SIAM J. CONTROL OPTIM.
Vol. 61, No. 3, pp. 1861–1880

A DISCRETE-TIME SWITCHING SYSTEM ANALYSIS OF Q-LEARNING*

DONGHWAN LEE†, JIANGHAI HU‡, AND NIAO HE§

Abstract. This paper develops a novel control-theoretic framework to analyze the nonasymptotic convergence of Q-learning. We show that the dynamics of asynchronous Q-learning with a constant step size can be naturally formulated as a discrete-time stochastic affine switching system. In particular, for a given Q-function parameter, Q, the greedy policy, $\pi_Q(s) := \arg\max_a Q(s,a)$, in the Q-learning update plays the role of the switching policy, and is the key connection between the switching system and Q-learning. Then, the evolution of the Q-learning estimation error is over- and under-estimated by trajectories of two simpler dynamical systems. Based on these two systems, we derive a new finite-time error bound of asynchronous Q-learning when a constant step size is used. In addition, the new analysis sheds light on the overestimation phenomenon of Q-learning.

Key words. Q-learning, switched linear system, stochastic approximation

MSC codes. 68Q25, 68R10, 68U05

DOI. 10.1137/22M1489976

1. Introduction.

First introduced by Watkins and Dayan [20], Q-learning is one of the most fundamental and important reinforcement learning algorithms. The theoretical behavior of Q-learning has been extensively studied over the years. Classical analysis of Q-learning mostly focused on asymptotic convergence of asynchronous Q-learning [18, 8] and synchronous Q-learning [3]. Substantial advances have been made recently in the guarantee of their finite-time convergence; see [17, 10, 5, 1, 2, 19, 15, 13, 4].

To list a few, Szepesvári in [17] gave the first nonasymptotic analysis of asynchronous Q-learning under an independent and identically distributed (i.i.d.) sampling setting. [5] first provided the nonasymptotic analysis for both synchronous and asynchronous Q-learning with polynomial and linear step sizes under a single trajectory Markovian sampling setting. Recently, [19] established the best known bound for synchronous Q-learning under a rescaled linear step size. In a subsequent work, [15] derived a matching bound for asynchronous Q-learning under the Markovian setting using a similar decaying step size. The sample complexity is further improved with a refined analysis based on constant step size in [13] and [4].

* Received by the editors April 12, 2022; accepted for publication (in revised form) December 7, 2022; published electronically June 26, 2023.
https://doi.org/10.1137/22M1489976
Funding: This material is based upon work supported by the National Research Foundation under Grant NRF-2021R1F1A1061613. The work was supported by the BK21 FOUR from the Ministry of Education (Republic of Korea) This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (2022-0-00469). This work was is supported by ETH research grant, NCCR Automation, and Swiss National Science Foundation Project Funding 200021_207343. This work was supported by the National Science Foundation under Grants 2014816 and 2038410.

† Department of Electrical and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (donghwan@kaist.ac.kr).

‡ Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 USA (jianghai@purdue.edu).

§ Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland (niao.he@inf.ethz.ch).

Existing results for the most part treat the Q-learning dynamics as a special case of general nonlinear stochastic approximation schemes with Markovian noises. In a different line of work, [12] discovered a close connection between Q-learning and continuous-time switching systems. The switching system perspective captures unique features of Q-learning dynamics and encapsulates a wide spectrum of Q-learning algorithms including asynchronous Q-learning, averaging Q-learning [12], and Q-learning with function approximation, etc. However, existing O.D.E. analysis of such continuous-time switching systems yields only asymptotic convergences of Q-learning algorithms and requires diminishing step sizes. Obtaining a finite-time convergence analysis would require a departure of the switching systems from the continuous-time domain to the discrete-time domain, which remains an open and challenging question.

In this paper, we aim to close this gap and provide a new finite-time error bound of Q-learning through the lens of discrete-time switching systems. In particular, we focus on asynchronous Q-learning with constant step sizes for solving a discounted Markov decision process with finite state and action spaces. We first show that asynchronous Q-learning with a constant step size can be naturally formulated as a stochastic discrete-time affine switching system. This allows us to transform the convergence analysis into a stability analysis of the switching system. However, its stability analysis is nontrivial due to the presence of the affine term and the noise term. The main breakthrough in our analysis lies in developing upper and lower comparison systems whose trajectories over- and under-estimate the original system's trajectory. The lower comparison system is a stochastic linear system, while the upper comparison system is a stochastic linear switching system [14], both of which have a much simpler structure than the original system or general nonlinear systems. Our finite-time error bound of Q-learning follows immediately by combining the error bounds of the stochastic linear system (i.e., lower comparison system, which has no affine term) and the error system (i.e., difference of the two comparison systems, which has no noise term). Comparing this to existing analyses based on nonlinear stochastic approximation schemes, our analysis seems more intuitive and builds on simple systems. It also sheds new light on the overestimation phenomenon in Q-learning due to the maximization bias [7].

Last, we emphasize that our goal is to provide new insights and an analysis framework to lay out a strong theoretical foundation for Q-learning via its unique connection to discrete-time switching systems, rather than improving existing convergence rates. In particular, as opposed to classical ODE analysis/stochastic approximation approaches, the proposed strategy adopts the idea of formulating the Q-learning algorithm as a stochastic affine switching system, and directly conducting analysis in discrete time, which is new in the literature. The switching system model of Q-learning in this paper allows us to use already well-established tools in control theory such as Lyapunov analysis, which make the analysis easier and more familiar to researchers in control community. Therefore, we expect that, such a control-theoretic analysis could promote more research activities of people with control backgrounds for reinforcement learning, further stimulate the synergy between control theory and reinforcement learning, and open up opportunities to the design of new reinforcement learning algorithms and refined analysis for Q-learning, such as double Q-learning [7], distributed Q-learning [9], and speedy Q-learning [1].

Moreover, the proposed analysis follows a particularly clean and simple strategy. The core idea that leads to the simplicity is identifying two simpler dynamical systems: a "lower comparison system" that is a stochastic linear system and "upper comparison

system" that is a stochastic switching system, which have favorable structures that are easily analyzed via control theory; stability of a linear system can be used to derive a finite-time error bound for Q-learning. Overall, we view our analysis technique as a complement rather than a replacement of existing techniques for Q-learning analysis. Moreover, our approach based on the comparison systems could be of independent interest to the finite-time stability analysis of more general switching systems.

The overall paper consists of the following parts: section 2 provides preliminary discussions including basics of Markov decision process, switching system, Q-learning, and useful definitions and notations used throughout the paper; section 3 provides the main results of the paper, including the switched system models of Q-learning, upper and lower comparison systems, and the finite-time error bounds; we conclude in section 4 with a discussion on potential extensions of this work.

2. Preliminaries.

2.1. Markov decision problem. We consider the infinite-horizon discounted Markov decision problem (MDP), where a decision making agent sequentially takes actions to maximize cumulative discounted rewards in environments called the Markov decision process. The Markov decision process is a mathematical model of dynamical systems with the state-space $\mathcal{S} := \{1, 2, \dots, |\mathcal{S}|\}$ and action-space $\mathcal{A} := \{1, 2, \dots, |\mathcal{A}|\}$. In a Markov decision process, the decision maker selects an action $a \in \mathcal{A}$ with the current state s, then the state transits to the next state s' with probability $P(s'|s, a)$, and the transition incurs a reward $r(s, a, s')$. For convenience, we consider a deterministic reward function and simply write $r(s_k, a_k, s_{k+1}) =: r_k, k \in \{0, 1, \dots\}$.

A deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, maps a state $s \in \mathcal{S}$ to an action $\pi(s) \in \mathcal{A}$. The objective of the MDP is to find a deterministic optimal policy, π^*, such that the cumulative discounted rewards over infinite-time horizons are maximized, i.e.,

$$\pi^* := \arg\max_{\pi \in \Theta} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \middle| \pi\right],$$

where $\gamma \in [0, 1)$ is the discount factor, Θ is the set of all admissible deterministic policies, $(s_0, a_0, s_1, a_1, \dots)$ is a state-action trajectory generated by the Markov chain indicated by the policy π, and $\mathbb{E}[\cdot|\pi]$ is an expectation conditioned on the policy π. The Q-function under policy π is defined as

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \middle| s_0 = s, a_0 = a, \pi\right], \quad s \in \mathcal{S}, a \in \mathcal{A},$$

and the optimal Q-function is defined as $Q^*(s, a) = Q^{\pi^*}(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Once Q^* is known, then an optimal policy can be retrieved by the greedy policy $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a)$. Throughout, we assume that the MDP is ergodic so that the stationary state distribution exists and the MDP is well posed.

