^p_{\theta,\omega}\right)}{\beta^p_*} < \infty.$$

B.6 Proofs related to the queueing model examples

B.6.1 Proof of Proposition 2

Proof. We define the normalized rates as

$$\tilde{\lambda} = \frac{\lambda}{\lambda+\theta_1+\theta_2}, \quad \tilde{\theta}_i = \frac{\theta_i}{\lambda+\theta_1+\theta_2}, \tag{B.68}$$

for $i = 1, 2$. From the choice of parameter space Θ, we have $\tilde{\lambda} \leq 0.5 - 0.5\delta$, $\theta_1 + \theta_2 \geq 0.5 + 0.5\delta$, and $\theta_1 \geq 0.25 + 0.25\delta$. To prove geometric ergodicity, from the discussions of Section 3.2, it suffices to show that there exists a finite set $C^g_{\theta,t}$, constants $b^g_{\theta,t} > 0$, $\gamma^g_{\theta,t} \in (0,1)$, and a function $V^g_{\theta,t} : \mathcal{X}_t \to [1, +\infty)$ satisfying

$$\Delta V^g_{\theta,t}(\boldsymbol{x}) \leq -\left(1-\gamma^g_{\theta,t}\right) V^g_{\theta,t}(\boldsymbol{x}) + b^g_{\theta,t}\mathbb{I}_{C^g_{\theta,t}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}_t. \tag{B.69}$$

Take $V^g_{\theta,t}(\boldsymbol{x}) = \exp(a^g_{\theta,t}\|\boldsymbol{x}\|_1)$ for some $a^g_{\theta,t} > 0$. For $i \geq 1$ and $\boldsymbol{x} = (i,1,1)$,

$$P^t_\theta V^g_{\theta,t}(i,1,1) = \tilde{\lambda} V^g_{\theta,t}(i+1,1,1) + \tilde{\theta}_1 V^g_{\theta,t}(i,0,1) + \tilde{\theta}_2 V^g_{\theta,t}(i,1,0),$$

where P^t_θ is the corresponding transition kernel. Thus,

$$\begin{aligned}
&P^t_\theta V^g_{\theta,t}(i,1,1) - (1-\gamma^g_{\theta,t}) V^g_{\theta,t}(i,1,1) \\
&= \tilde{\lambda}\exp\left(a^g_{\theta,t}(i+3)\right) + (\tilde{\theta}_1+\tilde{\theta}_2)\exp\left(a^g_{\theta,t}(i+1)\right) - (1-\gamma^g_{\theta,t})\exp\left(a^g_{\theta,t}(i+2)\right) \\
&= \exp\left(a^g_{\theta,t}(i+1)\right)\left(\tilde{\lambda}\exp(2a^g_{\theta,t}) + \tilde{\theta}_1 + \tilde{\theta}_2 - (1-\gamma^g_{\theta,t})\exp(a^g_{\theta,t})\right).
\end{aligned}$$

Take $\tilde{a}_{\theta,t} = \exp(a^g_{\theta,t})$. We need to find $\tilde{a}_{\theta,t} > 1$ and $0 < \gamma^g_{\theta,t} < 1$ such that

$$\tilde{\lambda}\tilde{a}^2_{\theta,t} - (1-\gamma^g_{\theta,t})\tilde{a}_{\theta,t} + \tilde{\theta}_1 + \tilde{\theta}_2 < 0. \tag{B.70}$$

Take $\tilde{a}_{\theta,t} = (1-\delta)^{-1} > 1$ and

$$\tilde{\gamma}_{\theta,t} := 1 - \gamma^g_{\theta,t} = \frac{1}{2}\left(1 + (1-\tilde{\lambda})(1-\delta) + \tilde{\lambda}\,(1-\delta)^{-1}\right).$$

We need to have $\tilde{\gamma}_{\theta,t} < 1$ which follows from the stability condition $\tilde{\lambda} \leq 0.5 - 0.5\delta$ as below:

$$\begin{aligned}
\tilde{\gamma}_{\theta,t} &= \frac{1}{2} + \frac{1}{2}\left((1-\tilde{\lambda})(1-\delta) + \frac{\tilde{\lambda}}{1-\delta}\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(1 - \delta - \tilde{\lambda}(1-\delta) + \frac{\tilde{\lambda}}{1-\delta}\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(1 - \delta + \tilde{\lambda}\frac{1-(1-\delta)^2}{1-\delta}\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(1 - \delta + \tilde{\lambda}\frac{\delta(2-\delta)}{1-\delta}\right) \\
&\leq \frac{1}{2} + \frac{1}{2}\left(1 - \delta + \frac{\delta(2-\delta)}{2}\right) \\
&= 1 - \frac{\delta^2}{4} \\
&< 1.
\end{aligned}$$

We now verify (B.70):

$$\begin{aligned}
\tilde{\lambda}\tilde{a}^2_{\theta,t} - (1-\gamma^g_{\theta,t})\tilde{a}_{\theta,t} + \tilde{\theta}_1 + \tilde{\theta}_2 &= \frac{\tilde{\lambda}}{(1-\delta)^2} - \frac{1}{2(1-\delta)} - \frac{1-\tilde{\lambda}}{2} - \frac{\tilde{\lambda}}{2(1-\delta)^2} + 1 - \tilde{\lambda} \\
&= \frac{\tilde{\lambda}}{2(1-\delta)^2} + \frac{1-\tilde{\lambda}}{2} - \frac{1}{2(1-\delta)} \\
&= \frac{1}{2(1-\delta)^2}\left(\tilde{\lambda} + (1-\tilde{\lambda})(1-\delta)^2 - (1-\delta)\right) \\
&= \frac{\delta}{2(1-\delta)^2}\left(\delta - 1 - \tilde{\lambda}\delta + 2\tilde{\lambda}\right) \\
&= \frac{\delta}{2(1-\delta)^2}\left(\tilde{\lambda}\,(2-\delta) + \delta - 1\right) \\
&< 0,
\end{aligned}$$

where the last line follows from $\tilde{\lambda} \leq 0.5 - 0.5\delta < (1-\delta)/\,(2-\delta)$.

For $\boldsymbol{x} = (i, 0, 1)$ and $i \geq 1$, we have

$$P^t_\theta V^g_{\theta,t}(i,0,1) = \tilde{\lambda} V^g_{\theta,t}(i,1,1) + \tilde{\theta}_1 V^g_{\theta,t}(i-1,0,1) + \tilde{\theta}_2 V^g_{\theta,t}(i-1,1,0),$$

and

$$
\begin{aligned}
&P_\theta^t V_{\theta,t}^g(i,0,1) - (1-\gamma_{\theta,t}^g) V_{\theta,t}^g(i,0,1) \\
&= \tilde{\lambda} \exp\left(a_{\theta,t}^g (i+2)\right) + (\tilde{\theta}_1 + \tilde{\theta}_2) \exp\left(a_{\theta,t}^g i\right) - (1-\gamma_{\theta,t}^g) \exp\left(a_{\theta,t}^g (i+1)\right) \\
&= \exp\left(a_{\theta,t}^g i\right) \left(\tilde{\lambda} \exp(2a_{\theta,t}^g) + \tilde{\theta}_1 + \tilde{\theta}_2 - (1-\gamma_{\theta,t}^g) \exp(a_{\theta,t}^g)\right),
\end{aligned}
$$

which results in the same conditions as previously discussed. When $\boldsymbol{x} = (i,1,0)$ and $i \geq t$ also same argument holds.

