RL-QN: A Reinforcement Learning Framework for Optimal Control of Queueing Systems

BAI LIU, Massachusetts Institute of Technology
QIAOMIN XIE, University of Wisconsin-Madison
EYTAN MODIANO, Massachusetts Institute of Technology

With the rapid advance of information technology, network systems have become increasingly complex and hence the underlying system dynamics are often unknown or difficult to characterize. Finding a good network control policy is of significant importance to achieve desirable network performance (e.g., high throughput or low delay). In this work, we consider using model-based reinforcement learning (RL) to learn the optimal control policy for queueing networks so that the average job delay (or equivalently the average queue backlog) is minimized. Traditional approaches in RL, however, cannot handle the unbounded state spaces of the network control problem. To overcome this difficulty, we propose a new algorithm, called RL for Queueing Networks (RL-QN), which applies model-based RL methods over a finite subset of the state space while applying a known stabilizing policy for the rest of the states. We establish that the average queue backlog under RL-QN with an appropriately constructed subset can be arbitrarily close to the optimal result. We evaluate RL-QN in dynamic server allocation, routing, and switching problems. Simulation results show that RL-QN minimizes the average queue backlog effectively.

CCS Concepts: • **Networks → Network performance analysis;**

Additional Key Words and Phrases: Queueing networks, reinforcement learning

ACM Reference format:
Bai Liu, Qiaomin Xie, and Eytan Modiano. 2022. RL-QN: A Reinforcement Learning Framework for Optimal Control of Queueing Systems. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 7, 1, Article 2 (August 2022), 35 pages.
https://doi.org/10.1145/3529375

1 INTRODUCTION

The rapid growth of information technology has resulted in increasingly complex network systems and poses challenges in obtaining 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 [54]. In these cases, only the overlay

This work was funded by National Science Foundation (NSF) grants CNS-1524317 and CNS-1907905 and by Office of Naval Research (ONR) grant N00014-20-1-2119.
Authors' addresses: B. Liu and E. Modiano, Massachusetts Institute of Technology, Cambridge, MA 02139; emails: {bailiu, modiano}@mit.edu; Q. Xie, University of Wisconsin-Madison, Madison, WI 53706; email: qiaomin.xie@wisc.edu.

2376-3639/2022/08-ART2 $15.00
https://doi.org/10.1145/3529375

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 [57] and Drift-plus-Penalty [45]) aim at stabilizing the system. However, existing optimization methods for long-term performance metrics (e.g., queue backlog and delay) only work for specific structures [15, 37], while the development for general setting remains insufficient.

To overcome the above challenges, it is desirable to apply inference and learning schemes. A natural approach is **reinforcement learning** (**RL**), which learns and reinforces a good decision policy by interacting with the environment and receiving feedbacks. RL methods provide a framework for the design of learning policies for general networks. There have been two main lines of work on RL methods: model-free RL (e.g., Q-learning [62], policy gradient [56]) and model-based RL (e.g., UCRL [27], PSRL [48]). In this work, we focus on the model-based approach.

1.1 Related Work

Existing methods for reducing the average job delay (or equivalently reducing the average queue backlog) of general networks can be classified into three types: equivalent constraint, Lyapunov drift, and **Markov decision process** (**MDP**) [18]. However, most equivalent constraint approaches can only be applied to single-hop networks [18]. Therefore, we focus on discussing the latter two classes of algorithms, as they are more universal and can be applied to multi-hop network systems.

We first discuss the Lyapunov drift approach, which generally does not require learning the dynamics for network problems with hidden dynamics. The Lyapunov drift approach has been widely applied in numerous network control problems. For instance, in dynamic server allocation problems, a Lyapunov-drift-based algorithm named the **longest connected queue** (**LCQ**) can stabilize the queue backlog without knowing the underlying dynamics (e.g., arrival rates, channel statistics) [22, 58]. For multiclass routing networks, the Backpressure algorithm was designed under the Lyapunov drift framework and only requires the observation of queue backlogs [3, 46, 57]. Similarly, Lyapunov drift methods that do not require learning the network dynamics have been proposed for switch scheduling [29, 40, 53] and inventory control problems [19, 44], and so on. Most Lyapunov drift algorithms are shown to be throughput optimal [57], i.e., they can stabilize the system whenever the system is stabilizable. Also, the Lyapunov drift approach usually only requires solving a linear programming problem and does not suffer from the "curse of dimensionality", which makes it applicable to large-scale queueing systems. However, without learning the system dynamics, the Lyapunov drift approach generally cannot guarantee minimum queueing delay. For instance, for dynamic server allocation problems, the optimal policy has only been developed for highly symmetric systems (i.e., uniform external arrival rates, uniform connectivity probabilities, and uniform success rates) [22]. Another well-known example is switching system, for which the Maximum Matching policy has been shown to be close to the optimal policy when the system is in the heavy traffic regime [37]. In Sections 5.1, 5.1, and 5.4, we conduct numerical experiments and show that our approach significantly outperforms the Lyapunov drift methods.

The MDP approach models the queueing system control problem as an MDP that aims at minimizing the long-term average queue backlog. Classical algorithms to solve MDP problems include value iteration and policy iteration [7]. Note that although the MDP approach can minimize the average job delay, it can only be applied to networks with finite buffer size, which is unrealistic for many practical network models. Therefore, the MDP approach fails to minimize the average job delay of general stochastic networks and is typically applied to MDPs of small scale [65, 66] or special structures [24, 51]. However, traditional MDP approach requires explicit knowledge of

the network dynamics. Since the network problems studied in this work have hidden dynamics, it is necessary to learn the dynamics when applying the MDP approach. A related field is RL. The majority of RL algorithms apply approximations to reduce the computation complexity, including state representation [20, 38], value function approximation [4, 28], and pruning [36]. RL methods have been applied to a wide range of network control problems, like routing [11, 17, 50, 67], spectrum access [23, 43, 61], switch scheduling [13], state probing [14] and so on. Although these RL methods achieve satisfactory performance in simulation, they lack performance guarantees.

We consider model-based RL methods, which tend to be more analytically tractable. Conventional model-based RL methods like UCRL [27] and PSRL [48] only work for finite-state-space systems, yet queueing systems are usually modeled to have unbounded buffer sizes. The work in [34] assumes that the system has an admission control scheme to keep queue backlogs finite. The resulting system can be modeled as an MDP with finite states, where UCRL can be applied. In [59], the authors modify the PSRL algorithm to deal with MDPs with large state space, yet the algorithm requires the MDP to have a finite bias span, which is unrealistic for problems that aim at minimizing the average cost with a countably infinite state space.

In summary, existing methods either cannot guarantee optimal delay (the Lyapunov drift approach), or lack performance guarantees (heuristic RL) or can only deal with finite state spaces (the MDP approach and theoretical RL). In this article, we aim at developing an algorithm that achieves provable optimality for networks with countably infinite state spaces. To the best of our knowledge, our algorithm is the first to achieve minimum average queue backlog in general networks with unbounded buffers. Our analysis also offers a possible roadmap to solving general MDPs with countably infinite state spaces.

1.2 Our Contributions

We apply model-based RL to queueing networks with unbounded state spaces and unknown dynamics. Our approach leverages the fact that for a vast class of stable queueing systems, the probability of the queue backlog being large is relatively small. This observation motivates us to focus on learning control policies over a finite subset of states that the system visits with high probability. Our main contributions are summarized as follows.

- We propose a model-based RL algorithm that can deal with unbounded state spaces. In particular, we introduce an auxiliary system with the state space bounded by a threshold U. Our approach employs a piecewise policy: for states below the threshold, a model-based RL algorithm is used; for all other states, a simple baseline algorithm is applied.
- We establish that the episodic average queue backlog under the proposed algorithm can be arbitrarily close to the optimum with a large threshold U. In particular, by applying Lyapunov analysis, we characterize the gap to the optimal performance as a function of the threshold U. In addition, our proof technique may be of independent interest for analyzing the convergence of other RL algorithms for queueing networks.
- Simulation results on dynamic server allocation, routing, and network switching problems corroborate the validity of our theoretical guarantees. In particular, the proposed algorithm effectively achieves a small average queue backlog, with the gap to optimum diminishing as the threshold U increases.

The article is organized as follows. In Section 2, we formulate the problem and introduce notations. Section 3 gives an outline of our approach, presents required assumptions and the proposed algorithm. We conduct theoretical performance analysis regarding convergence and optimality of the proposed algorithm in Section 4. We evaluate the algorithm in various settings and the numerical results are given in Section 5.

2 SYSTEM MODEL

We consider a discrete time network with the general topology of a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{L}$ is the set of links. Each node maintains one or more queues for undelivered packets, and each queue has an unbounded buffer. We denote by D the number of queues. Nodes represent the locations where data packets are generated, relayed, and/or processed in the network. Links represent the communication links.

During each time slot, external data packets arrive at nodes where they are either processed and then depart the system or are relayed. For relayed packets, the communication links can be stochastic, i.e., data transmissions between nodes can fail. We assume the underlying dynamics are partially or fully unknown. This model captures a large class of queueing networks that involve routing, scheduling, and switching.

We suppose that the system has the Markovian property: The probability distributions of the queue backlogs in the next time slot only depend on the current queue backlogs and the control decision. In other words, the system can be modeled as a countable-state-space MDP $\mathcal{M}$ with average cost as follows:

— State space $\mathcal{S}$: the set of queue backlog vectors, i.e., $\mathcal{S} \triangleq \mathbb{N}^D$.
— Action space $\mathcal{A}$: the set of feasible control decisions, which are specified by the problem setting. In this article, we only consider finite action space.
— State-transition probability $p(\boldsymbol{Q}' \mid \boldsymbol{Q}, a)$: the probability of transitioning into state $\boldsymbol{Q}'$ from state $\boldsymbol{Q}$ with action a. For simplicity, we assume that the magnitude that a queue backlog can change during one time slot is upper bounded by a constant W. Let $\mathcal{R}(\boldsymbol{Q})$ denote the set of one-step reachable queue backlog vectors from state $\boldsymbol{Q}$.
— Cost function $c(\boldsymbol{Q})$: the total backlog of D queues, i.e., $c(\boldsymbol{Q}) \triangleq \sum_{i=1}^{D} Q_i$.

As discussed in Section 1.1, a large number of queueing networks have been studied extensively and various stabilizing policies that aim at keeping the average total queue backlog finite have been developed. Here, we take a step further to go beyond stability and our goal is to minimize the average queue backlog.

For readers' convenience, we summarize the notations used in this article in Table 1.

3 OUR APPROACH

Classical model-based RL fits a model of state transition kernel to observed data and then solves the dynamic programming problem on the estimated system. The challenge of applying such an approach to countable-state MDPs arises from both of estimating model parameters and solving the estimated MDP, due to the fact that the state space is unbounded.

Here, we introduce an auxiliary system $\tilde{\mathcal{M}}$ with a bounded state space. We only apply RL techniques on the constructed bounded state space, while simply apply a *known* stabilizing policy π_0 (cf. Assumption 1) to the rest of the states. We show that the performance gap between the proposed algorithm and the optimal policy can be made arbitrarily small by designing appropriate $\tilde{\mathcal{M}}$.

3.1 Overview

We first provide an overview of our approach and defer a detailed description of the algorithm to Section 3.3. Our RL method operates in an episodic manner.

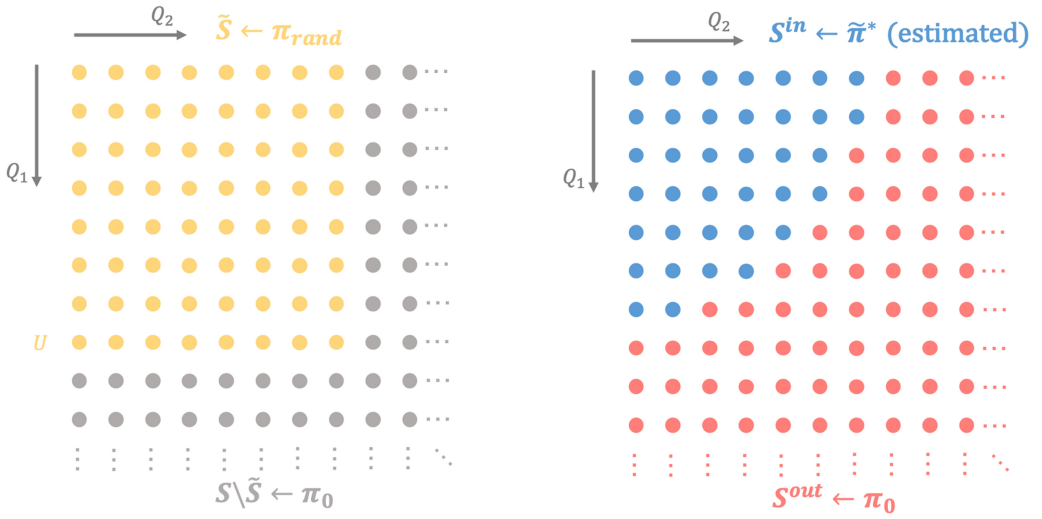
We apply a decaying ϵ-greedy method to decide whether an episode should conduct exploration or exploitation. For each episode, we perform exploration (i.e., apply some random policy to collect diverse samples) with probability ϵ_i, and we conduct exploitation (i.e., using current estimates of dynamics to compute an estimated optimal policy) with probability $1 - \epsilon_i$. At the beginning of

Table 1. Notations

D	The number of queues in the queueing network
$\boldsymbol{Q}$	The D-dimensional queue backlog vector of the queueing network
Q_i	The queue backlog of node i
$Q_{\max}$	The maximum entry of $\boldsymbol{Q}$
$\mathcal{R}(\boldsymbol{Q})$	The set of one-step reachable queue backlog vectors from $\boldsymbol{Q}$
W	The magnitude that a queue backlog can change during one time slot
U	The buffer size of the auxiliary system
$\mathcal{M}$	The countable-state-space MDP of the original queueing network
$\tilde{\mathcal{M}}$	The MDP of the auxiliary system
$\mathcal{S}$	The state space of $\mathcal{M}$
$\tilde{\mathcal{S}}$	The state space of $\tilde{\mathcal{M}}$
$\mathcal{A}$	The action space of $\mathcal{M}$
$p(\cdot \mid \cdot, \cdot)$	The state-transition kernel of $\mathcal{M}$
$\tilde{p}(\cdot \mid \cdot, \cdot)$	The state-transition kernel of $\tilde{\mathcal{M}}$
π_0	The known stabilizing policy of the queueing network
π_{rand}	The random policy that selects an action in $\mathcal{A}$ uniformly
π^*	The stationary policy that minimizes the expected average total queue backlog of $\mathcal{M}$
$\tilde{\pi}^*$	The stationary policy that minimizes the expected average total queue backlog of $\tilde{\mathcal{M}}$
ρ^*	The expected average queue backlog of $\mathcal{M}$ when applying π^*
$\tilde{\rho}^*$	The expected average queue backlog of $\tilde{\mathcal{M}}$ when applying $\tilde{\pi}^*$
$\tilde{\Phi}^*(\cdot)$	The Lyapunov function with negative drift regarding $\tilde{\pi}^*$
β	The order of $\tilde{\Phi}^*(\cdot)$
$\mathcal{S}^{\text{in}}$	The set of $\boldsymbol{Q}$'s with $\tilde{\Phi}^*(\boldsymbol{Q}) \leqslant (U - W)^\beta$
$\mathcal{S}^{\text{out}}$	$\mathcal{S} \setminus \mathcal{S}^{\text{in}}$
$p^{\pi+\pi'}(\cdot)$	The stationary probability distribution in $\mathcal{M}$ with π applied to $\mathcal{S}^{\text{in}}$ and π' applied to $\mathcal{S}^{\text{out}}$
$\mathbb{E}_{\pi}[\cdot]$	The expection of a random variable in $\mathcal{M}$ with π applied to $\mathcal{S}$
$\mathbb{E}_{\pi+\pi'}[\cdot]$	The expection of a random variable in $\mathcal{M}$ with π applied to $\mathcal{S}^{\text{in}}$ and π' applied to $\mathcal{S}^{\text{out}}$
$\tilde{\mathbb{E}}_{\tilde{\pi}}[\cdot]$	The expection of a random variable in $\tilde{\mathcal{M}}$ with $\tilde{\pi}$ applied to $\tilde{\mathcal{S}}$

the learning process, we tend to explore the system to collect samples, and thus ϵ_i is relatively large when i is small. As the learning process goes on, we gradually obtain enough samples and are close to the optimal policy, and thus ϵ_i is decreased to exploit the learned policy. The scheme has been applied in network control [2, 23, 55] to achieve a tradeoff between exploration and exploitation.

For an exploration episode, we apply a randomized policy π_{rand} that takes action uniformly to obtain samples for the estimation of state-transition probabilities in $\mathcal{M}$. However, since $\mathcal{M}$ has a countably infinite state space, we instead estimate the model for an auxiliary system $\tilde{\mathcal{M}}$ with a bounded state space. The auxiliary system $\tilde{\mathcal{M}}$ has a threshold U: the system has the same dynamics as the real one, with the only difference that each queue has a bounded buffer size U. For each queue in $\tilde{\mathcal{M}}$, when its queue backlog reaches U, new packets to it will be dropped. Mathematically, the state space of $\tilde{\mathcal{M}}$ is given by $\tilde{\mathcal{S}} \triangleq \{\boldsymbol{Q} \in \mathcal{S} : Q_{\max} \leqslant U\}$ where $Q_{\max} \triangleq \max Q_i$. The auxiliary system $\tilde{\mathcal{M}}$ shares the same action space $\mathcal{A}$ and cost function $c(\boldsymbol{Q})$ as $\mathcal{M}$. With the introduction



(a) Exploration (b) Exploitation

Fig. 1. Schemetic illustration of our approach (when $D = 2$).

of $\tilde{\mathcal{M}}$, an exploration episode operates as in Figure 1(a): we only apply π_{rand} to states in $\tilde{\mathcal{S}}$, while simply applying the *known* stabilizing policy π_0 to other states.

