Reinforcement Learning in Network Control

by

Bai Liu

B.Eng, Tsinghua University (2017)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature redacted

Author .....
Department of Aeronautics and Astronautics
May 14, 2019

Signature redacted

Certified by .....
Eytan Modiano
Professor, Department of Aeronautics and Astronautics
Thesis Supervisor

Signature redacted

Accepted by .....
Sertac Karaman
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Reinforcement Learning in Network Control

by

Bai Liu

Submitted to the Department of Aeronautics and Astronautics
on May 14, 2019, in partial fulfillment of the
requirements for the degree of
Master of Science in Aeronautics and Astronautics

Abstract

With the rapid growth of information technology, network systems have become increasingly complex. In particular, designing network control policies requires knowledge of underlying network dynamics, which are often unknown, and need to be learned.

Existing reinforcement learning methods such as Q-Learning, Actor-Critic, etc. are heuristic and do not offer performance guarantees. In contrast, model-based learning methods offer performance guarantees, but can only be applied with bounded state spaces.

In the thesis, we propose to use model-based reinforcement learning. By applying Lyapunov analysis, our algorithm can be applied to queueing networks with unbounded state spaces. We prove that under our algorithm, the average queue backlog can get arbitrarily close to the optimal result. We also implement simulations to illustrate the effectiveness of our algorithm.

Thesis Supervisor: Eytan Modiano
Title: Professor, Department of Aeronautics and Astronautics

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor, Professor Eytan Modiano. When I joined MIT two years ago, I was completely a novice in scientific research. Under Eytan's mentoring, I have learned how to find important problems of the field, how to model and solve the problems, and how to do scientific writings elegantly. This work would not be possible without his guidance.

I would also like to extend my deepest appreciation to Professor Qiaomin Xie. Applying reinforcement learning techiques to network control problems is a new interdisciplinary topic and I have encountered great difficulty during the research process. Qiaomin is an expert in both fields and has enlightened me a lot. She also helps to proofread my thesis and her patience can nver be underestimatesd.

I am also thankful to my labmates in CNRG. I really enjoyed my discussions with Xinzhe Fu, from which I got enlightened a lot. I learned a lot from my weekly meetings with Qingkai Liang and Thomas Stahlbuhk. Especially, Qingkai was the pioneer in our group to explore reinforcement learning, who motivated me to work on this field. I also enjoyed chatting with Igor Kadota and Jianan Zhang on various topics ranging from research to life. My thanks also goes to other group members who help me in various ways: Vishrant Tripathi, Rajat Talak, Xinyu Wu, Georgia (Gina) Dimaki, Ertem Nusret Tas, Anurag Rai, Ruihao Zhu and Hyang-Won Lee.

I want to express my special thanks to Shaoxiong Wang, who is my roommate and my very best friend. There were multiple times that I encountered bottlenecks in research and his encouragement really helped a lot. I also want to thank my friends with whom I have had many unforgettable moments: Fengyi Li, Tianyi Peng, Hanshen Xiao, Yilun Zhou, Lei Xu, Zehao Yu, Xiaoyue Gong, Pengxiang Zhang, Ge Zhang, Yiliang Li etc.

I am especially grateful to my father, Lelin Liu, who nurtured me from infant. He put tremendous efforts on my education and shaped me who I am today. I cannot imagine how difficult the process was. He is always my greatest support and I really owe him a lot.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

With the rapid growth of information technology, the network systems have become increasingly complex, making it harder to obtain explicit knowledge of system dynamics. For instance, due to security or economic concerns, a number of network systems are built as overlay networks, e.g. caching overlays, routing overlays and security overlays [23]. In these cases, only the overlay part is fully controllable by the network administrator, while the underlay part remains uncontrollable and/or unobservable. The "black box" components make network control policy design challenging.

In addition to the challenges brought by unknown system dynamics, many of the current network control algorithms (e.g. MaxWeight [25] and Drift-plus-Penalty [18]) aim at stabilizing the system, instead of optimizing performances metrics such as queueing backlog or delay.

To overcome above challenges, it is desirable to apply inference and learning schemes. A natural solution is reinforcement learning, which optimizes the decision policy by repeatedly interacting with the environment and estimating the unknown dynamics from the received feedbacks. Reinforcement learning methods provide a framework that enables the design of learning policies for general networks. Reinforcement learning methods can be roughly divided into two types: model-free reinforcement learning (e.g. Q-learning [28], policy gradient [24]) and model-based

reinforcement learning (e.g. UCRL [13], PSRL [20]). Since model-based reinforcement learning methods offer explicit performance guarantees, we focus on model-based reinforcement learning framework in this work.

Almost all existing model-based reinforcement learning methods only work for finite-state-space systems. However, network systems are usually modeled to have unbounded buffer sizes. Therefore, we aim at designing a model-based reinforcement learning method, which is capable for optimizing countable-state MDPs with unknown dynamics.

1.2 Problem Formulation

We target at optimizing the average queue backlog of a general discrete-time queueing network system with possibly unknown dynamics.

The system consists of a set of nodes and links. Each node maintains one or more queues for the undelivered packets, and each queue has unbounded buffer size. The system may have arbitrary topology and operation scheme, and these dynamics can be partially or fully unknown to us.

To fit the problem into stochastic process framework, we only consider the discrete-time network systems with time-invariant stochastic schemes, i.e. under a fixed stochastic control policy, the increment/decrement of each queue backlog has i.i.d. distribution over time.

1.3 Related Works

1.3.1 Stochastic Network Optimization

MaxWeight algorithm is a widely-applied network control policy proposed by [25]. It can be applied to general multi-server networks with arbitrary topology and the servers can be interdependent. MaxWeight algorithm has been proved to be throughput-optimal (i.e. can stabilize the system whenever the system is stabilizable). Moreover, MaxWeight algorithm does not require explicit system dynamics but only the current

queue backlog, which enables it to be applied to complex systems. Extended from MaxWeight, the work in [18] considers the metric of faireness (i.e. to what extent can all traffic gets served). The authors introduced Drift-plus-Penalty algorithm and showed that the optimum regarding fairness can be approached arbitrarily with a trade-off on end-to-end delays.

Both MaxWeight algorithm and Drift-plus-Penalty algorithm work well for general network systems and have throughput performance guarentees. Yet our work goes beyond stabilizing queue backlog to optimize the queue backlog.

1.3.2 Overlay Network

To design control policies for overlay networks, an intuitive solution is: firstly estimating the parameters of the underlay components, then applying classic network control techniques based on the estimated dynamics. A number of different learning methods have been applied.

A popular method is probing, i.e. sending probe packets at a certain time intervals and collect tunnel information. For instance, the works in [10, 16] gathers direct and indirect path information by collecting traceroute and ping data. In [21], simulation results illustrate that the probing approach could achieve optimal throughput.

With the rapid development of machine learning techiques, reinforcement learning has become increasingly popular. In [22], the authors apply Q-learning algorithm in overlay non-cooperative multi-agent wireless sensor networks (WSNs) to achieve optimal mutual response between two agents. The work in [7] applies neural network in reinforcement learning and improves scalability compared with probe-based inference methods.

The probing methods usually work in the ad-hoc manner for different problem settings, while reinforcement learning methods applied to network control so far usually lack rigorous performance guaranteed. Our algorithm overcomes both issues: it works for general network with rigorous performance guarantees.

1.3.3 Model-Based Reinforcement Learning

We consider the model-based reinforcement learning methods that are developed from multi-armed bandit problems, since these algorithms tend to be more tractable in analysis.

UCRL (Upper Confidence Reinforcement Learning) is proposed in [13]. UCRL offers a mathematically rigorous reinforcement learning method that is able to solve Markov decision process with unknown parameters (e.g. transition probability, reward function). UCRL works in an episodic manner: at the beginning of each episode, we first estimate the parameters (e.g. simply using sample mean of history data) and calculate a confidence bound. We then construct a set that consists of all the MDPs whose parameters fall into the confidence bound. Finally, we select the most optimistic MDP (i.e. the one with the minimum average cost) and apply the optimistic solution during this episode. When the current episode meets the termination creteria, start the next episode and repeat the same procedure. Since the true MDP is inside the confidence set with high probablity, and the confidence interval decays with the learning progress, we asymptotically learn the true optimal policy. The work in [19] extends UCRL to continuous state sapce using Hölder continuity assumption.

PSRL (Posterior Sampling for Reinforcement Learning), proposed in [20], shares a similar scheme with UCRL. PSRL maintains a posterior (conditioned on the history data) distribution of parameters. At the beginning of each episode, instead of selecting the most optimistic MDP, PSRL now only samples an MDP from the maintained posterior distribution. PSRL harvests similar performance as UCRL, yet requires less computation.

However, nowadays, the buffer sizes of practical network systems tend to be large or even unbounded, for which it is hard to directly apply the original UCRL and PSRL due to heavy computation. A modified PSRL algorithm is proposed in [27], which can deal with MDPs with large state space. It requires the MDP to have finite bias span, which is unrealistic for the MDP problems with unbounded cost functions. Yet in our case, the cost function is just the queue backlog, which might grow to

infinity.

Our algorithm is inspired by the model-based reinforcement learning methods, yet we propose a new approach which can help us deal with large scale (or even countably infinite) network systems.

1.4 Our Contributions

There exist a number of works on network control that aim at stabilizing the queue backlog, yet the works on minimizing queue metrics (e.g. queue backlog, delay) remain insufficient. Our algorithm goes beyond stability and targets at optimality.

Even among the existing works on queue backlog optimization, most of them propose ad-hoc solutions for some specific scenarios. Our approach is applicable to a broad range of network problems (e.g. scheduling, routing) and does not require explict knowledge on the operation scheme.

Moreover, for MDP optimization problems with average cost criterion (in contrast to discounted cost criterion), almost none of existing methods are applicable to countably infinite state MDP. Specifically, the classical model-based reinforcement learning method (UCRL and PSRL) can only solve finite state MDPs. Our algorithm utilizes drift analysis tools and is able to solve countably infinite state MDPs.

1.5 Thesis Outline

Chapter 2 illustrates the mathematical model of the studied network system under MDP framework, followed by the required assumptions and discussions on them. Chapter 3 presents the proposed algorithm, performance analysis and simulation results. Final conclusions are given in Chapter 4.

Chapter 2

Model

In this chapter, we formulate the mathematical models for the targeted queueing systems. In Section 2.1, we model the system as a countable-state MDP (Markov decision process). However, directly solving the countable-state MDP is ususally infeasible. Therefore, we construct a corresponding truncated finite-state MDP to approximate the true MDP in Section 2.2. In Section 2.3 we state and discuss the assumptions.

2.1 Countable-State Markov Decision Process

As stated in Section 1.2, we consider a discrete-time network system with time-invariant stochastic schemes and aim at minimizing the average queue backlog with unbounded buffer sizes. The problem is especially suitable to be modeled as an MDP, with queue backlog vectors as states and the long-term average queue backlog as the objective function.

More specifically, the MDP M is modeled as follows:

- State space $\mathcal{S}$

  We denote the number of queues as D. We define the set of D-dimensional queue backlog vectors $\boldsymbol{Q}$ as the state space, i.e. $\mathcal{S} = \underbrace{\mathbb{N} \times \cdots \times \mathbb{N}}_{D}$.

- Action space $\mathcal{A}$

The exact form of action space depends on the problem setting. For instance, in server allocation problem where D parallel queues compete for the service of a single server [26], the action is the queue served by the server at each time slot and the action space is naturally the set of queue indexes. We define the action space as $\mathcal{A}$ and assume that $|\mathcal{A}| < \infty$.