Finally, (B.69) holds for

$$
\begin{aligned}
C_{\theta,t}^g &= \{(x_0, x_1, 0) : x_0 < t\} \cup \{(0,0,1)\}, \\
a_{\theta,t}^g &= -\log(1-\delta), \\
\gamma_{\theta,t}^g &= \frac{1}{2} - \frac{1}{2}\left((1-\tilde{\lambda})(1-\delta) + \tilde{\lambda}\,(1-\delta)^{-1}\right), \\
V_{\theta,t}^g(\boldsymbol{x}) &= \exp(a_{\theta,t}^g \|\boldsymbol{x}\|_1), \\
b_{\theta,t}^g &= \max_{\boldsymbol{x} \in C_{\theta,t}^g} \exp(a_{\theta,t}^g \|\boldsymbol{x}\|_1) \left(\exp(a_{\theta,t}^g) + 1\right),
\end{aligned}
$$

where the last line holds because $PV_{\theta,t}^g(\boldsymbol{x}) \leq V_{\theta,t}^g(\boldsymbol{y})$ for $\boldsymbol{y}$ such that $\|\boldsymbol{y}\|_1 = \|\boldsymbol{x}\|_1 + 1$. □

B.6.2 Proof of Proposition 3

Proof. In order to show polynomially ergodicity, we will verify (B.47). We define $V_{\theta,t}^p(\boldsymbol{x}) = \|\boldsymbol{x}\|_1^2$ and $\alpha_{\theta,t}^p = 1/2$, which is equal to $r/(r+1)$ for $r = 1$; r is defined in Assumption 1. For $\boldsymbol{x} = (i,0,1)$ and $i \geq 1$,

$$
P_\theta^t V_{\theta,t}^p(i,0,1) = \tilde{\lambda} V_{\theta,t}^p(i,1,1) + \tilde{\theta}_1 V_{\theta,t}^p(i-1,0,1) + \tilde{\theta}_2 V_{\theta,t}^p(i-1,1,0),
$$

in which $\tilde{\lambda}$, $\tilde{\theta}_1$, and $\tilde{\theta}_2$ are the normalized rates defined in (B.68). Thus,

$$
\begin{aligned}
&P_\theta^t V_{\theta,t}^p(i,0,1) - V_{\theta,t}^p(i,0,1) + \beta_{\theta,t}^p \sqrt{V_{\theta,t}^p(i,0,1)} \\
&= \tilde{\lambda}(i+2)^2 + (\tilde{\theta}_1 + \tilde{\theta}_2) i^2 - (i+1)^2 + \beta_{\theta,t}^p (i+1) \\
&= i(4\tilde{\lambda} - 2 + \beta_{\theta,t}^p) + 4\tilde{\lambda} - 1 + \beta_{\theta,t}^p.
\end{aligned}
$$

For $\beta^p_{\theta,t} = 1 - 2\tilde{\lambda}$, the right-hand side of above equation is non-positive for $i \geq \frac{2\tilde{\lambda}}{1-2\tilde{\lambda}}$. For $\boldsymbol{x} = (i, 1, 0)$ and $i \geq t$,

$$P^t_\theta V^p_{\theta,t}(i,1,0) = \tilde{\lambda} V^p_{\theta,t}(i,1,1) + \tilde{\theta}_1 V^p_{\theta,t}(i-1,0,1) + \tilde{\theta}_2 V^p_{\theta,t}(i-1,1,0).$$

Thus,

$$\begin{aligned}
&P^t_\theta V^p_{\theta,t}(i,1,0) - V^p_{\theta,t}(i,1,0) + \beta^p_{\theta,t}\sqrt{V^p_{\theta,t}(i,1,0)} \\
&= \tilde{\lambda}(i+2)^2 + (\tilde{\theta}_1 + \tilde{\theta}_2)i^2 - (i+1)^2 + \beta^p_{\theta,t}(i+1) \\
&= i(4\tilde{\lambda} - 2 + \beta^p_{\theta,t}) + 4\tilde{\lambda} - 1 + \beta^p_{\theta,t},
\end{aligned}$$

which is also non-positive under the same conditions as the previous case. For $i \geq 1$ and $\boldsymbol{x} = (i, 1, 1)$,

$$P^t_\theta V^p_{\theta,t}(i,1,1) = \tilde{\lambda} V^p_{\theta,t}(i+1,1,1) + \tilde{\theta}_1 V^p_{\theta,t}(i,0,1) + \tilde{\theta}_2 V^p_{\theta,t}(i,1,0).$$

Thus,

$$\begin{aligned}
&P^t_\theta V^p_{\theta,t}(i,1,1) - V^p_{\theta,t}(i,1,1) + \beta^p_{\theta,t}\sqrt{V^p_{\theta,t}(i,1,1)} \\
&= \tilde{\lambda}(i+3)^2 + (\tilde{\theta}_1 + \tilde{\theta}_2)(i+1)^2 - (i+2)^2 + \beta^p_{\theta,t}(i+2) \\
&= i(4\tilde{\lambda} - 2 + \beta^p_{\theta,t}) + 8\tilde{\lambda} - 3 + 2\beta^p_{\theta,t},
\end{aligned}$$

which is non-positive under the same conditions as the first case. Finally, (B.47) holds for

$$\begin{aligned}
C^p_{\theta,t} &= \{(x_0, x_1, 0) : x_0 < t\} \cup \left\{(x_0, x_1, x_2) : x_0 < \frac{2\tilde{\lambda}}{1-2\tilde{\lambda}}, x_1 + x_2 \geq 1\right\}, \\
\beta^p_{\theta,t} &= 1 - 2\tilde{\lambda}, \\
\alpha^p_{\theta,t} &= \frac{1}{2}, \\
V^p_{\theta,t}(\boldsymbol{x}) &= \|\boldsymbol{x}\|_1^2, \\
b^p_{\theta,t} &= \max_{\boldsymbol{x} \in C^p_{\theta,t}} (\|\boldsymbol{x}\|_1 + 1)^2),
\end{aligned}$$

where the last line holds because $PV^p_{\theta,t}(\boldsymbol{x}) \leq V^p_{\theta,t}(\boldsymbol{y})$ for $\boldsymbol{y}$ such that $\|\boldsymbol{y}\|_1 = \|\boldsymbol{x}\|_1 + 1$. □

B.6.3 Proof of Proposition 4

Proof. To show geometric ergodicity of the chain that follows π_ω, we verify (B.55). Take $a^g_{\theta,\omega} > 0$ and

$$V^g_{\theta,\omega}(\boldsymbol{x}) = \frac{\omega}{\omega+1}\exp\left(a^g_{\theta,\omega}\frac{x_1+1}{\omega}\right) + \frac{1}{\omega+1}\exp\left(a^g_{\theta,\omega}\left(x_2+1\right)\right). \tag{B.71}$$

First, we find $PV^g_{\theta,\omega}(\boldsymbol{x})$ for the function defined above. We have

$$PV^g_{\theta,\omega}(\boldsymbol{x}) = \mathbb{E}^{\pi_\omega}_{\boldsymbol{x}}\left[\frac{\omega}{\omega+1}\exp\left(a^g_{\theta,\omega}\frac{X_1(2)+1}{\omega}\right)\right] + \mathbb{E}^{\pi_\omega}_{\boldsymbol{x}}\left[\frac{1}{\omega+1}\exp\left(a^g_{\theta,\omega}\left(X_2(2)+1\right)\right)\right], \tag{B.72}$$

where $\boldsymbol{X}(2) = (X_1(2), X_2(2))$ is the state of the system at the second arrival, starting from state $\boldsymbol{x}$. To find the above expectations, we first find the corresponding transition probabilities. If the number of departures from server i during a fixed interval with length t is less than the total number of jobs in the queue of that server, the number of departures follows a Poisson distribution with parameter $\theta_i t$. Let $\mathbb{P}\left((x_1,x_2)\to(x_1',\mathcal{X})\right)$ be the probability of transitioning from a system with x_i jobs in server-queue pair i (just after the assignment of the arrival) to a queueing system with x_1' jobs in the first server-queue pair (just before the upcoming arrival). For $1 \le x_1' \le x_1$, we have

$$\begin{aligned}\mathbb{P}\left((x_1,x_2)\to(x_1',\mathcal{X})\right) &= \int_0^\infty \lambda\exp(-\lambda t)\frac{(\theta_1 t)^{x_1-x_1'}}{(x_1-x_1')!}\exp(-\theta_1 t)\,dt \\ &= \frac{\lambda}{\theta_1+\lambda}\left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1-x_1'},\end{aligned} \tag{B.73}$$

and

$$\mathbb{P}\left((x_1,x_2)\to(0,\mathcal{X})\right) = 1 - \sum_{i=1}^{x_1}\frac{\lambda}{\theta_1+\lambda}\left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1-i} = \left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1}. \tag{B.74}$$