For an exploitation episode, we first repartition the state space $\mathcal{S}$ due to technical reasons. We denote by $\tilde{\pi}^*$ the optimal policy for $\tilde{\mathcal{M}}$, and define a Lyanopuv function $\tilde{\Phi}^*(\cdot)$ (cf. Assumption 2). Denoting by β the order of $\tilde{\Phi}^*(\cdot)$, we partition $\mathcal{S}$ in the following manner:

$$\begin{cases} \mathcal{S}^{\text{in}} \triangleq \left\{ Q : \tilde{\Phi}^*(Q) \leqslant (U - W)^\beta \right\} \\ \mathcal{S}^{\text{out}} \triangleq \mathcal{S} \setminus \mathcal{S}^{\text{in}} \end{cases}.$$

As illustrated in Figure 1(b), during an exploitation episode, we first compute an estimated $\tilde{\pi}^*$ using the estimated dynamics obtained from exploration episodes. We then apply the estimated $\tilde{\pi}^*$ to states in $\mathcal{S}^{\text{in}}$ and π_0 to states in $\mathcal{S}^{\text{out}}$ throughout the episode.

We will show that the average queue backlog under our algorithm converges to the optimal average queue backlog ρ^* as we increase U. We divide the analysis into two stages: before and after $\tilde{\pi}^*$ is learned. For the first stage, our model-based RL approach applies ϵ-greedy exploration. We show that the proposed algorithm gradually obtains $\tilde{\pi}^*$ with high probability. For the second stage, by applying drift analysis on the Markov chain, we show that when $\tilde{\pi}^*$ is applied for states in $\mathcal{S}^{\text{in}}$ and π_0 for states in $\mathcal{S}^{\text{out}}$, the probability of queue backlog exceeding into $\mathcal{S}^{\text{out}}$ decays exponentially with U. In addition, whenever Q leaves $\mathcal{S}^{\text{in}}$, the expected accumulated queue backlog before Q returns back to $\mathcal{S}^{\text{in}}$ can be upper bounded as a polynomial term in U. Together, we show that the gap between our result and the optimal average queue backlog ρ^* is upper bounded by $O(\text{poly}(U)/\exp(U))$, which diminishes as U increases.

3.2 Assumptions

Our algorithm can be applied to a broad class of network problems. To establish rigorous performance guarantee, we need to impose some assumptions on $\mathcal{M}$ and $\tilde{\mathcal{M}}$. We will discuss these assumptions and argue that they are reasonable under many queueing networks.

Assumption on π_0: As introduced in Section 3.1, we only learn the optimal policy for an auxiliary system with a bounded state space, while for the rest of the states, we apply a *known* stabilizing policy π_0 to control the queue backlog. We quantify the stability of π_0 as follows:

Assumption 1. *There exists a* known *policy π_0, a Lyapunov function $\Phi_0 : \mathcal{S} \to \mathbb{R}_+$ and constants $a, \alpha, \epsilon_0, B_0 > 0$ such that the following properties hold:*

(1) $\Phi_0(\boldsymbol{Q}) \leqslant a Q_{\max}^{\alpha}$ *for each* $\boldsymbol{Q} \in \mathcal{S}$;
(2) $\mathbb{E}_{\pi_0}[\Phi_0(\boldsymbol{Q}(t+1)) - \Phi_0(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t) = \boldsymbol{Q}] \leqslant -\epsilon_0$ *for each* $\boldsymbol{Q}$ *such that* $Q_{\max} \geqslant B_0$.

The requirement of MDP respecting Lyapunov function under a stable policy is not restrictive. For a stabilizable stochastic network, there exists a policy π_0 under which the corresponding Markov chain is positive recurrent. Moreover, the property of positive recurrence is known to be equivalent to the existence of the so-called Lyapunov function [41]. These observations motivate us to focus on MDPs that satisfy Assumption 1. In practice, for stabilizable networks, numerous stable policies have been proposed, including dynamic server allocation problems [5, 15, 58], multiclass routing networks [9, 10, 26, 32, 33], switch scheduling [29, 53, 58], and inventory control [19, 44], which all satisfy Assumption 1.

Assumption on $\tilde{\pi}^*$: From the outline of performance analysis in Section 3.1, we need to show that the probability of queue backlog entering $\mathcal{S}^{\text{out}}$ decays exponentially with respect to U. We therefore make a natural assumption on the stability property of $\tilde{\pi}^*$. This assumption is necessary for the proof of Lemma 2. For clarity, we use $\tilde{\mathbb{E}}_{\tilde{\pi}}[\cdot]$ to denote the expectation with respect to the randomness of the auxilary system $\tilde{\mathcal{M}}$ under the policy $\tilde{\pi}$ (to distinguish from $\mathbb{E}[\cdot]$ for $\mathcal{M}$).

Assumption 2. *There exist a Lyapunov function $\tilde{\Phi}^* : \tilde{\mathcal{S}} \to \mathbb{R}_+$ and constants $\beta \geqslant 1, b_1, b_2, \tilde{B}^*, \tilde{\epsilon}^* > 0$, such that for any $U > 0$, the following properties hold:*

(1) $Q_{\max}^{\beta} \leqslant \tilde{\Phi}^*(\boldsymbol{Q}) \leqslant b_1 Q_{\max}^{\beta}$ *for each* $\boldsymbol{Q} \in \tilde{\mathcal{S}}$;
(2) $\max_{\boldsymbol{Q}' \in \mathcal{R}(\boldsymbol{Q})} \left|\tilde{\Phi}^*(\boldsymbol{Q}') - \tilde{\Phi}^*(\boldsymbol{Q})\right| \leqslant b_2(Q_{\max}^{\beta-1} + Q_{\max}^{'\beta-1})$ *for each* $\boldsymbol{Q} \in \tilde{\mathcal{S}}$;
(3) $\tilde{\mathbb{E}}_{\tilde{\pi}^*}[\tilde{\Phi}^*(\boldsymbol{Q}(t+1)) - \tilde{\Phi}^*(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t) = \boldsymbol{Q}] \leqslant -\tilde{\epsilon}^* \cdot Q_{\max}^{\beta-1}$ *for each* $\boldsymbol{Q} \in \tilde{\mathcal{S}}$ *such that* $Q_{\max} \geqslant \tilde{B}^*$.

The assumption is natural in the sense that optimal policies are stable and thus are likely to have good Lyapunov drift properties. For instance, if $\tilde{\Phi}^*(\boldsymbol{Q}) = \sum_{i=1}^{D} \omega_i Q_i^2$ has a negative drift for each $\boldsymbol{Q}$ satisfying $\sum_{i=1}^{D} Q_i \geqslant B$, one can easily show that $\omega_{min} \cdot Q_{\max}^2 \leqslant \tilde{\Phi}^*(\boldsymbol{Q}) \leqslant (\sum_{i=1}^{D} \omega_i) \cdot Q_{\max}^2$, $\max_{\boldsymbol{Q}' \in \mathcal{R}(\boldsymbol{Q})} \left|\tilde{\Phi}^*(\boldsymbol{Q}') - \tilde{\Phi}^*(\boldsymbol{Q})\right| \leqslant W \cdot (\sum_{i=1}^{D} \omega_i) \cdot (Q_{\max} + Q'_{\max})$, and has a negative drift for each $\boldsymbol{Q}$ with $Q_{\max} \geqslant B$. Possible methods to analyze Lyapunov drift properties from queueing stability can be found in [42].

Assumption on communication properties: Existing model-based RL algorithms with performance guarantees usually require the "diameter" (i.e., the upper bound for the shortest first hitting time between any two states) of the MDP to be finite [27]. However, it is unrealistic to assume the MDP diameter to be finite when the state space is unbounded. Instead, we make the following assumption (this assumption is used in the proof of Lemma 1) where $T_{\boldsymbol{Q} \to \boldsymbol{Q}'}$ denotes the first hitting time from $\boldsymbol{Q}$ to $\boldsymbol{Q}'$.

Assumption 3. *There exist constants $c, \gamma > 0$, such that for any $U > 0$ and every $\boldsymbol{Q}, \boldsymbol{Q}' \in \tilde{\mathcal{S}}$,*

$$\min_{\tilde{\pi}} \tilde{\mathbb{E}}_{\tilde{\pi}} \left[T_{\boldsymbol{Q} \to \boldsymbol{Q}'}\right] \leqslant c \|\boldsymbol{Q}' - \boldsymbol{Q}\|_1^{\gamma},$$

where $\tilde{\pi}$ is a policy that can be applied to $\tilde{S}$.

In other words, Assumption 3 states that there exists a policy $\tilde{\pi}$ such that the first hitting time between two states $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ in $\tilde{\mathcal{S}}$ is a polynomial function of $\|\boldsymbol{Q}' - \boldsymbol{Q}\|_1$. We emphasize that

the constant γ is independent of the value of U. Indeed, with larger U's, $\tilde{\mathbb{E}}_{\tilde{\pi}}[T_{Q\to Q'}]$ might grow larger, since the trajectory from Q to Q' are more likely to be longer. However, it is possible that the increment is bounded. For instance, certain $M/G/1$ queueing systems [52] and random walk models [30] have constant γ's even if the state space is unbounded. Directly verifying Assumption 3 might be computationally difficult. Theorem 11.3.11 from [42] offers a drift analysis approach for justifying Assumption 3.

We remark that the maximum first passage time of the MDP plays an essential role in analyzing the RL algorithms in infinite-horizon average-reward MDPs. Existing related works either use MDP diameter [27, 68], or span [6, 21, 49], or mixing time [1, 47, 63] to characterize the maximum first passage time of the MDP, where these metrics serve as coefficients in the gap to the optimum. The analysis in our article also relies on the assumption of reasonable bounds for the maximum first passage time. It would be of great interest to establish performance guarantees under relaxed assumptions.

Assumption on error tolerance: As the learning process proceeds, ideally the estimation for $\tilde{\mathcal{M}}$ becomes increasingly accurate. However, it is unrealistic for us to obtain the exact $\tilde{\mathcal{M}}$. Therefore, we assume that if the estimates for the state-transition kernels are accurately enough (i.e., within a certain error bound), the optimal policy to the estimated $\tilde{\mathcal{M}}$ is the same as $\tilde{\pi}^*$. The assumption is stated as follows.

Assumption 4. *There exists a $\Delta p > 0$, such that for any MDP $\tilde{\mathcal{M}}'$ with the same state space, action space and cost function as $\tilde{\mathcal{M}}$, if for each $Q \in \tilde{\mathcal{S}}$ and each $a \in \mathcal{A}$, we have*

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

then the optimal policy for $\tilde{\mathcal{M}}'$ is the same as the optimal policy $\tilde{\pi}^*$ for $\tilde{\mathcal{M}}$.

In this article, we focus on network control problems with finite action space. For many networks, when the dynamics (e.g., exogenous arrival rates, service rates, and channel capacities) vary slightly, the optimal policies remain the same. For instance, if the arrival rates of the switching networks vary yet remain heavy loads, Maximum Matching still remains a close-to-optimal policy [37].

3.3 Algorithm

We propose an algorithm called **RL for Queueing Networks** (**RL-QN**). RL-QN operates in an episodic manner: at the beginning of episode k, we uniformly draw a real number $\xi \in [0, 1]$ to decide whether to explore or exploit during this episode. The length of each episode depends on the observations.

- If $\xi \leqslant \epsilon_k \triangleq \ell/\sqrt{k}$ (where $\ell \in (0, 1]$ can be tuned to control the exploration frequency), we perform exploration during this episode. For states in $\tilde{\mathcal{S}}$, we apply the randomized policy π_{rand}. For states in $\mathcal{S} \setminus \tilde{\mathcal{S}}$, we apply π_0.
- If $\xi > \epsilon_k$, we enter the exploitation stage. We first calculate sample means to estimate the state-transition function $\tilde{p}$ of $\tilde{\mathcal{M}}$. We then apply value iteration on the estimated system $\tilde{M}_k$ to obtain an estimated optimal policy $\tilde{\pi}_k^*$. During this episode, we apply $\tilde{\pi}_k^*$ for states in $\mathcal{S}^{\text{in}}$ and π_0 otherwise.
- When the number of visits to states in $\mathcal{S}^{\text{in}}$ exceeds $L_k = L \cdot \sqrt{k}$ (where L is a positive constant and can be tuned to adjust the sampling rate), RL-QN enters episode $k + 1$ and repeat the process above.

ALGORITHM 1: The RL-QN algorithm

Input: $\mathcal{A}, U > W, 0 < \ell \leqslant 1, L > 0, K > 0$
for episodes $k \leftarrow 1, 2, \ldots, K$ **do**
  Set $L_k \leftarrow L \cdot \sqrt{k}$, $\epsilon_k \leftarrow \ell/\sqrt{k}$ and uniformly draw $\xi \in [0, 1]$.
  if $\xi \leqslant \epsilon_k$ **then**
    Set
$$\pi_k(\boldsymbol{Q}) = \begin{cases} \pi_{\text{rand}}(\boldsymbol{Q}), & \text{for } \boldsymbol{Q} \in \tilde{\mathcal{S}} \\ \pi_0(\boldsymbol{Q}), & \text{for } \boldsymbol{Q} \in \mathcal{S} \setminus \tilde{\mathcal{S}} \end{cases}.$$
  else
    For each $\boldsymbol{Q}, \boldsymbol{Q}' \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$, let $\tilde{p}_k(\boldsymbol{Q}' \mid \boldsymbol{Q}, a) = \tilde{P}(\boldsymbol{Q}, a, \boldsymbol{Q}')/N(\boldsymbol{Q}, a)$ for $N(\boldsymbol{Q}, a) > 0$ and $\tilde{p}_k(\boldsymbol{Q}' \mid \boldsymbol{Q}, a) = 1/|\mathcal{R}(\boldsymbol{Q})|$ otherwise.
    Solve the estimated MDP $\tilde{M}_k$ and obtain the estimated optimal policy $\tilde{\pi}_k^*$.
    Set
$$\pi_k(\boldsymbol{Q}) = \begin{cases} \tilde{\pi}_k^*(\boldsymbol{Q}), & \text{for } \boldsymbol{Q} \in \mathcal{S}^{\text{in}} \\ \pi_0(\boldsymbol{Q}), & \text{for } \boldsymbol{Q} \in \mathcal{S}^{\text{out}} \end{cases}.$$
  end if
  while visits to states in $\mathcal{S}^{\text{in}}$ is smaller than L_k **do**
    Take the action $a_t = \pi_k(\boldsymbol{Q}(t))$ for the real system.
    Observe the next state $\boldsymbol{Q}(t+1)$.
    if $\boldsymbol{Q}(t) \in \tilde{\mathcal{S}}$ **then**
      Increase $N(\boldsymbol{Q}(t), a_t)$ by 1.
      Increase $\tilde{P}(\boldsymbol{Q}(t), a_t, \min\{U, \boldsymbol{Q}\})$ by 1.
    end if
    $t \leftarrow t + 1$.
  end while
end for
Output: estimated optimal policy $\tilde{\pi}_K^*$

The details are presented in Algorithm 1. We use $\min\{U, \boldsymbol{Q}\}$ to denote a vector with the ith coordinate being $\min\{U, Q_i\}$.

4 PERFORMANCE ANALYSIS

We analyze the performance of our algorithm from both exploration and exploitation perspectives, under Assumptions 1–4. We first prove that RL-QN can learn $\tilde{\pi}^*$ within finite episodes with high probability, which implies that RL-QN explores different states sufficiently to obtain an accurate estimation of $\tilde{\mathcal{M}}$ (cf. Theorem 1). We then show that RL-QN exploits the estimated optimal policy and achieves a performance close to the optimal result of ρ^* (cf. Theorem 2).

In this article, we focus on MDPs such that all states are accessible from each other under the following policies: (a) $\pi_{\text{rand}} + \pi_0$: applying π_{rand} to $\tilde{\mathcal{S}}$ and π_0 to $\mathcal{S} \setminus \tilde{\mathcal{S}}$; (b) $\tilde{\pi}^* + \pi_0$: applying $\tilde{\pi}^*$ to $\mathcal{S}^{\text{in}}$ and π_0 to $\mathcal{S}^{\text{out}}$; and (c) π^* : applying (a truncated version of) π^* to $\tilde{\mathcal{S}}$ in $\tilde{\mathcal{M}}$. That is, the corresponding Markov chains under the above policies are irreducible.

4.1 Convergence to the Optimal Policy

The following theorem states that, with arbitrarily high probability, RL-QN learns $\tilde{\pi}^*$ within a finite number of episodes.