- State-transition function p (*can be unknown to us*)

  We define that, when taking action a at state $\boldsymbol{Q}$, the probability of transiting to state $\boldsymbol{Q}'$ as $p\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right)$.

  We assume that the number of new arrived and served packets during each time slot are both bounded. Therefore, for every $\boldsymbol{Q}(t)$, there exists a constant W such that

$$\|\boldsymbol{Q}(t+1) - \boldsymbol{Q}(t)\|_\infty \leqslant W.$$

  We also define the set of states within the one-step reachable region of $\boldsymbol{Q}$ as

$$\mathcal{R}\left(\boldsymbol{Q}, a\right) \triangleq \left\{\boldsymbol{Q}' \in \mathcal{S} : p\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right) > 0\right\},$$

  and $R = \max_{\boldsymbol{Q}\in\mathcal{S}, a\in\mathcal{A}} \left|\mathcal{R}\left(\boldsymbol{Q}, a\right)\right|$.

- Cost function $c\left(\boldsymbol{Q}\right)$

  Since we aim at minimizing the average queue backlog, we define the cost function as $c\left(\boldsymbol{Q}\right) = \sum_i Q_i$. We denote the optimal average queue backlog as ρ^*, and the corresponding optimal policy as π^*.

2.2 Truncated Markov Decision Process

Model-based reinforcement learning techiques usually operate in episodic manner: for each episode the system dynamics are estimated and an approximated optimal policy is obtained based on the learned dynamics. However, there is no effective solution for general countable-state MDPs with optimal average cost (in contrast to discounted

cost) criteria. Therefore, we introduce truncation scheme to our algorithm.

We imagine a truncated queueing system with threshold U: the system has exact dynamics as the real one, with the only difference that each queue has buffer size U. In the truncated system, for each queue, when the queue backlog reaches U, new packets to the queue will get dropped.

The truncated queueing system can be modeled as a finite-state MDP $\tilde{M}$ with state space of $\tilde{\mathcal{S}} \triangleq \{0, 1, \cdots, U\}^D$. $\tilde{M}$ shares the same action space $\mathcal{A}$ and cost function $c(\boldsymbol{Q})$ as M. For this case, we denote the optimal average queue backlog in $\tilde{M}$ as $\tilde{\rho}^*$, and the corresponding optimal policy as $\tilde{\pi}^*$.

The state-transition function $\tilde{p}$ needs to be modified. For simplicity, we first define a mapping $TR(\cdot) : \mathcal{S} \to \tilde{\mathcal{S}}$ that describes the packet dropping scheme in the truncated system:

$$\tilde{\boldsymbol{Q}} = TR(\boldsymbol{Q}) \triangleq \left\{\min\{U, Q_i\}\right\}_{i=1}^{D}.$$

By the definition of the truncated system, for each $\boldsymbol{Q}' \in \tilde{\mathcal{S}}$, $\tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right)$ is, when action a is taken at state $\boldsymbol{Q}$, the probability to transfer to states $\boldsymbol{Q}'' \in \mathcal{S}$ such that $TR(\boldsymbol{Q}'') = \boldsymbol{Q}'$. Specifically, we define that

$$\mathcal{S}\left(\tilde{\boldsymbol{Q}}\right) \triangleq \left\{\boldsymbol{Q} \in \mathcal{S} : TR(\boldsymbol{Q}) = \tilde{\boldsymbol{Q}}\right\}.$$

Then for each $\boldsymbol{Q}' \in \tilde{\mathcal{S}}$ and any $a \in \mathcal{A}$, the state-transition function $\tilde{p}$ can be expressed as

$$\tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right) = \sum_{\boldsymbol{Q}'' \in \mathcal{S}(\boldsymbol{Q}')} p\left(\boldsymbol{Q}'' \mid \boldsymbol{Q}, a\right).$$

2.3 Preliminaries

In this sectiom, we introduce the required assumptions for performance analysis. We further illustrate that our assumptions are natural under the queueing network settings.

2.3.1 Existence of a Known Stabilizing Policy

We first need to control the performance degradation brought by the unboundedness of the state space.

As introduced in Section 1.3.1, a large number of queueing systems can be stablized by stochastic control policies (e.g. MaxWeight) that does not require the knowledge of system dynamics. Stabilizing policies usually have negative Lyapunov drifts. By applying Theorem 3 in [3], we can upper bound the probability for queue backlog to grow large.

Therefore, we define that $Q_{max} \triangleq \max_i Q_i$ and make a natural assumption as follows to control the unboundedness of the state space.

Assumption 1. *There exists a* known *policy* π_0*, a Lyapunov function* $\Phi_0(\boldsymbol{Q}) \leqslant aQ_{max}^{\alpha}$ *with* $a, \alpha > 0$ *and* $\epsilon_0, B_0 > 0$*, such that for any* $\boldsymbol{Q}(t) \in \mathcal{S}$*, when* $Q_{max}(t) \geqslant B_0$*, we have*

$$\mathbb{E}_{\pi_0}\left[\Phi_0\left(\boldsymbol{Q}(t+1)\right) - \Phi_0\left(\boldsymbol{Q}(t)\right) \mid \boldsymbol{Q}(t)\right] \leqslant -\epsilon_0.$$

A broad class of queueing systems have been proven to have π_0 as Assumption 1. For instance, stabilizing policies are proposed for dynamic server allocation problem [1, 8, 26], multiclass routing network [5, 12, 15, 14, 4], inventory control [17, 9] etc., all with linear or quadratic forms of Lyapunov functions.

2.3.2 Lyapunov Drift Under the Optimal Policy

Denote the optimal policy of the truncated system as $\tilde{\pi}^*$. We further assume that $\tilde{\pi}^*$ has negative drift under sub-quadric Lyapunov function.

Assumption 2. *For any* $U > 0$*, under* $\tilde{\pi}^*$ *there exists a Lyapunov function* $b_1 Q_{max}^{\beta} \leqslant \tilde{\Phi}^*(\boldsymbol{Q}) \leqslant b_2 Q_{max}^{\beta}$ *with* $b_1, b_2 > 0$ *and* $0 < \beta < 2$ *and* $\tilde{\epsilon}^*, \tilde{B}^* > 0$*, such that for any* $\boldsymbol{Q}(t) \in \tilde{\mathcal{S}}$*, when* $Q_{max}(t) \geqslant \tilde{B}^* > 0$*, we have*

$$\mathbb{E}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*\left(\boldsymbol{Q}(t+1)\right) - \tilde{\Phi}^*\left(\boldsymbol{Q}(t)\right) \mid \boldsymbol{Q}(t)\right] \leqslant -\tilde{\epsilon}^*.$$

We further assume that there esists $b_3 > 0$, such that for any $\boldsymbol{Q}(t) \in \tilde{\mathcal{S}}$,

$$\tilde{\Phi}^*\left(\boldsymbol{Q}(t+1)\right) - \tilde{\Phi}^*\left(\boldsymbol{Q}(t)\right) \leqslant b_3 U^{\max\{\beta-1,0\}}.$$

2.3.3 First Hitting Time Under the Optimal Policy

Due to mathematical requirements, it is required for us to impose some restrictions over the communication properties on the truncated MDP.

Assumption 3. *In the truncated system $\tilde{M}$, there exists $c > 0$, such that for any $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathcal{S}^{in}$, we have*

$$\min_{\tilde{\pi}} \mathbb{E}\left[T^{\tilde{\pi}}_{\boldsymbol{Q}\to\boldsymbol{Q}'}\right] \leqslant c\|\boldsymbol{Q}' - \boldsymbol{Q}\|_1^{\gamma},$$

where $\tilde{\pi}$ is the policy applied to $\tilde{S}$.

2.3.4 Error Tolerence for MDP Estimation

As the learning process proceeds, the estimation for $\tilde{M}$ becomes increasingly accurate. However, the parameters to estimate are real numbers, and it is impossible for us to obtain the exact $\tilde{M}$ (due to the density of real numbers).

To simplify our analysis, we make the assumption that if we estimate the state-transition function accurate enough (i.e. within a certain error bound), the solution to the estimated MDP is the same as $\tilde{\pi}^*$. The assumption is as follows.

Assumption 4. *There exists a $\Delta p > 0$, such that for any finite-state MDP M' with the same state space, action space and cost function as $\tilde{M}$, if the condition that*

$$\left\|\tilde{p}\left(\cdot \mid \boldsymbol{Q}, a\right) - p'\left(\cdot \mid \boldsymbol{Q}, a\right)\right\|_1 \leqslant \Delta p,$$

holds for each $(\boldsymbol{Q}, a)$, then the optimal policy to M' is also $\tilde{\pi}^*$.

Notice that in most queueing networks, when system dynamics (e.g. exogenous arrival rates, service rates, channel capacities) varies slightly, the optimal policy remains the same. Therefore, the assumption is reasonable for queueing systems.

Chapter 3

Main Results

In this chapter, we present our algorithm and performance analysis, which are the main results of our work.

We illustrate our algorithm in Section 3.1. In Section 3.2, We present our performance results from both exploration and exploitation perspectives. Section 3.2.1 serves as the exploration part, in which we discussed the number of episodes it takes for our algorithm to obtain $\tilde{\pi}^*$. While in Section 3.2.2, we turn to exploitation performance, showing that as learning process proceeds, PDGRL utilizes the learned (sub-)optimal policies increasingly frequently and harvests the average queue backlog close to the optimal one. Proofs are given in Section 3.4.

3.1 Algorithm

We propose PDGRL (Piecewise Decaying ϵ-Greedy Reinforcement Learning) alsorithm. For simplicity, we partition $\mathcal{S}$ into $\mathcal{S}^{in} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : Q_{max} \leqslant U\}$ and $\mathcal{S}^{out} \triangleq \mathcal{S} \setminus \mathcal{S}^{in}$.

PDGRL operates in episodic manner: at the beginning of episode k, we uniformly draw a real number $\xi \in [0, 1]$. If $\xi \leqslant l/\sqrt{k}$ ($\triangleq \epsilon_k$) (where $0 < l \leqslant 1$), we do exploration during the episode by applying purely random policy π_{rand} (i.e. selecting actions uniformly) for states in $\mathcal{S}^{in}$, while still apply π_0 for the rest states. If $\xi > \epsilon_k$, we enter exploitation stage: we first estimate the parameters of $\tilde{M}$ using sample means, then solve the estimated system and obtain a sub-optimal policy $\tilde{\pi}_k$. For the

rest of the episode, we apply $\tilde{\pi}_k$ for states in $\mathcal{S}^{in}$ and π_0 otherwise. When visits to states in $\mathcal{S}^{in}$ exceed $L_k = L \cdot \sqrt{k}$ (where $L > 0$), PDGRL enters episode $k+1$ and repeat the process above.

The detailed algorithm is as Algorithm 1.