Assume $1+x_1 \leq \omega(1+x_2)$, which results in the new arrival being assigned to the first server. For the first term in (B.72), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}^{\pi_\omega}\left[\exp\left(a_{\theta,\omega}^g \frac{X_1(2)}{\omega}\right)\right] \\
&= \sum_{i=0}^{x_1+1} \mathbb{P}\left((x_1+1, x_2) \to (i, \mathcal{X})\right) \exp\left(a_{\theta,\omega}^g \frac{i}{\omega}\right) \\
&= \left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1} + \sum_{i=1}^{x_1+1} \exp\left(a_{\theta,\omega}^g \frac{i}{\omega}\right) \frac{\lambda}{\theta_1+\lambda} \left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1-i} \\
&= \left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1} + \frac{\lambda}{\theta_1+\lambda} \exp\left(a_{\theta,\omega}^g \frac{x_1+1}{\omega}\right) \frac{1-\exp\left(-a_{\theta,\omega}^g \frac{x_1+1}{\omega}\right)\left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1}}{1-\exp\left(-\frac{a_{\theta,\omega}^g}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}}, \\
&< \left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1} + \frac{\lambda}{\theta_1+\lambda} \exp\left(a_{\theta,\omega}^g \frac{x_1+1}{\omega}\right) \frac{1}{1-\exp\left(-\frac{a_{\theta,\omega}^g}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}}. \qquad \text{(B.75)}
\end{aligned}$$

Similarly, for the second term in (B.72), we have

$$\mathbb{E}_{\boldsymbol{x}}^{\pi_\omega}\left[\exp\left(a_{\theta,\omega}^g X_2(2)\right)\right] \leq \left(\frac{\theta_2}{\theta_2+\lambda}\right)^{x_2} + \frac{\lambda}{\theta_2+\lambda} \exp\left(a_{\theta,\omega}^g x_2\right) \frac{1}{1-\exp\left(-a_{\theta,\omega}^g\right)\frac{\theta_2}{\theta_2+\lambda}}. \qquad \text{(B.76)}$$

To satisfy (B.55), for some $0 < \gamma_{\theta,\omega}^g < 1$ and all but finitely many $\boldsymbol{x}$, the following should hold,

$$PV_{\theta,\omega}^g(\boldsymbol{x}) \leq \gamma_{\theta,\omega}^g V_{\theta,\omega}^g(\boldsymbol{x}),$$

or from (B.71) and (B.72),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}^{\pi_\omega}\left[\omega \exp\left(a_{\theta,\omega}^g \frac{X_1(2)+1}{\omega}\right)\right] + \mathbb{E}_{\boldsymbol{x}}^{\pi_\omega}\left[\exp\left(a_{\theta,\omega}^g \left(X_2(2)+1\right)\right)\right] \\
&\leq \gamma_{\theta,\omega}^g \left(\omega \exp\left(a_{\theta,\omega}^g \frac{x_1+1}{\omega}\right) + \exp\left(a_{\theta,\omega}^g \left(x_2+1\right)\right)\right).
\end{aligned}$$

Notice that

$$\omega\left(\frac{\theta_1}{\theta_1+\lambda}\right)^{x_1+1} + \left(\frac{\theta_2}{\theta_2+\lambda}\right)^{x_2} \leq c_R R + 1.$$

From (B.75) and (B.76), it suffices to have

$$(c_R R+1)\exp(c_R R a^g_{\theta,\omega}) + \frac{\omega\frac{\lambda}{\theta_1+\lambda}\exp\left(a^g_{\theta,\omega}\frac{x_1+2}{\omega}\right)}{1-\exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}} + \frac{\frac{\lambda}{\theta_2+\lambda}\exp\left(a^g_{\theta,\omega}(x_2+1)\right)}{1-\exp\left(-a^g_{\theta,\omega}\right)\frac{\theta_2}{\theta_2+\lambda}}$$
$$\leq \gamma^g_{\theta,\omega}\left(\omega\exp\left(a^g_{\theta,\omega}\frac{x_1+1}{\omega}\right)+\exp\left(a^g_{\theta,\omega}(x_2+1)\right)\right). \tag{B.77}$$

Define

$$\zeta_{1,\theta,\omega} = \frac{\frac{\lambda}{\theta_1+\lambda}}{1-\exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}}, \qquad \zeta_{2,\theta,\omega} = \frac{\frac{\lambda}{\theta_2+\lambda}}{1-\exp\left(-a^g_{\theta,\omega}\right)\frac{\theta_2}{\theta_2+\lambda}}.$$

Simplifying (B.77), we need the following to hold

$$(c_R R+1)\exp(c_R R a^g_{\theta,\omega}) + \omega\exp\left(a^g_{\theta,\omega}\frac{x_1+1}{\omega}\right)\left(\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)-\gamma^g_{\theta,\omega}\right)$$
$$+\exp\left(a^g_{\theta,\omega}(x_2+1)\right)\left(\zeta_{2,\theta,\omega}-\gamma^g_{\theta,\omega}\right)\leq 0. \tag{B.78}$$

As $\zeta_{i,\theta,\omega}<1$, there exists $\gamma^g_{\theta,\omega}$ such that

$$\zeta_{2,\theta,\omega} < \gamma^g_{\theta,\omega} < 1.$$

From the assumption $1+x_1 \leq \omega(1+x_2)$ and the above equation, (B.78) can be further simplified as

$$(c_R R+1)\exp(c_R R a^g_{\theta,\omega}) + \exp\left(a^g_{\theta,\omega}\frac{x_1+1}{\omega}\right)\left(\omega\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)+\zeta_{2,\theta,\omega}-(\omega+1)\gamma^g_{\theta,\omega}\right)\leq 0. \tag{B.79}$$

For the above to hold outside a finite set, we need to have

$$\frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right)+\frac{\zeta_{2,\theta,\omega}}{1+\omega} < \gamma^g_{\theta,\omega}. \tag{B.80}$$

Define

$$\zeta_3 = \frac{1}{1+\delta}, \quad \zeta_4 = \frac{1-0.5\delta}{1-\delta}. \tag{B.81}$$

Note that $\zeta_3 < 1$ and $\zeta_4 > 1$. Defining function $f(y) := 1+\zeta_4 y - \exp(y)$, we note that for $y \leq \log\zeta_4$, $f(y)>0$, where $\log\zeta_4$ is the maximizer of $f(y)$. Similarly, taking $g(y) := 1-\zeta_3 y -$

$\exp(-y)$, for $y \leq -\log\zeta_3$, $g(y) > 0$, where $-\log\zeta_3$ is the maximizer of $g(y)$. Thus, we conclude that for $a^g_{\theta,\omega} \leq \min\left(-\omega\log\zeta_3, -\log\zeta_3, \omega\log\zeta_4\right)$,

$$\exp(-y) \leq 1 - \zeta_3 y \quad \text{holds for} \quad y \leq \max\left(\frac{a^g_{\theta,\omega}}{\omega}, a^g_{\theta,\omega}\right), \tag{B.82}$$

$$\exp(y) \leq 1 + \zeta_4 y \quad \text{holds for} \quad y \leq \frac{a^g_{\theta,\omega}}{\omega}. \tag{B.83}$$

To guarantee the existence of $0 < \gamma^g_{\theta,\omega} < 1$ that satisfies (B.80), we need to ensure the left-hand side of (B.80) is strictly less than 1. Using the bounds found in (B.82) and (B.83) and the definition of $\zeta_{1,\theta,\omega}$ and $\zeta_{2,\theta,\omega}$, we simplify (B.80) to get

$$\frac{\frac{\lambda}{1+\omega}\left(\omega + a^g_{\theta,\omega}\zeta_4\right)}{\lambda + \frac{\theta_1 a^g_{\theta,\omega}\zeta_3}{\omega}} + \frac{\frac{\lambda}{1+\omega}}{\lambda + \theta_2 a^g_{\theta,\omega}\zeta_3} < 1,$$

which is equivalent to

$$a^g_{\theta,\omega}\zeta_3\theta_2\left(\lambda\zeta_4 - \frac{\zeta_3\theta_1(1+\omega)}{\omega}\right) < \lambda\zeta_3\left(\theta_1+\theta_2\right) - \lambda^2\zeta_4. \tag{B.84}$$