Theorem 1. *Suppose Assumption 4 holds. For each $\delta \in (0,1)$, there exists $k^* < \infty$ such that* RL-QN *learns $\tilde{\pi}^*$ (i.e. $\tilde{\pi}^*_{k^*} = \tilde{\pi}^*$) within k^* episodes with probability at least $1-\delta$.*

Proof. We give the detailed proof in Appendix A. Let us now sketch the main steps in our analysis. By analyzing the mixing property of the underlying Markov chain of applying π_{rand} to $\tilde{\mathcal{S}}$ and π_0 to $\mathcal{S} \setminus \tilde{\mathcal{S}}$, we first show that there exists a constant K_0 such that for episode $k \geqslant K_0$, we are able to obtain $\Theta(\sqrt{k})$ samples for each (Q,a) pair.

We then use a result in [64] to compute the number of samples required for each (Q,a) pair to ensure $\tilde{\pi}^*_k = \tilde{\pi}^*$, under Assumption 4. Finally, we show that by choosing the exploration probability of $\epsilon_k = \ell/\sqrt{k}$, exploration episodes occur infinitely often, which implies that we can reach the number of required exploration episodes within finite number of episodes. □

Theorem 1 indicates that RL-QN explores (i.e., samples) state-transition functions of each state-action pair (Q,a) in $\tilde{\mathcal{M}}$ sufficiently.

4.2 Gap to Optimum

Theorem 1 states the sufficient exploration aspect of RL-QN. In RL, the tradeoff between exploration and exploitation is of significant importance to the algorithm performance. In this section, we show that RL-QN also exploits the learned policy such that the episodic average queue backlog is bounded and can get arbitrarily close to ρ^*, the optimal average queue backlog of $\mathcal{M}$, as we increase U.

We denote the timestep at the end of episode k by t_k and the actual length of episode k by L'_k, i.e., $L'_k = t_k - t_{k-1}$ with $t_0 \triangleq 0$. We use π^{in}_k to represent the policy applied to $\mathcal{S}^{\text{in}}$ during episode k and $p^{\tilde{\pi}+\pi_0}(\cdot)$ to denote the stationary distribution of states when applying $\tilde{\pi}$ to states in $\mathcal{S}^{\text{in}}$ and π_0 to states in $\mathcal{S}^{\text{out}}$.

By Theorem 1, RL-QN learns $\tilde{\pi}^*$ with high probability. Note that as the exploration probability decays by $1/\sqrt{k}$, the probability of utilizing the learned policy converges to 1 as the episodes increase. Hence, the episodic average queue backlog when $\pi^{\text{in}}_k = \tilde{\pi}^*$ constitutes a large proportion of the overall expected average queue backlog. Therefore, the key step to upper bound the expected average queue backlog is to upper bound the episodic average queue backlog when $\pi^{\text{in}}_k = \tilde{\pi}^*$. We prove that it can be upper bounded with respect to $\tilde{\rho}^*$, the optimal average queue backlog of $\tilde{\mathcal{M}}$, as stated in Lemma 1. We first define $\mathcal{S}^{\text{in}}_{\text{bd}}$ as the set of states in $\mathcal{S}^{\text{in}}$ that is possible to exit into $\mathcal{S}^{\text{out}}$ at the next time slot, i.e., $\mathcal{S}^{\text{in}}_{\text{bd}} \triangleq \{Q \in \mathcal{S}^{\text{in}} : \mathcal{R}(Q) \cap \mathcal{S}^{\text{out}} \neq \emptyset\}$.

Lemma 1. *Under Assumptions 1–3, we have*

$$\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\sum_i Q_i(t)}{L'_k}\right] = \tilde{\rho}^* + p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}^{\text{in}}_{\text{bd}}\right)\cdot O\left(U^{1+\max\{2\alpha,\gamma\}}\right).$$

Proof. We first define the accumulated regret regarding $\tilde{\rho}^*$ for a given episode k with $\pi^{\text{in}}_k = \tilde{\pi}^*$ as $\sum_{t=t_{k-1}+1}^{t_k}(\sum_i Q_i(t) - \tilde{\rho}^*)$. We then define $\mathcal{S}^{\text{in}}_{\text{in}} \triangleq \mathcal{S}^{\text{in}} \setminus \mathcal{S}^{\text{in}}_{\text{bd}}$ and decompose the expected accumulated regret into three parts according to state position: (i) $Q \in \mathcal{S}^{\text{in}}_{\text{in}}$, (ii) $Q \in \mathcal{S}^{\text{in}}_{\text{bd}}$, and (iii) $Q \in \mathcal{S}^{\text{out}}$.

For the first part (i), we use Bellman equation analysis to show that, every time Q enters $\mathcal{S}^{\text{in}}_{\text{in}}$ from $\mathcal{S}^{\text{in}}_{\text{bd}}$ and returns back to $\mathcal{S}^{\text{in}}_{\text{bd}}$, the expected accumulated regret is upper bounded by the span of the solutions to the Bellman equations. We then use Proposition 5.5.1 in [7] to obtain a polynomial upper bound for the span of the solution to the Bellman equation under Assumption 3.

The second part (ii) is trivially upper bounded by DU.

For the third part (iii), we prove that the time it takes to return back $\mathcal{S}^{\text{in}}$ is polynomial in U under Assumption 1 (using techniques as the proof of Theorem 1.1 in Chapter 5 of [12]). Therefore, by Theorem 6.3.4 in [25], every time $\boldsymbol{Q}$ enters $\mathcal{S}^{\text{out}}$ from $\mathcal{S}^{\text{in}}_{\text{bd}}$ and returns back to $\mathcal{S}^{\text{in}}_{\text{bd}}$, the incurred expected accumulated regret is upper bounded by a polynomial function of U.

The detailed proof is given in Appendix B. □

We then proceed to upper bound $p^{\tilde{\pi}^*+\pi_0}(\mathcal{S}^{\text{in}}_{\text{bd}})$, as stated in the following lemma.

Lemma 2. *Under Assumption 2, we have*

$$p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}^{\text{in}}_{\text{bd}}\right) = O(\exp(-U)).$$

Proof. We show that under Assumption 2, we can construct a linear Lyapunov function with a negative drift for states with large $\boldsymbol{Q}_{\max}$. By applying similar techniques as Lemma 1 in [8], we establish an upper bound for the tail probability of Lyapunov values, which decays exponentially.

The detailed proof is given in Appendix C. □

With Theorem 1, Lemmas 1 and 2, we establish the following main result of this article.

Theorem 2. *Suppose Assumptions 1–4 hold. When applying* RL-QN *to $\mathcal{M}$, the asympototic episodic average queue backlog is upper bounded as follows:*

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\sum_i Q_i(t)}{L'_k}\right] \le \rho^* + O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right). \tag{1}$$

Proof. With Lemmas 1 and 2, we show that when $k \to \infty$, for each episode with $\pi^{\text{in}}_k = \tilde{\pi}^*$, the expected average queue backlog is upper bounded by $\tilde{\rho}^* + O(U^{1+\max\{2\alpha,\gamma\}}/\exp(U))$. We next show that the expected average queue backlog contributed by episodes where $\pi^{\text{in}}_k \neq \tilde{\pi}^*$ becomes negligible as $k \to \infty$. By applying similar techniques as the proof of Lemmas 1 and 2, we then obtain that $\tilde{\rho}^* = \rho^* + O(U^{1+\gamma}/\exp(U))$.

The detailed proof is given in Appendix D. □

Theorem 2 provides an upper bound on the performance guarantee of RL-QN with respect to the threshold parameter U: by increasing U, the long-term episodic average queue backlog approaches ρ^* exponentially fast. Recall that the episodic length L'_k increases to ∞ as $k \to \infty$. We conjecture that the same upper bound hold for the overall average queue backlog regarding the time horizon T, i.e.,

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^{T}\sum_i Q_i(t)\right]}{T} \le \rho^* + O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right). \tag{2}$$

We note that the result of Theorem 2 does not imply Equation (2) directly. A rigorous proof of the conjecture (2) seems difficult with current techniques. We leave as an interesting future direction to investigate if Equation (2) holds.

4.3 Complexity Analysis

Here, we present the complexity analysis. Our algorithm requires solving an estimated MDP for each exploitation episode. Since episode k has length of at least $L \cdot \sqrt{k}$, and

$$\sum_{k=1}^{K} \sqrt{k} \geqslant \frac{2}{3}K^{\frac{3}{2}},$$

we have that given a time horrzon T, the number of episodes is upper bounded as

$$K_T \leqslant \left(\frac{3T}{2L}\right)^{\frac{2}{3}}.$$

It has been shown that to solve a general MDP, the time complexity is at least polynomial in the size of the state space [35]. The concrete time complexity depends on the solution method and its parameters. In our setting, the state space $\mathcal{S}^{in}$ has a size of $\Theta(U^N)$. Therefore, the time complexity of our algorithm is

$$O\Big(\big(T/L\big)^{\frac{2}{3}} \cdot \text{poly}\big(U^N\big)\Big).$$

Therefore, for a given system with fixed number of nodes, the time complexity grows at most polynomially in the threshold U. When applying our algorithm to problems of larger scales (larger N), the complexity suffers from "curse of dimensionality". However, our algorithm remains computationally feasible in practice for the following reasons. First, although the "curse of dimensionality" persists, our algorithm substantially simplifies the original MDP with unbounded state space to an MDP with finite state space. For instance, the system in Section 5.1 cannot be optimized by traditional methods but is optimized efficiently under our algorithm. Second, we only require solving the estimated MDP sparsely, i.e., no more than $(3T/2L)^{2/3}$ times. Choosing a relatively large L can greatly reduce the time complexity. Third, solving the MDP is independent of other steps of our algorithm, which allows us to employ appropriate methods to solve MDPs according to different application scenarios, computational capacities and performance requirements. For instance, since we aim at optimizing the average cost of the MDP, undiscounted value iteration/policy iteration methods should be applied. However, we apply discounted methods in our numerical experiments and obtain the same solution with significantly less computation time. Also, we may utilize the special structure of the MDP [24, 31] or apply approximation methods [16, 17, 60] to alleviate the computational cost. Our numerical results in Section 5.1 further validates the conclusion and shows that our algorithm is computationally feasible in practice.

5 NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of RL-QN for three classes of queueing systems: server allocation, routing, and switching.

5.1 Server Allocation (Two Nodes)

We consider a dynamic server allocation problem: exogenous packets arrive to two nodes according to Bernoulli process with rate λ_1 and λ_2, respectively. At each time slot, the central server selects one node to serve. The head of line job in the selected queue i completes the required service and leaves the system with probability p_i. The system model and parameters are illustrated in Figure 2.

According to [58], whenever the condition

$$\frac{\lambda_1}{p_1} + \frac{\lambda_2}{p_2} < 1,$$

is satisfied, one stabilizing policy is to always serve the node with the **longest queue** (**LQ**). Therefore, we can use LQ policy as π_0. To evaulate whether this problem satisfies Assumption 3, we apply π_0 to the truncated state space $\tilde{\mathcal{S}}$ with $U = 10$ and simulate the system to collect the hitting times, as shown in Figure 3. We can see that as the state distance grows, the average first hitting time grows sublinearly, which is obviously upper bounded by linear growth and thus indicates that Assumption 3 is satisfied. Moreover, our numerical experiments show that our algorithm can

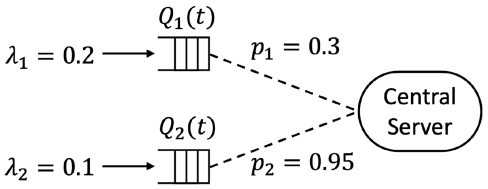


Fig. 2. System model of the dynamic server allocation model with two nodes.

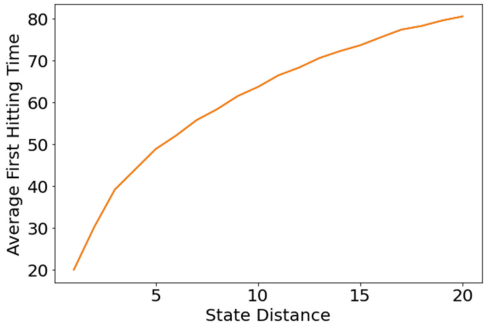


Fig. 3. Simulation results for the dynamic server allocation model of two nodes.

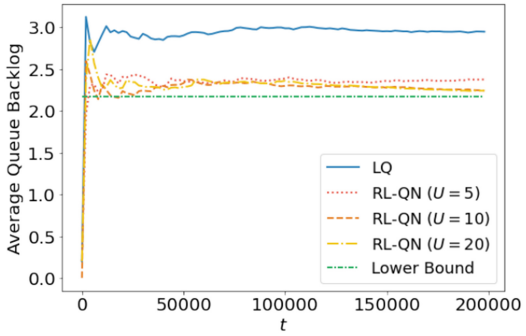


Fig. 4. Average queue backlog under different policies for the server allocation model with two nodes.

optimize the average queue backlog for various settings even if we cannot provably verify Assumption 3. Therefore, the applicability of our algorithm may not be sensitive to Assumption 3, which helps to show that our algorithm is practical.

We simulate RL-QN for U = 5, U = 10, and U = 20. The results are shown in Figure 4. To obtain a lower bound on the average queue backlog, we assume that the system dynamics, i.e., arrival rates and success rates, are known. We then optimize the MDP and obtain its average queue backlog. This average queue backlog is guaranteed to be a lower bound since, in practice, the system dynamics are unknown and errors that occur during the learning process can downgrade the performance. From Figure 4, we see that the LQ policy stabilizes the queue backlog, yet its average queue backlog is far from the lower bound. All of our RL-QN methods outperform π_0. When U = 5, the average queue backlog converges to 2.38, while for U = 10 and U = 20, the average queue backlog becomes 2.24. This indicates that RL-QN achieves better performance with a larger threshold parameter U, as implied by Theorem 2. Moreover, since the cases with U = 10 and U = 20 achieve similar performance, it is very likely that in practice a small U achieves satisfactory performance, which makes our algorithm more computationally feasible.

We then study the time complexity of our algorithm. We apply three different MDP solvers in our algorithm: value iteration, relative value iteration, and policy iteration, all with discount factor 0.99. Figure 5 shows the relationship between the relative computation time (i.e., the actual computation time of RL-QN divided by the actual computation time of LQ) and the average queue backlog. Even when $U = 20$, the computation time is less than three times of the LQ policy, which shows that increasing U has relatively small impact on the computation time. From Figure 6, we

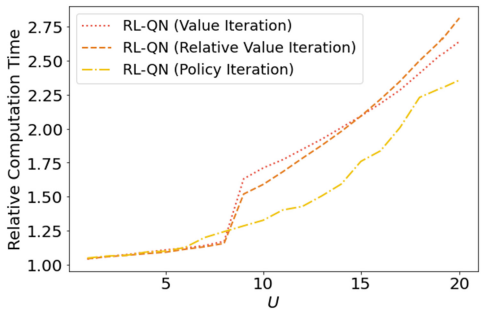


Fig. 5. The relationship between U and the relative computation time.

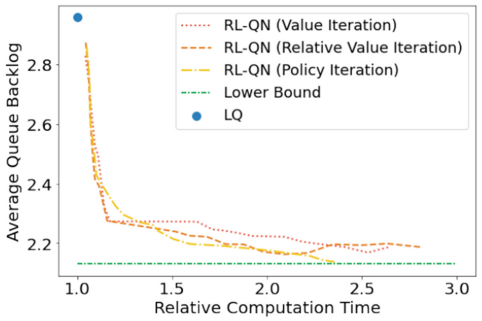


Fig. 6. The relationship between the relative computation time and the average queue backlog.

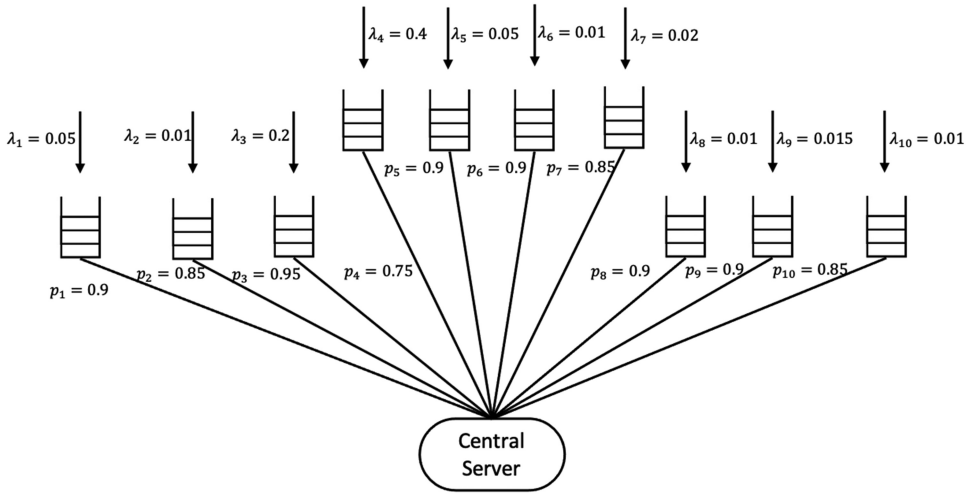


Fig. 7. System model of the dynamic server allocation model with 10 nodes.

see that the average queue backlog drops significantly even with a relatively small U. The average queue backlog is reduced by 23.1% when $U = 8$, while the computation time is only 1.17 times of the traditional non-ML LQ policy. Therefore, our algorithm is computationally efficient in practice.

5.2 Server Allocation (10 Nodes)

We then consider a dynamic server allocation problem with greater scale: exogenous packets arrive to ten nodes according to Bernoulli process with rate λ_i, $i = 1, \ldots, 10$ respectively. At each time slot, the central server selects one node to serve. The head of line job in the selected queue i completes the required service and leaves the system with probability p_i. The system model and parameters are illustrated in Figure 7.

