A Unified Switching System Perspective and Convergence Analysis of Q-Learning Algorithms

Donghwan Lee
Korea Advanced Institute of Science and Technology
donghwan@kaist.ac.kr

Niao He
UIUC & ETH Zurich
niao.he@inf.ethz.ch

Abstract

This paper develops a novel and unified framework to analyze the convergence of a large family of Q-learning algorithms from the switching system perspective. We show that the nonlinear ODE models associated with Q-learning and many of its variants can be naturally formulated as *affine switching systems*. Building on their asymptotic stability, we obtain a number of interesting results: (i) we provide a simple ODE analysis for the convergence of *asynchronous Q-learning* under relatively weak assumptions; (ii) we establish the first convergence analysis of the averaging Q-learning algorithm, and (iii) we derive a new sufficient condition for the convergence of Q-learning with linear function approximation.

1 Introduction

Reinforcement learning (RL) addresses the optimal control problem for unknown systems through experiences [30]. Q-learning, originally introduced by Watkins [36], is one of the most popular and fundamental reinforcement learning algorithms for unknown systems described by Markov decision processes. The convergence of Q-learning has been extensively studied in the literature and proven via several different approaches, including the original proof [36], the stochastic approximation and contraction mapping-based approach [14, 33, 2, 9, 32, 9, 1, 38], and the ODE (ordinary differential equation) approach [4].

The ODE approach analyzes the convergence of general stochastic recursions by examining stability of the associated ODE model [3, 17, 4] and has been used as a convenient analysis tool to prove convergence of many RL algorithms, especially the temporal difference (TD) learning algorithm [29] and its variants [23, 31, 19, 10]. However, its application to Q-learning has been limited due to the presence of the max-operator, which makes the associated ODE model a complex nonlinear system. In contrast, the associated ODE of TD learning for policy evaluation is a linear system, whose asymptotic stability is much easier to analyze in general. While [4] gave the convergence proof of Q-learning based on a nonlinear ODE model, to the authors' knowledge, substantial analysis is required to prove the stability of the corresponding nonlinear ODE [5] by using the max-norm contraction of the Bellman operator. In addition, the result in [4] only applies to synchronous Q-learning, where every state-action pair is visited at each iteration, instead of the commonly used asynchronous Q-learning. Last but not least, the stability analysis does not immediately extend to other Q-learning variants, such as double Q-learning [11], averaging Q-learning [19], and Q-learning with linear function approximation, etc.

In this paper, we provide a simple and unified framework to analyze Q-learning and its variants through switched linear system (SLS) models [21] of the associated ODE. SLSs are an important class of nonlinear hybrid systems, where the system dynamics matrix switches within a finite set of subsystem matrices (or modes) according to a switching signal. The study of SLSs has attracted tremendous attention in the past years and their stability behaviors have been well established in the

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

literature; see [22] and [21] for comprehensive surveys. Our main contributions are summarized as follows:

1. For a number of Q-learning algorithms such as the asynchronous Q-learning, we show that the nonlinear ODE models associated with these algorithms can be characterized as affine switching systems with a state-feedback switching policy.
2. We construct both upper and lower comparison systems of the corresponding affine switching systems, and prove their asymptotic stability based on existing control theory and comparison principles. As a result of the Borkar and Meyn theorem [4], we obtain the asymptotic convergence of these Q-learning algorithms.
3. We extend the approach to analyze the averaging Q-learning [19]. To our best knowledge, this is the first convergence analysis of averaging Q-learning in the literature.
4. We also examine Q-learning with linear function approximation and derive a new sufficient condition to ensure its convergence based on the switching system theory. We show that, under specific assumptions, our new diagonal dominating condition is weaker than the well-known Melo's condition provided in [24].

Related Work. There exists few work on the non-asymptotic convergence rate of these classical Q-learning algorithms such as synchronous Q-learning [34, 9], asynchronous Q-learning [32, 9, 27], Q-learning with linear function approximation [6], etc. Most of the analyses build on completely different techniques and whether these finite-time bounds are sharp or not remains an open question. On the other hand, there is growing interest on designing variants of Q-learning algorithms with improved performance guarantees, e.g., [8, 1, 20, 15, 18], to name a few. Different from these lines of work, the goal of this paper is to establish an initial connection between switching systems and a family of Q-learning algorithms and provide a unified convergence analysis technique. This could potentially open up new opportunities to the development of a tight non-asymptomatic analysis for Q-learning algorithms and the design of new RL algorithms.

It is worth mentioning that several recent work established the analysis of reinforcement learning algorithms based on their connections to control theory. For example, [28] provided the finite sample bound of TD learning based on Lyapunov stability theory for linear ODE. [6] extended the analysis to Q-learning with linear function approximation. [13] explored the connection between temporal difference (TD) learning and the Markov jump linear systems (MJLS). Note that MJLS cannot be used to characterize the nonlinear dynamics of Q-learning. Instead, we resort to linear switching systems with state-feedback switching policies. Our new ODE approach based on linear switching systems can be used as a viable alternative to prove the stability of the associated ODE of various reinforcement learning algorithms as well as their asymptotic (and potentially non-asymptotic) convergence. Finally, we remark that an earlier work [23] also exploited the stability of linear switching system for Greedy-GQ to establish the boundedness of iterates.