Algorithm 1 The PDGRL algorithm

Input: $\mathcal{A}$, $U > 2W + (B'/b_1)^{1/\beta}$, $l > 0$, $L > 0$
Initialization: $t \leftarrow 1$, $N(\cdot,\cdot) \leftarrow 0$, $\tilde{P}(\cdot,\cdot) \leftarrow 0$
for episodes $k \leftarrow 1, 2, \cdots$ **do**
  Set $L_k \leftarrow L \cdot \sqrt{k}$, $\epsilon_k \leftarrow l/\sqrt{k}$ and uniformly draw $\xi \in [0,1]$.
  if $\xi \leqslant \epsilon_k$ **then**
    $\pi_k^{in} \leftarrow \pi_{rand}$.
  else
    For each $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$, estimate that $\tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right) = \tilde{P}\left(\boldsymbol{Q}, a, \boldsymbol{Q}'\right)/N\left(\boldsymbol{Q}, a\right)$ for $N\left(\boldsymbol{Q}, a\right) > 0$ and $\tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, a\right) = 1/\left|\mathcal{R}\left(\boldsymbol{Q}, a\right)\right|$ otherwise.
    Solve the estimated MDP $\tilde{M}_k$ and obtain the estimated optimal policy $\tilde{\pi}_k$.
    $\pi_k^{in} \leftarrow \tilde{\pi}_k$.
  end if
  while visits to states in $\mathcal{S}^{in}$ is smaller that L_k **do**
    Take $a_t = \pi_k^{in}\left(\boldsymbol{Q}(t)\right)$ for $\boldsymbol{Q}(t) \in \mathcal{S}^{in}$ and $a_t = \pi_0\left(\boldsymbol{Q}(t)\right)$ for $\boldsymbol{Q}(t) \in \mathcal{S}^{out}$.
    Implement a_t to the real system and observe the next state $\boldsymbol{Q}(t+1)$.
    if $\boldsymbol{Q}(t) \in \mathcal{S}^{in}$ **then**
      $N\left(\boldsymbol{Q}(t), a_t\right) \leftarrow N\left(\boldsymbol{Q}(t), a_t\right) + 1$.
      $\tilde{P}\left(\boldsymbol{Q}(t), a_t, TR\left(\boldsymbol{Q}(t+1)\right)\right) \leftarrow \tilde{P}\left(\boldsymbol{Q}(t), a_t, TR\left(\boldsymbol{Q}(t+1)\right)\right) + 1$.
    end if
    $t \leftarrow t+1$.
  end while
end for
Output: estimated optimal policy $\tilde{\pi}_k$

3.2 Performance Analysis

We illustrate the performance of our algorithm from both exploration and exploitation perspectives. We first prove that PDGRL can learn $\tilde{\pi}^*$ with arbitrarily high probability, which illutrates that PDGRL explores different states sufficiently to obtain an accurate estimation of $\tilde{M}$. We then show that PDGRL exploits the estimated optimal policy and has a tight gap to the true optimal result ρ^*.

3.2.1 Convergence to the Optimal Policy (Exploration)

We define $p^{\pi+\pi'}(\boldsymbol{Q})$ as the stationary probability of $\boldsymbol{Q}$ under the policy that applies π to states in $\mathcal{S}^{in}$ and π' to states in $\mathcal{S}^{out}$.

The following theorem shows that, with arbitrarily high probability, PDGRL learns $\tilde{\pi}^*$ within finite number of episodes (see Section 3.4.1 for the proof).

Theorem 1. *For any* $0 < \delta < 1$, PDGRL *learns* $\tilde{\pi}^*$ *within* $k^* < \infty$ *episodes with probability at least* $1 - \delta$. *Specifically,* k^* *is upper bounded as*

$$k^* \leqslant \frac{2}{\delta}\left(K_0 + J^* + \frac{\pi^2}{6} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0 + J^* + 1)^{2J^*+2}}{J^*!}\right),$$

where

$$J^* = \left\lceil \frac{4|\mathcal{A}| \log \frac{2^{r+1}U^D|\mathcal{A}|}{\delta}}{L\sqrt{K_0}(\Delta p)^2 \cdot \min_{\boldsymbol{Q}\in\mathcal{S}^{in}} p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} \right\rceil,$$

and K_0 *is a constant.*

Note that Theorem 1 only provides a loose upper bound for k^*. By applying a tighter inequality in Eq (A.12), we expect Theorem 1 to have a much tighter upper bound.

3.2.2 Average Queue Backlog (Exploitation)

Theorem 1 indicates that PDGRL explores (i.e. samples) state-transition functions of each $(\boldsymbol{Q}, a)$ in $\tilde{M}$ sufficiently. The following theorem shows that PDGRL makes a balanced trade-off between exploration and exploitation (see Section 3.4.2 for the proof). We define t_k as the starting time of the k^{th} episode.

Theorem 2. *Applying* PDGRL *to* M, *the expected average queue backlog is upper bounded as*

$$\lim_{K\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^{t_K}\sum_i Q_i(t)\right]}{t_K} = \tilde{\rho}^* + \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right).$$

Theorem 2 gives us an asymptotically optimal result regarding the threshold parameter U: by increasing U, the long-term average queue backlog approaches $\tilde{\rho}^*$ exponentially fast.

3.3 Numerical Experiments

3.3.1 Problem Setting

We consider a simple server allocation problem: exogenous packets arrive to two nodes according to Bernoulli process with rate λ_1 and λ_2 respectively. Both nodes have unbounded buffers. At each time slot, a central server need to select one of the two queues to serve. The selected queue i is served successfully with probability p_i. Specifically, the system model and parameters are as Figure 3-1.

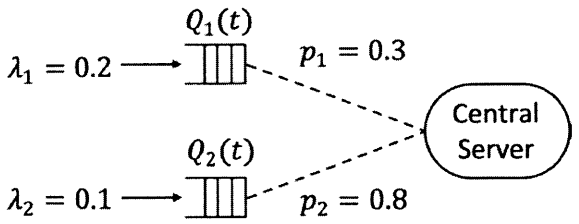

Figure 3-1: System model

According to [26], whenever $\lambda_1/p_1 + \lambda_2/p_2 < 1$, a stabilizing policy is to always serve the node with the longest connected queue (LCQ). Therefore, we can use LCQ policy as π_0. Note that in our setting, the channels are always connected, π_0 is actually serving the node with the longest queue (LQ).

On the other hand, according to $c\mu$-rule in [8], the optimal policy π^* that minimizes the average queue backlog is to select the node with the largest successful transmission rate among all the nonempty queues.

In the model depicted in Figure 3-1, π^* is to serve node 2 whenever it is nonempty. However, since node 1 has larger arrival rate and smaller successful transmission rate, queue in node 1 is easier to get queued up. Therefore, we would expect π_0 to serve node 1 more frequently and there exists a gap to the result under π^*.

3.3.2 Results

When conducting simulation, we compare the performances under four policies: π_0 (LCQ), PDGRL, π^* (true optimal policy) and $\tilde{\pi}^* + \pi_0$ (applying π^* for $\boldsymbol{Q} \in \mathcal{S}^{in}$ and π_0 otherwise). Note that the $\tilde{\pi}^* + \pi_0$ policy is exactly the best policy PDGRL can learn. We simulate it to study the convergence rate of PDGRL.

We first implement the simulation under $U = 5$, and the result is as Figure 3-2.

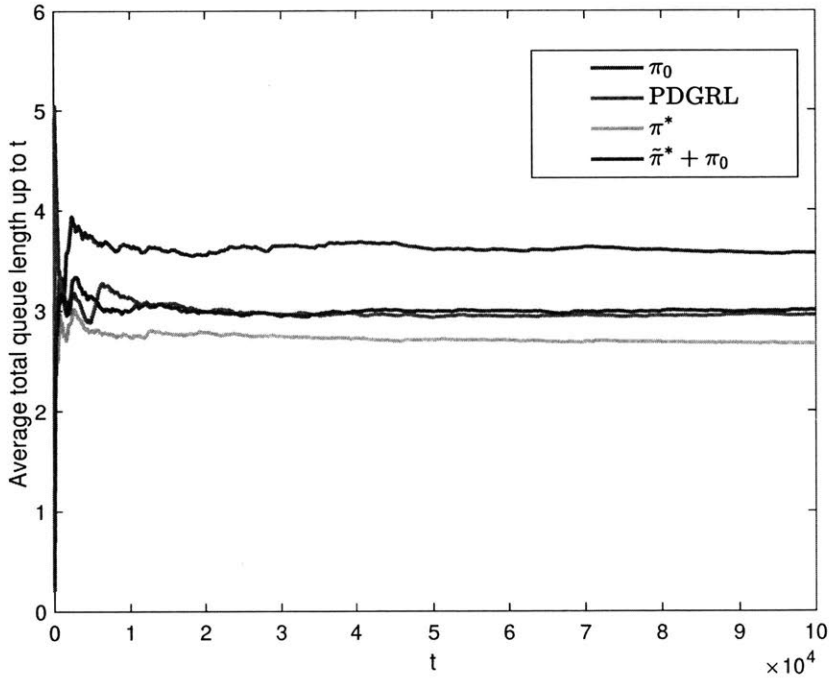

Figure 3-2: Simulation results under $U = 5$

Figure 3-2 shows that PDGRL beats π_0, and quickly converges to $\tilde{\pi}^*+\pi_0$. However, the gap to π^* still exists.

From Theorem 2, we know that when U grows, the average queue backlog of PDGRL approaches the optimal result exponentially fast. We then set $U = 10$ and repeat the simulation.

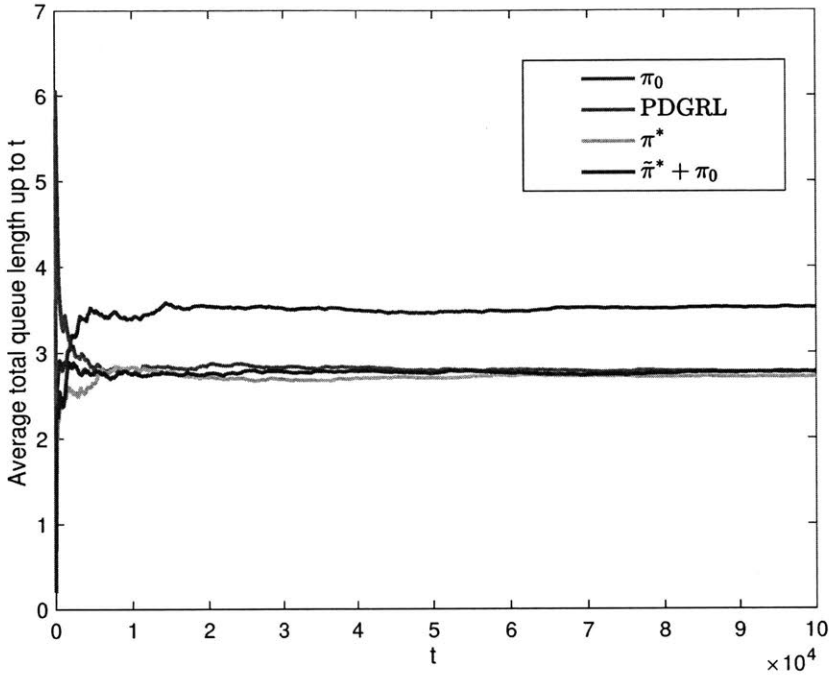

Figure 3-3: Simulation results under $U = 10$

From Figure 3-2, we can see that now PDGRL still converges to $\tilde{\pi}^* + \pi_0$ fast, and the gap between PDGRL and π^* almost diminishes, as indicated by Theorem 2.

3.4 Appendices

3.4.1 Proof of Theorem 1

We denote $N_k(\boldsymbol{Q}, a)$ as the number of times that $(\boldsymbol{Q}, a)$ is selected during episode k. The following lemma illustrate that after after a certain number of episodes, π_{rand} samples every $(\boldsymbol{Q}, a)$ sufficiently with relatively large (e.g. greater than 1/2) probability (see Appendix A.1 for the proof).

Lemma 1. *Under algorithm 1, there exists $K_0 > 0$ such that for any $k \geqslant K_0$,*

$$\Pr\left\{N_k(\boldsymbol{Q}, a) > \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q}) \cdot L_k}{2|\mathcal{A}|} \mid \pi_k^{in} = \pi_{rand}\right\} \geqslant \frac{1}{2},$$

for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$.