As discussed in Section 3.1, if we directly apply classical methods to solve the estimated MDP, the computational complexity is at least $\Theta(U^N)$, which grows exponentially with the number of nodes N. Therefore, for the ten-node dynamic server allocation model, we prefer to utilize the special structures of the MDP or apply approximation methods to reduce the computational cost.

It has been shown in [15] that when all nodes are always connected to the central server, the policy to minimize the average job delay is to serve the node with the largest service rate p_i among

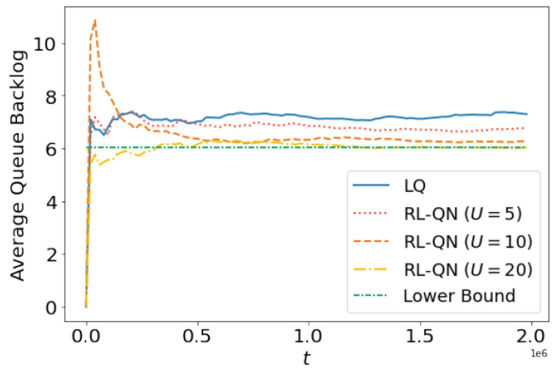


Fig. 8. Simulation results for the server allocation model with 10 nodes.

all non-empty nodes. Therefore, for each exploitation episode, instead of solving the estimated MDP $\tilde{\mathcal{M}}_k$ with classical methods, we directly obtain the optimal solution: serving the node with the largest estimated p_i among all non-empty nodes. We emphasize that our algorithm does not depend on how the estimated MDPs are solved. The above computational trick utilizes the special structure of the server allocation model to speed up the computation, and other techniques for efficiently solving the MDPs can also potentially be employed.

The results are shown in Figure 8. From the figure, we observe that RL-QN outperforms LQ. The LQ reaches an average queue backlog of 7.6, while RL-QN reaches 7 when $U = 5$. When U is beyond 10, RL-QN reaches an average queue backlog of 6, which is close to the optimal result.

5.3 Routing

We consider a simple routing problem: exogenous packets arrive at the source node s according to Bernoulli process with rate $\lambda = 0.85$. Nodes 1 and 2 are two intermediate nodes and can serve at most one packet during each time slot, with probability p_1 and p_2, respectively. Node d is the destination node. At each time slot, node s has to choose between routes $s \to 1 \to d$ and $s \to 2 \to d$ to transit new exogenous packets. Specifically, the system model and parameters are shown in Figure 9.

The parameters (p_1, p_2) are queue-dependent here:

$$(p_1, p_2) = \begin{cases} (0.9, 0.1), & Q_2(t) \leqslant 5 \\ (0.1, 0.9), & Q_2(t) > 5 \end{cases}.$$

For each $\lambda < 0.9$, an intuitive stabilizing policy is to always use the fixed path $s \to 1 \to d$, while never choose $s \to 2 \to d$. Therefore, we can use the policy that always routes through $s \to 1 \to d$ as π_0. However, it is possible that we could split the external arrivals into the two routes to fully utilize the service capacities of both nodes 1 and 2, and achieve better performance.

We simulate RL-QN for $U = 10$. The results are plotted in Figure 10, which shows that RL-QN outperforms the fixed path stabilizing policy and converges to the optimum quickly.

5.4 Switching

We consider a 2×2 input-queued switch as illustrated in Figure 11. Data packets arriving at input i destined for output j are stored at input port i, in queue $Q_{i,j}$, thus there are four queues in total. We consider the case where the new data packets are arriving at queue (i, j) at rate $\lambda_{i,j}$ for $1 \leq i, j \leq 2$, according to a Bernoulli process. That is, for each time slot, the number of packets arriving at queue

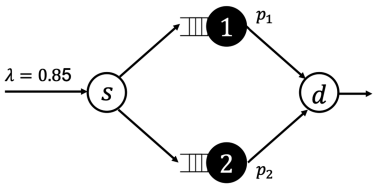


Fig. 9. System model of the routing network with two nodes.

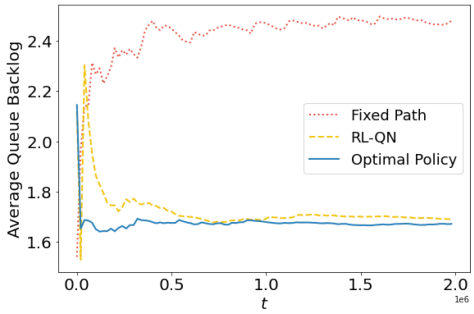


Fig. 10. Simulation results of the routing network with two nodes.

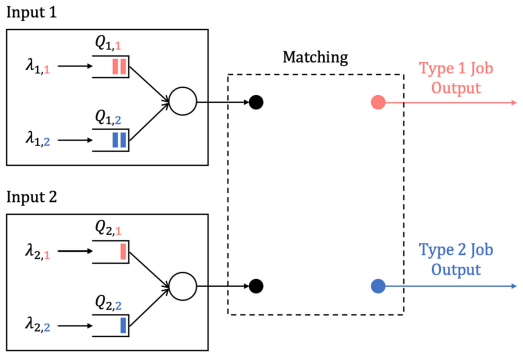
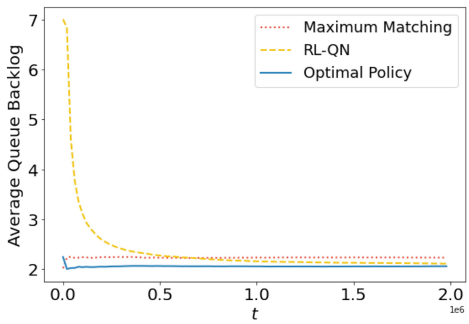
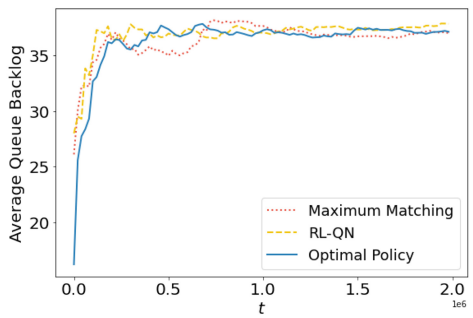


Fig. 11. System model of a 2×2 input-queued cell-switch.



(a) $\lambda_{1,1} = \lambda_{1,2} = 0.4$, $\lambda_{2,1} = 0.2$, $\lambda_{2,2} = 0.1$



(b) $\lambda_{1,1} = \lambda_{1,2} = \lambda_{2,1} = \lambda_{2,2} = 0.49$

Fig. 12. Simulation results for the input-queued switch model.

$Q_{i,j}$ is a Bernoulli random variable with mean $\lambda_{i,j}$. The server then selects a matching between the inputs and outputs to transmit packets. If input i is connected with output j, then a buffered packet is removed from the input queue $Q_{i,j}$ and sent to output j.

According to [39], whenever the condition that $\sum_i \lambda_{i,j} < 1, \sum_j \lambda_{i,j} < 1$ is satisfied, the Maximum Matching algorithm, which selects the matching that maximizes the total queue backlog of the connected channels is stabilizing; hence, we use it as π_0.

We implement the simulation for $U = 5$ under two different settings of arrival traffic. The results are shown in Figure 12. When $\lambda_{1,1} = \lambda_{1,2} = 0.4$, $\lambda_{2,1} = 0.2$, $\lambda_{2,2} = 0.1$, we can see that our algorithm outperforms the stabilizing policy π_0 and converges to $\tilde{\pi}^* + \pi_0$. When $\lambda_{1,1} = \lambda_{1,2} = \lambda_{2,1} = \lambda_{2,2} = 0.49$, the system is under heavy traffic. In [37], it was shown that the Maximum Matching policy π_0 is close to the optimal policy that minimizes the average queue backlog. Our simulation result (Figure 12(b)) indicates that our algorithm achieves comparable performance as the near-optimal policy π_0.

6 CONCLUSION

In this work, we apply a model-based RL framework to general queueing networks with unbounded state spaces. We propose the RL-QN algorithm, which applies an ϵ-greedy exploration scheme. By leveraging Lyapunov analysis, we show that the average queue backlog of the proposed approach can get arbitrarily close to the optimal average queue backlog under the optimal policy. The proposed RL-QN algorithm requires the knowledge of a stable policy. An interesting future direction is to investigate this problem when such information is not available.

APPENDICES

A PROOF OF THEOREM 1

In this section, we prove Theorem 1. As outlined in Section 4.1, our proof consists of three steps, which are presented in the subsections to follow.

A.1 Sufficient Exploration for $\tilde{\mathcal{S}} \times \mathcal{A}$

For each $(Q, a) \in \tilde{\mathcal{S}} \times \mathcal{A}$, we denote by $N_k(Q, a)$ the number of times that (Q, a) is encountered during episode k. For simplicity, we use $\tilde{\pi}_k$ to denote the policy applied to $\tilde{\mathcal{S}}$ during episode k and $p^{\text{rand}}(\cdot)$ to denote the stationary distribution of states when applying π_{rand} to states in $\tilde{\mathcal{S}}$ and π_0 to states in $\mathcal{S} \setminus \tilde{\mathcal{S}}$. The following lemma illustrates that after a certain number of episodes, π_{rand} samples each $(Q, a) \in \tilde{\mathcal{S}} \times \mathcal{A}$ sufficiently with a relatively large probability (e.g., greater than 1/2).

Lemma 3. *Under the setting of Theorem 1 and Algorithm 1, there exists $K_0 > 0$ such that for each $k \geqslant K_0$,*

$$\Pr\left\{ \frac{N_k^{\text{rand}}(Q, a)}{L_k} \leqslant \frac{p^{\text{rand}}(Q)}{2|\mathcal{A}|}, \forall Q \in \tilde{\mathcal{S}}, \forall a \in \mathcal{A} \right\} \geqslant \frac{1}{2}.$$

Proof. Here, we only consider the episodes that $\tilde{\pi}_k = \pi_{\text{rand}}$. (That is, in this proof episode k is understood as the episode in which the policy π_{rand} is executed for the kth time.) Recall that under the policy that applies π_{rand} to states in $\tilde{\mathcal{S}}$ and π_0 to states in $\mathcal{S} \setminus \tilde{\mathcal{S}}$, the corresponding Markov chain is positive recurrent. For each $Q \in \mathcal{S}$, define $N_k^{\text{rand}}(Q)$ as the number of times that Q is visited during episode k. For an irreducible positive recurrent Markov chain a on countable state space, we have the mixing property that for every $Q \in \mathcal{S}$,

$$\lim_{k \to \infty} \frac{N_k^{\text{rand}}(Q)}{L_k'} = p^{\text{rand}}(Q) \qquad w.p.1. \tag{3}$$

Note that under $\pi_{\text{rand}} + \pi_0$, for every $Q \in \tilde{\mathcal{S}}$, we take each $a \in \mathcal{A}$ with equal probability $1/|\mathcal{A}|$. According to strong law of large number, we have

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

for every $Q \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$.

Since both Equations (3) and (4) converge to constants, the multiplication rule for limit holds. That is, for every $Q \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$, we have

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

Note that almost sure convergence implies convergence in probability. Hence, for each $(Q,a) \in \tilde{\mathcal{S}} \times \mathcal{A}$ and $\epsilon, \xi > 0$, there exists a finite constant such that for each k larger than the constant,

$$\Pr\left\{\left|\frac{N_k^{\text{rand}}(Q,a)}{L'_k} - \frac{p^{\text{rand}}(Q)}{|\mathcal{A}|}\right| \geqslant \epsilon\right\} \leqslant \xi.$$

Since $L'_k \geq L_k$, we have

$$\begin{aligned}\Pr\left\{\frac{N_k^{\text{rand}}(Q,a)}{L_k} \leqslant \frac{p^{\text{rand}}(Q)}{2|\mathcal{A}|}\right\} &\leqslant \Pr\left\{\frac{N_k^{\text{rand}}(Q,a)}{L'_k} \leqslant \frac{p^{\text{rand}}(Q)}{2|\mathcal{A}|}\right\} \\ &\leqslant \Pr\left\{\left|\frac{N_k^{\text{rand}}(Q,a)}{L'_k} - \frac{p^{\text{rand}}(Q)}{|\mathcal{A}|}\right| \geqslant \frac{p^{\text{rand}}(Q)}{2|\mathcal{A}|}\right\}.\end{aligned}$$

By setting $\epsilon = p^{\text{rand}}(Q)/(2|\mathcal{A}|)$, $\xi = 1/(2|\tilde{\mathcal{S}}||\mathcal{A}|)$ and taking a union bound over $\tilde{\mathcal{S}}$ and $\mathcal{A}$, we have that there exists some constant $K_0 < \infty$ such that for each $k \geqslant K_0$,

$$\Pr\left\{\frac{N_k^{\text{rand}}(Q,a)}{L_k} \leqslant \frac{p^{\text{rand}}(Q)}{2|\mathcal{A}|}, \forall Q \in \tilde{\mathcal{S}} \text{ and } \forall a \in \mathcal{A}\right\} \leqslant \frac{1}{2\left|\tilde{\mathcal{S}}\right||\mathcal{A}|} \cdot \left|\tilde{\mathcal{S}}\right| \cdot |\mathcal{A}| = \frac{1}{2}.$$

We complete the proof. □

A.2 Sample Requirement for Learning $\tilde{\pi}^*$

Define $R \triangleq \max_{Q\in\tilde{\mathcal{S}}, a\in\mathcal{A}} |\mathcal{R}(Q,a)|$. We have the following lemma on the number of samples required for each $(Q,a) \in \tilde{\mathcal{S}}, a \in \mathcal{A}$ to estimate the model of the auxiliary system $\tilde{\mathcal{M}}$ sufficiently accurate.

Lemma 4. *Given $\delta > 0$, if for each $(Q,a) \in \tilde{\mathcal{S}} \times \mathcal{A}$, the number of samples $N(Q,a)$ satisfies*

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

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

Proof. Our proof is based on the following lemma from [64].

Lemma 5 (Theorem 2.1 in [64]). *For a probability distribution p over n_1 distinct events, we obtain the empirical distribution $\hat{p}$ based on n_2 samples from p. Then the L^1- deviation between p and $\hat{p}$ is*

upper bounded as

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

In our case, for a given $(\boldsymbol{Q},a)\in\tilde{\mathcal{S}}\times\mathcal{A}$, the true distribution over next state $\mathcal{R}(\boldsymbol{Q})$ is given by $\tilde{p}\left(\cdot\mid\boldsymbol{Q},a\right)$. We denote the empirical distribution by $\hat{p}\left(\cdot\mid\boldsymbol{Q},a\right)$. Note that $|\mathcal{R}\left(\boldsymbol{Q}\right)|\leqslant R$. We set

$$n_2 \geqslant \frac{2}{(\Delta p)^2}\cdot\log\frac{2^{R+1}(U+1)^D|\mathcal{A}|}{\delta}.$$

By Lemma 5, we have that for each $(\boldsymbol{Q},a)\in\tilde{\mathcal{S}}\times\mathcal{A}$,

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

Taking a union bound over each $\boldsymbol{Q}\in\tilde{\mathcal{S}}$ and $a\in\mathcal{A}$, we obtain

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

This completes the proof. □

A.3 Proof of Theorem 1

We are now ready to prove Theorem 1.

Proof. Define k^* as the number of required episodes for RL-QN to learn $\tilde{\pi}^*$. Based on Lemmas 3 and 4, we show that as the learning process proceeds, each $(\boldsymbol{Q},a)\in\tilde{\mathcal{S}}\times\mathcal{A}$ will be sampled sufficiently to learn the optimal policy for $\tilde{\mathcal{M}}$. We define the event

$$B_k \triangleq \left\{k\geqslant K_0, \tilde{\pi}_k=\pi_{\text{rand}}, N_k(\boldsymbol{Q},a)>\frac{p^{\text{rand}}\left(\boldsymbol{Q}\right)\cdot L_k}{2|\mathcal{A}|}, \forall\boldsymbol{Q}\in\tilde{\mathcal{S}}, a\in\mathcal{A}\right\}.$$

When B_k is true, at least $p^{\text{rand}}\left(\boldsymbol{Q}\right)L\sqrt{K_0}/(2|\mathcal{A}|)$ samples are obtained for each $(\boldsymbol{Q},a)$. Therefore, a sufficient condition to obtain the required number of samples for each $(\boldsymbol{Q},a)$ as Lemma 4 is that B_k occurs for

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

times.