2.2. Switching system. Since the switching system is a special form of nonlinear systems, we first consider the general nonlinear system

$$x_{k+1} = f(x_k), \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{1, 2, \dots\}, \tag{2.1}$$

where $x_k \in \mathbb{R}^n$ is the state and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. An important concept in dealing with the nonlinear system is the equilibrium point. A point $x = x^*$ in the state-space $\mathbb{R}^n$ is said to be an equilibrium point of (2.1) if it has the property

that whenever the state of the system starts at x^*, it will remain at x^* [11]. For (2.1), the equilibrium points are the real roots of the equation $f(x) = x$. The equilibrium point x^* is said to be globally asymptotically stable if for any initial state $x_0 \in \mathbb{R}^n$, $x_k \to x^*$ as $k \to \infty$.

Next, let us consider the particular nonlinear system, the *linear switching system*,

$$x_{k+1} = A_{\sigma_k} x_k, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\}, \tag{2.2}$$

where $x_k \in \mathbb{R}^n$ is the state, $\sigma \in \mathcal{M} := \{1, 2, \ldots, M\}$ is called the mode, $\sigma_k \in \mathcal{M}$ is called the switching signal, and $\{A_\sigma, \sigma \in \mathcal{M}\}$ are called the subsystem matrices. The switching signal can be either arbitrary or controlled by the user under a certain switching policy. Especially, a state-feedback switching policy is denoted by $\sigma_k = \sigma(x_k)$. A more general class of systems is the *affine switching system*

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},$$

where $b_{\sigma_k} \in \mathbb{R}^n$ is the additional input vector, which also switches according to σ_k. Due to the additional input b_{σ_k}, its stabilization becomes much more challenging.

2.3. Revisiting Q-learning. We now briefly review the standard Q-learning and its convergence. Recall the Q-learning update:

$$\begin{aligned} Q_{k+1}(s_k, a_k) = {} & Q_k(s_k, a_k) \\ & + \alpha_k(s_k, a_k) \left\{ r_k + \gamma \max_{u \in \mathcal{A}} Q_k(s_{k+1}, u) - Q_k(s_k, a_k) \right\}, \end{aligned}$$

where $0 \le \alpha_k(s, a) \le 1$, is called the learning rate or step size associated with the state-action pair (s, a) at iteration k. This value is assumed to be zero if $(s, a) \ne (s_k, a_k)$. If

$$\sum_{k=0}^{\infty} \alpha_k(s, a) = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2(s, a) < \infty,$$

and every state-action pair is visited infinitely often, then the iterate is guaranteed to converge to Q^* with probability one [16]. Note that the state-action pair can be visited arbitrarily, which is more general than stochastic visiting rules.

In this paper, we focus on the following setting: $\{(s_k, a_k)\}_{k=0}^{\infty}$ are i.i.d. samples under a behavior policy β, where the behavior policy is the policy by which the reinforcement learning agent actually behaves to collect experiences. For simplicity, we assume that the state at each time is sampled from the state distribution p and, in this case, the state-action distribution at each time is identically given by

$$d(s, a) = p(s)\beta(a|s), \quad (s, a) \in \mathcal{S} \times \mathcal{A}.$$

2.4. Assumptions and definitions. Throughout, we make the following standard assumptions.

Assumption 2.1. $d(s, a) > 0$ holds for all $s \in \mathcal{S}, a \in \mathcal{A}$.

Assumption 2.2. The step size is a constant $\alpha \in (0, 1)$.

Assumption 2.3. The reward is bounded as follows:

$$\max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s, a, s')| =: R_{\max} \le 1.$$

Assumption 2.4. The initial iterate Q_0 satisfies $\|Q_0\|_\infty \le 1$.

Remark 2.5. All the assumptions are standard and widely used in the reinforcement learning literature. All these assumptions will be used throughout this paper for the convergence proofs. Assumption 2.1 guarantees that every state-action pair is visited infinitely often with probability one for sufficient exploration. This assumption corresponds to the sufficient exploration condition in the standard Q-learning analysis [8]: every state-action pair (s,a) is visited infinitely often. Moreover, this assumption is used when the state-action visit distribution is given. It has also been considered in [13] and [4]. The work in [2] considers another deterministic exploration condition, called the cover time condition, which states that there is a certain time period, within which all the state-action pairs are expected to be visited at least once. Slightly different cover time conditions have been used in [5] and [13] for convergence rate analysis. Assumption 2.3 is required to ensure the boundedness of Q-learning iterates, which is applied in almost all reinforcement learning algorithms. The unit bounds imposed on $R_{\max}$ and Q_0 are just for simplicity of analysis. The constant step size in Assumption 2.2 has been also studied in [2] and [4] using different approaches.

The following quantities will be frequently used in this paper; hence, we define them for convenience.

DEFINITION 2.6.

1. *Maximum state-action visit probability:*

$$d_{\max} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s,a) \in (0,1).$$

2. *Minimum state-action visit probability:*

$$d_{\min} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s,a) \in (0,1).$$

3. *Exponential decay rate:*

$$\rho := 1 - \alpha d_{\min}(1-\gamma). \tag{2.3}$$

Under Assumption 2.2*, the decay rate satisfies* $\rho \in (0,1)$.

Throughout the paper, we will use the following compact vector and matrix notations for dynamical system representations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix}, R := \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix}, Q := \begin{bmatrix} Q(\cdot,1) \\ \vdots \\ Q(\cdot,|\mathcal{A}|) \end{bmatrix},$$

$$D_a := \begin{bmatrix} d(1,a) & & \\ & \ddots & \\ & & d(|\mathcal{S}|,a) \end{bmatrix}, D := \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{|\mathcal{A}|} \end{bmatrix}, \tag{2.4}$$

where $P_a = P(\cdot|a,\cdot) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, $Q(\cdot,a) \in \mathbb{R}^{|\mathcal{S}|}, a \in \mathcal{A}$, and $R_a(s) := \mathbb{E}[r(s,a,s')|s,a]$. Note that $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$, $R \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and $D \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$. In this notation, the Q-function is encoded as a single vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, which enumerates $Q(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The single value $Q(s,a)$ can be written as $Q(s,a) = (e_a \otimes e_s)^T Q$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are sth basis vectors (all components are 0

except for the sth component which is 1) and ath basis vectors, respectively. Note also that under Assumption 2.1, D is a nonsingular diagonal matrix with strictly positive diagonal elements.

For any stochastic policy, $\pi : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, where $\Delta_{|\mathcal{A}|}$ is the set of all probability distributions over $\mathcal{A}$, we define the corresponding action transition matrix as

$$\Pi^\pi := \begin{bmatrix} \pi(1)^T \otimes e_1^T \\ \pi(2)^T \otimes e_2^T \\ \vdots \\ \pi(|S|)^T \otimes e_{|\mathcal{S}|}^T \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \tag{2.5}$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$. Then, it is well known that $P\Pi^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ is the transition probability matrix of the state-action pair under policy π. If we consider a deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, the stochastic policy can be replaced with the corresponding one-hot encoding vector $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{A}|}$, where $e_a \in \mathbb{R}^{|\mathcal{A}|}$, and the corresponding action transition matrix is identical to (2.5) with π replaced with $\vec{\pi}$. For any given $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, denote the greedy policy w.r.t. Q as

$$\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a) \in \mathcal{A}. \tag{2.6}$$

We will frequently use the following shorthand:

$$\Pi_Q := \Pi^{\pi_Q}.$$

We note that this notation, $\Pi_Q := \Pi^{\pi_Q}$, will play an important role in the derivation of the switching system model in this paper. In particular, the matrix appears in the system parameters, and switches as the greedy policy $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a) \in \mathcal{A}$ is changed according to Q.

The boundedness of Q-learning iterates [6] plays an important role in our analysis.

LEMMA 2.7 (boundedness of Q-learning iterates [6]). *If the step size is less than one, then for all* $k \geq 0$,

$$\|Q_k\|_\infty \leq Q_{\max} := \frac{\max\{R_{\max}, \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} Q_0(s,a)\}}{1-\gamma}.$$

From Assumptions 2.3 *and* 2.4, *we can easily see that* $Q_{\max} \leq \frac{1}{1-\gamma}$.

3. Finite-time analysis of Q-learning from switching system theory. In this section, we study a discrete-time switching system model of Q-learning and establish its finite-time convergence based on the stability analysis of switching systems. We consider a version of Q-learning given in Algorithm 3.1. Compared to the original Q-learning, the step size, α, does not depend on the state-action pair and is constant in this paper. Moreover, the output of Algorithm 3.1 is the average, $\tilde{Q}_k = \frac{1}{k}\sum_{i=0}^{k-1} Q_k, k \geq 1$, instead of the final iteration Q_k.