We also have the following lemma on the number of samples for each $(\boldsymbol{Q}, a)$ required to estimate $\tilde{M}$ accurate enough (see Appendix A.2 for the proof).

Lemma 2. *If for any $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$*

$$N(\boldsymbol{Q}, a) \geqslant \frac{2}{(\Delta p)^2} \cdot \log \frac{2^{r+1}(U+1)^D|\mathcal{A}|}{\delta},$$

then with probability at least $1 - \delta/2$, the optimal solution of the estimated truncated MDP is exactly $\tilde{\pi}^*$.

Based on Lemma 1 and Lemma 2, we are able to prove that as the learning process proceeds, each $(\boldsymbol{Q}, a)$ will be sampled sufficiently for $\tilde{M}$ to be estimated accurately enough. The following theorem provides an upper bound for the expected number of required episodes k^* (See Appendix A.3 for the proof).

Lemma 3.

$$\mathbb{E}[k^*] \leqslant K_0 + J^* + \frac{\pi^2}{6} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0 + J^* + 1)^{2J^*+2}}{J^*! \cdot l^{2J^*+2}} \triangleq K(J^*),$$

where $J^* = \left\lceil \frac{4|\mathcal{A}| \log \frac{2^{R+1}(U+1)^D|\mathcal{A}|}{\delta}}{L\sqrt{K_0}(\Delta p)^2 \cdot \min_{\boldsymbol{Q} \in \mathcal{S}^{in}} p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} \right\rceil$.

Using Lemma 3 and apply Markov's inequality, we have a probablistic upper bound for k^*:

$$k^* \leqslant \frac{2\mathbb{E}[k^*]}{\delta} \leqslant \frac{2K(J^*)}{\delta}, \tag{3.1}$$

with probability at least $1 - \delta/2$.

By taking a union bound over the events of Lemma 2 and Eq (3.1), we have that with probability at least $1 - \delta$,

$$k^* \leqslant \frac{2}{\delta}\left(K_0 + J^* + \frac{\pi^2}{6} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0 + J^* + 1)^{2J^*+2}}{J^*!}\right),$$

which completes the proof of Theorem 1.

3.4.2 Proof of Theorem 2

From the design of PDGRL and Theorem 1, we observe that after a certain number of episodes, PDGRL selects $\tilde{\pi}$ as π_k^{in} with high probability. In the following lemma, we provide an upper bound for the expected average queue backlog under the piecewise policy that applies $\tilde{\pi}^*$ to states inside $\mathcal{S}^{in}$ and π_0 to states outside $\mathcal{S}^{out}$ (since Lemma 4 plays core role in analyzing the performance, we place the proof in Section 3.4.3). Define the beginning of episode k as t_k and the length of episode k as L'_k, the following lemma holds.

Lemma 4. *The expected episodic backlog conditioned on $\pi_k^{in}(\cdot) = \tilde{\pi}^*(\cdot)$ is upper bounded as follows.*

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_k}^{t_k+L'_k-1}\sum_i Q_i(t)}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right] = \tilde{\rho}^* + \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right).$$

To analyze the overall expected queue backlog, we need to further consider two possible cases: we may never learn $\tilde{\pi}^*$, and even if we have successfully learned $\tilde{\pi}^*$, π_{rand} may be selected as π_k^{in} with small but positive probability. The following lemma provides an upper bound for the overall expected queue backlog (See Appendix A.4 for the proof).

Lemma 5. *Under* PDGRL, *the overall expected queue backlog is upper bounded as follows.*

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{\sum_{t=1}^{t_K}\sum_i Q_i(t)}{t_K}\right] = \tilde{\rho}^* + \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)} + \delta U^{1+2\alpha}\right),$$

where L'_k is the actual episode length of episode k, i.e. L_k plus the time spent in $\mathcal{S}^{out}$.

By taking $\delta = U^{-2\alpha-1}\cdot\exp\left(-U^{\min\{\beta,2-\beta\}}\right)$, we have an upper bound for the overall expected queue backlog as follows.

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{\sum_{t=1}^{t_K}\sum_i Q_i(t)}{t_K}\right] = \tilde{\rho}^* + \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right),$$

which completes the proof.

3.4.3 Proof of Lemma 4

For simplicity, we partition $\mathcal{S}$ as follows.

$$\begin{cases} \mathcal{Z}_{in}^{in} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : Q_{max} \leqslant U - 2W\} \\ \mathcal{Z}_{bd}^{in} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : U - 2W + 1 \leqslant Q_{max} \leqslant U - W\} \\ \mathcal{Z}_{bd}^{out} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : U - W + 1 \leqslant Q_{max} \leqslant U\} \\ \mathcal{Z}_{out}^{out} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : Q_{max} \geqslant U + 1\} \end{cases}$$

We further define that $\mathcal{Z}^{in} = \mathcal{Z}_{in}^{in} \cup \mathcal{Z}_{bd}^{in}$ and $\mathcal{Z}^{out} = \mathcal{Z}_{bd}^{out} \cup \mathcal{Z}_{out}^{out}$.

We define the regret of episode k as $\sum_{t=t_k}^{t_k+L'_k-1} \left(\sum_i Q_i(t) - \tilde{\rho}^*\right)$. We also define $\mathcal{T}_k^{in}$ and $\mathcal{T}_k^{out}$ as the set of time slots that $\boldsymbol{Q}(t)$ is in $\mathcal{Z}^{in}$ and $\mathcal{Z}^{out}$ during episode k. We then can decompose average episodic regret as follows.

$$\begin{aligned} &\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right] \\ = &\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{in}}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right] + \qquad (3.2) \\ &\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{out}}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right]. \qquad (3.3) \end{aligned}$$

We obtain an upper bound for Eq (3.2) in the following lemma (see Appendix A.5 for the proof).

Lemma 6.

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{in}}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right] = p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right) \cdot \mathcal{O}\left(U^{D+\gamma}\right).$$

For Eq (3.3), we also obtain an upper bound in the following lemma (see Appendix

A.6 for the proof).

Lemma 7.

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'} \mid \pi_k^{in}=\tilde{\pi}^*\right] = p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot\mathcal{O}\left(U^{D+2\alpha}\right).$$

We further propose the following lemma to upper bound $p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right)$ and $p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)$ (see Appendix A.7 for the proof).

Lemma 8.

$$p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right)+p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)=\mathcal{O}\left(\exp\left(-U^{\min\{\beta,2-\beta\}}\right)\right).$$

By combining Lemma 6, Lemma 7 and Lemma 8, we have

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_k}^{t_k+L_k'-1}\sum_i^D Q_i(t)}{L_k'} \mid \pi_k^{in}=\tilde{\pi}^*\right] = \tilde{\rho}^*+\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right),$$

which completes the proof.

Chapter 4

Conclusion

In this work, we apply model-based reinforcement learning framework to general queueing networks with unbounded state space. We propose PDGRL algorithm, which applies ϵ-greedy exploration scheme. We then use Lyapunov analysis and prove that the average queue backlog can get arbitrarily close to the minimal average queue backlog under oracle policy. Numerical experiment results are consistent with our analysis.

Appendix A

Proofs

A.1 Proof of Lemma 1

In the prof, we only discuss the episodes that applies π_{rand}.

Under Assumption 1 and Assumption 3, by applying Foster-Lyapunov theorem, we can show that under the policy that applies π_{rand} to states in $\mathcal{S}^{in}$ and π_0 to states in $\mathcal{S}^{out}$, the corresponding Markov chain is positive recurrent with stationary distribution $p^{\pi_{rand}+\pi_0}$.

Define $N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})$ as the numebr of times that $\boldsymbol{Q}$ is selected during episode k. For an irreducible positive recurrent Markov chain on countable state space, we have the mixing property that for any given

$$\lim_{L'_k \to \infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{L'_k} = p^{\pi_{rand}+\pi_0}(\boldsymbol{Q}) \qquad w.p.1, \tag{A.1}$$

for each $\boldsymbol{Q} \in \mathcal{S}$, where L'_k is the actual length of episode k.

Since $L'_k \geqslant L_k = L \cdot \sqrt{k}$, Eq (A.1) can be further expressed as

$$\lim_{k \to \infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{L'_k} = p^{\pi_{rand}+\pi_0}(\boldsymbol{Q}) \qquad w.p.1. \tag{A.2}$$

Since under $\pi_{rand}+\pi_0$, for each $\boldsymbol{Q} \in \mathcal{S}^{in}$, we take each $a \in \mathcal{A}$ with equal probability

$\frac{1}{|\mathcal{A}|}$, then according to strong law of large number, we have

$$\lim_{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} = \frac{1}{|\mathcal{A}|} \qquad w.p.1, \tag{A.3}$$

for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$.

Also note that according to Eq (A.2), $N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q}) \to \infty$ as $k \to \infty$ for $p^{\pi_{rand}+\pi_0}(\boldsymbol{Q}) > 0$, Eq (A.3) can be further expressed as

$$\lim_{k\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} = \frac{1}{|\mathcal{A}|} \qquad w.p.1. \tag{A.4}$$

Since both Eq (A.2) and Eq (A.4) are almost sure convergence to constants, the multiplication rule for limit holds, i.e.

$$\begin{aligned}
\lim_{k\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L'_k} &= \lim_{k\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{L'_k} \cdot \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} \\
&= \lim_{k\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{L'_k} \cdot \lim_{k\to\infty} \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q})} \\
&= \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{|\mathcal{A}|} \qquad w.p.1,
\end{aligned}$$

for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$.

Note that almost sure convergence indicates convergence in probability, for any $\epsilon > 0$, we have

$$\lim_{k\to\infty} \Pr\left\{ \left| \frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L'_k} - \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{|\mathcal{A}|} \right| \geqslant \epsilon \right\} = 0, \tag{A.5}$$

for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$.

Given a $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$, by taking $\epsilon = \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{2|\mathcal{A}|}$ in Eq (A.5), we have that

there exists $K_0 < \infty$ such that when $k \geqslant K_0$,

$$\begin{aligned}
&\Pr\left\{\frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L_k} \leqslant \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{2|\mathcal{A}|}\right\} \\
\leqslant &\Pr\left\{\frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L_k'} \leqslant \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{2|\mathcal{A}|}\right\} \\
\leqslant &\Pr\left\{\left|\frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L_k'} - \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{|\mathcal{A}|}\right| \geqslant \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{2|\mathcal{A}|}\right\} \\
\leqslant &\frac{1}{2|\mathcal{S}^{in}||\mathcal{A}|},
\end{aligned} \tag{A.6}$$

for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$.

By taking a union bound over $\mathcal{S}^{in}$ and $\mathcal{A}$ in (A.6), we have that when $k \geqslant K_0$, for each $\boldsymbol{Q} \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$,

$$\Pr\left\{\frac{N_k^{\pi_{rand}+\pi_0}(\boldsymbol{Q},a)}{L_k} \leqslant \frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}{2|\mathcal{A}|}\right\} \leqslant \frac{1}{2|\mathcal{S}^{in}||\mathcal{A}|} \cdot \left|\mathcal{S}^{in}\right| \cdot |\mathcal{A}| = \frac{1}{2},$$

which completes the proof.