We denote by m^* the number of episodes needed for B_k to occur for J^* times. Define $E_{k_1,k_2,\ldots,k_n}$ as the event that for episodes $k = k_1, k_2, \ldots, k_n$, the event B_k does NOT occur. For $n \geqslant 1$, we have

$$\begin{aligned}
&\Pr\{m^* \geqslant K_0 + J^* + n\} \\
&= \Pr\left\{\text{from episode } K_0 \text{ to } K_0 + J^* + n - 1,\ B_k \text{ does not occur for at least } n \text{ times}\right\} \\
&= \Pr\left\{\bigcup_{K_0 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0 + J^* + n - 1} E_{k_1,k_2,\ldots,k_n}\right\} \\
&\leqslant \sum_{K_0 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0 + J^* + n - 1} \Pr\left\{E_{k_1,k_2,\ldots,k_n}\right\}. \qquad (5)
\end{aligned}$$

By applying Lemma 3, we have that for each $k \geqslant K_0$,

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

Therefore, for any $K_0 \leqslant k_1 < k_2 < \cdots < k_n \leqslant K_0 + J^* + n - 1$, we have

$$\Pr\left\{E_{k_1,k_2,\ldots,k_n}\right\} \leqslant \prod_{i=1}^{n}\left(1 - \frac{\ell}{2\sqrt{k_i}}\right) \leqslant \left(1 - \frac{\ell}{2\sqrt{K_0 + J^* + n - 1}}\right)^n. \qquad (6)$$

Inserting Equation (6) into (5) yields

$$\begin{aligned}
\Pr\{m^* \geqslant K_0 + J^* + n\} &\leqslant \binom{J^* + n}{n} \cdot \left(1 - \frac{\ell}{2\sqrt{K_0 + J^* + n - 1}}\right)^n \\
&= \frac{(J^* + n) \cdot (J^* + n - 1) \cdots (n + 1)}{J^*!} \cdot \left(1 - \frac{\ell}{2\sqrt{K_0 + J^* + n - 1}}\right)^n \\
&\leqslant \frac{(J^* + n)^{J^*}}{J^*!} \cdot \left(1 - \frac{\ell}{2\sqrt{K_0 + J^* + n - 1}}\right)^n \\
&\leqslant \frac{(J^* + n)^{J^*}}{J^*!} \cdot \exp\left(-\frac{n\ell}{2\sqrt{K_0 + J^* + n - 1}}\right) \qquad (7) \\
&< \frac{(J^* + n)^{J^*}}{J^*!} \cdot \frac{(2J^* + 4)! \cdot 4^{J^*+2} \cdot (K_0 + J^* + n - 1)^{J^*+2}}{(n\ell)^{2J^*+4}} \qquad (8) \\
&\leqslant \frac{4^{J^*+2} \cdot (2J^* + 4)! \cdot (K_0 + J^*)^{2J^*+2}}{J^*! \cdot \ell^{2J^*+4}} \cdot \frac{1}{n^2},
\end{aligned}$$

where Equation (7) follows from the fact that for $x \in (0, 1)$, $1 - x \leq e^{-x}$ and Equation (8) is obtained by the fact that $\exp(u) = \sum_{k=0}^{\infty}(u)^k/k! > u^{2J^*+4}/(2J^* + 4)!$ holds for $u > 0$.

We then have

$$\begin{aligned}
\mathbb{E}[k^*] \leqslant \mathbb{E}[m^*] &= \sum_{i=1}^{\infty} \Pr\{m^* \geqslant i\} = K_0 + J^* + \sum_{n=1}^{\infty} \Pr\{m^* \geqslant K_0 + J^* + n\} \\
&\leq K_0 + J^* + \frac{4^{J^*+2} \cdot (2J^* + 4)! \cdot (K_0 + J^*)^{2J^*+2}}{J^*! \cdot \ell^{2J^*+4}} \sum_{n=1}^{\infty} \frac{1}{n^2} \\
&\leq K_0 + J^* + \frac{\pi^2 \cdot 4^{J^*+2} \cdot (2J^* + 4)! \cdot (K_0 + J^*)^{2J^*+2}}{6 \cdot J^*! \cdot \ell^{2J^*+4}} \triangleq K(J^*).
\end{aligned}$$

By Markov's inequality, we have

$$k^* \leqslant \frac{2\mathbb{E}\left[k^*\right]}{\delta} \leqslant \frac{2K(J^*)}{\delta}, \tag{9}$$

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

By taking a union bound over the events of Lemma 4 and Equation (9), we have that with probability at least $1 - \delta$,

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

which completes the proof. □

B PROOF OF LEMMA 1

This section is devoted to the proof of Lemma 1.

Proof. We only discuss an episode k with $\pi_k^{\text{in}} = \tilde{\pi}^*$ here.

We define $\sum_{t=t_{k-1}+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)$ as the regret of episode k. Let t_k' and t_k'' denote the first and last time slot such that $\boldsymbol{Q}(t) \in \mathcal{S}_{\text{bd}}^{\text{in}}$ for $t \in \{t_{k-1}, \ldots, t_k\}$. For the simplicity, we first include the regret incurred at time step t_{k-1} into the episodic regret analysis and subtract it in the end.

We decompose average episodic regret as follows:

$$\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L_k'}\right]$$

$$= \lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_k'}^{t_k''}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L_k'}\right] + \tag{10}$$

$$\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_{k-1}}^{t_k'-1}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right) + \sum_{t=t_k''+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right) - \left(\sum_i Q_i(t_{k-1}) - \tilde{\rho}^*\right)}{L_k'}\right]. \tag{11}$$

Upper bound of Equation (10): We partition the time slots from t_k' to t_k'' into $\mathcal{T}_k^{\text{in}}$, $\mathcal{T}_k^{\text{bd}}$ and $\mathcal{T}_k^{\text{out}}$, which denote the set of time slots that $\boldsymbol{Q}(t)$ is in $\mathcal{S}_{\text{in}}^{\text{in}}$, $\mathcal{S}_{\text{bd}}^{\text{in}}$, and $\mathcal{S}^{\text{out}}$, respectively.

To analyze the regret associated with $\mathcal{T}_k^{\text{in}}$, we define an "in process" unit as the process that $\boldsymbol{Q}(t)$ leaves $\mathcal{S}_{\text{bd}}^{\text{in}}$, enters $\mathcal{S}_{\text{in}}^{\text{in}}$, stays in $\mathcal{S}_{\text{in}}^{\text{in}}$ for some time and finally returns back to $\mathcal{S}_{\text{bd}}^{\text{in}}$. Then, the process during $\mathcal{T}_k^{\text{in}}$ can be decomposed into multiple "in process" units. An "in process" unit is said to start from $\boldsymbol{Q}^{\text{in}} \in \mathcal{S}_{\text{bd}}^{\text{in}}$ if $\boldsymbol{Q}^{\text{in}}$ is its last state before entering $\mathcal{S}_{\text{in}}^{\text{in}}$. We use $N_k^{\text{in}}(\boldsymbol{Q}^{\text{in}})$ to denote the number of times that an "in process" unit starts from $\boldsymbol{Q}^{\text{in}}$. The accumulated regret during the ith "in process" unit starting from $\boldsymbol{Q}^{\text{in}}$ is denoted by $R_{k,i}^{\text{in}}(\boldsymbol{Q}^{\text{in}})$.

Similarly, we define "out process" units to decompose $\mathcal{T}_k^{\text{out}}$. An "out process" unit is said to start from $\boldsymbol{Q}^{\text{out}} \in \mathcal{S}_{\text{bd}}^{\text{in}}$ if $\boldsymbol{Q}^{\text{out}}$ is its last state before entering $\mathcal{S}^{\text{out}}$. We define $N_k^{\text{out}}(\boldsymbol{Q}^{\text{out}})$ and $R_{k,i}^{\text{out}}(\boldsymbol{Q}^{\text{out}})$ in similar manners.

Now we can further decompose Equation (10) as follows:

$$\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_k'}^{t_k''}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'}\right]=\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{\mathcal{Q}^{\text{in}}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\sum_{i=1}^{N_k^{\text{in}}(\mathcal{Q}^{\text{in}})}R_{k,i}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)}{L_k'}\right]+ \tag{12}$$

$$\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{\mathcal{Q}^{\text{out}}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\sum_{i=1}^{N_k^{\text{out}}(\mathcal{Q}^{\text{out}})}R_{k,i}^{\text{out}}\left(\mathcal{Q}^{\text{out}}\right)}{L_k'}\right]+ \tag{13}$$

$$\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t\in\mathcal{T}_k^{\text{bd}}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'}\right]. \tag{14}$$

For Equation (12), we use $Y_{k,i}^{\text{in}}(\mathcal{Q}^{\text{in}})$ to denote the length of the time interval between the starting time of the ith and $(i+1)$−th "in process" units that start from $\mathcal{Q}^{\text{in}}$. By the Markovian property of the system, for a given $\mathcal{Q}^{\text{in}}$, $Y_{k,i}^{\text{in}}(\mathcal{Q}^{\text{in}})$'s are i.i.d. and $R_{k,i}^{\text{in}}(\mathcal{Q}^{\text{in}})$'s are also i.i.d. Then according to the renewal reward theorem, for every $\mathcal{Q}^{\text{in}}\in\mathcal{S}_{\text{bd}}^{\text{in}}$, we have

$$\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{i=1}^{N_k^{\text{in}}(\mathcal{Q}^{\text{in}})}R_{k,i}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)}{L_k'}\right]=\frac{\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)\right]}{\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[Y_{k,1}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)\right]}. \tag{15}$$

Note that we inherently use the fact that as $k\to\infty$, $L_k'\to\infty$ since $L_k'\geqslant L\sqrt{k}$.

However, it is not straightforward to compute $\mathbb{E}_{\tilde{\pi}^*+\pi_0}[Y_{k,1}^{\text{in}}(\mathcal{Q}^{\text{in}})]$ directly. Note that every time when an "in process" unit starting from $\mathcal{Q}^{\text{in}}$ occurs, $\mathcal{Q}^{\text{in}}$ must be visited. Therefore, we have the bound that for every $\mathcal{Q}^{\text{in}}\in\mathcal{S}_{\text{bd}}^{\text{in}}$,

$$\frac{1}{\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[Y_{k,1}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)\right]}\leqslant\frac{1}{\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\text{Interval between visits to }\mathcal{Q}^{\text{in}}\right]}=p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Q}^{\text{in}}\right). \tag{16}$$

By inserting Equations (15) and (16) into Equation (12), we have

$$\begin{aligned}\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{\mathcal{Q}^{\text{in}}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\sum_{i=1}^{N_k^{\text{in}}(\mathcal{Q}^{\text{in}})}R_{k,i}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)}{L_k'}\right]&\leqslant\sum_{\mathcal{Q}^{\text{in}}\in\mathcal{S}_{\text{bd}}^{\text{in}}}p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{Q}^{\text{in}}\right)\cdot\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{in}}\left(\mathcal{Q}^{\text{in}}\right)\right]\\&\leqslant p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}_{\text{bd}}^{\text{in}}\right)\cdot\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{in}}\left(\mathcal{Q}\right)\right].\end{aligned} \tag{17}$$

We provide an upper bound for $\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\mathbb{E}_{\tilde{\pi}^*+\pi_0}[R_{k,1}^{\text{in}}(\mathcal{Q})]$, as stated in the following lemma (see Appendix B.1 for the proof).

Lemma 6. *Under Assumptions 1–3, we have*

$$\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{in}}\left(\mathcal{Q}\right)\right]\leqslant cDU^{1+\gamma}.$$

For Eqaution (13), by following a similar argument, we have that

$$\lim_{k\to\infty}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{\mathcal{Q}^{\text{out}}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\sum_{i=1}^{N_k^{\text{out}}(\mathcal{Q}^{\text{out}})}R_{k,i}^{\text{out}}\left(\mathcal{Q}^{\text{out}}\right)}{L_k'}\right]\leqslant p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}_{\text{bd}}^{\text{in}}\right)\cdot\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{out}}\left(\mathcal{Q}\right)\right]. \tag{18}$$

The following lemma gives an upper bound for $\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}} \mathbb{E}_{\tilde{\pi}^*+\pi_0}[R_{k,1}^{\text{out}}(\mathcal{Q})]$ (see Appendix B.2 for the proof).

Lemma 7. *Under Assumptions 1–3, we have*

$$\max_{\mathcal{Q}\in\mathcal{S}_{\text{bd}}^{\text{in}}} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R_{k,1}^{\text{out}}(\mathcal{Q})\right] \leqslant \frac{2a^2DU^{2\alpha+1}}{\epsilon_0^2}.$$

For Eqaution (14), since under $\tilde{\pi}^* + \pi_0$, the Markov chain is positive recurrent, we have that

$$\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t\in\mathcal{T}_k^{\text{bd}}}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'}\right] \leqslant (DU-\tilde{\rho}^*)\cdot \lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\left|\mathcal{T}_k^{\text{bd}}\right|}{L_k'}\right] \leqslant p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}_{\text{bd}}^{\text{in}}\right)\cdot DU. \tag{19}$$

By combining Equations (17)–(19), Lemmas 6 and 7, we upper bound Equation (10) as follows:

$$\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_k'}^{t_k''}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)}{L_k'}\right] \leqslant p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}_{\text{bd}}^{\text{in}}\right)\cdot O\left(U^{1+\max\{2\alpha,\gamma\}}\right). \tag{20}$$

Upper bound of Equation (11): From the episode termination criteria of RL-QN, we have $\mathcal{Q}(t_{k-1})\in\mathcal{S}^{\text{in}}$. If $\mathcal{Q}(t_{k-1})\in\mathcal{S}_{\text{bd}}^{\text{in}}$, we have that $t_k' = t_{k-1}$, and therefore have that

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_{k-1}}^{t_k'-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right) \mid \mathcal{Q}(t_{k-1})\in\mathcal{S}_{\text{bd}}^{\text{in}}\right] = 0.$$

If $\mathcal{Q}(t_{k-1})\in\mathcal{S}_{\text{in}}^{\text{in}}$, then from $t = t_{k-1}$ to $t = t_k' - 1$, $\mathcal{Q}(t)\in\mathcal{S}_{\text{in}}^{\text{in}}$. By applying techniques in the proof of Lemma 6, we have

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_{k-1}}^{t_k'-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right) \mid \mathcal{Q}(t_{k-1})\in\mathcal{S}_{\text{in}}^{\text{in}}\right] \leqslant cDU^{1+\gamma}.$$

Therefore, we have

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_{k-1}}^{t_k'-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\right] \leqslant cDU^{1+\gamma}. \tag{21}$$

Since we also have that $\mathcal{Q}(t_k)\in\mathcal{S}^{\text{in}}$, by following a similar argument, we have

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_k''+1}^{t_k}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)\right] \leqslant cDU^{1+\gamma}. \tag{22}$$

By combining Equations (21) and (22), together with the fact that $L'_k \geqslant L_k$, we have an upper bound for Equation (11) as follows:

$$\begin{aligned}
&\lim_{k\to\infty} \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\frac{\sum_{t=t_{k-1}}^{t'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)+\sum_{t=t''_k+1}^{t_k}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)-\left(\sum_i Q_i(t_{k-1})-\tilde{\rho}^*\right)}{L'_k}\right]\\
&\leqslant \lim_{k\to\infty}\frac{1}{L_k}\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_{k-1}}^{t'_k-1}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)+\sum_{t=t''_k+1}^{t_k}\left(\sum_i Q_i(t)-\tilde{\rho}^*\right)-\left(\sum_i Q_i(t_{k-1})-\tilde{\rho}^*\right)\right]\\
&\leqslant \lim_{k\to\infty}\frac{2cDU^{1+\gamma}+\tilde{\rho}^*}{L\cdot\sqrt{k}}=0. \qquad (23)
\end{aligned}$$

By summing up Equations (20) and (23), we complete the proof. □

B.1 Proof of Lemma 6

Proof. From Proposition 5.5.1 in [7], when applying $\tilde{\pi}^*$ to $\tilde{\mathcal{M}}$, there exists a function $\tilde{h}^* : \tilde{\mathcal{S}} \to \mathbb{R}$ such that the following Bellman equation holds:

$$\tilde{\rho}^* + \tilde{h}^*(\mathcal{Q}) = \sum_i Q_i + \sum_{\mathcal{Q}'\in\mathcal{S}^{\text{in}}} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right)\cdot \tilde{h}^*\left(\mathcal{Q}'\right), \qquad \forall \mathcal{Q}\in\mathcal{S}^{\text{in}}. \qquad (24)$$

By the construction of $\tilde{\mathcal{M}}$ in Section 3, for each $\mathcal{Q}\in\mathcal{S}^{\text{in}}_{\text{in}}$, $\mathcal{Q}'\in\mathcal{S}^{\text{in}}$ and $a\in\mathcal{A}$, we have

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

Therefore, for each $\mathcal{Q}\in\mathcal{S}^{\text{in}}_{\text{in}}$, Equation (24) can be rewritten as

$$\tilde{\rho}^* + \tilde{h}^*(\mathcal{Q}) = \sum_i Q_i + \sum_{\mathcal{Q}'\in\mathcal{S}^{\text{in}}} p\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right)\cdot \tilde{h}^*\left(\mathcal{Q}'\right). \qquad (25)$$

Fix an $\hat{\mathcal{Q}}\in\mathcal{S}^{\text{in}}_{\text{bd}}$, we now analyze $R^{\text{in}}_{k,1}\left(\hat{\mathcal{Q}}\right)$. We denote the start and end time slot of this "in process" unit as t_s and t_e, respectively. Note that $\mathcal{Q}(t)\in\mathcal{S}^{\text{in}}_{\text{in}}$ for $t\in[t_s,t_e]$. Using Equation (25), we have

$$\begin{aligned}
\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R^{\text{in}}_{k,1}\left(\hat{\mathcal{Q}}\right)\right] &= \mathbb{E}\left[\sum_{t=t_s}^{t_e}\left(\sum_i Q_i-\tilde{\rho}^*\right)\mid \mathcal{Q}(t_s-1)=\hat{\mathcal{Q}}\right]\\
&= \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_s}^{t_e}\left(\tilde{h}^*(\mathcal{Q}(t))-\mathbb{E}\left[\tilde{h}^*(\mathcal{Q}(t+1))\mid\mathcal{Q}(t)\right]\right)\mid\mathcal{Q}(t_s-1)=\hat{\mathcal{Q}}\right]\\
&= \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\tilde{h}^*(\mathcal{Q}(t_s))-\mathbb{E}\left[\tilde{h}^*(\mathcal{Q}(t_e+1))\mid\mathcal{Q}(t_e)\right]\mid\mathcal{Q}(t_s-1)=\hat{\mathcal{Q}}\right]\\
&\quad+\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\sum_{t=t_s}^{t_e-1}\left(\tilde{h}^*(\mathcal{Q}(t+1))-\mathbb{E}\left[\tilde{h}^*(\mathcal{Q}(t+1))\mid\mathcal{Q}(t)\right]\right)\mid\mathcal{Q}(t_s-1)=\hat{\mathcal{Q}}\right]\\
&= \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\tilde{h}^*(\mathcal{Q}(t_s))-\tilde{h}^*(\mathcal{Q}(t_e+1))\mid\mathcal{Q}(t_s-1)=\hat{\mathcal{Q}}\right]. \qquad (26)
\end{aligned}$$