3.1. Q-learning as a stochastic affine switching system. Using the notation introduced, the update in Algorithm 3.1 can be rewritten as

$$Q_{k+1} = Q_k + \alpha\{DR + \gamma DP\Pi_{Q_k} Q_k - DQ_k + w_k\}, \tag{3.1}$$

Algorithm 3.1 Q-learning with a constant step size.

Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ randomly such that $\|Q_0\|_\infty \le 1$.
Set $\tilde{Q}_0 = Q_0$
Sample $s_0 \sim p$
for iteration $k = 0, 1, \ldots$ **do**
    Sample $a_k \sim \beta(\cdot|s_k)$
    Sample $s'_k \sim P(\cdot|s_k, a_k)$ and $r_k = r(s_k, a_k, s'_k)$
    Update $Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha\{r_k + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k)\}$
    Update $\tilde{Q}_{k+1} = \tilde{Q}_k + \frac{1}{k+1}(Q_k - \tilde{Q}_k)$
end for

where

$$\begin{aligned} w_k =& (e_{a_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{s_k})(e_{s'_k})^T \Pi_{Q_k} Q_k \\ & - (e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^T Q_k - (DR + \gamma DP \Pi_{Q_k} Q_k - DQ_k), \end{aligned} \tag{3.2}$$

and (s_k, a_k, r_k, s'_k) is the sample in the kth time step.

Remark 3.1. Note that in Algorithm 3.1, (s_k, a_k, s'_k) is sampled from the joint distribution

$$P(s'_k|s_k, a_k)p(s_k)\beta(a_k|s_k) = P(s'_k|s_k, a_k)d(s_k, a_k),$$

which is represented by the matrix multiplication, DP, in (3.1). By the definition of matrix D in (2.4), it is a diagonal matrix whose diagonal entries are an enumeration of $d(s, a) = p(s)\beta(a|s), (s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, it is easy to see that an entry of DP is a joint distribution of a certain $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Moreover, from the definition of matrix Π^π in (2.5) and the greedy policy in (2.6), the multiplication $\Pi_{Q_k} Q_k$ in (3.2) represents that max operator in the Q-function update in (3.1).

In more details, a vector form of the Q-function update in (3.1) can be written as

$$Q_{k+1} = Q_k + \alpha((e_{a_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{s_k})(e_{s_{k'}})^T \Pi_{Q_k} Q_k - (e_{a_k} \otimes e_{s_k})^T Q_k). \tag{3.3}$$

Taking the conditional expectation conditioned on Q_k leads to the mean dynamic

$$\mathbb{E}[Q_{k+1}|Q_k] = Q_k + \alpha(DR + \gamma DP \Pi_{Q_k} Q_k - DQ_k), \tag{3.4}$$

where $D = \mathbb{E}[(e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^T|Q_k]$ and $DP = \mathbb{E}[(e_{a_k} \otimes e_{s_k})(e_{s_{k'}})^T|Q_k]$. Adding the right-hand side of (3.4) to (3.3) and subtracting it from (3.3), we obtain (3.1).

Moreover, by definition, the noise term has the zero mean conditioned on Q_k, i.e., $\mathbb{E}[w_k|Q_k] = 0$. Recall the definitions of $\pi_Q(s)$ and Π_Q. Invoking the optimal Bellman equation $(\gamma DP \Pi_{Q^*} - D)Q^* + DR = 0$, (3.1) can be further rewritten as

$$\begin{aligned} (Q_{k+1} - Q^*) =& \{I + \alpha(\gamma DP \Pi_{Q_k} - D)\}(Q_k - Q^*) + \alpha\gamma DP(\Pi_{Q_k} - \Pi_{Q^*})Q^* + \alpha w_k, \end{aligned} \tag{3.5}$$

which is a linear switching system with an extra affine term, $\gamma DP(\Pi_{Q_k} - \Pi_{Q^*})Q^*$, and stochastic noise w_k. For notational simplicity, given any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, define

$$A_Q := I + \alpha(\gamma DP \Pi_Q - D), \quad b_Q := \alpha\gamma DP(\Pi_Q - \Pi_{Q^*})Q^*.$$

Using the notation, the Q-learning iteration can be concisely represented as the *stochastic affine switching system*

$$Q_{k+1} - Q^* = A_{Q_k}(Q_k - Q^*) + b_{Q_k} + \alpha w_k, \tag{3.6}$$

where A_{Q_k} and b_{Q_k} switch among matrices from $\{I + \alpha(\gamma DP\Pi^\pi - D) : \pi \in \Theta\}$ and vectors from $\{\alpha\gamma DP(\Pi^\pi - \Pi^{\pi^*})Q^* : \pi \in \Theta\}$ respectively. Note that in the switching system in (3.6), the switching signal is not arbitrary, and the switching signal follows a switching rule associated with the greedy policy $\pi_{Q_k}(s) := \arg\max_{a\in\mathcal{A}} Q_k(s,a) \in \mathcal{A}$, which changes according to Q_k.

Therefore, the convergence of Q-learning is now reduced to analyzing the stability of the above switching system. A main obstacle in proving the stability arises from the presence of the affine and stochastic terms. Without these terms, we can easily establish the exponential stability of the corresponding deterministic switching system under an arbitrary switching policy. Specifically, we have the following result.

PROPOSITION 3.2. *For arbitrary $H_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, k \geq 0$, the linear switching system*

$$Q_{k+1} - Q^* = A_{H_k}(Q_k - Q^*), \quad Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

is exponentially stable with

$$\|Q_k - Q^*\|_\infty \leq \rho^k \|Q_0 - Q^*\|_\infty, \quad k \geq 0,$$

where ρ is defined in (2.3).

The above result follows immediately from the key fact that $\|A_Q\|_\infty \leq \rho$, which is formally stated in the next lemma.

LEMMA 3.3. *For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$\|A_Q\|_\infty \leq \rho.$$

Here the matrix norm $\|A\|_\infty := \max_{1\leq i\leq m} \sum_{j=1}^n |A_{ij}|$ and A_{ij} is the element of A in the ith row and jth column.

Proof. Note the following identities

$$\begin{aligned}
\sum_j |[A_Q]_{ij}| &= \sum_j |[I - \alpha D + \alpha\gamma DP\Pi_Q]_{ij}| \\
&= [I - \alpha D]_{ii} + \sum_j [\alpha\gamma DP\Pi_Q]_{ij} \\
&= 1 - \alpha[D]_{ii} + \alpha\gamma[D]_{ii} \sum_j [P\Pi_Q]_{ij} \\
&= 1 - \alpha[D]_{ii} + \alpha\gamma[D]_{ii} \\
&= 1 + \alpha[D]_{ii}(\gamma - 1),
\end{aligned}$$

where the second line is due to the fact that A_Q is a nonnegative matrix. Taking the maximum over i, we have

$$\begin{aligned}
\|A_Q\|_\infty &= \max_{i\in\{1,2,\ldots,|\mathcal{S}||\mathcal{A}|\}} \{1 + \alpha[D]_{ii}(\gamma - 1)\} \\
&= 1 - \alpha \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s,a)(1-\gamma) \\
&= \rho,
\end{aligned}$$

which completes the proof. ∎

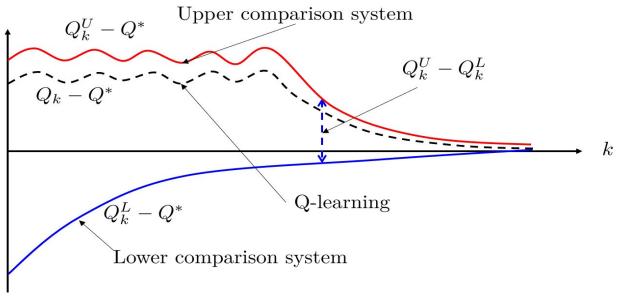

FIG. 1. *Overview of the proposed analysis.*

However, because of the additional affine term and stochastic noises in the original switching system (3.6), it is not obvious how to directly derive its finite-time convergence. To circumvent the difficulty with the affine term, we will resort to two simpler comparison systems, whose trajectories upper and lower bound that of the original system, and can be more easily analyzed. These systems will be called the upper and lower comparison systems depicted in Figure 1, which capture important behaviors of Q-learning. The upper comparison system, denoted by Q_k^U, upper bounds Q-learning iterate Q_k, while the lower comparison system, denoted by Q_k^L, lower bounds Q_k. The construction of these comparison systems is partly inspired by [12] and exploits the special structure of the Q-learning algorithm. Unlike [12], here we focus on the discrete-time domain directly and a finite-time analysis. To address the difficulty with the stochastic noise, we introduce a two-phase analysis: the first phase captures the noise effect of the lower comparison system, while the second phase captures the difference between the two comparison systems when the noise effect vanishes.