A.2 Proof of Lemma 2

According to [29], for a probabiliy distribution over n_1 distinct events, the L^1- deviation of the true distribution $\tilde{p}$ and the empirical distribution $\hat{p}$ based on n_2 samples from the true distribution $\tilde{p}$ is upper bounded as

$$\Pr\left\{\left\|\tilde{p}\left(\cdot \mid \boldsymbol{Q},a\right) - \hat{p}\left(\cdot \mid \boldsymbol{Q},a\right)\right\|_1 \geqslant \epsilon\right\} \leqslant (2^{n_1}-2)\exp\left(-\frac{n_2\epsilon^2}{2}\right).$$

By definition, for each $(\boldsymbol{Q},a)$, $n_1 = \left|\mathcal{R}(\boldsymbol{Q},a)\right| \leqslant R$. By taking $n_2 \geqslant \frac{2}{(\Delta p)^2} \cdot$

$\log \frac{2^{R+1}(U+1)^D|\mathcal{A}|}{\delta}$, we have

$$\begin{aligned}
&\Pr\left\{\sum_{\boldsymbol{Q}'\in\mathcal{R}(\boldsymbol{Q},a)}\left|\tilde{p}(\boldsymbol{Q}'\mid\boldsymbol{Q},a)-\hat{p}(\boldsymbol{Q}'\mid\boldsymbol{Q},a)\right|\geqslant\Delta p\right\}\\
\leqslant&(2^R-2)\cdot\exp\left(-\frac{(\Delta p)^2}{2}\cdot\frac{2}{(\Delta p)^2}\cdot\log\frac{2^{R+1}(U+1)^D|\mathcal{A}|}{\delta}\right)\\
\leqslant&\frac{\delta}{2(U+1)^D|\mathcal{A}|}.
\end{aligned}$$

By taking a union bound over each $\boldsymbol{Q}\in\mathcal{S}^{in}$ and $a\in\mathcal{A}$, we have

$$\begin{aligned}
&\Pr\left\{\text{there exists }(\boldsymbol{Q},a)\text{ such that }\left\|\tilde{p}\left(\cdot\mid\boldsymbol{Q},a\right)-\hat{p}\left(\cdot\mid\boldsymbol{Q},a\right)\right\|_1\geqslant\Delta p\right\}\\
\leqslant&\frac{\delta}{2(U+1)^D|\mathcal{A}|}\cdot(U+1)^D\cdot|\mathcal{A}|=\frac{\delta}{2},
\end{aligned}$$

which completes the proof.

A.3 Proof of Lemma 3

We define event

$$B_k\triangleq\left\{\pi_k^{in}=\pi_{rand},N_k(\boldsymbol{Q},a)>\frac{p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})\cdot L_k}{2|\mathcal{A}|},\forall\boldsymbol{Q}\in\mathcal{S}^{in},a\in\mathcal{A}\right\}.$$

From Lemma 1, when $k\geqslant K_0$, at least $p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})\cdot L\cdot\sqrt{K_0}/(2|\mathcal{A}|)$ samples can be obtained for each $(\boldsymbol{Q},a)$ if B_k is true. Therefore, a sufficient condition to obtain J samples for each $(\boldsymbol{Q},a)$ is that B_k occurs for $J^*\triangleq\left\lceil\frac{2|\mathcal{A}|J}{L\sqrt{K_0}\cdot\min_{\boldsymbol{Q}\in\mathcal{S}^{in}}p^{\pi_{rand}+\pi_0}(\boldsymbol{Q})}\right\rceil$ times.

Denote m^* as the number of episodes needed for B_k to occur for J^* times when $k\geqslant K_0$. Then we have

$$\mathbb{E}\left[k^*\right]\leqslant\mathbb{E}\left[m^*\right].$$

For $n \geqslant 1$, we have

$$\begin{aligned}
&\Pr\{m^* \geqslant K_0 + J^* + n\} \\
&= \Pr\{\text{from episode } K_0+1 \text{ to } K_0 + J^* + n,\ \pi_{rand} \text{ is not selected for at least } n \text{ times}\} \\
&= \Pr\left\{ \bigcup_{K_0+1 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0+J^*+n} E_{k_1,k_2,\cdots,k_n} \right\} \\
&\leqslant \sum_{K_0+1 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0+N+n} \Pr\left\{E_{k_1,k_2,\cdots,k_n}\right\}. \qquad \text{(A.7)}
\end{aligned}$$

where $E_{k_1,k_2,\cdots,k_n}$ is defined as the event that during episodes $k_1, k_2, \cdots, k_n$, B_k does NOT occur.

By applying Lemma 1, we have

$$\Pr\{B_k\} \geqslant \frac{1}{2} \cdot \Pr\{\pi_{rand} \text{ is selected at episode } k\} = \frac{l}{2\sqrt{k}}.$$

Therefore, for any $K_0 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0 + N + n$, we have

$$\Pr\left\{E_{k_1,k_2,\cdots,k_n}\right\} \leqslant \prod_{i=1}^{n} \left(1 - \frac{l}{2\sqrt{k_i}}\right) \leqslant \left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right)^n. \qquad \text{(A.8)}$$

By inserting Eq (A.8) into Eq (A.7), we have

$$\begin{aligned}
&\Pr\{m^* \geqslant K_0 + J^* + n\} \\
\leqslant &\binom{J^*+n}{n} \cdot \left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right)^n \\
= &\frac{(J^*+n)\cdot(J^*+n-1)\cdots(n+1)}{J^*!} \cdot \left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right)^n \\
\leqslant &\frac{1}{J^*!} \cdot (J^*+n)^{J^*} \cdot \left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right)^n. \qquad \text{(A.9)}
\end{aligned}$$

Since for $x > 0$, natural logarithm can be upper bounded as

$$\log x \leqslant x - 1.$$

We therefore have

$$n \log \left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right) \leqslant -\frac{nl}{2\sqrt{K_0 + J^* + n}}$$

which indicates

$$\left(1 - \frac{l}{2\sqrt{K_0 + J^* + n}}\right)^n \leqslant \exp\left(-\frac{nl}{2\sqrt{K_0 + J^* + n}}\right) \tag{A.10}$$

By inserting Eq (A.10) into Eq (A.9), we have

$$\Pr\left\{m^* \geqslant K_0 + J^* + n\right\} \leqslant \frac{1}{J^*!} \cdot \frac{(J^* + n)^{J^*}}{\exp\left(\frac{nl}{2\sqrt{K_0 + J^* + n}}\right)}.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[k^*\right] \leqslant & \mathbb{E}\left[m^*\right] \\
= & \sum_{i=1}^{\infty} \Pr\left\{m^* \geqslant i\right\} \\
= & K_0 + J^* + \sum_{n=1}^{\infty} \Pr\left\{m^* \geqslant K_0 + J^* + n\right\} \\
\leqslant & K_0 + J^* + \frac{1}{N!} \sum_{n=1}^{\infty} \frac{(J^* + n)^{J^*}}{\exp\left(\frac{nl}{2\sqrt{K_0 + J^* + n}}\right)}.
\end{aligned} \tag{A.11}$$

Since for $u > 0$, we have

$$\exp(u) = \sum_{k=0}^{\infty} \frac{u^k}{k!} > \frac{u^{2J^*+4}}{(2J^* + 4)!}. \tag{A.12}$$

Therefore, for $n \geqslant 1$

$$\begin{aligned}
\frac{(J^*+n)^{J^*}}{\exp\left(\frac{nl}{2\sqrt{K_0+J^*+n}}\right)} <& (J^*+n)^{J^*} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0+J^*+n)^{J^*+2}}{(nl)^{2J^*+4}} \\
<& (K_0+J^*+n)^{J^*} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0+J^*+n)^{J^*+2}}{(nl)^{2J^*+4}} \\
=& (2J^*+4)! \cdot 4^{J^*+2} \cdot \left(1+\frac{K_0+J^*}{nl}\right)^{2J^*+2} \cdot \frac{1}{n^2} \\
\leqslant& (2J^*+4)! \cdot 4^{J^*+2} \cdot \left(1+\frac{K_0+J^*}{l}\right)^{2J^*+2} \cdot \frac{1}{n^2}. \qquad \text{(A.13)}
\end{aligned}$$

Insert Eq (A.13) into Eq (A.11), we therefore have

$$\begin{aligned}
\mathbb{E}[k^*] \leqslant& K_0 + J^* + \frac{1}{J^*!} \sum_{n=1}^{\infty} \frac{(J^*+n)^{J^*}}{\exp\left(\frac{nl}{2\sqrt{K_0+J^*+n}}\right)} \\
<& K_0 + J^* + \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0+J^*+l)^{2J^*+2}}{J^*! \cdot l^{2J^*+2}} \sum_{n=1}^{\infty} \frac{1}{n^2} \\
=& K_0 + J^* + \frac{\pi^2}{6} \cdot \frac{(2J^*+4)! \cdot 4^{J^*+2} \cdot (K_0+J^*+1)^{2J^*+2}}{J^*! \cdot l^{2J^*+2}},
\end{aligned}$$

which completes the proof.

A.4 Proof of Lemma 5

During episode k, $\boldsymbol{Q}$ can exit into $\mathcal{S}^{out}$ for at most L_k times, and each time the expected regret is uniformly upper bounded by $\mathcal{O}\left(U^{1+2\alpha}\right)$ from Eq (A.49). Therefore, for any policy $\tilde{\pi}$ that is applied to $\mathcal{S}^{in}$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right) \mid \pi_k^{in}=\tilde{\pi}\right] \leqslant & L_k\cdot DU + L_k\cdot\mathcal{O}\left(U^{1+2\alpha}\right)\\ =&\mathcal{O}\left(\sqrt{k}\cdot U^{1+2\alpha}\right).\end{aligned} \tag{A.14}$$

From Lemma 3, we know that there exists a $k^*<\infty$ such that $\pi_k^{in}(\cdot)\neq\tilde{\pi}^*(\cdot)$ when $k\geqslant k^*$ with probability at most $1-(1-\delta)(1-\epsilon_k)\leqslant\delta+\epsilon_k$.

Also, by applying Lemma 4, there exists a $k_1<\infty$ such that when $k\geqslant k_1$, we have

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{in}=\tilde{\pi}^*\right] \leqslant & \frac{3}{2}\cdot\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right)\\ =&\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right).\end{aligned}$$

Therefore, for any $k\geqslant k_2\triangleq\max\{k^*,k_1\}$, the overall expected episodic regret can

be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\right]\\
=&\Pr\left\{\pi_k^{in}(\cdot)=\tilde{\pi}^*(\cdot)\right\}\cdot\mathbb{E}\left[\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\mid\pi_k^{in}(\cdot)=\tilde{\pi}^*(\cdot)\right]\\
&+\Pr\left\{\pi_k^{in}(\cdot)\neq\tilde{\pi}^*(\cdot)\right\}\cdot\mathcal{O}\left(\sqrt{k}\cdot U^{1+2\alpha}\right)\\
\leqslant&L'_k\cdot\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right)+(\delta+\epsilon_k)\cdot\mathcal{O}\left(\sqrt{k}\cdot U^{1+2\alpha}\right)\\
=&L'_k\cdot\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right)+\mathcal{O}\left(\delta\cdot\sqrt{k}\cdot U^{1+2\alpha}+U^{1+2\alpha}\right).
\end{aligned}
$$

We finally can bound the expected average regret as

$$
\begin{aligned}
&\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{t=1}^{t_K}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{t_K}\right]\\
&\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{t=1}^{t_{K+1}-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{t_{K+1}-1}\right]\\
=&\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{k=1}^{K}\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{\sum_{k=1}^{K}L'_k}\right]\\
=&\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{k=1}^{k_2-1}\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{\sum_{k=1}^{K}L'_k}\right]+\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{k=k_2}^{K}\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{\sum_{k=1}^{K}L'_k}\right]\\
\leqslant&\lim_{K\to\infty}\frac{\mathbb{E}\left[\sum_{k=1}^{k_2-1}\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\right]}{\sum_{k=1}^{K}L_k}+\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{k=k_2}^{K}L'_k\cdot\mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right)}{\sum_{k=1}^{K}L'_k}\right]\\
&+\lim_{K\to\infty}\frac{\sum_{k=k_2}^{K}\mathcal{O}\left(\delta\cdot\sqrt{k}\cdot U^{1+2\alpha}+U^{1+2\alpha}\right)}{\sum_{k=1}^{K}L_k}. \qquad \text{(A.15)}
\end{aligned}
$$