To make sure there exists $a^g_{\theta,\omega} > 0$ that satisfies (B.84), the right-hand side of (B.84) needs to be positive, which follows as below:

$$\begin{aligned}
\lambda\zeta_3\left(\theta_1+\theta_2\right) - \lambda^2\zeta_4 &= \lambda\left(\frac{\theta_1+\theta_2}{1+\delta} - \lambda\frac{1-0.5\delta}{1-\delta}\right) \\
&= \lambda(\theta_1+\theta_2+\lambda)\left(\frac{1-\tilde{\lambda}}{1+\delta} - \tilde{\lambda}\frac{1-0.5\delta}{1-\delta}\right) \\
&= \lambda(\theta_1+\theta_2+\lambda)\left(\frac{1}{1+\delta} - \tilde{\lambda}\left(\frac{1}{1+\delta} + \frac{1-0.5\delta}{1-\delta}\right)\right) \\
&\geq \lambda(\theta_1+\theta_2+\lambda)\left(\frac{1}{1+\delta} - \frac{1-\delta}{2}\left(\frac{1}{1+\delta} + \frac{1-0.5\delta}{1-\delta}\right)\right) \\
&= \frac{\delta}{4}\lambda(\theta_1+\theta_2+\lambda)
\end{aligned} \tag{B.85}$$

where $\tilde{\lambda}$, $\tilde{\theta}_1$, and $\tilde{\theta}_2$ are the normalized rates defined in (B.68) and we have used the stability condition $\tilde{\lambda} \leq 0.5 - 0.5\delta$. We further simplify the left-hand side of (B.84) as

$$\zeta_3\theta_2\left(\lambda\zeta_4 - \frac{\zeta_3\theta_1(1+\omega)}{\omega}\right) < \theta_2\lambda\zeta_3\zeta_4 < \frac{1-0.5\delta}{1-\delta^2}(\theta_1+\theta_2+\lambda)\lambda.$$

From the above equation and (B.85), $a^g_{\theta,\omega}$ needs to satisfy

$$a^g_{\theta,\omega} \leq \frac{\delta(1-\delta^2)}{8(1-0.5\delta)}.$$

Finally, we take $a^g_{\theta,\omega}$ as

$$a^g_{\theta,\omega} = \min\left(-\omega\log\zeta_3, -\log\zeta_3, \omega\log\zeta_4, \frac{\delta(1-\delta^2)}{8(1-0.5\delta)}\right).$$

After finding an appropriate $a^g_{\theta,\omega}$, we can choose $0 < \gamma^g_{\theta,\omega} < 1$ such that (B.80) holds or

$$\gamma^g_{\theta,\omega} \geq \frac{1}{2}\left(1 + \frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) + \frac{\zeta_{2,\theta,\omega}}{1+\omega}\right).$$

Moreover, from (B.79) a lower bound $x^{g_1}_{1,\theta,\omega}$ for x_1 is derived; In other words,(B.79) holds for $x_1 > x^{g_1}_{1,\theta,\omega}$. From (B.78), we can find the corresponding $x^{g_1}_{2,\theta,\omega}$ and take $\boldsymbol{x}^{g_1}_{\theta,\omega} = (x^{g_1}_{1,\theta,\omega}, x^{g_1}_{2,\theta,\omega})$. By repeating the same arguments when $1 + x_1 < \omega(1 + x_2)$, we finally conclude that

$$\Delta V^g_{\theta,\omega}(\boldsymbol{x}) \leq -\left(1-\gamma^g_{\theta,\omega}\right)V^g_{\theta,\omega}(\boldsymbol{x}) + b^g_{\theta,\omega}\mathbb{I}_{C^g_{\theta,\omega}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X},$$

for

$$V^g_{\theta,\omega}(\boldsymbol{x}) = \frac{\omega}{\omega+1}\exp\left(a^g_{\theta,\omega}\frac{x_1+1}{\omega}\right) + \frac{1}{\omega+1}\exp\left(a^g_{\theta,\omega}\left(x_2+1\right)\right),$$

$$a^g_{\theta,\omega} = \min\left(\omega\log(1+\delta), \log(1+\delta), \omega\log\frac{1-0.5\delta}{1-\delta}, \log\frac{1-0.5\delta}{1-\delta}, \frac{\delta(1-\delta^2)}{4c_R R(1-0.5\delta)}\right),$$

$$C^g_{\theta,\omega} = \{(x_1,x_2)\in\mathcal{X} : x_i \le \max\left(x^{g_j}_{i,\theta,\omega}, 0\right), i,j=1,2\},$$

$$\gamma^g_{\theta,\omega} = \frac{1}{2} + \frac{1}{2}\max\left(\zeta_{1,\theta,\omega}, \zeta_{2,\theta,\omega}, \frac{\zeta_{1,\theta,\omega}\omega}{1+\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) + \frac{\zeta_{2,\theta,\omega}}{1+\omega}, \frac{\zeta_{1,\theta,\omega}\omega}{1+\omega} + \frac{\zeta_{2,\theta,\omega}}{1+\omega}\exp\left(a^g_{\theta,\omega}\right)\right),$$

$$b^g_{\theta,\omega} = \max_{\boldsymbol{x}\in C^g_{\theta,\omega}}\left(\frac{2\omega}{\omega+1}\exp\left(a^g_{\theta,\omega}\frac{x_1+2}{\omega}\right) + \frac{2}{\omega+1}\exp\left(a^g_{\theta,\omega}\left(x_2+2\right)\right)\right),$$

$$\zeta_{1,\theta,\omega} = \frac{\frac{\lambda}{\theta_1+\lambda}}{1-\exp\left(-\frac{a^g_{\theta,\omega}}{\omega}\right)\frac{\theta_1}{\theta_1+\lambda}},$$

$$\zeta_{2,\theta,\omega} = \frac{\frac{\lambda}{\theta_2+\lambda}}{1-\exp\left(-a^g_{\theta,\omega}\right)\frac{\theta_2}{\theta_2+\lambda}},$$

$$x^{g_1}_{1,\theta,\omega} = \frac{\omega}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega})}{(\omega+1)\gamma^g_{\theta,\omega} - \omega\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) - \zeta_{2,\theta,\omega}},$$

$$x^{g_1}_{2,\theta,\omega} = \frac{1}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega}) + \omega\exp\left(a^g_{\theta,\omega}\frac{x^{g_1}_{1,\theta,\omega}+1}{\omega}\right)\left(\zeta_{1,\theta,\omega}\exp\left(\frac{a^g_{\theta,\omega}}{\omega}\right) - \gamma^g_{\theta,\omega}\right)}{\gamma^g_{\theta,\omega} - \zeta_{2,\theta,\omega}},$$

$$x^{g_2}_{2,\theta,\omega} = \frac{1}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega})}{(\omega+1)\gamma^g_{\theta,\omega} - \omega\zeta_{1,\theta,\omega} - \zeta_{2,\theta,\omega}\exp\left(a^g_{\theta,\omega}\right)},$$

$$x^{g_2}_{1,\theta,\omega} = \frac{\omega}{a^g_{\theta,\omega}}\log\frac{(c_R R+1)\exp(c_R R a^g_{\theta,\omega}) + \exp\left(a^g_{\theta,\omega}(x^{g_2}_{2,\theta,\omega}+1)\right)\left(\zeta_{2,\theta,\omega}\exp\left(a^g_{\theta,\omega}\right) - \gamma^g_{\theta,\omega}\right)}{\omega\left(\gamma^g_{\theta,\omega} - \zeta_{1,\theta,\omega}\right)}.$$