3.2. Lower comparison system. Consider the stochastic linear system

$$Q_{k+1}^L - Q^* = A_{Q^*}(Q_k^L - Q^*) + \alpha w_k, \quad Q_0^L - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{3.7}$$

where the stochastic noise w_k is the same as the original system (3.5). We call it the *lower comparison system*.

PROPOSITION 3.4. *Suppose $Q_0^L - Q^* \le Q_0 - Q^*$, where $\le$ is used as the elementwise inequality. Then,*

$$Q_k^L - Q^* \le Q_k - Q^*$$

for all $k \ge 0$.

Proof. The proof is done by an induction argument. Suppose the result holds for some $k \ge 0$. Then,

$$\begin{aligned}
&(Q_{k+1} - Q^*) \\
&\quad = A_{Q^*}(Q_k - Q^*) + (A_{Q_k} - A_{Q^*})(Q_k - Q^*) + b_{Q_k} + \alpha w_k \\
&\quad = A_{Q^*}(Q_k - Q^*) + \alpha\gamma DP(\Pi_{Q_k} - \Pi_{Q^*})Q_k + \alpha w_k \\
&\quad \ge A_{Q^*}(Q_k - Q^*) + \alpha w_k \\
&\quad \ge A_{Q^*}(Q_k^L - Q^*) + \alpha w_k \\
&\quad = Q_{k+1}^L - Q^*,
\end{aligned}$$

where the first inequality is due to $DP(\Pi_{Q_k} - \Pi_{Q^*})Q_k \geq DP(\Pi_{Q^*} - \Pi_{Q^*})Q_k = 0$ and the second inequality is due to the hypothesis $Q_k^L - Q^* \leq Q_k - Q^*$ and the fact that A_{Q^*} is a nonnegative matrix (all elements are nonnegative). The proof is completed by induction. ∎

Remark 3.5. Rearranging terms, the original system (3.6) can be written as

$$Q_{k+1} - Q^* = A_{Q^*}(Q_k - Q^*) + \underbrace{\alpha\gamma DP(\Pi_{Q_k} - \Pi_{Q^*})Q_k}_{=:h_{Q_k}} + \alpha w_k,$$

where one can easily prove that $h_{Q_k} \geq 0$ using the definition of Π_{Q_k}, i.e., $\Pi_{Q_k}Q_k \geq \Pi_{Q^*}Q_k$. Intuitively, removing this nonnegative bias term, h_{Q_k}, leads to the lower comparison system. The vector h_{Q_k} represents a portion of the gap between the original and lower systems incurred at a single time step.

Note that the mean dynamics of the lower comparison system is simply a linear system. By Proposition 3.2, we have the exponential stability of the mean dynamics:

$$\|\mathbb{E}[Q_k^L] - Q^*\|_\infty \leq \rho^k \|Q_0^L - Q^*\|_\infty \quad \forall k \geq 0. \tag{3.8}$$

Furthermore, we can conclude that A_{Q^*} is Schur, i.e., the magnitude of all its engenvalues is strictly less than one, and from the Lyapunov theory for linear systems, there exists a positive definite matrix $M \succ 0$ and $\beta \in (0,1)$ such that

$$A_{Q^*}^T M A_{Q^*} \preceq \beta M.$$

The parameter $\beta \in (0,1)$ determines the convergence speed of the state to the origin, and it depends on the structure of the matrix A_{Q^*}. We prove that in our case, an upper bound on β can be expressed in terms of ρ. In fact, we can set $\beta = (\rho + \epsilon)^2$ for arbitrary $\epsilon > 0$ such that $\beta \in (0,1)$.

PROPOSITION 3.6. *For any $\epsilon > 0$ such that $\rho + \epsilon \in (0,1)$, there exists the corresponding positive definite $M \succ 0$ such that*

$$A_{Q^*}^T M A_{Q^*} = (\rho + \epsilon)^2 (M - I)$$

and

$$\lambda_{\min}(M) \geq 1,\ \lambda_{\max}(M) \leq \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\rho}{\rho+\epsilon}\right)^2}.$$

The above result can be easily verified by setting

$$M = \sum_{k=0}^{\infty} \left(\frac{1}{\rho + \epsilon}\right)^{2k} (A_{Q^*}^k)^T A_{Q^*}^k.$$

We defer the detailed proof to Appendix 5.1. Based on this result, we can derive a finite-time error bound for the lower comparison system.

THEOREM 3.7. *Under Assumptions* 2.1–2.4*, for any $N \geq 0$, it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{k=0}^{N-1} Q_k^L - Q^*\right\|_\infty\right] \leq \sqrt{\frac{32\alpha|\mathcal{S}|^2|\mathcal{A}|^2}{d_{\min}(1-\gamma)^3} + \frac{1}{N}\frac{2|\mathcal{S}|^2|\mathcal{A}|^2}{\alpha d_{\min}(1-\gamma)}\mathbb{E}[\|Q_0 - Q^*\|_\infty^2]}. \tag{3.9}$$

Proof. Define the Lyapunov function $V(x) = x^T M x$, the history set

$$\mathcal{F}_k := \{Q_0^L, Q_0, w_0, Q_1^L, Q_1, w_1, \ldots, Q_{k-1}^L, Q_{k-1}, w_{k-1}, Q_k^L, Q_k\},$$

and also denote $A = A_{Q^*}$ for simplicity of the presentation, where the matrix M satisfies the conditions in Proposition 3.6. Then, we have

$$\begin{aligned}
&\mathbb{E}[V(Q_{k+1}^L - Q^*)|\mathcal{F}_k] \\
&\quad = \mathbb{E}[(A(Q_k^L - Q^*) + \alpha w_k)^T M (A(Q_k^L - Q^*) + \alpha w_k)|\mathcal{F}_k] \\
&\quad = \mathbb{E}[(Q_k^L - Q^*)^T A^T M A (Q_k^L - Q^*) + \alpha^2 w_k^T M w_k|\mathcal{F}_k] \\
&\quad \le (\rho+\epsilon)^2 V(Q_k^L - Q^*) - (\rho+\epsilon)^2 \|Q_k^L - Q^*\|^2 + \lambda_{\max}(M)\alpha^2 \mathbb{E}[w_k^T w_k|\mathcal{F}_k].
\end{aligned}$$

Here $\epsilon > 0$ is such that $\rho + \epsilon < 1$. The first inequality comes from Proposition 3.6. The second equality is due to the fact that

$$\begin{aligned}
&\mathbb{E}[w_k|\mathcal{F}_k] \\
&\quad = \mathbb{E}[(e_{a_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{s_k})(e_{s_{k'}})^T \Pi_{Q_k} Q_k \\
&\qquad - (e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^T Q_k - (DR + \gamma DP\Pi_{Q_k} Q_k - DQ_k)|\mathcal{F}_k] \\
&\quad = \mathbb{E}[(e_{a_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{s_k})(e_{s_{k'}})^T \Pi_{Q_k} Q_k - (e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^T Q_k|Q_k] \\
&\qquad - (DR + \gamma DP\Pi_{Q_k} Q_k - DQ_k) \\
&\quad = DR + \gamma DP\Pi_{Q_k} Q_k - DQ_k - (DR + \gamma DP\Pi_{Q_k} Q_k - DQ_k) \\
&\quad = 0.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\alpha w_k^T M A (Q_k^L - Q^*)|\mathcal{F}_k] &= \mathbb{E}[\alpha w_k^T M A (Q_k^L - Q^*)|Q_k^L, Q_k] \\
&= \alpha \mathbb{E}[w_k^T|Q_k] M A (Q_k^L - Q^*) \\
&= 0.
\end{aligned}$$