For the first term in Eq (A.15), since the numerator is finite, while the denominator

grows to infinity as $K \to \infty$, we have

$$\lim_{K\to\infty} \frac{\mathbb{E}\left[\sum_{k=1}^{k_2-1}\sum_{t=t_k}^{t_k+L'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\right]}{\sum_{k=1}^{K} L_k} = 0. \tag{A.16}$$

For the second term in Eq (A.15), since the big-O term holds for every $k \geqslant k_2$, we simply have

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{\sum_{k=k_2}^{K} L'_k \cdot \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right)}{\sum_{k=1}^{K} L'_k}\right] = \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}\right). \tag{A.17}$$

Since the sum of square roots can be bounded as

$$\frac{2n^{\frac{3}{2}}}{3} \leqslant \sum_{i=1}^{n}\sqrt{i} \leqslant \frac{2(n+1)^{\frac{3}{2}}}{3},$$

the third term in Eq (A.15) can be further bounded as

$$\begin{aligned}
\lim_{K\to\infty} \frac{\sum_{k=k_2}^{K}\mathcal{O}\left(\delta\sqrt{k}U^{1+2\alpha}+U^{1+2\alpha}\right)}{\sum_{k=1}^{K} L_k} &\leqslant \lim_{K\to\infty} \frac{\mathcal{O}\left(\frac{2(K+1)^{\frac{3}{2}}}{3}\cdot\delta U^{1+2\alpha}+KU^{1+2\alpha}\right)}{L\cdot\frac{2K^{\frac{3}{2}}}{3}} \\
&= \lim_{K\to\infty}\mathcal{O}\left(\left(1+\frac{1}{K}\right)^{\frac{3}{2}}\cdot\delta U^{1+2\alpha}+\frac{U^{1+2\alpha}}{\sqrt{K}}\right) \\
&= \mathcal{O}\left(\delta U^{1+2\alpha}\right).
\end{aligned} \tag{A.18}$$

By inserting Eq (A.16), Eq (A.17) and Eq (A.18) into Eq (A.15), we have

$$\lim_{K\to\infty}\mathbb{E}\left[\frac{\sum_{t=1}^{t_K}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{t_K}\right] = \mathcal{O}\left(\frac{U^{D+\max\{2\alpha,\gamma\}}}{\exp\left(U^{\min\{\beta,2-\beta\}}\right)}+\delta U^{1+2\alpha}\right),$$

which completes the proof.

A.5 Proof of Lemma 6

By Proposition 5.5.1 in [2], when applying $\tilde{\pi}^*$ to $\tilde{M}$, there exists $\tilde{h}^*(\cdot)$ such that the for each $\boldsymbol{Q} \in \mathcal{Z}^{in}$, the following Bellman equation holds:

$$\tilde{\rho}^* + \tilde{h}^*(\boldsymbol{Q}) = \sum_i Q_i + \sum_{\boldsymbol{Q}' \in \mathcal{S}^{in}} \tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, \tilde{\pi}^*(\boldsymbol{Q})\right) \cdot \tilde{h}^*\left(\boldsymbol{Q}'\right). \tag{A.19}$$

Note that Eq (A.19) works for $\tilde{M}$, and we extend it to M for analysis afterwards. By the truncation scheme in Section 2.2, for each $\boldsymbol{Q} \in \mathcal{Z}^{in}$, $\boldsymbol{Q}' \in \mathcal{S}^{in}$ and $a \in \mathcal{A}$, we have

$$\tilde{p}\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, \tilde{\pi}^*(\boldsymbol{Q})\right) = p\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, \tilde{\pi}^*(\boldsymbol{Q})\right).$$

Therefore, Eq (A.19) can be rewritten as

$$\tilde{\rho}^* + \tilde{h}^*(\boldsymbol{Q}) = \sum_i Q_i + \sum_{\boldsymbol{Q}' \in \mathcal{S}^{in}} p\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, \tilde{\pi}^*(\boldsymbol{Q})\right) \cdot \tilde{h}^*\left(\boldsymbol{Q}'\right), \tag{A.20}$$

for each $\boldsymbol{Q} \in \mathcal{Z}^{in}$.

During episode k, $\boldsymbol{Q}$ may enter $\mathcal{Z}^{in}$ and leave $\mathcal{Z}^{in}$ for multiple times. We define the process starting from $\boldsymbol{Q}$ entering $\mathcal{Z}^{in}$ to leaving $\mathcal{Z}^{in}$ as an "enter and leave" process. For an "enter and leave" process, we define $\boldsymbol{Q}^{en}$ and $\boldsymbol{Q}^{le}$ as its first and last state in $\mathcal{Z}^{in}$. We classify the "enter and leave" processes according to $(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le})$ pairs. Note that according to the setting in Section 2.2, we have $\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le} \in \mathcal{Z}^{in}_{bd}$.

We denote that, during $\mathcal{T}^{in}_k$, "enter and leave" processes with $(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le})$ occur for $N_k(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le})$ times, and when they occur for the i^{th} time, the start and end time slots are $t^{en}_{k,i}(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le})$ and $t^{le}_{k,i}(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le})$ separately.In addition to "enter and leave" processes, it is posssible that at the beginning or at the end of episode k, $\boldsymbol{Q} \in \mathcal{Z}^{in}$. For episode k, define $\mathcal{T}^{in}_{k,0}$ as the set of time slots that $\boldsymbol{Q} \in \mathcal{Z}^{in}$ before the first "enter and leave" process starts, $\mathcal{T}^{in}_{k,\infty}$ as the set of time slots that $\boldsymbol{Q} \in \mathcal{Z}^{in}$ after the last

"enter and leave" process ends. We therefore can make the following decomposition.

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_k^{in}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\mid \pi_k^{in}=\tilde{\pi}^*\right]$$

$$=\mathbb{E}\left[\sum_{t\in\mathcal{T}_k^{in}}\left(\tilde{h}^*\left(\boldsymbol{Q}(t)\right)-\mathbb{E}\left[\tilde{h}^*\left(\boldsymbol{Q}(t+1)\right)\mid \boldsymbol{Q}(t)\right]\right)\mid \pi_k^{in}=\tilde{\pi}^*\right]$$

$$=\mathbb{E}\left[\sum_{\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}\in\mathcal{Z}_{bd}^{in}}\sum_{i=1}^{N_k(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}\underbrace{\sum_{t=t_{k,i}^{en}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}^{t_{k,i}^{le}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}\left(\tilde{h}^*\left(\boldsymbol{Q}(t)\right)-\mathbb{E}\left[\tilde{h}^*\left(\boldsymbol{Q}(t+1)\right)\mid \boldsymbol{Q}(t)\right]\right)}_{\triangleq H_i(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}\right]+ \tag{A.21}$$

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{k,0}^{in}\cup\mathcal{T}_{k,\infty}^{in}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\mid \pi_k^{in}=\tilde{\pi}^*\right] \tag{A.22}$$

We then proceed to bound the value of (3.2) and (3.3) over L'_k (conditioned on $\pi_k^{in}=\tilde{\pi}^*$) separately.

For (A.21), we have the following lemma (see Appendix A.8 for the proof).

Lemma 9. *For every* $\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}\in\mathcal{Z}_{bd}^{in}$,

$$\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{i=1}^{N_k(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}H_i(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]\leqslant p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}^{en}\right)\cdot cDU^{1+\gamma}.$$

For Eq (A.22), similar to Eq (A.39) in the proof of Lemma 9, it can be upper bounded as follows.

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{k,0}^{in}\cup\mathcal{T}_{k,\infty}^{in}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\mid \pi_k^{in}=\tilde{\pi}^*\right]\leqslant 4cNU^{1+\gamma}. \tag{A.23}$$

By combining Lemma 9 and Eq (A.23), we have

$$\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{in}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]\\
\leqslant&\sum_{\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}\in\mathcal{Z}_{bd}^{in}} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}^{en}\right)\cdot cDU^{1+\gamma}+\lim_{k\to\infty}\frac{4cNU^{1+\gamma}}{\mathbb{E}\left[L'_k\right]}\\
=&\left|\mathcal{Z}_{bd}^{in}\right|\cdot p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right)\cdot cDU^{1+\gamma}+0\\
=&p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right)\cdot\mathcal{O}\left(U^{D+\gamma}\right),
\end{aligned}\tag{A.24}$$

where (A.24) holds because $\left|\mathcal{Z}_{bd}^{in}\right|=(U-W)^D-(U-2W)^D=\mathcal{O}\left(U^{D-1}\right)$.

A.6 Proof of Lemma 7

We define the set of time slots that $\boldsymbol{Q}\in\mathcal{Z}_{bd}^{out}$ as $\mathcal{T}_{bd,k}^{out}$ and the set of time slots that $\boldsymbol{Q}\in\mathcal{Z}_{out}^{out}$ as $\mathcal{T}_{out,k}^{out}$. We therefore have the following decomposition.

$$\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]\\
=&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_{bd,k}^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]+ &\text{(A.25)}\\
&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_{out,k}^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]. &\text{(A.26)}
\end{aligned}$$

For Eq (A.25), we have

$$\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_{bd,k}^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]\\
\leqslant&\lim_{k\to\infty}\mathbb{E}\left[\frac{DU\cdot\mathcal{T}_{bd,k}^{out}}{L'_k}\mid \pi_k^{in}=\tilde{\pi}^*\right]=p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot DU.
\end{aligned}\tag{A.27}$$

For Eq (A.26), we have the following analysis.

During episode k, $\boldsymbol{Q}$ may exit $\mathcal{S}^{in}$ and return back to $\mathcal{S}^{in}$ for multiple times. We define the process from the time that $\boldsymbol{Q}$ just exits $\mathcal{S}^{in}$ to the time that $\boldsymbol{Q}$ is just about to return back to $\mathcal{S}^{in}$ as an "exit and return" process. For an "exit and return" process, we define $\boldsymbol{Q}^{ex}$ as the last state before exiting and $\boldsymbol{Q}^{re}$ as the first state after returning back. We classify the "exit and return" processes according to $(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ pairs. Note that according to the setting in Section 2.2, we have $\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re} \in \mathcal{Z}_{bd}^{out}$.

We denote that, during $\mathcal{T}_k^{out}$, "exit and return" processes with $(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ occur for $N_k'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ times, and when they occur for the i^{th} time, the start and end time slots are $t_{k,i}^{ex}(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ and $t_{k,i}^{re}(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ separately.

We define $R_i\left(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re}\right)$ as the regret from $t_{k,i}^{ex}(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ to $t_{k,i}^{re}(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$. We therefore can decompose $\mathbb{E}\left[\sum_{t\in\mathcal{T}_{out,k}^{out}}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)\right]$ as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{T}_{out,k}^{out}}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)\right] \\
=&\mathbb{E}\left[\sum_{\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re}\in\mathcal{Z}_{bd}^{out}} \sum_{i=1}^{N_k'(\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re})} R_i\left(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re}\right)\right] \qquad \text{(A.28)}
\end{aligned}$$

We then proceed to bound the value of (A.28) over L_k' (conditioned on $\pi_k^{in} = \tilde{\pi}^*$). We have the following lemma (see Appendix A.9 for the proof).