□

B.6.4 Proof of Proposition 5

Proof. Define $V^p_{\theta,\omega}(\boldsymbol{x}) = \frac{x_1^2}{\omega} + x_2^2$, and $\alpha^p_{\theta,\omega} = 1/2$. Assume that $x_1 = 0$ and $x_2 > (1-\omega)/\omega$; which means the new job will be assigned to the first server. The transition probabilities of the discrete-time chain sampled at Poisson arrivals is given in (B.73) and (B.74), and we calculate

$PV^p_{\theta,\omega}(\boldsymbol{x})$ as

$$\begin{aligned} PV^p_{\theta,\omega}(\boldsymbol{x}) &= \frac{\lambda}{\omega(\lambda+\theta_1)} + \sum_{i=1}^{x_2} i^2 \frac{\lambda}{\lambda+\theta_2}\left(\frac{\theta_2}{\theta_2+\lambda}\right)^{x_2-i} \\ &< c_R R + \sum_{i=1}^{x_2} i^2 \frac{\lambda}{\lambda+\theta_2}\left(\frac{\theta_2}{\theta_2+\lambda}\right)^{x_2-i}. \end{aligned} \tag{B.86}$$

We define $d_i := \theta_i/(\theta_i+\lambda)$ for $i = 1, 2$ and

$$\begin{aligned} &\sum_{i=1}^{x_2} i^2 \frac{\lambda}{\lambda+\theta_2}\left(\frac{\theta_2}{\theta_2+\lambda}\right)^{x_2-i} \\ &= \frac{1}{(1-d_2)^2}\left(-d_2^{x_2}\left(d_2+d_2^2\right) + d_2^2\left(x_2^2+2x_2+1\right) + d_2\left(-2x_2^2-2x_2+1\right) + x_2^2\right) \\ &= \frac{1}{(1-d_2)^2}\left(\left(1-d_2^{x_2}\right)\left(d_2+d_2^2\right) + x_2^2\left(d_2^2-2d_2+1\right) + x_2\left(2d_2^2-2d_2\right)\right) \\ &= x_2^2 - \frac{2d_2}{1-d_2}x_2 + \frac{\left(1-d_2^{x_2}\right)\left(d_2+d_2^2\right)}{(1-d_2)^2}. \end{aligned} \tag{B.87}$$

From (B.86),

$$PV^p_{\theta,\omega}(\boldsymbol{x}) - V^p_{\theta,\omega}(\boldsymbol{x}) + \beta^p_{\theta,\omega}x_2 < \left(-\frac{2d_2}{1-d_2} + \beta^p_{\theta,\omega}\right)x_2 + \frac{\left(1-d_2^{x_2}\right)\left(d_2+d_2^2\right)}{(1-d_2)^2} + c_R R.$$

Outside a finite set, we need the above equation to be non-positive; which is equivalent to

$$\left(-2 + \beta^p_{\theta,\omega}\frac{1-d_2}{d_2}\right)x_2 + \frac{\left(1-d_2^{x_2}\right)\left(1+d_2\right)}{1-d_2} + c_R R\frac{1-d_2}{d_2} \le 0.$$

As $d_2 < 1$,

$$\frac{1-d_2^y}{1-d_2} = 1 + d_2 + \ldots + d_2^{y-1} \le y \qquad \text{for } \; y \ge 1. \tag{B.88}$$

Thus,

$$\begin{aligned} &\left(-2 + \beta^p_{\theta,\omega}\frac{1-d_2}{d_2}\right)x_2 + \frac{\left(1-d_2^{x_2}\right)\left(1+d_2\right)}{1-d_2} + c_R R\frac{1-d_2}{d_2} \\ &\le \left(d_2 - 1 + \beta^p_{\theta,\omega}\frac{1-d_2}{d_2}\right)x_2 + c_R R\frac{1-d_2}{d_2}. \end{aligned}$$

By taking $\beta^p_{\theta,\omega} \le d_2/2$, it suffices for the following to be non-positive,

$$-\frac{1-d_2}{2}x_2 + c_R R\frac{1-d_2}{d_2} \le 0,$$

which holds for $x_2 \geq 2c_R R/d_2$. Thus, for $x_1 = 0$ and $x_2 \geq \max\left(2c_R R(\lambda+\theta_2)/\theta_2, (1-\omega)/\omega\right) = 2c_R R(\lambda+\theta_2)/\theta_2$, (B.61) holds. The case of $x_2 = 0$ and non-zero x_1 follows same arguments and (B.61) holds for $\beta^p_{\theta,\omega} \leq d_1/2\sqrt{\omega}$, $x_2 = 0$, and $x_1 \geq \max\left(2c_R R(\lambda+\theta_1)/\theta_1, \omega - 1\right) = 2c_R R(\lambda+\theta_1)/\theta_1$. We now consider the case of $x_1, x_2 > 0$ and $x_1 + 1 \leq \omega(x_2+1)$, and note that

$$\sqrt{V^p_{\theta,\omega}(\boldsymbol{x})} = \sqrt{\frac{x_1^2}{\omega} + x_2^2} \leq \sqrt{\frac{(x_1+1)^2}{\omega} + (x_2+1)^2} \leq \sqrt{\omega+1}(x_2+1).$$

Hence, it suffices to find finite set $C^p_{\theta,\omega}$, constants $b^p_{\theta,\omega}$ and $\beta^p_{\theta,\omega} > 0$, such that the following holds for $V^p_{\theta,\omega}(\boldsymbol{x}) = \frac{x_1^2}{\omega} + x_2^2$,

$$\Delta V^p_{\theta,\omega}(\boldsymbol{x}) \leq -\sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) + b^p_{\theta,\omega}\mathbb{I}_{C^p_{\theta,\omega}}(\boldsymbol{x}).$$

As $x_1 + 1 \leq \omega(x_2+1)$, the new arrival is assigned to the first queue and we find $\Delta V^p_{\theta,\omega}(\boldsymbol{x}) + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1)$ using the same calculations as (B.87).

$$\begin{aligned}
&\Delta V^p_{\theta,\omega}(\boldsymbol{x}) + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) \\
&= \frac{1}{\omega}\left((x_1+1)^2 - \frac{2d_1}{1-d_1}(x_1+1) + \frac{\left(1-d_1^{x_1+1}\right)\left(d_1+d_1^2\right)}{(1-d_1)^2} - x_1^2\right) \\
&- \frac{2d_2}{1-d_2}x_2 + \frac{\left(1-d_2^{x_2}\right)\left(d_2+d_2^2\right)}{(1-d_2)^2} + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) \\
&= \frac{x_1}{\omega}\left(2 - \frac{2d_1}{1-d_1}\right) + \frac{1-3d_1}{\omega(1-d_1)} + \frac{\left(1-d_1^{x_1+1}\right)\left(d_1+d_1^2\right)}{\omega(1-d_1)^2} \qquad \text{(B.89)} \\
&+ (x_2+1)\left(-\frac{2d_2}{1-d_2} + \sqrt{\omega+1}\beta^p_{\theta,\omega}\right) + \frac{2d_2}{1-d_2} + \frac{\left(1-d_2^{x_2}\right)\left(d_2+d_2^2\right)}{(1-d_2)^2}. \qquad \text{(B.90)}
\end{aligned}$$

We next consider two different cases based on the value of d_1 and analyze them separately.

One. $0.8 \leq d_1 < 1$: We first notice that the coefficient of x_1 in (B.89) is negative, as $d_1 > 1/2$.