Subtracting $V(Q_k^L - Q^*)$ from both sides and using $\lambda_{\min}(M) \ge 1$ in Proposition 3.6 lead to

$$\begin{aligned}
&\mathbb{E}[V(Q_{k+1}^L - Q^*)|\mathcal{F}_k] - V(Q_k^L - Q^*) \\
&\quad \le (\rho+\varepsilon)^2 V(Q_k^L - Q^*) - V(Q_k^L - Q^*) - (\rho+\varepsilon)^2 \|Q_k^L - Q^*\|^2 \\
&\qquad + \alpha^2 \lambda_{\max}(M) \mathbb{E}[w_k^T w_k|\mathcal{F}_k] \\
&\quad = ((\rho+\varepsilon)^2 - 1) V(Q_k^L - Q^*) - (\rho+\varepsilon)^2 \|Q_k^L - Q^*\|^2 \\
&\qquad + \alpha^2 \lambda_{\max}(M) \mathbb{E}[w_k^T w_k|\mathcal{F}_k] \\
&\quad \le ((\rho+\varepsilon)^2 - 1) \|Q_k^L - Q^*\|^2 - (\rho+\varepsilon)^2 \|Q_k^L - Q^*\|^2 \\
&\qquad + \alpha^2 \lambda_{\max}(M) \mathbb{E}[w_k^T w_k|\mathcal{F}_k] \\
&\quad = -\|Q_k^L - Q^*\|^2 + \alpha^2 \lambda_{\max}(M) \mathbb{E}[w_k^T w_k|\mathcal{F}_k],
\end{aligned}$$

where the last inequality uses the facts, $(\rho+\varepsilon)^2 - 1 < 0$ and $\lambda_{\min}(M) \ge 1$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}[V(Q_{k+1}^L - Q^*)|\mathcal{F}_k] - V(Q_k^L - Q^*) \\
&\qquad \le -\|Q_k^L - Q^*\|^2 + \alpha^2 \lambda_{\max}(M) \mathbb{E}[w_k^T w_k|\mathcal{F}_k].
\end{aligned}$$

Taking the expectation $\mathbb{E}[\cdot]$ on both sides and rearranging terms yield

$$\begin{aligned}&\mathbb{E}[\|Q_k^L - Q^*\|^2]\\ &\quad\le \mathbb{E}[V(Q_k^L - Q^*)] - \mathbb{E}[V(Q_{k+1}^L - Q^*)] + \alpha^2\lambda_{\max}(M)\mathbb{E}[w_k^T w_k].\end{aligned}$$

Next, we show that the variance of w_k is bounded as follows:

$$\mathbb{E}[w_k^T w_k|\mathcal{F}_k] \le W := \frac{16|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}.$$

This is because

$$\begin{aligned}\|w_k\|_\infty &\le \|((e_a \otimes e_s) - D)r_k\|_\infty\\ &\quad + \gamma\|(e_a \otimes e_s)(e_{s'})^T - DP\|_\infty\|\Pi_{Q_k}\|_\infty\|Q_k\|_\infty\\ &\quad + \|((e_a \otimes e_s)(e_a \otimes e_s)^T - D)\|_\infty\|Q_k\|_\infty\\ &\le 2R_{\max} + 2\gamma Q_{\max} + 2Q_{\max}\\ &\le \frac{4}{1-\gamma},\end{aligned}$$

where the last inequality comes from Assumptions 2.3–2.4 and Lemma 2.7. Hence, $\mathbb{E}[w_k^T w_k|\mathcal{F}_k] \le W$.

Summing both sides from $k=0$ to $k=N-1$ and dividing by $N>0$ leads to

$$\begin{aligned}&\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}[\|Q_k^L - Q^*\|^2]\\ &\quad\le \alpha^2\lambda_{\max}(M)W + \frac{\lambda_{\max}(M)}{N}\mathbb{E}[\|Q_0^L - Q^*\|^2],\end{aligned}$$

where we use $\lambda_{\min}(M)\|x\|_2^2 \le V(x) \le \lambda_{\max}(M)\|x\|_2^2$. We use the bound $\lambda_{\max}(M) \le \frac{|\mathcal{S}||\mathcal{A}|}{1-(\frac{\rho}{\rho+\epsilon})^2}$ in Proposition 3.6, let $Q_0^L = Q_0$, and set $\varepsilon = \frac{1-\rho}{2}$ so that $\rho+\varepsilon = \frac{1+\rho}{2} \in (0,1)$ to have

$$\begin{aligned}&\sum_{k=0}^{N-1}\frac{1}{N}\mathbb{E}[\|Q_k^L - Q^*\|^2]\\ &\quad\le \frac{\alpha^2|\mathcal{S}||\mathcal{A}|W}{1-\left(\frac{\rho}{\rho+\varepsilon}\right)^2} + \frac{1}{N}\frac{|\mathcal{S}||\mathcal{A}|}{1-\left(\frac{\rho}{\rho+\varepsilon}\right)^2}\mathbb{E}[\|Q_0 - Q^*\|^2]\\ &\quad\le \alpha^2\frac{(1+\rho)|\mathcal{S}||\mathcal{A}|W}{1-\rho} + \frac{1}{N}\frac{(1+\rho)|\mathcal{S}||\mathcal{A}|}{1-\rho}\mathbb{E}[\|Q_0 - Q^*\|^2]\\ &\quad\le \frac{2\alpha|\mathcal{S}||\mathcal{A}|W}{d_{\min}(1-\gamma)} + \frac{1}{N}\frac{2|\mathcal{S}||\mathcal{A}|}{\alpha d_{\min}(1-\gamma)}\mathbb{E}[\|Q_0 - Q^*\|^2].\end{aligned}$$

Taking the square root on both sides, using the subadditivity of the square root, and combining with the relations

$$\begin{aligned}&\sqrt{\sum_{k=0}^{N-1}\frac{1}{N}\mathbb{E}[\|Q_k^L - Q^*\|_2^2]}\\ &\quad\ge \sum_{k=0}^{N-1}\frac{1}{N}\mathbb{E}[\|Q_k^L - Q^*\|_2] \ge \sum_{k=0}^{N-1}\frac{1}{N}\mathbb{E}[\|Q_k - Q^*\|_\infty]\end{aligned}$$

and

$$\|Q_0 - Q^*\|_2^2 \leq |\mathcal{S}||\mathcal{A}|\|Q_0 - Q^*\|_\infty^2,$$

which applies the concavity of the square root function and Jensen's inequality, we further have

$$\sum_{k=0}^{N-1} \frac{1}{N}\mathbb{E}[\left\|Q_k^L - Q^*\right\|_\infty] \leq \sqrt{\frac{32\alpha|\mathcal{S}|^2|\mathcal{A}|^2}{d_{\min}(1-\gamma)^3} + \frac{1}{N}\frac{2|\mathcal{S}|^2|\mathcal{A}|^2}{\alpha d_{\min}(1-\gamma)}\mathbb{E}[\|Q_0 - Q^*\|_\infty^2]}. \tag{3.10}$$

Using the Jensen inequality again yields the desired result. ∎

Before closing this subsection, we provide a simple example which shows the case that the gap between the lower comparison system and the original system is tight.

Example 3.8. Consider an MDP with $\mathcal{S} = \{1\}$, $\mathcal{A} = \{1\}$, $\gamma = 0.9$, where a reward is one at every time instances. In this case, the optimal policy is defined with $\pi^*(1) = 1$, and the corresponding optimal Q-function is $Q^* = \frac{1}{1-\gamma}$. The overall system is deterministic. In this case, $D = P = 1$ and $\Pi_Q = 1$ for any $Q \in \mathbb{R}$. Then, we have $A_Q = 1 + \alpha(0.9 - 1), b_Q = 0$, and the switching system in (3.6) is given as

$$Q_{k+1} - Q^* = (1 - 0.1\alpha)(Q_k - Q^*).$$

On the other hand, since $A_{Q^*} = 1 + \alpha(0.9 - 1)$, the lower system in (3.7) is the same as the original system, i.e.,

$$Q_{k+1}^L - Q^* = (1 - 0.1\alpha)(Q_k^L - Q^*).$$

Therefore, with $Q_0 = Q_0^L$, the lower bound is tight in the sense that $Q_k - Q^* = Q_k^L - Q^*$ for all $k \geq 0$.

3.3. Upper comparison system. Now, let us consider the stochastic linear switching system

$$Q_{k+1}^U - Q^* = A_{Q_k}(Q_k^U - Q^*) + \alpha w_k, \quad Q_0^U - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{3.11}$$

where the stochastic noise w_k is kept the same as the original system. We will call it the *upper comparison system*.

PROPOSITION 3.9. *Suppose $Q_0^U - Q^* \geq Q_0 - Q^*$, where $\geq$ is used as the elementwise inequality. Then,*

$$Q_k^U - Q^* \geq Q_k - Q^*$$

for all $k \geq 0$.