Lemma 10. *For every* $\boldsymbol{Q}^{en}, \boldsymbol{Q}^{le} \in \mathcal{Z}_{bd}^{out}$,

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^{N_k'(\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re})} R_i\left(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re}\right)}{L_k'} \mid \pi_k^{in} = \tilde{\pi}^*\right] \leqslant p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}^{ex}\right) \cdot \frac{2a^2DW(U+W)^{1+2\alpha}}{\epsilon_0^2}.$$

By combining Lemma Eq 10, Eq (A.28) and Eq (A.27), we have

$$\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{t\in\mathcal{T}_k^{out}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'}\mid \pi_k^{in}=\tilde{\pi}^*\right]\\
\leqslant& p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot DU+\sum_{\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re}\in\mathcal{Z}_{bd}^{out}} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}^{ex}\right)\cdot\frac{2a^2DW(U+W)^{1+2\alpha}}{\epsilon_0^2}\\
=&p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot DU+p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot\left|\mathcal{Z}_{bd}^{out}\right|\cdot\frac{2a^2DW(U+W)^{1+2\alpha}}{\epsilon_0^2}\\
=&p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)\cdot\mathcal{O}\left(U^{D+2\alpha}\right),
\end{aligned}\tag{A.29}$$

where Eq (A.29) holds because $\left|\mathcal{Z}_{bd}^{out}\right|=U^D-(U-W)^D=\mathcal{O}\left(U^{D-1}\right)$.

A.7 Proof of Lemma 8

In Assumption 2, we define a Lyapunov function $\tilde{\Phi}^*(\cdot)$ on $\mathcal{S}^{in}$. To extend it to $\mathcal{S}$, we define $\Phi'(\cdot)$ as follows

$$\Phi'\left(\boldsymbol{Q}\right)=\begin{cases}\tilde{\Phi}^*\left(\boldsymbol{Q}\right) & \text{if } \boldsymbol{Q}\in\mathcal{S}^{in}\\ 0 & \text{if } \boldsymbol{Q}\in\mathcal{S}^{out}\end{cases}.\tag{A.30}$$

In the following lemma, we prove that $\Phi'(\cdot)$ has similar drift properties as $\tilde{\Phi}^*(\cdot)$ does (see Appendix A.10 for the proof).

Lemma 11. *When* $\boldsymbol{Q}(t)\in\left\{\boldsymbol{Q}\in\mathcal{S}:\tilde{B}^*\leqslant Q_{max}\leqslant U\right\}$,

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi'(\boldsymbol{Q}(t+1))-\Phi'(\boldsymbol{Q}(t))\mid\boldsymbol{Q}(t)\right]\leqslant-\tilde{\epsilon}^*.$$

It has been proven in [3] that for a Markov chain with negative Lyapunov drifts, the probability for Lyapunovvalues to grow large decays exponentially, as the following Lemma states.

Lemma 12 (Theorem 3 in [3]). *Given a nonnegative Lyapunov function* $\Phi(\cdot)$, *if for*

$\Phi(\boldsymbol{Q}) > B$, *we have*

$$\mathbb{E}\left[\Phi\left(\boldsymbol{Q}(t+1)\right) - \Phi\left(\boldsymbol{Q}(t)\right) \mid \boldsymbol{Q}(t)\right] \leqslant -\epsilon. \tag{A.31}$$

Then for $m \geqslant 0$, *the stationary distribution for the states that satisfy* $\Phi(\boldsymbol{Q}) > B$ *can be upper bounded as follows.*

$$\Pr\left\{\Phi\left(\boldsymbol{Q}(t)\right) > B + m\right\} \leqslant \frac{1}{\left(1+\frac{\epsilon}{V}\right)^{\frac{m}{2V}+1}}, \tag{A.32}$$

where V *is the maximum value of drifts.*

In our case, when $\Phi'(\boldsymbol{Q}) > c_2\left(\tilde{B}^*\right)^{\beta} \triangleq B'$, we have $Q_{max} > \tilde{B}^*$ and $\Phi'(\cdot)$ has negative drift upper bounded by $-\tilde{\epsilon}^*$. Also, by Assumption 2, $V \leqslant b_3 U^{\max\{\beta-1,0\}}$. By applying Lemma 12, we have

$$\begin{aligned}
&p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right) + p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right) \\
=&\Pr\left\{U - 2W + 1 \leqslant Q_{max} \leqslant U\right\} \\
\leqslant&\Pr\left\{b_1 Q_{max}^{\beta} \geqslant b_1(U-2W)^{\beta}, Q_{max} \leqslant U\right\} \\
\leqslant&\Pr\left\{\Phi'(\boldsymbol{Q}) \geqslant b_1(U-2W)^{\beta}\right\} \\
=&\Pr\left\{\Phi'(\boldsymbol{Q}) \geqslant B' + b_1(U-2W)^{\beta} - B'\right\} \\
\leqslant&\frac{1}{\left(1+\frac{\tilde{\epsilon}^*}{V}\right)^{\frac{b_1(U-2W)^{\beta}-B'}{2V}+1}}
\end{aligned} \tag{A.33}$$

Since for $x > 0$, the following inequality holds.

$$\log(1+x) \geqslant \frac{x}{1+x}$$

We therefore have

$$\begin{aligned}
&\left(\frac{b_1(U-2W)^{\beta}-B'}{2V}+1\right)\cdot\log\left(1+\frac{\tilde{\epsilon}^*}{V}\right)\\
\geqslant&\frac{b_1(U-2W)^{\beta}-B'}{2V}\cdot\frac{\tilde{\epsilon}^*}{V+\tilde{\epsilon}^*}\\
\geqslant&\frac{b_1(U-2W)^{\beta}-B'}{2b_3U^{\max\{\beta-1,0\}}}\cdot\frac{\tilde{\epsilon}^*}{b_3U^{\max\{\beta-1,0\}}+\tilde{\epsilon}^*}\\
=&\left(U^{\min\{2-\beta,\beta\}}\right)
\end{aligned}\tag{A.34}$$

By inserting (A.34) into (A.33), we have

$$p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{in}\right)+p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Z}_{bd}^{out}\right)=\mathcal{O}\left(\exp\left(-U^{\min\{2-\beta,\beta\}}\right)\right)\tag{A.35}$$

which completes the proof.

A.8 Proof of Lemma 9

Define $Y_i^{in}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})$ as the time interval between the starting time of the i^{th} and $(i+1)^{th}$ "enter and leave" process with $(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})$. By the Markovian property of the system, $Y_i^{in}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})$'s are i.i.d. and $H_i(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})$'s are also i.i.d.

Since $L_k'\geqslant L\sqrt{k}$ and $L_k'\to\infty$ as $k\to\infty$, then according to the renewal reward throrem, for every $\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}\in\mathcal{Z}_{bd}^{in}$, we have

$$\lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{i=1}^{N_k(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}H_i(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}{L_k'}\mid\pi_k^{in}=\tilde{\pi}^*\right]=\frac{\mathbb{E}\left[H_1(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right]}{\mathbb{E}\left[Y_1^{in}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right]}.\tag{A.36}$$

However, directly computing $\mathbb{E}\left[Y_1^{in}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right]$ is not straightforward. We have the bound that for every $\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}\in\mathcal{Z}_{bd}^{in}$,

$$\mathbb{E}\left[Y_1^{in}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right]\geqslant\mathbb{E}\left[\text{Interval between visits to }\boldsymbol{Q}^{en}\right].\tag{A.37}$$

Also, From Assumption 3, Assumption 1 and Foster-Lyapunov Throrem, under

$\tilde{\pi}^* + \pi_0$ the Markov chain is positive recurrent. Therefore we have

$$p^{\tilde{\pi}^*+\pi_0}(\boldsymbol{Q}^{en}) = \frac{1}{\mathbb{E}\left[\text{Interval between visits to } \boldsymbol{Q}^{en}\right]}. \tag{A.38}$$

Inserting Eq A.37 and A.38 into A.36, we have

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^{N_k(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})} H_i(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}{L'_k} \mid \pi_k^{in} = \tilde{\pi}^*\right] \leqslant p^{\tilde{\pi}^*+\pi_0}(\boldsymbol{Q}^{en}) \cdot \mathbb{E}\left[H_1(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right].$$

For $H_1(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})$,

$$\begin{aligned}
&\mathbb{E}\left[H_1(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})\right] \\
=&\mathbb{E}\left[\sum_{t=t_{k,1}^{en}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}^{t_{k,1}^{le}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})} \left(\tilde{h}^*\left(\boldsymbol{Q}(t)\right) - \mathbb{E}\left[\tilde{h}^*\left(\boldsymbol{Q}(t+1)\right) \mid \boldsymbol{Q}(t)\right]\right)\right] \\
=&\mathbb{E}\left[\tilde{h}^*\left(\boldsymbol{Q}(t_{k,1}^{en}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}))\right) - \mathbb{E}\left[\tilde{h}^*\left(t_{k,1}^{le}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})+1\right) \mid \boldsymbol{Q}(t)\right]\right] \\
&+\mathbb{E}\left[\sum_{t=t_{k,1}^{en}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})}^{t_{k,1}^{le}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})-1} \left(\tilde{h}^*(\boldsymbol{Q}(t+1)) - \mathbb{E}\left[\tilde{h}^*(\boldsymbol{Q}(t+1)) \mid \boldsymbol{Q}(t)\right]\right)\right] \\
=&\mathbb{E}\left[\tilde{h}^*\left(\boldsymbol{Q}(t_{k,1}^{en}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le}))\right)\right] - \mathbb{E}\left[\tilde{h}^*\left(t_{k,1}^{le}(\boldsymbol{Q}^{en},\boldsymbol{Q}^{le})+1\right)\right]
\end{aligned} \tag{A.39}$$

Define that $H \triangleq \max_{\boldsymbol{Q}\in\mathcal{Z}^{in}}\left|\tilde{h}^*(\boldsymbol{Q})\right|$. From the analysis following the proof of Proposition 5.5.1 in [2], $\tilde{h}^*(\boldsymbol{Q}') - \tilde{h}^*(\boldsymbol{Q})$ can be interpreted as the minimum of the expected cost to reach $\boldsymbol{Q}'$ from $\boldsymbol{Q}$ for the first time, when the cost is defined as $c(\boldsymbol{Q}) - \tilde{\rho}^*$. Apply Assumption 3, and note that $c(\boldsymbol{Q}) \leqslant NU$, we thus have that for each $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathcal{Z}^{in}$,

$$\tilde{h}^*(\boldsymbol{Q}') - \tilde{h}^*(\boldsymbol{Q}) \leqslant \mathbb{E}\left[T_{\boldsymbol{Q}\to\boldsymbol{Q}'}^{\tilde{\pi}^*}\right] \cdot NU \leqslant cNU^{1+\gamma}. \tag{A.40}$$

By inserting (A.40) into (A.39), we complete the proof of Lemma 9.

A.9 Proof of Lemma 10

Define $Y_i'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ as the time interval between the starting time of the i^{th} and $(i+1)^{th}$ "exit and return" process with $(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$. By the Markovian property of the system, $Y_i'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$'s are i.i.d. and $R_i(\boldsymbol{Q}^{en}, \boldsymbol{Q}^{re})$'s are also i.i.d.