For $x_1 \geq 1$, (B.89) is equal to

$$
\begin{aligned}
&\frac{1}{\omega(1-d_1)}\left((2-4d_1)x_1+1-3d_1+(d_1+d_1^2)\sum_{i=0}^{x_1}d_1^i\right)\\
&=\frac{1}{\omega(1-d_1)}\left((2-4d_1)(x_1-1)+d_1^3(1+d_1)\sum_{i=0}^{x_1-2}d_1^i+d_1(1+d_1)^2+3-7d_1\right)\\
&\leq\frac{1}{\omega(1-d_1)}\left((2-4d_1)(x_1-1)+d_1^3(1+d_1)(x_1-1)+d_1(1+d_1)^2+3-7d_1\right)\\
&=\frac{1}{\omega(1-d_1)}\left((d_1^4+d_1^3-4d_1+2)(x_1-1)+d_1(1+d_1)^2+3-7d_1\right)\\
&=\frac{-d_1^3-2d_1^2-2d_1+2}{\omega}(x_1-1)+\frac{-d_1^2-3d_1+3}{\omega}\\
&<0,
\end{aligned}
$$

where the third line follows from (B.88), and the last line from the fact that when $0.8 \leq d_1 < 1$, both terms $-d_1^3-2d_1^2-2d_1+2$ and $-d_1^2-3d_1+3$ are negative. Next, we notice that (B.90) is equal to

$$
\begin{aligned}
&x_2\left(-\frac{2d_2}{1-d_2}+\sqrt{\omega+1}\beta_{\theta,\omega}^p\right)+\sqrt{\omega+1}\beta_{\theta,\omega}^p+\frac{(1-d_2^{x_2})(d_2+d_2^2)}{(1-d_2)^2}\\
&\leq x_2\left(-\frac{2d_2}{1-d_2}+\sqrt{\omega+1}\beta_{\theta,\omega}^p\right)+\frac{d_2+d_2^2}{1-d_2}x_2+\sqrt{\omega+1}\beta_{\theta,\omega}^p\\
&=x_2\left(-\frac{2d_2}{1-d_2}+\frac{d_2+d_2^2}{1-d_2}+\sqrt{\omega+1}\beta_{\theta,\omega}^p\right)+\sqrt{\omega+1}\beta_{\theta,\omega}^p\\
&=x_2\left(-d_2+\sqrt{\omega+1}\beta_{\theta,\omega}^p\right)+\sqrt{\omega+1}\beta_{\theta,\omega}^p,
\end{aligned}
$$

where the second line follows from (B.88). Taking $\beta_{\theta,\omega}^p \leq d_2/2\sqrt{\omega+1}$, we get

$$
x_2\left(-\frac{2d_2}{1-d_2}+\sqrt{\omega+1}\beta_{\theta,\omega}^p\right)+\sqrt{\omega+1}\beta_{\theta,\omega}^p+\frac{(1-d_2^{x_2})(d_2+d_2^2)}{(1-d_2)^2}\leq-\frac{d_2}{2}x_2+\frac{d_2}{2},
$$

which is non-positive for $x_2 \geq 1$. Finally, when $0.8 \leq d_1 < 1$, $x_1, x_2 > 0$, and $x_1+1 \leq \omega(x_2+1)$, (B.61) holds for $\beta_{\theta,\omega}^p \leq d_2/2\sqrt{\omega+1}$.

Two. $d_1 < 0.8$: Taking $\beta_{\theta,\omega}^p \leq \frac{d_2}{\sqrt{\omega+1}(1-d_2)}$, we note that the coefficient of x_2 in (B.90) is

negative. Thus, from $x_1 + 1 \leq \omega(x_2 + 1)$, (B.89) and (B.90),

$$\begin{aligned}
&\Delta V^p_{\theta,\omega}(\boldsymbol{x}) + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) \\
&\leq \frac{x_1+1}{\omega}\left(2 - \frac{2d_1}{1-d_1}\right) - \frac{1}{\omega} + \frac{\left(1-d_1^{x_1+1}\right)(d_1+d_1^2)}{\omega(1-d_1)^2} \\
&+ \frac{x_1+1}{\omega}\left(-\frac{2d_2}{1-d_2} + \sqrt{\omega+1}\beta^p_{\theta,\omega}\right) + \frac{2d_2}{1-d_2} + \frac{(1-d_2^{x_2})(d_2+d_2^2)}{(1-d_2)^2} \\
&< \frac{x_1+1}{\omega}\left(2 - \frac{2d_1}{1-d_1} - \frac{2d_2}{1-d_2} + \sqrt{\omega+1}\beta^p_{\theta,\omega}\right) + \frac{2d_2}{1-d_2} + \frac{d_1+d_1^2}{\omega(1-d_1)^2} + \frac{d_2+d_2^2}{(1-d_2)^2}.
\end{aligned} \tag{B.91}$$

As $d_i = \tilde{\theta}_i/(\tilde{\theta}_i + \tilde{\lambda})$ in terms of the normalized rates, we get

$$2 - \frac{2d_1}{1-d_1} - \frac{2d_2}{1-d_2} = 2 - \frac{2\tilde{\theta}_1}{\tilde{\lambda}} - \frac{2\tilde{\theta}_1}{\tilde{\lambda}} = \frac{-2(\tilde{\theta}_1 + \tilde{\theta}_2 - \tilde{\lambda})}{\tilde{\lambda}},$$

which is negative from the stability condition. For $\beta^p_{\theta,\omega} \leq \frac{\tilde{\theta}_1+\tilde{\theta}_2-\tilde{\lambda}}{\tilde{\lambda}\sqrt{\omega+1}}$, from (B.91) we get

$$\begin{aligned}
&\Delta V^p_{\theta,\omega}(\boldsymbol{x}) + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) \\
&< \frac{-(\tilde{\theta}_1 + \tilde{\theta}_2 - \tilde{\lambda})}{\omega\tilde{\lambda}}(x_1+1) + \frac{2d_2}{1-d_2} + \frac{d_1+d_1^2}{\omega(1-d_1)^2} + \frac{d_2+d_2^2}{(1-d_2)^2} \\
&= \frac{-(\tilde{\theta}_1 + \tilde{\theta}_2 - \tilde{\lambda})}{\omega\tilde{\lambda}}(x_1+1) + \frac{2\tilde{\theta}_2}{\tilde{\lambda}} + \frac{\tilde{\theta}_1(2\tilde{\theta}_1+\tilde{\lambda})}{\omega\tilde{\lambda}^2} + \frac{\tilde{\theta}_2(2\tilde{\theta}_2+\tilde{\lambda})}{\tilde{\lambda}^2},
\end{aligned}$$

which is non-positive for

$$x_1 + 1 \geq \frac{\tilde{\theta}_1(2\tilde{\theta}_1+\tilde{\lambda}) + \omega\tilde{\theta}_2(2\tilde{\theta}_2+3\tilde{\lambda})}{\tilde{\lambda}(\tilde{\theta}_1+\tilde{\theta}_2-\tilde{\lambda})}.$$

As $d_1 < 0.8$, we can see that $\tilde{\lambda} > \tilde{\theta}_1/4$; thus,

$$\frac{\tilde{\theta}_1(2\tilde{\theta}_1+\tilde{\lambda}) + \omega\tilde{\theta}_2(2\tilde{\theta}_2+3\tilde{\lambda})}{\tilde{\lambda}(\tilde{\theta}_1+\tilde{\theta}_2-\tilde{\lambda})} < \frac{4\tilde{\theta}_1(2\tilde{\theta}_1+\tilde{\lambda}) + 4\omega\tilde{\theta}_2(2\tilde{\theta}_2+3\tilde{\lambda})}{\tilde{\theta}_1(\tilde{\theta}_1+\tilde{\theta}_2-\tilde{\lambda})} < \frac{4c_R R(1+2\tilde{\lambda})}{\delta} \leq 4c_R R,$$

where we have used the fact that $\tilde{\theta}_1 \geq \tilde{\theta}_2$, $\omega \leq c_R R$, $\tilde{\theta}_1 + \tilde{\theta}_2 - \tilde{\lambda} \geq \delta$, and $\tilde{\lambda} \leq 0.5 - 0.5\delta$ and it suffices for x_1 to be greater than or equal to $4c_R R$. For $x_1 < 4c_R R$, (B.89) can be upper bounded as

$$\frac{8c_R R}{\omega} + \frac{1-3d_1}{\omega(1-d_1)} + \frac{d_1+d_1^2}{\omega(1-d_1)^2} \leq \frac{8c_R R}{\omega} + \frac{2}{\omega(1-d_1)^2} < \frac{8c_R R + 50}{\omega},$$