On the other hand, for each $(\boldsymbol{Q}, \boldsymbol{Q}') \in \mathcal{S} \times \mathcal{S}$, by following the analysis for the proof of Proposition 5.5.1 in [7], we can upper bound $\tilde{h}^*(\boldsymbol{Q}') - \tilde{h}^*(\boldsymbol{Q})$ as follows:

$$\tilde{h}^*\left(\boldsymbol{Q}'\right) - \tilde{h}^*(\boldsymbol{Q}) = \min_{\tilde{\pi}} \tilde{\mathbb{E}}_{\tilde{\pi}}\left[\sum_{t=1}^{T_{\boldsymbol{Q}\to\boldsymbol{Q}'}}\left[\sum_i Q_i(t) - \tilde{\rho}^*\right]\right] \overset{(a)}{\leqslant} \min_{\tilde{\pi}} \tilde{\mathbb{E}}_{\tilde{\pi}}\left[T_{\boldsymbol{Q}\to\boldsymbol{Q}'}\right] \cdot DU \overset{(b)}{\leqslant} cDU^{1+\gamma}, \quad (27)$$

where (a) comes from the fact that $\sum_i Q_i(t) \leqslant DU$ for $\boldsymbol{Q} \in \tilde{\mathcal{S}}$, and the second inequality (b) holds under Assumption 3. By inserting Equation (27) into Equation (26), we complete the proof. □

B.2 Proof of Lemma 7

Proof. Fix an $\hat{\boldsymbol{Q}} \in \mathcal{S}^{\text{in}}_{\text{bd}}$, we now turn to analyze $R^{\text{out}}_{k,1}(\hat{\boldsymbol{Q}})$. From the definition of $\mathcal{S}^{\text{in}}$, we have $\hat{Q}_{\max} \leqslant U - W$. We denote the length of this "out process" unit as τ. We then have

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R^{\text{out}}_{k,1}(\hat{\boldsymbol{Q}})\right] \leqslant \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[D \cdot (U - W + W\tau) \cdot \tau\right] = D(U-W)\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau] + DW\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\tau^2\right], \quad (28)$$

where the inequality comes from the fact that the backlog of each queue can be no larger than $(U - W + W\tau)$ during this "out process" unit. We upper bound $\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau]$ and $\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau^2]$ separately.

Upper bound of $\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau]$: We use the following lemma from [12].

Lemma 8 (Theorem 1.1 in Chapter 5 of [12]). *For an irreducible Markov chain with countable state space, if there exists $\epsilon > 0$, a finite state set $\mathcal{F}$ and a nonnegative Lyapunov function $\Phi(\cdot)$, such that $\mathbb{E}[\Phi(\boldsymbol{Q}(t+1)) \mid \boldsymbol{Q}(t) = \boldsymbol{Q}] < \infty$ for each $\boldsymbol{Q} \notin \mathcal{F}$ and $\mathbb{E}[\Phi(\boldsymbol{Q}(t+1)) - \Phi(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t) = \boldsymbol{Q}] \leqslant -\epsilon$ for each $\boldsymbol{Q} \notin \mathcal{F}$, we then have that for each $\boldsymbol{Q} \notin \mathcal{F}$, $\mathbb{E}[T_{\boldsymbol{Q}\to\mathcal{F}}] \leqslant \Phi(\boldsymbol{Q})/\epsilon$, where $T_{\boldsymbol{Q}\to\mathcal{F}}$ is the first hitting time of the set $\mathcal{F}$ when starting from state $\boldsymbol{Q}$.*

We can interpret τ as the hitting time of the set $\mathcal{S}^{\text{in}}$ when the Markov chain starts from state $\boldsymbol{Q} \in \mathcal{S}^{\text{out}}$. From the selection of U in Algorithm 1, each state in $\mathcal{S}^{\text{out}}$ has negative Lyapunov drift regarding Φ_0. Also, from Assumption 1, the value of Lyapunov function Φ_0 for the beginning state of the "out process" unit can be upper bounded as $a(U - W + W)^\alpha = aU^\alpha$. By Lemma 8, it follows that

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau] \leqslant \frac{aU^\alpha}{\epsilon_0} \triangleq T_0. \quad (29)$$

Upper bound of $\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\tau^2]$: We use the following lemma from [25].

Lemma 9 (Theorem 6.3.4 in [25]). *In a Markov chain with a countable state space, for a nonempty state set $\mathcal{B}$, and a state s, if there exists a constant C such that $\mathbb{E}[T_{s\to\mathcal{B}}] \leqslant C \cdot F^*_{s,\mathcal{B}}$, then for $p \geqslant 1$, we have $\mathbb{E}[T^p_{s\to\mathcal{B}}] \leqslant p! \cdot C^p \cdot F^*_{s,\mathcal{B}}$, where $F^*_{s,\mathcal{B}} \triangleq \Pr\{s_1 \in \mathcal{B} \mid s_0 = s\} + \sum_{n=2}^{\infty} \Pr\{s_1 \notin \mathcal{B}, \ldots, s_{n-1} \notin \mathcal{B}, s_n \in \mathcal{B} \mid s_0 = s\}$.*

In our case, under $\tilde{\pi}^* + \pi_0$, the Markov chain is positive recurrent. Thus, for each $\boldsymbol{Q} \in \mathcal{S}^{\text{out}}$ and each $\boldsymbol{Q}' \in \mathcal{S}^{\text{in}}$, we have $F^*_{\boldsymbol{Q},\boldsymbol{Q}'} = 1$. Note that $F^*_{\boldsymbol{Q},\mathcal{S}^{\text{in}}}$ is upper bounded by 1 since it is a probability. Also, $F^*_{\boldsymbol{Q},\mathcal{S}^{\text{in}}} \geqslant F^*_{\boldsymbol{Q},\boldsymbol{Q}'}$, we thus have $F^*_{\boldsymbol{Q},\mathcal{S}^{\text{in}}} = 1$. Therefore, $\mathbb{E}[\tau] \leqslant T_0 = T_0 \cdot F^*_{\boldsymbol{Q},\mathcal{S}^{\text{in}}}$. By applying Lemma 9, we have that

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\tau^2\right] \leqslant 2 \cdot T_0^2 \cdot F^*_{\boldsymbol{Q},\mathcal{S}^{\text{in}}} = 2T_0^2. \quad (30)$$

Inserting Equations (29) and (30) into Equation (28) gives

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[R^{\text{out}}_{k,1}\left(\hat{\boldsymbol{Q}}\right)\right] \leqslant D(U-W)T_0 + 2DWT_0^2 \leqslant 2D(U - W + W)T_0^2 = \frac{2a^2DU^{2\alpha+1}}{\epsilon_0^2},$$

which completes the proof. □

C PROOF OF LEMMA 2

Proof. For the ease of exposition, we first define a "linear" Lyapunov function $\Phi'(\cdot)$ as

$$\Phi'(Q) = \begin{cases} \left[\tilde{\Phi}^*(Q)\right]^{\frac{1}{\beta}}, & \text{for } Q \in \mathcal{S}^{\text{in}} \\ U - W, & \text{for } Q \in \mathcal{S}^{\text{out}} \end{cases},$$

which has the following properties (see Appendix C.1 for the proof).

Lemma 10. *Under Assumption 2, there exist constants $V, \epsilon' > 0$ such that for each $U > W$, the following properties hold:*

(1) *For each $Q \in \mathcal{S}$, $\max_{Q' \in \mathcal{R}(Q)} |\Phi'(Q') - \Phi'(Q)| \leqslant V$;*
(2) *For each $Q \in \mathcal{S}^{\text{in}}$ such that $Q_{\max} \geqslant \tilde{B}^*$, we have*

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

With Lemma 10, we obtain the following properties on the distribution of $\Phi'(\cdot)$ (see Appendix C.2 for the proof).

Lemma 11. *Suppose Assumption 2 holds. Define $m^* \triangleq \lfloor (U + V - W - b_1^{1/\beta}\tilde{B}^*)/(2V) \rfloor$, we then have that for $m = 1, 2, \ldots, m^*$,*

$$\begin{aligned} & p^{\tilde{\pi}^*+\pi_0}\left(\left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - (2m-1)V\right\}\right) \\ \leqslant\ & \frac{V}{V+\tilde{\epsilon}^*} \cdot p^{\tilde{\pi}^*+\pi_0}\left(\left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - (2m+1)V\right\}\right). \end{aligned}$$

Lemma 11 implies that

$$\begin{aligned} & p^{\tilde{\pi}^*+\pi_0}\left(\left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - V\right\}\right) \\ \leqslant\ & \left(\frac{V}{V+\epsilon^*}\right)^{m^*} \cdot p^{\tilde{\pi}^*+\pi_0}\left(\left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - (2m^*+1)V\right\}\right) \leqslant \left(\frac{V}{V+\epsilon^*}\right)^{m^*}. \end{aligned}$$

Note that our goal is to obtain the stationary probability of the state being in $\mathcal{S}^{\text{in}}_{\text{bd}}$. Recall that $\mathcal{S}^{\text{in}} \triangleq \left\{Q : \tilde{\Phi}^*(Q) \leqslant (U-W)^\beta\right\}$ and $\mathcal{S}^{\text{in}}_{\text{bd}} \triangleq \{Q \in \mathcal{S}^{\text{in}} : \mathcal{R}(Q) \cap \mathcal{S}^{\text{out}} \neq \emptyset\}$. Therefore,

$$\max_{Q' \in \mathcal{R}(Q)} \tilde{\Phi}^*\left(Q'\right) > (U-W)^\beta, \qquad \forall Q \in \mathcal{S}^{\text{in}}_{\text{bd}}.$$

By the definition of Φ' and Lemma 10, we have $\Phi'(Q) > U - W - V$. Therefore,

$$\mathcal{S}^{\text{in}}_{\text{bd}} \subseteq \left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - V\right\},$$

and the following inequality holds

$$p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}^{\text{in}}_{\text{bd}}\right) \leqslant p^{\tilde{\pi}^*+\pi_0}\left(\left\{Q \in \mathcal{S}^{\text{in}} : \Phi'(Q) > U - W - V\right\}\right) \leqslant \left(\frac{V}{V+\epsilon^*}\right)^{m^*}.$$

By taking natural logarithm on both sides and applying the fact that $\log(1+x) \geqslant x/(1+x)$ for $x > 0$, we have

$$\log p^{\tilde{\pi}^*+\pi_0}\left(\mathcal{S}^{\text{in}}_{\text{bd}}\right) \leqslant -m^* \log\left(1 + \frac{\tilde{\epsilon}^*}{V}\right) \leqslant -m^* \cdot \frac{\tilde{\epsilon}^*}{V+\tilde{\epsilon}^*} = -\left\lfloor \frac{U + V - W - b_1^{1/\beta}\tilde{B}^*}{2V} \right\rfloor \cdot \frac{\tilde{\epsilon}^*}{V+\tilde{\epsilon}^*} = O(-U),$$

which completes the proof. □

C.1 Proof of Lemma 10

Proof. Consider a fixed $\mathcal{Q} \in \mathcal{S}^{\text{in}}$. From the definition of $\mathcal{S}^{\text{in}}$ and W, we have $\max_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})} Q'_{\max} \leqslant U$. Hence $\mathcal{Q}' \in \tilde{\mathcal{S}}$ for each $\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})$. By the definition of function $\Phi'(\cdot)$, we have

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi'\left(\mathcal{Q}(t+1)\right) - \Phi'\left(\mathcal{Q}(t)\right) \mid \mathcal{Q}(t) = \mathcal{Q}\right] \\
&= -\left[\tilde{\Phi}^*\left(\mathcal{Q}\right)\right]^{\frac{1}{\beta}} + \sum_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right) \cdot \Phi'\left(\mathcal{Q}'\right). \qquad (31)
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right) \cdot \Phi'\left(\mathcal{Q}'\right) \\
&= \sum_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q}) \cap \mathcal{S}^{\text{in}}} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right) \cdot \left[\tilde{\Phi}^*\left(\mathcal{Q}'\right)\right]^{\frac{1}{\beta}} + \sum_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q}) \cap \mathcal{S}^{\text{out}}} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right) \cdot (U - W) \\
&\leqslant \sum_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})} \tilde{p}\left(\mathcal{Q}' \mid \mathcal{Q}, \tilde{\pi}^*(\mathcal{Q})\right) \cdot \left[\tilde{\Phi}^*\left(\mathcal{Q}'\right)\right]^{\frac{1}{\beta}}, \qquad (32)
\end{aligned}$$

where the inequality comes from the fact that $[\tilde{\Phi}^*(\mathcal{Q}')]^{1/\beta} \geqslant U - W$ for $\mathcal{Q}' \in \mathcal{S}^{\text{out}}$.

Combining Equations (31) and (32) yields

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi'\left(\mathcal{Q}(t+1)\right) - \Phi'\left(\mathcal{Q}(t)\right) \mid \mathcal{Q}(t) = \mathcal{Q}\right] \\
&\leqslant \tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\left(\tilde{\Phi}^*\left(\mathcal{Q}(t+1)\right)\right)^{\frac{1}{\beta}} \mid \mathcal{Q}(t) = \mathcal{Q}\right] - \left(\tilde{\Phi}^*\left(\mathcal{Q}\right)\right)^{\frac{1}{\beta}} \\
&\overset{(a)}{\leqslant} \left(\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*\left(\mathcal{Q}(t+1)\right) \mid \mathcal{Q}(t) = \mathcal{Q}\right]\right)^{\frac{1}{\beta}} - \left(\tilde{\Phi}^*\left(\mathcal{Q}\right)\right)^{\frac{1}{\beta}} \\
&\overset{(b)}{\leqslant} \frac{\left(\tilde{\Phi}^*\left(\mathcal{Q}\right)\right)^{\frac{1}{\beta}-1}}{\beta} \cdot \left(\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\tilde{\Phi}^*\left(\mathcal{Q}(t+1)\right) \mid \mathcal{Q}(t) = \mathcal{Q}\right] - \tilde{\Phi}^*\left(\mathcal{Q}\right)\right) \\
&\overset{(c)}{\leqslant} -\frac{\left(\tilde{\Phi}^*\left(\mathcal{Q}\right)\right)^{\frac{1}{\beta}-1}}{\beta} \cdot \tilde{\epsilon}^* \cdot Q_{\max}^{\beta-1} \overset{(d)}{\leqslant} -\frac{b_1^{\frac{1}{\beta}-1}}{\beta} \cdot \tilde{\epsilon}^* \triangleq -\epsilon',
\end{aligned}$$

where (a) follows from Jensen's inequality, (b) holds because $f(x) = x^{1/\beta}$ is concave and thus $f(y) - f(x) \leqslant f'(x)(y - x)$, (c) and (d) come from Assumption 2.

Next we discuss the upper bound of $\max_{\mathcal{Q}' \in \mathcal{R}(\mathcal{Q})} |\Phi'(\mathcal{Q}') - \Phi'\left(\mathcal{Q}\right)|$.

If $\mathcal{Q}, \mathcal{Q}' \in \mathcal{S}^{\text{in}}$ with $Q_{\max} = 0$ or $Q'_{\max} = 0$, we have $|\Phi'(\mathcal{Q}') - \Phi'\left(\mathcal{Q}\right)| \leqslant b_1^{1/\beta} W$.

If $\mathcal{Q}, \mathcal{Q}' \in \mathcal{S}^{\text{in}}$ with $Q_{\max} > 0$ and $Q'_{\max} > 0$, we have $\Phi'(\mathcal{Q}') = (\tilde{\Phi}^*(\mathcal{Q}'))^{\frac{1}{\beta}}$ and $\Phi'\left(\mathcal{Q}\right) = (\tilde{\Phi}^*(\mathcal{Q}))^{\frac{1}{\beta}}$. Note that for the concave function $f(x) = x^{1/\beta}$, it holds that

$$\left|f(y) - f(x)\right| \leqslant \max\{f'(x), f'(y)\} \cdot \left|y - x\right|.$$

Without losing generality, we assume that $Q'_{\max} > Q_{\max}$. Thus

$$\begin{aligned}\left|\Phi'\left(\boldsymbol{Q}'\right)-\Phi'\left(\boldsymbol{Q}\right)\right| &\leqslant \max\left\{\left(\tilde{\Phi}^*\left(\boldsymbol{Q}\right)\right)^{\frac{1}{\beta}-1}/\beta,\ \left(\tilde{\Phi}^*\left(\boldsymbol{Q}'\right)\right)^{\frac{1}{\beta}-1}/\beta\right\}\cdot\left|\tilde{\Phi}^*\left(\boldsymbol{Q}'\right)-\tilde{\Phi}^*\left(\boldsymbol{Q}\right)\right| \\ &\overset{(a)}{\leqslant} \frac{1}{\beta\cdot Q_{\max}^{\beta-1}}\cdot 2b_2 Q_{\max}^{'\beta-1} \leqslant \frac{2b_2}{\beta}\left(1+\frac{W}{Q_{\max}}\right)^{\beta-1} \leqslant \frac{2b_2}{\beta}(1+W)^{\beta-1}, \end{aligned} \tag{33}$$

where (a) follows from Assumption 2.