Since $L_k' \geqslant L\sqrt{k}$ and $L_k' \to \infty$ as $k \to \infty$, then according to the renewal reward throrem, for every $\boldsymbol{Q}^{en}, \boldsymbol{Q}^{re} \in \mathcal{Z}_{bd}^{out}$ we have

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^{N_k'(\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re})} R_i(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})}{L_k'} \mid \pi_k^{in} = \tilde{\pi}^*\right] = \frac{\mathbb{E}\left[R_1(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right]}{\mathbb{E}\left[Y_1'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right]}. \tag{A.41}$$

However, directly computing $\mathbb{E}\left[Y_1'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right]$ is not straightforward. We have the bound that for every $\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re} \in \mathcal{Z}_{bd}^{out}$,

$$\mathbb{E}\left[Y_1'(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right] \geqslant \mathbb{E}\left[\text{Interval between visits to } \boldsymbol{Q}^{ex}\right]. \tag{A.42}$$

Also, From Assumption 3, Assumption 1 and Foster-Lyapunov Throrem, under $\tilde{\pi}^* + \pi_0$ the Markov chain is positive recurrent. Therefore we have

$$p^{\tilde{\pi}^*+\pi_0}(\boldsymbol{Q}^{ex}) = \frac{1}{\mathbb{E}\left[\text{Interval between visits to } \boldsymbol{Q}^{ex}\right]}. \tag{A.43}$$

Inserting Eq A.42 and A.43 into A.41, we have

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^{N_k'(\boldsymbol{Q}^{ex},\boldsymbol{Q}^{re})} R_i(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})}{L_k'} \mid \pi_k^{in} = \tilde{\pi}^*\right] = p^{\tilde{\pi}^*+\pi_0}(\boldsymbol{Q}^{ex}) \cdot \mathbb{E}\left[R_1(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right]. \tag{A.44}$$

We now come to bound $R_1(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$. Define $\tau(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$ as the time spent in the "exit and return" processes with $(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$. Define $\mathcal{S}_{bd}^{out} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : U+1 \leqslant Q_{max} \leqslant U+W\}$. According to the proof of Theorem 1.1 in Chapter 5 of [6], for any $(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})$, we have the uniform upper bound that

$$\mathbb{E}\left[\tau(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right] \leqslant \frac{\max_{\boldsymbol{Q}\in\mathcal{S}_{bd}^{out}} \Phi_0(\boldsymbol{Q})}{\epsilon_0} \leqslant \frac{a(U+W)^\alpha}{\epsilon_0} \triangleq T_0. \tag{A.45}$$

Theorem 6.3.4 in [11] states that for a nonempty state set $\mathcal{B}$, and a state s, If there exists a constant C such that

$$\mathbb{E}\left[T_{s\to\mathcal{B}}\right] \leqslant C \cdot F^*_{s,\mathcal{B}}. \tag{A.46}$$

Then for $p \geqslant 1$

$$\mathbb{E}\left[T^p_{s\to\mathcal{B}}\right] \leqslant p! \cdot C^p \cdot F^*_{s,\mathcal{B}} \tag{A.47}$$

where $F^*_{s,\mathcal{B}} \triangleq \sum_{n=1}^{\infty} \Pr\{s_v \notin \mathcal{B}, 0 < v < n, s_n \in \mathcal{B} \mid s_0 = s\}$.

Therefore, we can bound $\mathbb{E}\left[\tau^2\left(\boldsymbol{Q}^{out}, \boldsymbol{Q}^{in}\right)\right]$ for any $\left(\boldsymbol{Q}^{out}, \boldsymbol{Q}^{in}\right)$ in the manner that

$$\mathbb{E}\left[\tau^2\left(\boldsymbol{Q}^{out}, \boldsymbol{Q}^{in}\right)\right] \leqslant 2T_0^2. \tag{A.48}$$

For $R_1\left(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re}\right)$, the queue length can grow to at most $D \cdot \left(U + W\tau(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[R_1\left(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re}\right)\right] \leqslant & \mathbb{E}\left[D \cdot \left(U + W\tau(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right) \cdot \tau(\boldsymbol{Q}^{ex}, \boldsymbol{Q}^{re})\right] \\
\leqslant & D(U+W)T_0 + 2DWT_0^2 \\
\leqslant & 2DW(U+W)T_0^2 \\
\leqslant & \frac{2a^2DW(U+W)^{1+2\alpha}}{\epsilon_0^2}
\end{aligned} \tag{A.49}$$

By inserting (A.49) into (A.44), we complete the proof of Lemma 10.

A.10 Proof of Lemma 11

From Assumption 2, we know that in the virtual truncated system $\tilde{M}$, we have

$$\mathbb{E}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*(\boldsymbol{Q}(t+1)) - \tilde{\Phi}^*(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t)\right] \leqslant -\tilde{\epsilon}^*, \tag{A.50}$$

when $\boldsymbol{Q}(t) \in \left\{\boldsymbol{Q} \in \mathcal{S} : \tilde{B}^* \leqslant Q_{max} \leqslant U\right\}$.

We then turn to the real system M. Since the state-transition functions remain

exactly the same $\boldsymbol{Q} \in \mathcal{Z}^{in}$, we have

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi'(\boldsymbol{Q}(t+1)) - \Phi'(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t)\right] = \mathbb{E}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*(\boldsymbol{Q}(t+1)) - \tilde{\Phi}^*(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t)\right]. \tag{A.51}$$

For $\boldsymbol{Q}(t) \in \mathcal{Z}_{bd}^{out}$, notice that it is possible that $\boldsymbol{Q}(t+1)$ is in $\mathcal{S}^{out}$, which makes (A.51) no longer hold. In this case, we need more intricate calculation:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi'(\boldsymbol{Q}(t+1)) - \Phi'(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t)\right] \\
=& -\tilde{\Phi}^*(\boldsymbol{Q}(t)) + \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot \Phi'\left(\boldsymbol{Q}\right) \\
=& -\tilde{\Phi}^*(\boldsymbol{Q}(t)) + \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{in}} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot \tilde{\Phi}^*\left(\boldsymbol{Q}\right) \\
&+ \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{out}} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot 0 \qquad \text{(A.52)} \\
\leqslant& -\tilde{\Phi}^*(\boldsymbol{Q}(t)) + \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{in}} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot \tilde{\Phi}^*\left(\boldsymbol{Q}\right) \\
&+ \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{out}} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot \tilde{\Phi}^*\left(TR\left(\boldsymbol{Q}\right)\right) \qquad \text{(A.53)} \\
=& -\tilde{\Phi}^*(\boldsymbol{Q}(t)) + \sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)} p\left(\boldsymbol{Q} \mid \boldsymbol{Q}(t), \tilde{\pi}^*(\boldsymbol{Q}(t))\right) \cdot \tilde{\Phi}^*\left(TR\left(\boldsymbol{Q}\right)\right), \qquad \text{(A.54)}
\end{aligned}$$

where (A.52) comes from (A.30), (A.53) holds because $\tilde{\Phi}^*(\cdot)$ is nonnegative and (A.54) comes from the property that for $\boldsymbol{Q} \in \mathcal{S}^{in}$, $TR\left(\boldsymbol{Q}\right) \equiv \boldsymbol{Q}$.

We further have

$$\sum_{\boldsymbol{Q}\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)} p\left(\boldsymbol{Q}\mid\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cdot\tilde{\Phi}^*\left(TR\left(\boldsymbol{Q}\right)\right)$$

$$=\sum_{\boldsymbol{Q}'\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{in}}\tilde{\Phi}^*\left(\boldsymbol{Q}'\right)\cdot\sum_{\boldsymbol{Q}\in\mathcal{S}(\boldsymbol{Q}')}p\left(\boldsymbol{Q}\mid\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right) \tag{A.55}$$

$$=\sum_{\boldsymbol{Q}'\in\mathcal{R}\left(\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cap\mathcal{S}^{in}}\tilde{p}\left(\boldsymbol{Q}'\mid\boldsymbol{Q}(t),\tilde{\pi}^*(\boldsymbol{Q}(t))\right)\cdot\tilde{\Phi}^*\left(\boldsymbol{Q}'\right) \tag{A.56}$$

$$=\mathbb{E}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*(\boldsymbol{Q}(t+1))\mid\boldsymbol{Q}(t)\right], \tag{A.57}$$

where (A.55) is obtained by rewriting the summation according to the values of $TR\left(\boldsymbol{Q}\right)$ and (A.56) comes from the definition in Section 2.2.

By combining (A.51), (A.54) and (A.57), we complete the proof of Lemma 11.

Bibliography

[1] JS Baras, D-J Ma, and AM Makowski. K competing queues with geometric service requirements and linear costs: The μc-rule is always optimal. *Systems & control letters*, 6(3):173–180, 1985.

[2] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 2017.

[3] Dimitris Bertsimas, David Gamarnik, John N Tsitsiklis, et al. Geometric bounds for stationary distributions of infinite markov chains via lyapunov functions. 1998.

[4] Dimitris Bertsimas, David Gamarnik, John N Tsitsiklis, et al. Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. *The Annals of Applied Probability*, 11(4):1384–1428, 2001.

[5] Dimitris Bertsimas, Ioannis Ch Paschalidis, and John N Tsitsiklis. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability*, pages 43–75, 1994.

[6] Pierre Brémaud. Lyapunov functions and martingales. In *Markov Chains*, pages 167–193. Springer, 1999.

[7] Olivier Brun, Lan Wang, and Erol Gelenbe. Big Data for Autonomic Intercontinental Overlays. *IEEE Journal on Selected Areas in Communications*, 34(3):575–583, 2016.

[8] C Buyukkoc, P Varaiya, and J Walrand. The $c\mu$ rule revisited. *Advances in applied probability*, 17(1):237–238, 1985.

[9] Jim G Dai and Wuqin Lin. Maximum pressure policies in stochastic processing networks. *Operations Research*, 53(2):197–218, 2005.

[10] Junghee Han, D Watson, and F Jahanian. Topology aware overlay networks. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, pages 2554–2565. IEEE, 2005.

[11] Zhenting Hou and Qingfeng Guo. *Homogeneous denumerable Markov processes.* Springer Science & Business Media, 2012.

[12] Carlos Humes Jr, J Ou, and PR Kumar. The delay of open markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability. *Mathematics of Operations Research*, 22(4):921–954, 1997.

[13] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[14] PR Kumar and Sean P Meyn. Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 40(2):251–260, 1995.

[15] Sunil Kumar and PR Kumar. Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 39(8):1600–1611, 1994.

[16] Z Li and P Mohapatra. QRON: QoS-aware routing in overlay networks. *IEEE Journal on Selected Areas in ...*, 2004.

[17] Michael J Neely and Longbo Huang. Dynamic product assembly and inventory control for maximum profit. In *49th IEEE Conference on Decision and Control (CDC)*, pages 2805–2812. IEEE, 2010.

[18] Michael J Neely, Eytan Modiano, and Chih-Ping Li. Fairness and optimal stochastic control for heterogeneous networks. *IEEE/ACM Transactions On Networking*, 16(2):396–409, 2008.

[19] Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2012.

[20] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[21] Anurag Rai, Rahul Singh, and Eytan Modiano. A Distributed Algorithm for Throughput Optimal Routing in Overlay Networks. *arXiv.org*, December 2016.

[22] Sajee Singsanga, Wipawee Hattagam, and Ewe Hong Tat. Packet forwarding in overlay wireless sensor networks using NashQ reinforcement learning. In *2010 Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 85–90. IEEE, 2010.

[23] R K Sitaraman and M Kasbekar. Overlay networks: An akamai perspective. ..., 2014.

[24] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[25] Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *29th IEEE Conference on Decision and Control*, pages 2130–2132. IEEE, 1990.

[26] 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.

[27] Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.

[28] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[29] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.