where in the last inequality we have used $d_1 < 0.8$. From (B.90) and taking $\beta^p_{\theta,\omega} \le d_2/2\sqrt{\omega+1}$,

$$\begin{aligned}
&\Delta V^p_{\theta,\omega}(\boldsymbol{x}) + \sqrt{\omega+1}\beta^p_{\theta,\omega}(x_2+1) \\
&\le \frac{8c_R R + 50}{\omega} + \left(-\frac{2d_2}{1-d_2} + \frac{d_2}{2}\right)(x_2+1) + \frac{2d_2}{1-d_2} + \frac{(1-d_2^{x_2})\,(d_2+d_2^2)}{(1-d_2)^2} \\
&\le \left(-\frac{2d_2}{1-d_2} + \frac{d_2}{2} + \frac{d_2+d_2^2}{1-d_2}\right)x_2 + \frac{d_2}{2} + \frac{8c_R R + 50}{\omega} \\
&= -\frac{d_2}{2}x_2 + \frac{d_2}{2} + \frac{8c_R R + 50}{\omega},
\end{aligned}$$

which is negative for

$$x_2 \ge 1 + \frac{16c_R R + 100}{\omega d_2}.$$

Finally, when $x_1 + 1 \le \omega(x_2+1)$ and $x_1, x_2 > 0$, (B.61) holds for $\beta^p_{\theta,\omega} \le \frac{1}{\sqrt{\omega+1}} \min\left(\frac{\tilde\theta_2}{2(\tilde\theta_2+\tilde\lambda)}, \tilde\theta_1 + \tilde\theta_2 - \tilde\lambda\right)$, $x_1 \ge 4c_R R$, and $x_2 \ge 1 + \frac{16c_R R+100}{\omega d_2}$. Repeating the same arguments when $x_1, x_2 > 0$ and $x_1 + 1 > \omega(x_2+1)$, (B.61) holds for $\beta^p_{\theta,\omega} \le \frac{1}{\sqrt{\omega+1}} \min\left(\frac{\tilde\theta_1}{2(\tilde\theta_1+\tilde\lambda)}, \tilde\theta_1 + \tilde\theta_2 - \tilde\lambda\right)$, $x_1 \ge 1 + \frac{\omega(16c_R R^2+100)}{d_1}$, and $x_2 \ge 4c_R R^2$. Finally, (B.61) holds with

$$\begin{aligned}
V^p_{\theta,\omega}(\boldsymbol{x}) &= \frac{x_1^2}{\omega} + x_2^2, \\
C^p_{\theta,\omega} &= \left\{(x_1,x_2) \in \mathcal{X} : x_i \le \left(16c_R^2 R^{3-i} + 101 c_R R\right)\frac{\lambda+\theta_i}{\theta_i}, i = 1,2\right\}, \\
\beta^p_{\theta,\omega} &= \min\left(\frac{\tilde\theta_2}{2(\tilde\theta_2+\tilde\lambda)\sqrt{\omega+1}}, \frac{\tilde\theta_1+\tilde\theta_2-\tilde\lambda}{\sqrt{\omega+1}}, \frac{\tilde\theta_2}{2(\tilde\theta_2+\tilde\lambda)}, \frac{\tilde\theta_1}{2(\tilde\theta_1+\tilde\lambda)\sqrt{\omega}}\right), \\
b^p_{\theta,\omega} &= (\beta^p_{\theta,\omega}+1)\max_{\boldsymbol{x}\in C^p_{\theta,\omega}}\left(\frac{(x_1+1)^2}{\omega} + (x_2+1)^2\right), \\
\alpha^p_{\theta,\omega} &= \frac{1}{2},
\end{aligned}$$

where the fourth line holds since $PV^p_{\theta,\omega}(\boldsymbol{x}) \le V^p_{\theta,\omega}(\boldsymbol{y})$ for $\boldsymbol{y} = (y_1, y_2)$ such that $y_i = x_i + 1$ for $i = 1, 2$. □

BIBLIOGRAPHY

[1] Yasin Abbasi-Yadkori and Csaba Szepesvari. Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm. *arXiv preprint arXiv:1406.3926*, 2014.

[2] Saghar Adler, Mehrdad Moharrami, and Vijay Subramanian. Learning a discrete set of optimal allocation rules in queueing systems with unknown service rates. *arXiv preprint arXiv:2202.02419*, 2022.

[3] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(12):1249–1259, 1989.

[4] Rajeev Agrawal and Demosthenis Teneketzis. Certainty equivalence control with forcing: Revisited. *Systems & control letters*, 13(5):405–412, 1989.

[5] Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(12):1249–1259, 1989.

[6] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

[7] Shipra Agrawal and Randy Jia. Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 743–744, 2019.

[8] Shipra Agrawal and Randy Jia. Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management. *Operations Research*, 70(3):1646–1664, 2022.

[9] Nima Akbarzadeh and Aditya Mahajan. On learning Whittle index policy for restless bandits with scalable regret. *arXiv preprint arXiv:2202.03463*, 2022.

[10] Aristotle Arapostathis, Vivek S Borkar, Emmanuel Fernández-Gaucherand, Mrinal K Ghosh, and Steven I Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

[11] Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

[12] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[13] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[14] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

[15] Gabriel R Bitran and Sriram Dasu. A review of open queueing network models of manufacturing systems. *Queueing systems*, 12:95–133, 1992.

[16] V. Borkar and P. Varaiya. Adaptive control of Markov chains, I: Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957, 1979.

[17] VS Borkar. The kumar-becker-lin scheme revisited. *Journal of optimization theory and applications*, 66:289–309, 1990.

[18] Rolando Cavazos-Cadena. Necessary conditions for the optimality equation in average-reward Markov decision processes. *Applied Mathematics and Optimization*, 19(1):97–112, 1989.

[19] Rolando Cavazos-Cadena. Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs. *Kybernetika*, 25(3):145–156, 1989.

[20] K Mani Chandy and Charles H Sauer. Approximate methods for analyzing queueing network models of computing systems. *ACM Computing Surveys (CSUR)*, 10(3):281–317, 1978.

[21] Xinyun Chen, Yunan Liu, and Guiyu Hong. An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research*, 2023.

[22] Tuhinangshu Choudhury, Gauri Joshi, Weina Wang, and Sanjay Shakkottai. Job dispatching policies for queueing systems with unknown service rates. *arXiv preprint arXiv:2106.04707*, 2021.

[23] Tuhinangshu Choudhury, Gauri Joshi, Weina Wang, and Sanjay Shakkottai. Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, page 181–190. Association for Computing Machinery, 2021.

[24] Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric MDPs with exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 1855–1863. PMLR, 2021.

[25] Asaf Cohen, Vijay Subramanian, and Yili Zhang. Learning-based optimal admission control in a single-server queuing system. *Stochastic Systems*, 2024.

[26] Guy L Curry and Richard M Feldman. *Manufacturing systems modeling and analysis*. Springer Science & Business Media, 2010.

[27] Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

[28] Rick Durrett. *Probability: Theory and examples*, volume 49. Cambridge university press, 2019.

[29] Anthony Ephremides, Pravin Varaiya, and Jean Walrand. A simple dynamic routing problem. *IEEE Transactions on Automatic Control*, 25(4):690–693, 1980.

[30] Lloyd Fisher and Sheldon M Ross. An example in denumerable decision processes. *The Annals of Mathematical Statistics*, 39(2):674–675, 1968.

[31] Daniel Freund, Thodoris Lykouris, and Wentao Weng. Efficient decentralized multi-agent learning in asymmetric queuing systems. In *Conference on Learning Theory*, pages 4080–4084. PMLR, 2022.

[32] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[33] Clement Gehring, Masataro Asai, Rohan Chitnis, Tom Silver, Leslie Kaelbling, Shirin Sohrabi, and Michael Katz. Reinforcement learning for classical planning: Viewing heuristics as dense reward generators. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, pages 588–596, 2022.

[34] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

[35] Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: A reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

[36] Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized Markov decision processes. In *Conference on Learning Theory*, pages 861–898. PMLR, 2015.

[37] Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[38] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.