If $\boldsymbol{Q},\boldsymbol{Q}'\in\mathcal{S}^{\text{out}}$, from the definition of $\Phi'(\cdot)$, we have $\left|\Phi'\left(\boldsymbol{Q}'\right)-\Phi'\left(\boldsymbol{Q}\right)\right|=0$.

If $\boldsymbol{Q}\in\mathcal{S}^{\text{in}},\boldsymbol{Q}'\in\mathcal{S}^{\text{out}}$, we have

$$\left|\Phi'\left(\boldsymbol{Q}'\right)-\Phi'\left(\boldsymbol{Q}\right)\right| = U-W-\left[\tilde{\Phi}^*\left(\boldsymbol{Q}\right)\right]^{\frac{1}{\beta}} \leqslant \left[\tilde{\Phi}^*\left(\boldsymbol{Q}'\right)\right]^{\frac{1}{\beta}}-\left[\tilde{\Phi}^*\left(\boldsymbol{Q}\right)\right]^{\frac{1}{\beta}},$$

which can be upper bounded in the same way as Equation (33). The case of $\boldsymbol{Q}\in\mathcal{S}^{\text{out}},\boldsymbol{Q}'\in\mathcal{S}^{\text{in}}$ has the same upper bound.

Therefore, we have that for each $\boldsymbol{Q}$ and each $\boldsymbol{Q}'\in\mathcal{R}\left(\boldsymbol{Q}\right)$,

$$\left|\Phi'\left(\boldsymbol{Q}'\right)-\Phi'\left(\boldsymbol{Q}\right)\right| \leqslant \max\left\{b_1^{\frac{1}{\beta}}W, \frac{2b_2}{\beta}(1+W)^{\beta-1}\right\} \triangleq V.$$

□

C.2 Proof of Lemma 11

Proof. We define a series of Lyapunov functions: $\Phi_m\left(\boldsymbol{Q}\right)\triangleq\max\{U-W-2mV,\Phi'\left(\boldsymbol{Q}\right)\},\forall\boldsymbol{Q}\in\mathcal{S}$, where $m=1,2,\ldots,m^*$. Note that $\Phi_m(\cdot)$ is finite. Since the Markov chain under policy $\tilde{\pi}^*+\pi_0$ is positive recurrent, the mean drift of Φ_m in steady-state must be 0, i.e.,

$$\mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi_m\left(\boldsymbol{Q}(t+1)\right)-\Phi_m\left(\boldsymbol{Q}(t)\right)\right]=0.$$

Define $\Delta\Phi_m\left(\boldsymbol{Q}\right)\triangleq\mathbb{E}_{\tilde{\pi}^*+\pi_0}[\Phi_m\left(\boldsymbol{Q}(t+1)\right)-\Phi_m\left(\boldsymbol{Q}(t)\right)\mid\boldsymbol{Q}(t)=\boldsymbol{Q}]$ as the conditional mean drift of Φ_m. Next, we decompose the mean drift of Φ_m:

$$\begin{aligned} 0 =\ & \mathbb{E}_{\tilde{\pi}^*+\pi_0}\left[\Phi_m\left(\boldsymbol{Q}(t+1)\right)-\Phi_m\left(\boldsymbol{Q}(t)\right)\right] \\ =\ & \sum_{\boldsymbol{Q}\in\mathcal{S}} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}\right)\cdot\Delta\Phi_m\left(\boldsymbol{Q}\right) \\ =\ & \sum_{\boldsymbol{Q}\in\mathcal{S}^{\text{in}}:\Phi'\left(\boldsymbol{Q}\right)\leqslant U-W-(2m+1)V} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}\right)\cdot\Delta\Phi_m\left(\boldsymbol{Q}\right) && (34) \\ &+ \sum_{\boldsymbol{Q}\in\mathcal{S}^{\text{in}}:U-W-(2m+1)V<\Phi'\left(\boldsymbol{Q}\right)\leqslant U-W-(2m-1)V} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}\right)\cdot\Delta\Phi_m\left(\boldsymbol{Q}\right) && (35) \\ &+ \sum_{\boldsymbol{Q}\in\mathcal{S}^{\text{in}}:\Phi'\left(\boldsymbol{Q}\right)>U-W-(2m-1)V} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}\right)\cdot\Delta\Phi_m\left(\boldsymbol{Q}\right) && (36) \\ &+ \sum_{\boldsymbol{Q}\in\mathcal{S}^{\text{out}}} p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q}\right)\cdot\Delta\Phi_m\left(\boldsymbol{Q}\right). && (37) \end{aligned}$$

The key here is to derive the upper bounds of the $\Delta\Phi_m\left(\boldsymbol{Q}\right)$'s in Equations (34)–(37) separately, so that we could analyze the stationary probability of the corresponding regions.

For $\boldsymbol{Q}\in\left\{\boldsymbol{Q}\in\mathcal{S}^{\text{in}}:\Phi'\left(\boldsymbol{Q}\right)\leqslant U-W-(2m+1)V\right\}$: Given $\boldsymbol{Q}(t)=\boldsymbol{Q}$, together with the definition of V, we have that $\Phi'\left(\boldsymbol{Q}(t+1)\right)\leqslant U-W-(2m+1)V+V=U-W-2mV$ and thus $\Phi_m\left(\boldsymbol{Q}(t+1)\right)=$

$\Phi_m(\boldsymbol{Q}(t)) = U - W - 2mV$. Therefore, we have

$$\Delta\Phi_m(\boldsymbol{Q}) = 0. \tag{38}$$

For $\boldsymbol{Q} \in \left\{\boldsymbol{Q} \in \mathcal{S}^{\text{in}} : U - W - (2m+1)V < \Phi'(\boldsymbol{Q}) \leqslant U - W - (2m-1)V\right\}$**:** Given $\boldsymbol{Q}(t) = \boldsymbol{Q}$, by analyzing all the possible relationships among $\Phi'(\boldsymbol{Q}(t))$, $\Phi'(\boldsymbol{Q}(t+1))$ and $U - W - 2mV$, it is not hard to verify that

$$\left|\Phi_m(\boldsymbol{Q}(t+1)) - \Phi_m(\boldsymbol{Q}(t))\right| \leqslant \left|\Phi'(\boldsymbol{Q}(t+1)) - \Phi'(\boldsymbol{Q}(t))\right| \leqslant V,$$

which gives us that

$$\Delta\Phi_m(\boldsymbol{Q}) \leqslant V. \tag{39}$$

For $\boldsymbol{Q} \in \left\{\boldsymbol{Q} \in \mathcal{S}^{\text{in}} : \Phi'(\boldsymbol{Q}) > U - W - (2m-1)V\right\}$**:** Given $\boldsymbol{Q}(t) = \boldsymbol{Q}$, we have $\Phi'(\boldsymbol{Q}(t+1)) > U - W - (2m-1)V - V = U - W - 2mV$, which indicates $\Delta\Phi_m(\boldsymbol{Q}) = \Delta\Phi'(\boldsymbol{Q})$. Since $Q_{\max} \geqslant \Phi'(\boldsymbol{Q})/b_1^{1/\beta} > (U - W - (2m^* - 1)V)/b_1^{1/\beta} \geqslant \tilde{B}^*$, by applying Lemma 10, we have

$$\Delta\Phi_m(\boldsymbol{Q}) = \Delta\Phi'(\boldsymbol{Q}) \leqslant -\epsilon'. \tag{40}$$

For $\boldsymbol{Q} \in \mathcal{S}^{\text{out}}$**:** Since $\Phi_m(\boldsymbol{Q})$ has reached maximum, the drift conditioned on $\boldsymbol{Q}(t) = \boldsymbol{Q}$ cannot be positive, i.e.,

$$\Delta\Phi_m(\boldsymbol{Q}) \leqslant 0. \tag{41}$$

By inserting Equations (38)–(41) into Equations (34)–(37), respectively, we have that

$$0 \leqslant p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q} \in \mathcal{S}^{\text{in}} : U - W - (2m+1)V < \Phi'(\boldsymbol{Q}) \leqslant U - W - (2m-1)V\right) \cdot V - p^{\tilde{\pi}^*+\pi_0}\left(\boldsymbol{Q} \in \mathcal{S}^{\text{in}} : \Phi'(\boldsymbol{Q}) > U - W - (2m-1)V\right) \cdot \tilde{\epsilon}^*,$$

which is equivalent to the statement in Lemma 11 after simple algebraic calculations. □

D PROOF OF THEOREM 2

In this section, we prove Theorem 2. The proof is build on Theorem 1 and Lemmas 1–2.

Proof. By combining Lemmas 1 and 2, we have that

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

Thus there exists a $k_1 < \infty$ such that when $k \geqslant k_1$,

$$\mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L'_k} \mid \pi_k^{\text{in}} = \tilde{\pi}^*\right] - O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right) \leqslant \frac{1}{2} \cdot O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right). \tag{42}$$

To analyze the overall expected queue backlog, we need to further consider three possible cases where $\pi_k^{\text{in}} \neq \tilde{\pi}^*$ for some $k > k^*$:

(i) We have not learned $\tilde{\pi}^*$ at $k = k^*$, which happens with probability at most δ according to Theorem 1;
(ii) We have learned $\tilde{\pi}^*$ when $k = k^*$, but obtain some bad samples and fail to estimate $\tilde{\mathcal{M}}$ in following episodes with probability at most $\delta/2$ (according to Lemma 4);
(iii) π_{rand} may be selected as π_k^{in}, which occurs with probability $\ell/\sqrt{k}$.

By taking union bound over the above three events, we have that

$$\Pr\left\{\pi_k^{\text{in}} \neq \tilde{\pi}^*\right\} \leqslant \delta + \frac{\delta}{2} + \epsilon_k = \frac{3\delta}{2} + \frac{\ell}{\sqrt{k}}, \qquad \forall k > k^*. \tag{43}$$

During episode k, the number of visits to $\mathcal{S}^{\text{in}}$ is L_k, and the associated expected regret is upper bounded by $L_k DU$. Additionally, the Markov chain can enter $\mathcal{S}^{\text{out}}$ from $\mathcal{S}^{\text{in}}$ at most L_k times, and each time the associated expected regret is uniformly upper bounded by $O(U^{1+2\alpha})$ from Lemma 7. Therefore, we have

$$\mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L_k'} \mid \pi_k^{\text{in}} \neq \tilde{\pi}^*\right] \leqslant \frac{L_k \cdot DU + L_k \cdot O\left(U^{1+2\alpha}\right)}{L_k} - \tilde{\rho}^* = O\left(U^{1+2\alpha}\right). \tag{44}$$

By combining Equations (42)–(44), for each $k \geqslant k_2 \triangleq \max\{k^*, k_1\}$, we can upper bound the overall expected episodic regret as follows:

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\left(\sum_i Q_i(t) - \tilde{\rho}^*\right)}{L_k'}\right] &\leqslant \Pr\left\{\pi_k^{\text{in}} = \tilde{\pi}^*\right\} \cdot O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right) + \Pr\left\{\pi_k^{\text{in}} \neq \tilde{\pi}^*\right\} \cdot O\left(U^{1+2\alpha}\right) \\ &\leqslant O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)} + \delta U^{1+2\alpha} + \frac{U^{1+2\alpha}}{\sqrt{k}}\right).\end{aligned} \tag{45}$$

By taking $\delta = \exp(-U)$, we have

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\sum_i Q_i(t)}{L_k'}\right] \leqslant \tilde{\rho}^* + O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right). \tag{46}$$

We now have the overall expected queue backlog in comparison to $\tilde{\rho}^*$. The following lemma states the relationship between $\tilde{\rho}^*$ and ρ^* (see Appendix D.1 for the proof).

Lemma 12. *Under the setting of Theorem 2, we have*

$$\tilde{\rho}^* = \rho^* + O\left(\frac{U^{1+\gamma}}{\exp(U)}\right).$$

By replacing $\tilde{\rho}^*$ in Equation (46) with ρ^* using Lemma 12, we have an upper bound for the overall expected queue backlog as follows:

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\sum_{t=t_{k-1}+1}^{t_k}\sum_i Q_i(t)}{L_k'}\right] \leqslant \rho^* + O\left(\frac{U^{1+\max\{2\alpha,\gamma\}}}{\exp(U)}\right),$$

which completes the proof. □

D.1 Proof of Lemma 12

As defined in Table 1, ρ^* is the average total queue backlog when applying π^* to $\mathcal{M}$, while $\tilde{\rho}^*$ is the average total queue backlog when applying $\tilde{\pi}^*$ to $\tilde{\mathcal{M}}$. To bridge the gap between ρ^* and $\tilde{\rho}^*$, we define the average queue backlog when applying (a truncated version of) π^* to $\tilde{\mathcal{M}}$ as $\tilde{\rho}(\pi^*)$.

Since $\tilde{\pi}^*$ is the optimal policy to $\tilde{\mathcal{M}}$, we have

$$\tilde{\rho}^* \leqslant \tilde{\rho}(\pi^*). \tag{47}$$

Next we compare $\tilde{\rho}(\pi^*)$ and ρ^*. The analysis follows a similar argument as that for Lemmas 1 and 2.

We partition the state space of $\tilde{\mathcal{M}}$, i.e., $\tilde{\mathcal{S}}$, into $\mathcal{S}^{\text{in}}$ and $\tilde{\mathcal{S}}^{\text{out}} \triangleq \tilde{\mathcal{S}} \backslash \mathcal{S}^{\text{in}}$. We define $\tilde{\mathcal{T}}^{\text{in}}$ and $\tilde{\mathcal{T}}^{\text{out}}$ as the set of time slots that $\boldsymbol{Q}(t)$ is in $\mathcal{S}^{\text{in}}$ and $\tilde{\mathcal{S}}^{\text{out}}$, respectively. We then can decompose the expected average regret under the (truncated) policy π^* with respect to ρ^* as follows:

$$\lim_{T\to\infty} \frac{\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\sum_{t=1}^{T}\left(\sum_i Q_i(t) - \rho^*\right)\right]}{T}$$
$$= \lim_{T\to\infty} \frac{\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\sum_{t\in\tilde{\mathcal{T}}^{\text{in}}}\left(\sum_i Q_i(t) - \rho^*\right)\right]}{T} + \lim_{T\to\infty} \frac{\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\sum_{t\in\tilde{\mathcal{T}}^{\text{out}}}\left(\sum_i Q_i(t) - \rho^*\right)\right]}{T}. \tag{48}$$

Bounding the first term in Equation (48): Similar to the analysis in Lemma 1, we define "in process" units as the process that $\boldsymbol{Q}(t)$ leaves $\tilde{\mathcal{S}}^{\text{out}}$, enters $\mathcal{S}^{\text{in}}$, stays in $\mathcal{S}^{\text{in}}$ for some time and finally returns back to $\tilde{\mathcal{S}}^{\text{out}}$. Then, the process during $\tilde{\mathcal{T}}^{\text{in}}$ can be decomposed into multiple "in process" units. An "in process" unit is said to start from $\tilde{\boldsymbol{Q}}$ if $\tilde{\boldsymbol{Q}}$ is its last state before entering into $\mathcal{S}^{\text{in}}$ (i.e. $\tilde{\boldsymbol{Q}} \in \tilde{\mathcal{S}}^{\text{out}}$). The accumulated regret during the ith "in process" unit that starts from $\tilde{\boldsymbol{Q}}$ is denoted by $\tilde{R}_i(\tilde{\boldsymbol{Q}})$.

We use $\tilde{p}^{\pi^*}(\cdot)$ to denote the stationary distribution of states when applying π^* to $\tilde{\mathcal{M}}$. By applying renewal theorem analysis as in the proof of Lemma 1, we have that

$$\lim_{T\to\infty} \frac{\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\sum_{t\in\tilde{\mathcal{T}}^{\text{in}}}\left(\sum_i Q_i(t) - \rho^*\right)\right]}{T} \leqslant \tilde{p}^{\pi^*}\left(\tilde{\mathcal{S}}^{\text{out}}\right) \cdot \max_{\boldsymbol{Q}\in\tilde{\mathcal{S}}^{\text{out}}} \tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\tilde{R}_1\left(\tilde{\boldsymbol{Q}}\right)\right]. \tag{49}$$

Bounding the second term in Equation (48): Since the total queue backlog in $\tilde{M}$ is upper bounded by DU, we simply have that

$$\lim_{T\to\infty} \frac{\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\sum_{t\in\tilde{\mathcal{T}}^{\text{out}}}\left(\sum_i Q_i(t) - \rho^*\right)\right]}{T} \leqslant DU \cdot \lim_{T\to\infty} \frac{\mathbb{E}\left[\tilde{\mathcal{T}}^{\text{out}}\right]}{T} = \tilde{p}^{\pi^*}\left(\tilde{\mathcal{S}}^{\text{out}}\right) \cdot DU. \tag{50}$$

Bounding $\tilde{\mathbb{E}}_{\tilde{\pi}^*}[\tilde{R}_1(\tilde{\boldsymbol{Q}})]$ in Equation (49): According to Bellman's equation, when applying π^* to $\mathcal{M}$, there exists $h^*(\cdot)$ such that the for every $\boldsymbol{Q} \in \mathcal{S}$, we have $\rho^* + h^*(\boldsymbol{Q}) = \sum_i Q_i + \sum_{\boldsymbol{Q}'\in\mathcal{S}} p\left(\boldsymbol{Q}' \mid \boldsymbol{Q}, \pi^*(\boldsymbol{Q})\right) \cdot h^*(\boldsymbol{Q}')$. Note that Bellman's equation might not have solutions for countably infinite state space MDP (see Section 5.6 in Volome 2 of [7]). For simplicity, we assume the existence of $h^*(\cdot)$.