Proof. Suppose the result holds for some $k \geq 0$. Then,

$$\begin{aligned}(Q_{k+1} - Q^*) &= A_{Q_k}(Q_k - Q^*) + b_{Q_k} + \alpha w_k \\ &\leq A_{Q_k}(Q_k - Q^*) + \alpha w_k \\ &\leq A_{Q_k}(Q_k^U - Q^*) + \alpha w_k \\ &= Q_{k+1}^U - Q^*,\end{aligned}$$

where we used the fact that $b_{Q_k} = D(\gamma P \Pi_{Q_k} Q^* - \gamma P \Pi_{Q^*} Q^*) \leq D(\gamma P \Pi_{Q^*} Q^* - \gamma P \Pi_{Q^*} Q^*) = 0$ in the first inequality. The second inequality is due to the hypothesis

$Q_k^U - Q^* \geq Q_k - Q^*$ and the fact that A_{Q_k} is a nonnegative matrix. Then, the proof is completed by induction. ∎

Remark 3.10. In the original system (3.6), one can easily prove that $b_{Q_k} := \gamma DP(\Pi_{Q_k} - \Pi_{Q^*})Q^* \leq 0$ using the definition of Π_{Q_k}, i.e., $\Pi_{Q_k} Q^* \leq \Pi_{Q^*} Q^*$. Intuitively, removing this nonpositive bias term, b_{Q_k}, leads to the upper comparison system. The vector b_{Q_k} represents a portion of the gap between the original and upper systems incurred at a single time step.

Hence, the trajectory of the stochastic linear switching system (3.11) bounds that of the original system from above. Note that the system matrix, A_{Q_k}, switches according to the change of Q_k, which depends probabilistically on Q_k^U. Therefore, if we take the expectation on both sides of (3.11), it is not possible to separate A_{Q_k} and the state $Q_k^U - Q^*$ unlike the lower comparison system, making it much harder to analyze the stability of the upper comparison system.

To circumvent such a difficulty, we instead study the following error system by subtracting the lower comparison system (3.7) from the upper comparison system (3.11):

$$Q_{k+1}^U - Q_{k+1}^L = A_{Q_k}(Q_k^U - Q_k^L) + B_{Q_k}(Q_k^L - Q^*), \tag{3.12}$$

where

$$B_{Q_k} := A_{Q_k} - A_{Q^*} = \alpha\gamma DP(\Pi_{Q_k} - \Pi_{Q^*}).$$

Here, the stochastic noise, αw_k, is canceled out in the error system. Moreover, matrices (A_{Q_k}, B_{Q_k}) switch according to the external signal, Q_k, and $Q_k^L - Q^*$ can be seen as an external disturbance.

The key insight is as follows: if we can prove the stability of the error system, i.e., $Q_k^U - Q_k^L \to 0$ as $k \to \infty$, then since $Q_k^L \to Q^*$ as $k \to \infty$, we have $Q_k^U \to Q^*$ as well.

Example 3.11. Consider Example 3.8 again. The upper system in (3.11) is the same as the original system, i.e.,

$$Q_{k+1}^U - Q^* = (1 - 0.1\alpha)(Q_k^U - Q^*).$$

Therefore, with $Q_0 = Q_0^U$, the upper bound is tight in the sense that $Q_k^U - Q^* = Q_k - Q^*$ for all $k \geq 0$.

Example 3.12. Consider an MDP with $\mathcal{S} = \{1\}$, $\mathcal{A} = \{1, 2\}$, $\gamma = 0.9$, where the reward is one when $a = 1$ and zero otherwise. In this case, the optimal policy is defined with $\pi^*(1) = 1$, and the corresponding optimal Q-function is $Q^*(s, 1) = \frac{1}{1-\gamma} = 10, Q^*(s, 2) = \frac{\gamma}{1-\gamma} = 9$. We consider the behavior policy $\beta(\cdot|1) = [\ 0.5 \quad 0.5\ ]^T$. Then, we have

$$Q = \begin{bmatrix} Q(1,1) \\ Q(1,2) \end{bmatrix}, R = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, P = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, D = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

and $\Pi_Q = [\ 1 \quad 0\ ]$ if $Q(1,1) \geq Q(1,2)$ and $\Pi_Q = [\ 0 \quad 1\ ]$ otherwise. In this case,

$$A_Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \alpha \left(0.9 \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \Pi_Q - \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \right)$$

and

$$b_Q = 0.9 \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \left(\Pi_Q - \begin{bmatrix} 1 & 0 \end{bmatrix} \right) \begin{bmatrix} 10 \\ 9 \end{bmatrix}.$$

From the result, the gap, $Q_{k+1}^U - Q_{k+1}$, between the upper and original systems incurred at each time step is $b_Q = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$ when $Q(1,1) \geq Q(1,2)$, and $b_Q = -0.9\left[\begin{smallmatrix}0.5\\0.5\end{smallmatrix}\right]$ when $Q(1,1) < Q(1,2)$. Similar results can be obtained for the lower system.

3.4. Finite-time error bound of Q-learning. In this subsection, we provide a finite-time error bound of Q-learning, which is the main result of this paper. We obtain the following main result.

THEOREM 3.13. *Under Assumptions* 2.1–2.4*, for any* $N \geq 0$*, we have the following error bound for Q-learning iterates,*

$$\mathbb{E}[\|\tilde{Q}_N - Q^*\|_\infty] \leq \left(\frac{4\gamma d_{\max} + d_{\min}(1-\gamma)}{d_{\min}^{3/2}(1-\gamma)^{5/2}}|\mathcal{S}||\mathcal{A}|\right)\sqrt{32\alpha + \frac{1}{N}\frac{4}{\alpha}}, \tag{3.13}$$

where $\alpha \in (0,1)$ *is the constant step size and* $\tilde{Q}_N = \frac{1}{N}\sum_{i=0}^{N-1} Q_k$.

Proof. Taking the norm on both sides of the error system (3.12), we have for any $k \geq 0$

$$\begin{aligned}
&\|Q_{k+1}^U - Q_{k+1}^L\|_\infty \\
&\quad \leq \|A_{Q_k}\|_\infty \|Q_k^U - Q_k^L\|_\infty + \|B_{Q_k}\|_\infty \|Q_k^L - Q^*\|_\infty \\
&\quad \leq (\rho + \varepsilon)\|Q_k^U - Q_k^L\|_\infty + \|B_{Q_k}\|_\infty \|Q_k^L - Q^*\|_\infty \\
&\quad \leq (\rho + \varepsilon)\|Q_k^U - Q_k^L\|_\infty + 2\alpha\gamma d_{\max}\|Q_k^L - Q^*\|_\infty.
\end{aligned}$$

Here, the last inequality uses the fact that

$$\|B_{Q_k}\|_\infty \leq \alpha\gamma d_{\max}\|P(\Pi_{Q_k} - \Pi_{Q^*})\|_\infty \leq 2\alpha\gamma d_{\max}.$$

Rearranging terms leads to

$$\begin{aligned}
&(1-\rho-\varepsilon)\|Q_k^U - Q_k^L\|_\infty \\
&\quad \leq \|Q_k^U - Q_k^L\|_\infty - \|Q_{k+1}^U - Q_{k+1}^L\|_\infty + 2\alpha\gamma d_{\max}\|Q_k^L - Q^*\|_\infty, \quad k \geq 0.
\end{aligned}$$

Summing both sides from $k = 0$ to $k = N-1$, dividing by $N > 0$, and letting $Q_0^U = Q_0^L = Q_0$ lead to

$$\sum_{k=0}^{N-1} \frac{1}{N}\|Q_k^U - Q_k^L\|_\infty \leq \frac{2\alpha\gamma d_{\max}}{1-\rho-\varepsilon}\sum_{k=0}^{N-1}\frac{1}{N}\|Q_k^L - Q^*\|_\infty. \tag{3.14}$$

Next, we will express the left-hand side in terms of Q_k. By the triangle inequality, we have

$$\begin{aligned}
\|Q_k - Q^*\|_\infty &\leq \|Q^* - Q_k^L\|_\infty + \|Q_k - Q_k^L\|_\infty \\
&\leq \|Q^* - Q_k^L\|_\infty + \|Q_k^U - Q_k^L\|_\infty.
\end{aligned}$$