[39] Bruce Hajek. Optimal control of two interacting service stations. *IEEE transactions on automatic control*, 29(6):491–499, 1984.

[40] Mor Harchol-Balter. *Performance modeling and design of computer systems: Queueing theory in action*. Cambridge University Press, 2013.

[41] Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear Markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

[42] Arie Hordijk and Flora Spieksma. On ergodicity and recurrence properties of a Markov chain by an application to an open Jackson network. *Advances in applied probability*, 24(2):343–376, 1992.

[43] Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 1712–1732. PMLR, 2022.

[44] Soren F Jarner and Gareth O Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12(1):224–247, 2002.

[45] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for service systems with reusable resources. *Operations Research*, 2022.

[46] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[47] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[48] Niels Knudsen. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica: Journal of the Econometric Society*, pages 515–528, 1972.

[49] Subhashini Krishnasamy, PT Akhil, Ari Arapostathis, Rajesh Sundaresan, and Sanjay Shakkottai. Augmenting Max-Weight with explicit learning for wireless scheduling with switching costs. *IEEE/ACM Transactions on Networking*, 26(6):2501–2514, 2018.

[50] Subhashini Krishnasamy, Ari Arapostathis, Ramesh Johari, and Sanjay Shakkottai. On learning the $c\mu$ rule in single and parallel server networks. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 153–154, 2018.

[51] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations Research*, 69(1):315–330, 2021.

[52] David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost. *arXiv preprint arXiv:2011.06709*, 2020.

[53] P. R. Kumar and A. Becker. A new family of optimal adaptive controllers for Markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146, 1982.

[54] P. R. Kumar and Woei Lin. Optimal adaptive controllers for unknown Markov chains. *IEEE Transactions on Automatic Control*, 27(4):765–774, 1982.

[55] P. R. Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[56] Tze-Leung Lai and Sidney Yakowitz. Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, 40(7):1199–1209, 1995.

[57] Ronald Larsen. *Control of multiple exponential servers with application to computer systems*. PhD thesis, University of Maryland, 1981.

[58] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[59] Woei Lin and P. R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic control*, 29(8):696–703, 1984.

[60] Steven A Lippman. Applying a new device in the optimization of exponential queuing systems. *Operations research*, 23(4):687–710, 1975.

[61] Steven A Lippman and Shaler Stidham Jr. Individual versus social optimization in exponential congestion systems. *Operations Research*, 25(2):233–247, 1977.

[62] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4):3133–3174, 2019.

[63] Mufti Mahmud, Mohammed Shamim Kaiser, Amir Hussain, and Stefano Vassanelli. Applications of deep learning and reinforcement learning to biological data. *IEEE transactions on neural networks and learning systems*, 29(6):2063–2079, 2018.

[64] Ashok P. Maitra. *Dynamic programming for countable state systems*. PhD thesis, University of California, 1963.

[65] Armand M Makowski and Adam Shwartz. The Poisson equation for countable Markov chains: Probabilistic methods and interpretations. *Handbook of Markov Decision Processes: Methods and Applications*, pages 269–303, 2002.

[66] P. Mandl. Estimation and control in Markov chains. *Advances in Applied Probability*, 6(1):40–60, 1974.

[67] Hongzi Mao, Malte Schwarzkopf, Hao He, and Mohammad Alizadeh. Towards safe online reinforcement learning in computer systems. In *NeurIPS Machine Learning for Systems Workshop*, 2019.

[68] Peter Marbach, Atilla Eryilmaz, and Asu Ozdaglar. Asynchronous CSMA policies in multihop wireless networks with primary interference constraints. *IEEE Transactions on Information Theory*, 57(6):3644–3676, 2011.

[69] Antonio Massaro, Francesco De Pellegrini, and Lorenzo Maggi. Optimal trunk-reservation by policy learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 127–135. IEEE, 2019.

[70] Akshay Mete, Rahul Singh, Xi Liu, and PR Kumar. Reward biased maximum likelihood estimation for reinforcement learning. In *Learning for Dynamics and Control*, pages 815–827. PMLR, 2021.

[71] Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[72] Pinhas Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

[73] Michael J Neely, Scott T Rager, and Thomas F La Porta. Max-Weight learning algorithms for scheduling in unknown environments. *IEEE Transactions on Automatic Control*, 57(5):1179–1191, 2012.

[74] Rui Nian, Jinfeng Liu, and Biao Huang. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886, 2020.

[75] César Ojeda, Kostadin Cvejoski, Bodgan Georgiev, Christian Bauckhage, Jannis Schuecker, and Ramsés J Sánchez. Learning deep generative models for queuing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9214–9222, 2021.

[76] Pedro A Ortega and Daniel A Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[77] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[78] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*. PMLR, 2017.

[79] Reda Ouhamma, Debabrota Basu, and Odalric Maillard. Bilinear exponential family of MDPs: Frequentist regret bound with tractable exploration & planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9336–9344, 2023.

[80] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

[81] Marcel Panzer and Benedict Bender. Deep reinforcement learning in production systems: A systematic literature review. *International Journal of Production Research*, 60(13):4316–4341, 2022.

[82] Athanasios S Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.

[83] Benjamin Rolf, Ilya Jackson, Marcel Müller, Sebastian Lang, Tobias Reggelin, and Dmitry Ivanov. A review on reinforcement learning algorithms and applications in supply chain management. *International Journal of Production Research*, 61(20):7151–7179, 2023.

[84] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[85] Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. *arXiv preprint arXiv:2006.04353*, 2020.

[86] Albert N. Shiryaev. *Probability*. Springer, 1996.

[87] Rayadurgam Srikant and Lei Ying. *Communication networks: An optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[88] Thomas Stahlbuhk, Brooke Shrader, and Eytan Modiano. Learning algorithms for minimizing queue length regret. *IEEE Transactions on Information Theory*, 67(3):1759–1781, 2021.

[89] Shaler Stidham. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30(8):705–713, 1985.

[90] Shaler Stidham and Richard Weber. A survey of Markov decision models for control of networks of queues. *Queueing systems*, 13:291–314, 1993.

[91] Malcolm Strens. A Bayesian framework for reinforcement learning. *ICML*, pages 943–950, 2000.

[92] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[93] Wojciech Szpankowski and Vernon Rego. Yet another application of a binomial recurrence order statistics. *Computing*, 43(4):401–410, 1990.

[94] Dengwang Tang, Rahul Jain, Botao Hao, and Zheng Wen. Efficient online learning with offline datasets for infinite horizon mdps: A Bayesian approach. *arXiv preprint arXiv:2310.11531*, 2023.

[95] Ming Tang and Vincent WS Wong. Deep reinforcement learning for task offloading in mobile edge computing systems. *IEEE Transactions on Mobile Computing*, 21(6):1985–1997, 2020.

[96] Leandros Tassiulas and Anthony Ephremides. Jointly optimal routing and scheduling in packet ratio networks. *IEEE Transactions on Information Theory*, 38(1):165–168, 1992.

[97] Leandros Tassiulas and Anthony Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39(2):466–478, 1993.

[98] Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.

[99] Georgios Theocharous, Zheng Wen, Yasin Abbasi Yadkori, and Nikos Vlassis. Scalar posterior sampling with applications. *Advances in Neural Information Processing Systems*, 31, 2018.

[100] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[101] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[102] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[103] Neil Walton and Kuang Xu. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 161–198. INFORMS, 2021.

[104] Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6202–6209, 2020.

[105] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[106] Zixian Yang, R Srikant, and Lei Ying. Learning while scheduling in multi-server systems with unknown statistics: Max-Weight with discounted UCB. In *International Conference on Artificial Intelligence and Statistics*, pages 4275–4312. PMLR, 2023.

[107] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[108] Yili Zhang, Asaf Cohen, and Vijay G Subramanian. Learning-based optimal admission control in a single server queuing system. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–2. IEEE, 2022.

[109] Yueyang Zhong, John R Birge, and Amy Ward. Learning the scheduling policy in time-varying multiclass many server queues with abandonment. *Available at SSRN*, 2022.

[110] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.