When applying π^* to $\tilde{\mathcal{M}}$, using the definition of $\mathcal{S}^{\text{in}}$ we have that for each $\boldsymbol{Q} \in \mathcal{S}^{\text{in}}$, $\boldsymbol{Q}' \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$, $\tilde{p}(\boldsymbol{Q}' \mid \boldsymbol{Q}, \pi^*(\boldsymbol{Q})) = p(\boldsymbol{Q}' \mid \boldsymbol{Q}, \pi^*(\boldsymbol{Q}))$. Therefore, for each $\boldsymbol{Q} \in \mathcal{S}^{\text{in}}$, we have $\rho^* + h^*(\boldsymbol{Q}) = \sum_i Q_i + \sum_{\boldsymbol{Q}'\in\tilde{\mathcal{S}}} \tilde{p}(\boldsymbol{Q}' \mid \boldsymbol{Q}, \pi^*(\boldsymbol{Q})) \cdot h^*(\boldsymbol{Q}')$.

Following a similar argument as Lemma 6, we obtain

$$\tilde{\mathbb{E}}_{\tilde{\pi}^*}\left[\tilde{R}_1\left(\tilde{\boldsymbol{Q}}\right)\right] \leqslant cDU^{1+\gamma}. \tag{51}$$

Bounding $\tilde{p}^{\pi^*}(\tilde{\mathcal{S}}^{\text{out}})$: The proof follows the same line of argument as that for Lemma 2. Denote by $\Phi^*(\cdot)$ the Lyanopuv function in Assumption 2 for $U = \infty$. We consider a new Lyapunov function $\tilde{\Phi}'(\boldsymbol{Q}) \triangleq [\Phi^*(\boldsymbol{Q})]^{\frac{1}{\beta}}$, for each $\boldsymbol{Q} \in \tilde{\mathcal{S}}$. We define a mapping $TR : \mathcal{S} \to \tilde{\mathcal{S}}$ to represent the packet dropping scheme in the bounded system. In particular, $TR(\boldsymbol{Q}) \triangleq \{\min\{U, Q_i\}\}_{i=1}^{D}$. Note that

$\Phi^*(TR(\boldsymbol{Q})) \leq \Phi^*(\boldsymbol{Q}), \forall \boldsymbol{Q} \in \mathcal{S}$. We then have that

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

We now can apply the analysis from the proof of Lemma 10 to show show that there exist constants $V, \epsilon' > 0$ such that for any $U > 0$, the following properties hold:

(1) For each $\boldsymbol{Q} \in \tilde{\mathcal{S}}$, $\max_{\boldsymbol{Q}' \in \mathcal{R}(\boldsymbol{Q})} |\tilde{\Phi}'\left(\boldsymbol{Q}'\right) - \tilde{\Phi}'\left(\boldsymbol{Q}\right)| \leqslant V$;
(2) For each $\boldsymbol{Q} \in \mathcal{S}^{\text{in}}$ with $Q_{\max} \geqslant \tilde{B}^*$, we have $\tilde{\mathbb{E}}_{\pi^*}[\tilde{\Phi}'(\boldsymbol{Q}(t+1)) - \tilde{\Phi}'(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t) = \boldsymbol{Q}] \leqslant -\epsilon'$.

We define $m^* \triangleq \lfloor (U + V - W - b_1^{1/\beta}\tilde{B}^*)/(2V) \rfloor$ and consider a series of Lyapunov functions $\tilde{\Phi}_m\left(\boldsymbol{Q}\right) \triangleq \max\{U - W - 2mV, \tilde{\Phi}'\left(\boldsymbol{Q}\right)\}$ with $m = 1, 2, \ldots, m^*$. We decompose $\tilde{\mathcal{S}}$ into three sets:

$$\begin{cases} \left\{\boldsymbol{Q} \in \tilde{\mathcal{S}} : \tilde{\Phi}'\left(\boldsymbol{Q}\right) \leqslant U - W - (2m+1)V\right\} \\ \left\{\boldsymbol{Q} \in \tilde{\mathcal{S}} : U - W - (2m+1)V < \tilde{\Phi}'\left(\boldsymbol{Q}\right) \leqslant U - W - (2m-1)V\right\} \\ \left\{\boldsymbol{Q} \in \tilde{\mathcal{S}} : \tilde{\Phi}'\left(\boldsymbol{Q}\right) > U - W - (2m-1)V\right\} \end{cases}.$$

By following the analysis for Equations (38)–(40), we can bound the drifts in these regions, respectively, and show that

$$\tilde{p}^{\pi^*}\left(\boldsymbol{Q} \in \tilde{\mathcal{S}} : \tilde{\Phi}'\left(\boldsymbol{Q}\right) > U - W - (2m-1)V\right) \leqslant \frac{V \cdot \tilde{p}^{\pi^*}\left(\boldsymbol{Q} \in \tilde{\mathcal{S}} : \tilde{\Phi}'\left(\boldsymbol{Q}\right) > U - W - (2m+1)V\right)}{V + \tilde{\epsilon}^*},$$

where $m = 1, 2, \ldots, m^*$. Similar to Lemma 2, we obtain that $\tilde{p}^{\pi^*}(\tilde{\mathcal{S}}^{\text{out}}) = \exp(-U)$.

Inserting Equation (51) into Equations (49) and (50), together with Equation (47), gives

$$\tilde{\rho}^* \leqslant \tilde{\rho}(\pi^*) = \lim_{T \to \infty} \frac{\tilde{\mathbb{E}}\left[\sum_{t=1}^{T} \sum_i Q_i(t)\right]}{T} \leq \rho^* + \mathcal{O}\left(\frac{U^{1+\gamma}}{\exp\left(U\right)}\right),$$

which completes the proof.

REFERENCES

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. 2019. Politex: Regret bounds for policy iteration using expert prediction. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3692–3702.

[2] Afef Ben Hadj Alaya-Feki, Eric Moulines, and Alain LeCornec. 2008. Dynamic spectrum access with non-stationary multi-armed bandit. In *Proceedings of the 2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*. IEEE, 416–420.

[3] Baruch Awerbuch and Tom Leighton. 1993. A simple local-control approximation algorithm for multicommodity flow. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*. IEEE, 459–468.

[4] Leemon Baird. 1995. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 1995 Machine Learning*. Elsevier, 30–37.

[5] J. S. Baras, D.-J. Ma, and A. M. Makowski. 1985. K competing queues with geometric service requirements and linear costs: The μc-rule is always optimal. *Systems & Control Letters* 6, 3 (1985), 173–180.

[6] Peter L. Bartlett and Ambuj Tewari. 2012. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. arXiv:1205.2661. Retrieved from https://arxiv.org/abs/1205.2661.

[7] Dimitri Bertsekas. 2017. *Dynamic Programming and Optimal Control*. Vol. 1. Athena scientific Belmont, MA.

[8] Dimitris Bertsimas, David Gamarnik, John N. Tsitsiklis, et al. 1998. Geometric bounds for stationary distributions of infinite markov chains via lyapunov functions. https://www.researchgate.net/profile/Dimitris-Bertsimas/publication/37594830_Geometric_Bounds_for_Stationary_Distributions_of_Infinite_Markov_Chains_Via_Lyapunov_Functions/links/53f389350cf2155be34f3316/Geometric-Bounds-for-Stationary-Distributions-of-Infinite-Markov-Chains-Via-Lyapunov-Functions.pdf.

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

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

[11] Justin A. Boyan and Michael L. Littman. 1994. Packet routing in dynamically changing networks: A reinforcement learning approach. In *Proceedings of the Advances in Neural Information Processing Systems*. Citeseer, 671–678.

[12] Pierre Brémaud. 1999. Lyapunov functions and martingales. In *Proceedings of the Markov Chains*. Springer, 167–193.

[13] Timothy X. Brown. 2001. Switch packet arbitration via queue-learning. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. 1337–1344.

[14] Olivier Brun, Lan Wang, and Erol Gelenbe. 2016. Big data for autonomic intercontinental overlays. *IEEE Journal on Selected Areas in Communications* 34, 3 (2016), 575–583.

[15] C. Buyukkoc, P. Varaiya, and J. Walrand. 1985. The cμ rule revisited. *Advances in Applied Probability* 17, 1 (1985), 237–238.

[16] Rong-Rong Chen and Sean Meyn. 1999. Value iteration and optimization of multiclass queueing networks. *Queueing Systems* 32, 1 (1999), 65–97.

[17] Xiaoliang Chen, Jiannan Guo, Zuqing Zhu, Roberto Proietti, Alberto Castro, and S. J. Ben Yoo. 2018. Deep-RMSA: A deep-reinforcement-learning routing, modulation and spectrum assignment agent for elastic optical networks. In *Proceedings of the 2018 Optical Fiber Communications Conference and Exposition*. IEEE, 1–3.

[18] Ying Cui, Vincent K. N. Lau, Rui Wang, Huang Huang, and Shunqing Zhang. 2012. A survey on delay-aware resource control for wireless systems-large deviation theory, stochastic lyapunov drift, and distributed stochastic learning. *IEEE Transactions on Information Theory* 58, 3 (2012), 1677–1701.

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

[20] Tim de Bruin, Jens Kober, Karl Tuyls, and Robert Babuška. 2018. Integrating state representation learning into deep reinforcement learning. *IEEE Robotics and Automation Letters* 3, 3 (2018), 1394–1401.

[21] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. 2018. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1578–1586.

[22] Anand Ganti, Eytan Modiano, and John N. Tsitsiklis. 2007. Optimal transmission scheduling in symmetric communication models with intermittent connectivity. *IEEE Transactions on Information Theory* 53, 3 (2007), 998–1008.

[23] Ziwei Guan, Timothy Y. Hu, Joseph Palombo, Michael J. Liston, Donald J. Bucci, and Yingbin Liang. 2020. Robust dynamic spectrum access in adversarial environments. In *Proceedings of the ICC 2020-2020 IEEE International Conference on Communications*. IEEE, 1–7.

[24] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research* 19, 1 (2003), 399–468.

[25] Zhenting Hou and Qingfeng Guo. 2012. *Homogeneous Denumerable Markov Processes*. Springer Science & Business Media.

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

[27] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11, 51 Apr (2010), 1563–1600.

[28] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. 2020. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the Conference on Learning Theory*. PMLR, 2137–2143.

[29] Isaac Keslassy and Nick McKeown. 2001. Analysis of scheduling algorithms that provide 100% throughput in input-queued switches. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, Vol. 39. Citeseer, 593–602.

[30] M. Khantha and V. Balakrishnan. 1983. First passage time distributions for finite one-dimensional random walks. *Pramana* 21, 2 (1983), 111–122.

[31] Daphne Koller and Ron Parr. 2000. Policy iteration for factored MDPs. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*.
[32] P. R. Kumar and Sean P. Meyn. 1995. Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* 40, 2 (1995), 251–260.
[33] Sunil Kumar and P. R. Kumar. 1994. Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* 39, 8 (1994), 1600–1611.
[34] Qingkai Liang and Eytan Modiano. 2019. Optimal network control in partially-controllable networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE.
[35] Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. 1985. On the complexity of solving markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*.
[36] Dor Livne and Kobi Cohen. 2020. PoPS: Policy pruning and shrinking for deep reinforcement learning. *IEEE Journal of Selected Topics in Signal Processing* 14, 4 (2020), 789–801.
[37] Siva Theja Maguluri and R. Srikant. 2016. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stochastic Systems* 6, 1 (2016), 211–250.
[38] Odalric-Ambrym Maillard, Rémi Munos, and Daniil Ryabko. 2013. Selecting the state-representation in reinforcement learning. arXiv:1302.2552. Retrieved from https://arxiv.org/abs/1302.2552.
[39] Nick McKeown, Adisak Mekkittikul, Venkat Anantharam, and Jean Walrand. 1999. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications* 47, 8 (1999), 1260–1267.
[40] Adisak Mekkittikul and Nick McKeown. 1998. A practical scheduling algorithm to achieve 100% throughput in input-queued switches. In *Proceedings. IEEE INFOCOM'98, the Conference on Computer Communications. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, 792–799.
[41] Jean-François Mertens, Ester Samuel-Cahn, and Shmuel Zamir. 1978. Necessary and sufficient conditions for recurrence and transience of markov chains, in terms of inequalities. *Journal of Applied Probability* 15, 4 (1978), 848–851.
[42] Sean P. Meyn and Richard L. Tweedie. 2012. *Markov Chains and Stochastic Stability*. Springer Science & Business Media.
[43] Oshri Naparstek and Kobi Cohen. 2017. Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks. In *Proceedings of the GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 1–7.
[44] Michael J. Neely and Longbo Huang. 2010. Dynamic product assembly and inventory control for maximum profit. In *Proceedings of the 49th IEEE Conference on Decision and Control*. IEEE, 2805–2812.
[45] Michael J. Neely, Eytan Modiano, and Chih-Ping Li. 2008. Fairness and optimal stochastic control for heterogeneous networks. *IEEE/ACM Transactions On Networking* 16, 2 (2008), 396–409.
[46] Michael J. Neely, Eytan Modiano, and Charles E. Rohrs. 2003. Dynamic power allocation and routing for time varying wireless networks. In *Proceedings of the IEEE INFOCOM 200322nd Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, 745–755.
[47] Ronald Ortner. 2020. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research* 67 (2020), 115–128.
[48] Ian Osband, Daniel Russo, and Benjamin Van Roy. 2013. (More) efficient reinforcement learning via posterior sampling. In *Proceedings of the Advances in Neural Information Processing Systems*. 3003–3011.
[49] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. 2017. Learning unknown markov decision processes: A thompson sampling approach. *Advances in Neural Information Processing Systems* 30 (2017).
[50] Leonid Peshkin and Virginia Savova. 2002. Reinforcement learning for adaptive routing. In *Proceedings of the 2002 International Joint Conference on Neural Networks*. IEEE, 1825–1830.
[51] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. 2020. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems* 33 (2020), 2074–2086.
[52] Sheldon M. Ross and Sridhar Seshadri. 1999. Hitting time in an M/G/1 queue. *Journal of Applied Probability* 30, 3 (1999), 934–940.
[53] Devavrat Shah and Damon Wischik. 2012. Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse. *Annals of Applied Probability* 22, 1 (2012), 70–127.
[54] Ramesh K. Sitaraman, Mangesh Kasbekar, Woody Lichtenstein, and Manish Jain. 2014. Overlay networks: An akamai perspective. *Advanced Content Delivery, Streaming, and Cloud Services* 51, 4 (2014), 305–328.
[55] Shabnam Sodagari and T. Charles Clancy. 2011. An anti-jamming strategy for channel access in cognitive radio networks. In *Proceedings of the International Conference on Decision and Game Theory for Security*. Springer, 34–43.
[56] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the Advances in Neural Information Processing Systems*. 1057–1063.

[57] Leandros Tassiulas and Anthony Ephremides. 1990. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *Proceedings of the 29th IEEE Conference on Decision and Control*. IEEE, 2130–2132.
[58] Leandros Tassiulas and Anthony Ephremides. 1993. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory* 39, 2 (1993), 466–478.
[59] Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. 2017. Posterior sampling for large scale reinforcement learning. arXiv:1711.07979. Retrieved from https://arxiv.org/abs/1711.07979.
[60] Michael H. Veatch. 2010. *Approximate Linear Programming for Networks: Average Cost Bounds*. Technical Report. Working paper, Gordon College.
[61] Shangxing Wang, Hanpeng Liu, Pedro Henrique Gomes, and Bhaskar Krishnamachari. 2018. Deep reinforcement learning for dynamic multichannel access in wireless networks. *IEEE Transactions on Cognitive Communications and Networking* 4, 2 (2018), 257–265.
[62] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8, 3–4 (1992), 279–292.
[63] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. 2020. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10170–10180.
[64] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. 2003. Inequalities for the L1 deviation of the empirical distribution. Hewlett-Packard Labs, Technical Report HPL-2003-97R1.
[65] Edmund Meng Yeh. 2001. *Multiaccess and Fading in Communication Networks*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[66] Edmund M. Yeh. 2003. Throughput and delay optimal resource allocation in multiple access fading channels. In *Proceedings of the IEEE International Symposium on Information Theory*.
[67] Changhe Yu, Julong Lan, Zehua Guo, and Yuxiang Hu. 2018. DROM: Optimizing the routing in software-defined networks with deep reinforcement learning. *IEEE Access* 6 (2018), 64533–64539.
[68] Zihan Zhang and Xiangyang Ji. 2019. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems* 32 (2019).

Received September 2020; revised December 2021; accepted March 2022