The second inequality comes from

$$0 \leq Q_k - Q_k^L \leq Q_k^U - Q_k^L.$$

This leads to

$$\|Q_k - Q^*\|_\infty - \|Q^* - Q_k^L\|_\infty \leq \|Q_k^U - Q_k^L\|_\infty.$$

Combining this inequality with (3.14), one gets

$$\sum_{k=0}^{N-1} \frac{1}{N}(\|Q_k - Q^*\|_\infty - \|Q^* - Q_k^L\|_\infty) \leq \frac{4\gamma d_{\max}}{d_{\min}(1-\gamma)} \sum_{k=0}^{N-1} \frac{1}{N}\|Q_k^L - Q^*\|_\infty,$$

where we let $\varepsilon = \frac{1-\rho}{2}$ so that $\rho + \varepsilon = \frac{1+\rho}{2}$ and $\frac{1}{1-\rho-\varepsilon} = \frac{2}{1-\rho} = \frac{2}{\alpha d_{\min}(1-\gamma)}$. Rearranging terms again, taking the expectation on both sides, and combining it with (3.10), we obtain

$$\begin{aligned}
&\sum_{k=0}^{N-1} \frac{1}{N}\mathbb{E}[\|Q_k - Q^*\|_\infty] \\
&\quad \leq \frac{5d_{\max}}{d_{\min}(1-\gamma)} \sum_{k=0}^{N-1} \frac{1}{N}\mathbb{E}[\|Q_k^L - Q^*\|_\infty] \\
&\quad \leq \frac{5d_{\max}}{d_{\min}(1-\gamma)} \sqrt{\frac{32\alpha|\mathcal{S}|^2|\mathcal{A}|^2}{d_{\min}(1-\gamma)^3} + \frac{1}{N}\frac{2|\mathcal{S}|^2|\mathcal{A}|^2}{\alpha d_{\min}(1-\gamma)}}\mathbb{E}[\|Q_0 - Q^*\|_\infty^2].
\end{aligned}$$

From Lemma 2.7, $-\mathbf{1}Q_{\max} \leq Q_k \leq \mathbf{1}Q_{\max}$ holds for all $k \geq 0$, where $\mathbf{1}$ denotes a column vector where all elements are one. Applying $Q_{\max} = 1/(1-\gamma)$ from Lemma 2.7 together with Assumptions 2.4 and 2.3, $\|Q_0\|_\infty \leq 1$ from Assumption 2.4, the Jensen inequality, and, after simplifications, we can obtain the desired conclusion. ∎

Remark 3.14. Lyapunov theory [4] has been applied for the lower comparison system, which is a linear time-invariant system. On the other hand, the techniques used for the error system between the upper and lower comparison systems more resemble those used in the optimization community rather than leveraging the nature of a switching dynamical system. However, the switching system formulation captures essential behaviors of Q-learning algorithm, and itself is mainly used in combination with the lower comparison system in the overall derivation process.

3.5. Remarks. *Overestimation and maximization bias.* Our analysis provides an intuitive explanation of the well-known overestimation phenomenon in Q-learning [7]. In particular, $Q_k(s,a)$ tends to overestimate $Q^*(s,a)$ due to the maximization bias in the Q-learning updates. This becomes severe especially when the action-space is large. In particular, it can be problematic when the action spaces depending on states are heterogeneous and the current estimate Q_k is used for the exploration, e.g., the ε-greedy behavior policy; in this case, since $\arg\max_{a\in\mathcal{A}} Q_k(s,a)$ tends to choose actions with larger maximization biases, thus degrading the quality of exploration and leading to slower convergence. Moreover, the overestimation error could be amplified at each iteration k when it passes through the max operator.

In fact, assuming that the initial $Q_0(s,a) - Q^*(s,a)$ is a zero mean random variable, namely, $\mathbb{E}[Q_0 - Q^*] = 0$, we can easily see through our analysis that

$$\mathbb{E}[Q_k - Q^*] \geq 0 \quad \forall k \geq 0.$$

This is because the lower comparison system (which is a stochastic linear system) satisfies that

$$\mathbb{E}[Q_k^L - Q^*] = A_{Q^*}^k \mathbb{E}[Q_0^L - Q^*] + \sum_{i=0}^{k-1} A_{Q^*}^{k-1-i} \alpha \mathbb{E}[w_i] = 0$$

provided that $Q_0^L = Q_0$, namely, there exists no biases in the lower system state. On the other hand, since $Q_k^L - Q^* \leq Q_k - Q^*$, $\mathbb{E}[Q_k - Q^*] \geq 0$ holds. Moreover, for $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $Q_k^L < Q_k$ holds strictly, then $\mathbb{E}[Q_k(s,a)] > Q^*(s,a)$, which potentially explains the overestimation phenomenon.

3.6. Sample complexity. Based on the finite-time error bound on the Q-learning iterates in Theorem 3.13, we can derive an upper bound on the sample or iteration complexity of Q-learning: to find an ε-optimal solution such that $\mathbb{E}[\|\tilde{Q}_N - Q^*\|_\infty] < \varepsilon$, we need at most

$$\mathcal{O}\left(\frac{d_{\max}^4 |\mathcal{S}|^4 |\mathcal{A}|^4}{\varepsilon^4 \delta^4 d_{\min}^6 (1-\gamma)^{10}}\right)$$

samples. Moreover, if the state-action pair is sampled uniformly from $\mathcal{S} \times \mathcal{A}$, then $d(s,a) = \frac{1}{|\mathcal{S}||\mathcal{A}|}$ $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ and $d_{\min} = d_{\max} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$. In this case, the sample complexity becomes $\mathcal{O}(\frac{|\mathcal{S}|^6 |\mathcal{A}|^6}{\varepsilon^4 \delta^4 (1-\gamma)^{10}})$. The proof is given in Appendix 5.2.

The finite-time analysis of asynchronous Q-learning with constant step size was first considered in [2], and has been recently studied in [13] and the concurrent work [4]. Based on the cover time assumption, which is deterministic, [2] provides $\tilde{\mathcal{O}}(\frac{t_{\text{cover}}^3 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5 \varepsilon^2})$, where t_{cover} is the cover time and $\tilde{\mathcal{O}}$ ignores the polylogarithmic factors. The results in [13] provide $\tilde{\mathcal{O}}(\frac{1}{d_{\min}(1-\gamma)^5 \varepsilon^2} + \frac{t_{\text{mix}}}{d_{\min}(1-\gamma)})$ with a single Markovian trajectory, where t_{mix} is the mixing time. Note that the mixing time and cover time assumptions are adopted in [13]. The complexity $\tilde{\mathcal{O}}(\frac{1}{d_{\min}^3 (1-\gamma)^5 \varepsilon^2})$ is given in [4] with a single Markovian trajectory. Note that the bounds in [4] and [2] are the expected error bounds, and those in [13] are the concentration error bounds. Besides, [15] offers a sharper bound using a diminishing step size. Based on the analysis, we summarize advantages and limitations of the proposed approach. A limitation of the proposed method lies in that the corresponding sample complexity is not tighter than the existing approaches. On the other hand, the main advantage is the proposition of a unique switching system and control perspectives, which inherit simplicity, and provides additional insights on Q-learning.

4. Conclusion. In this paper, we introduced a novel control-theoretic framework based on discrete-time switching systems to derive finite-time error bounds of Q-learning algorithm. By sandwiching the dynamics of asynchronous Q-learning between two simpler stochastic (switched) linear systems, a new finite-time analysis of the Q-learning can be easily derived. We believe it is important to emphasize that the proposed control-theoretic analysis can be viewed as a new analysis which gives additional insights into Q-learning rather than a replacement or improvement of existing convergence rate analysis. The proposed analysis has simplicity, novelty, and more intuition. We expect that such a control-theoretic analysis could further stimulate the synergy between control theory and reinforcement learning, and open up opportunities for the design of new reinforcement learning algorithms and refined analysis for Q-learning. Moreover, our approach based on the comparison systems could be of independent interest to the finite-time stability analysis of more general switching systems.

As promising next steps, the proposed bounds can be further tightened, and the analysis can be extended to more general Markovian settings. The proposed analysis

framework can potentially be applied to derive finite-time error bounds for other variants of Q-learning, such as double Q-learning [7], averaging Q-learning [12], speedy Q-learning [1], and multiagent Q-learning [9], as well as their function approximation counterparts. We will leave these topics for future investigations.

5. Appendix.

5.1. Proof of Proposition 3.6.

Proof. For simplicity, denote $A = A_{Q^*}$. Consider matrix M such that

$$M = \sum_{k=0}^{\infty} \left(\frac{1}{\rho+\epsilon}\right)^{2k} (A^k)^T A^k. \tag{5.1}$$

Noting that

$$\begin{aligned}(\rho+\epsilon)^{-2} A^T M A + I &= \frac{1}{(\rho+\epsilon)^2} A^T \left(\sum_{k=0}^{\infty} \left(\frac{1}{\rho+\epsilon}\right)^{2k} (A^k)^T A^k\right) A + I \\ &= M,\end{aligned}$$

we have

$$(\rho+\epsilon)^{-2} A^T M A + I = M,$$

resulting in the desired conclusion. Next, it remains to prove the existence of M by proving its boundedness. Taking the norm on M leads to

$$\begin{aligned}\|M\|_2 &= \left\|I + (\rho+\varepsilon)^{-2} A^T A + (\rho+\varepsilon)^{-4} (A^2)^T A^2 + \cdots\right\|_2 \\ &\le \|I\|_2 + (\rho+\varepsilon)^{-2} \left\|A^T A\right\|_2 + (\rho+\varepsilon)^{-4} \left\|(A^2)^T A^2\right\|_2 + \cdots \\ &= \|I\|_2 + (\rho+\varepsilon)^{-2} \|A\|_2^2 + (\rho+\varepsilon)^{-4} \left\|A^2\right\|_2^2 + \cdots \\ &= 1 + |\mathcal{S}||\mathcal{A}|(\rho+\varepsilon)^{-2} \|A\|_\infty^2 + |S||A|(\rho+\varepsilon)^{-4} \left\|A^2\right\|_\infty^2 + \cdots \\ &= 1 - |\mathcal{S}||\mathcal{A}| + \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\rho}{\rho+\varepsilon}\right)^2}.\end{aligned}$$

Finally, we prove the bounds on the maximum and minimum eigenvalues. From the definition (5.1), $M \succeq I$ and, hence, $\lambda_{\min}(M) \ge 1$. On the other hand, one gets

$$\begin{aligned}\lambda_{\max}(M) &= \lambda_{\max}(I + (\rho+\epsilon)^{-2} A^T A \\ &\quad + (\rho+\epsilon)^{-4} (A^2)^T A^2 + \cdots) \\ &\le \lambda_{\max}(I) + (\rho+\epsilon)^{-2} \lambda_{\max}(A^T A) \\ &\quad + (\rho+\epsilon)^{-4} \lambda_{\max}((A^2)^T A^2) + \cdots \\ &= \lambda_{\max}(I) + (\rho+\epsilon)^{-2} \|A\|_2^2 + (\rho+\epsilon)^{-4} \|A^2\|_2^2 + \cdots \\ &\le 1 + |\mathcal{S}||\mathcal{A}|(\rho+\epsilon)^{-2} \|A\|_\infty^2 \\ &\quad + |\mathcal{S}||\mathcal{A}|(\rho+\epsilon)^{-4} \|A^2\|_\infty^2 + \cdots \\ &\le \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\rho}{\rho+\epsilon}\right)^2}.\end{aligned}$$

The proof is completed. ∎

5.2. Sample complexity.

PROPOSITION 5.1 (sample complexity). *To achieve*

$$\|\tilde{Q}_N - Q^*\|_\infty < \varepsilon$$

with probability at least $1-\delta$*, we need the number of samples/iterations to be at most*

$$\mathcal{O}\left(\frac{d_{\max}^4 |\mathcal{S}|^4 |\mathcal{A}|^4}{\varepsilon^4 \delta^4 d_{\min}^6 (1-\gamma)^{10}}\right).$$

Proof. For convenience, we first find a simplified overestimate on the right-hand side of (3.13) as

$$\begin{aligned}&\mathbb{E}[\|\tilde{Q}_N - Q^*\|_\infty] \\ &\quad \le \frac{20 d_{\max}|\mathcal{S}||\mathcal{A}|}{d_{\min}(1-\gamma)^2}\left(\sqrt{\frac{2\alpha}{d_{\min}(1-\gamma)}} + \sqrt{\frac{1}{N}\frac{2}{\alpha d_{\min}(1-\gamma)}}\right) =: C.\end{aligned}$$

Applying the Markov inequality

$$\mathbb{P}[\|\tilde{Q}_N - Q^*\|_\infty \ge \varepsilon] \le \frac{C}{\varepsilon},$$

we conclude that $\|\tilde{Q}_N - Q^*\|_\infty < \varepsilon$ with probability at least $1-\delta$, i.e.,

$$\mathbb{P}[\|\tilde{Q}_N - Q^*\|_\infty < \varepsilon] \ge 1-\delta,$$

where

$$\delta = \frac{1}{\varepsilon}\frac{20 d_{\max}|S||A|}{d_{\min}(1-\gamma)^2}\left(\sqrt{\frac{2\alpha}{d_{\min}(1-\gamma)}} + \sqrt{\frac{1}{N}\frac{2}{\alpha d_{\min}(1-\gamma)}}\right).$$

N, and α are appropriately chosen so that $\delta \in (0,1)$. One concludes that to satisfy $\|Q_N - Q^*\|_\infty < \varepsilon$ with probability at least $1-\delta$, we should have

$$\delta \ge \underbrace{\frac{1}{\varepsilon}\frac{20 d_{\max}|\mathcal{S}||\mathcal{A}|}{d_{\min}(1-\gamma)^2}\sqrt{\frac{2\alpha}{d_{\min}(1-\gamma)}}}_{\Phi_1} + \underbrace{\frac{1}{\varepsilon}\frac{20 d_{\max}|\mathcal{S}||\mathcal{A}|}{d_{\min}(1-\gamma)^2}\sqrt{\frac{1}{N}}\frac{2}{\alpha d_{\min}(1-\gamma)}}_{\Phi_2}$$

which is achieved if $\delta/2 \ge \Phi_1$ and $\delta/2 \ge \Phi_2$.

The first inequality is satisfied if

$$\alpha = \frac{\delta^2\varepsilon^2}{8}\frac{d_{\min}^3(1-\gamma)^5}{400 d_{\max}^2 |\mathcal{S}|^2 |cal A|^2} \tag{5.2}$$

and the second inequality holds if

$$N \ge \frac{3200 d_{\max}^2 |\mathcal{S}|^2 |\mathcal{A}|^2}{\alpha\varepsilon^2\delta^2 d_{\min}^3 (1-\gamma)^5}.$$

Plugging (5.2) into the last inequality, we can arrive at the desired conclusion. ∎

Acknowledgment. The authors wish to thank the referees for their careful reading and constructive suggestions.

REFERENCES

[1] M. G. Azar, R. Munos, M. Ghavamzadeh, and H. J. Kappen, *Speedy Q-learning*, in Proceedings of the 24th International Conference on Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2011, pp. 2411–2419.

[2] C. L. Beck and R. Srikant, *Error bounds for constant step-size Q-learning*, Systems Control Lett., 61 (2012), pp. 1203–1208.

[3] V. S. Borkar and S. P. Meyn, *The O.D.E. method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim., 38 (2000), pp. 447–469.

[4] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam, *A Lyapunov Theory for Finite-Sample Guarantees of Asynchronous Q-Learning and TD-Learning Variants*, preprint, arXiv:2102.01567, 2021.

[5] E. Even-Dar and Y. Mansour, *Learning rates for Q-learning*, J. Mach. Learn. Res., 5 (2003), pp. 1–25.

[6] A. Gosavi, *Boundedness of iterates in Q-learning*, Systems Control Lett., 55 (2006), pp. 347–349.

[7] H. V. Hasselt, *Double Q-learning*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2010, pp. 2613–2621.

[8] T. Jaakkola, M. I. Jordan, and S. P. Singh, *Convergence of stochastic iterative dynamic programming algorithms*, in Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, 1994, pp. 703–710.

[9] S. Kar, J. M. Moura, and H. V. Poor, *QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations*, IEEE Trans. Signal Process., 61 (2013), pp. 1848–1862.

[10] M. J. Kearns and S. P. Singh, *Finite-sample convergence rates for Q-learning and indirect algorithms*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 1999, pp. 996–1002.

[11] H. K. Khalil, *Nonlinear Systems*, Prentice Hall, Upper Saddle River, NJ, 2002.

[12] D. Lee and N. He, *A unified switching system perspective and convergence analysis of q-learning algorithms*, in 34th Conference on Neural Information Processing Systems, NeurIPS 2020, 2020.

[13] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, *Sample complexity of asynchronous Q-learning: sharper analysis and variance reduction*, IEEE Trans. Inform. Theory, 68 (2022), pp. 448-473.

[14] H. Lin and P. J. Antsaklis, *Stability and stabilizability of switched linear systems: A survey of recent results*, IEEE Trans. Automat. Control, 54 (2009), pp. 308–322.

[15] G. Qu and A. Wierman, *Finite-time analysis of asynchronous stochastic approximation and Q-learning*, in 33rd Conferece on Learning Theory, COLT2020, 2020.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[17] C. Szepesvári, *The asymptotic convergence-rate of Q-learning*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 1998, pp. 1064–1070.

[18] J. N. Tsitsiklis, *Asynchronous stochastic approximation and q-learning*, Mach. Learn., 16 (1994), pp. 185–202.

[19] M. J. Wainwright, *Stochastic Approximation with Cone-Contractive Operators: Sharp ℓ_∞-Bounds for Q-Learning*, preprint, arXiv:1905.06265, 2019.

[20] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Mach. Learn., 8 (1992), pp. 279–292.