2 Preliminaries: MDPs, switching systems, and stochastic approximation

2.1 Markov decision problem

We consider the infinite-horizon (discounted) Markov decision process (MDP) with state-space $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$, action-space $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, transition matrices $P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, a \in \mathcal{A}$, where $P_a(s, s')$ is the probability transiting from state s to the next state s' under action $a \in \mathcal{A}$, and random reward function $r_a(s, s')$. 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 goal is to find a deterministic optimal policy, π^*, such that the cumulative discounted rewards over infinite time horizons is maximized, i.e., $\pi^* := \arg\max_{\pi\in\Theta} \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r_{a_k}(s_k, s_{k+1}) \middle| \pi\right]$,where $\gamma \in [0, 1)$ is the discount factor, Θ is the set of all admissible deterministic policies, $(s_0, a_0, s_1, a_1, \ldots)$ is a state-action trajectory generated by the Markov chain under policy π. The Q-function under policy π is defined as

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{a_k}(s_k, s_{k+1}) \middle| s_0 = s, a_0 = a, \pi\right], \quad s \in \mathcal{S}, a \in \mathcal{A},$$

and the corresponding optimal Q-function is defined as $Q^*(s, a) = Q^{\pi^*}(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Once Q^* is known, then an optimal policy can be retrieved by $\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q^*(s, a)$.

2.2 Basics of nonlinear system theory

Consider the nonlinear system

$$\frac{d}{dt}x_t = f(x_t), \quad x_0 = z, \quad t \in \mathbb{R}_+, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. For simplicity, we assume that the solution to (1) exists and is unique. This holds true if f is globally Lipschitz continuous.

Lemma 1 ([16, Theorem 3.2]). *Consider the nonlinear system* (1) *and assume that f is globally Lipschitz continuous, i.e., $\|f(x) - f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^n$, for some $L > 0$ and norm $\|\cdot\|$, then it has a unique solution $x(t)$ for all $t \geq 0$ and $x_0 \in \mathbb{R}^n$.*

An important concept in dealing with the nonlinear system is the equilibrium point. A point $x = x^e$ in the state space is said to be an equilibrium point of (1) if it has the property that whenever the state of the system starts at x^e, it will remain at x^e [16]. For (1), the equilibrium points are the real roots of the equation $f(x) = 0$. The equilibrium point x^e is said to be globally asymptotically stable if for any initial state $x_0 \in \mathbb{R}^n$, $x_t \to x^e$ as $t \to \infty$. Now, we provide a vector comparison principle [35, 12, 26] for multi-dimensional ODE models, which plays a central role in the analysis below. We first introduce the quasi-monotone increasing function, which is a necessary prerequisite for the comparison principle.

Definition 1 (Quasi-monotone function). *A vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^n$ with $f := [f_1 \quad f_2 \quad \cdots \quad f_n]^T$ is said to be quasi-monotone increasing if $f_i(x) \leq f_i(y)$ holds for all $i \in \{1, 2, \ldots, n\}$ and $x, y \in \mathbb{R}^n$ such that $x_i = y_i$ and $x_j \leq y_j$ for all $j \neq i$.*

An example of a quasi-monotone increasing function f is $f(x) = Ax$ where A is a Metzler matrix, which implies the off-diagonal elements of A are nonnegative. The vector comparison principle is presented below. For completeness, we provide a different proof tailored to our interests in the Appendix.

Lemma 2 (Vector comparison principle [35, page 112], [12, Theorem 3.2]). *Suppose that $\overline{f}$ and $\underline{f}$ are globally Lipschitz continuous. Let v_t be a solution of the system $\frac{d}{dt}x_t = \overline{f}(x_t), x_0 \in \mathbb{R}^n, \forall t \geq 0$, assume that $\overline{f}$ is quasi-monotone increasing, and let v_t be a solution of the system*

$$\frac{d}{dt}v_t = \underline{f}(v_t), \quad v_0 < x_0, \quad \forall t \geq 0, \tag{2}$$

where $\underline{f}(v) \leq \overline{f}(v)$ holds for any $v \in \mathbb{R}^n$. Then, $v_t \leq x_t$ for all $t \geq 0$.

2.3 Switching system theory

Consider the particular nonlinear system, the *linear switching system*,

$$\frac{d}{dt}x_t = A_{\sigma_t}x_t, \quad x_0 = z \in \mathbb{R}^n, \quad t \in \mathbb{R}_+, \tag{3}$$

where $x_t \in \mathbb{R}^n$ is the state, $\sigma \in \mathcal{M} := \{1, 2, \ldots, M\}$ is called the mode, $\sigma_t \in \mathcal{M}$ 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 $\sigma(x_t)$. The global asymptotic stability of the switching system is guaranteed under a fundamental algebraic stability condition reported in [22].

Lemma 3 ([22, Theorem 8]). *The origin of the linear switching system* (3) *is the unique globally asymptotically stable equilibrium point under arbitrary switchings, σ_t, if and only if there exist a full column rank matrix , $L \in \mathbb{R}^{m\times n}$, $m \geq n$, and a family of matrices, $\bar{A}_\sigma \in \mathbb{R}^{m\times n}, \sigma \in \mathcal{M}$, with the so-called "strictly negative row dominating diagonal condition", i.e., for each $\bar{A}_\sigma, \sigma \in \mathcal{M}$, its elements satisfy*

$$[\bar{A}_\sigma]_{ii} + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[\bar{A}_\sigma]_{ij}| < 0, \quad \forall i \in \{1, 2, \ldots, m\},$$

such that the following matrix relation is satisfied: $LA_\sigma = \bar{A}_\sigma L, \quad \forall \sigma \in \mathcal{M}$.

2.4 ODE-based stochastic approximation

Because of its generality, the convergence analyses of many RL algorithms rely on the ODE approach [3, 17]. It analyzes convergence of general stochastic recursions by examining stability of the associated ODE model based on the fact that the stochastic recursions with diminishing step-sizes approximate the corresponding ODEs in the limit. One of the most popular approach is based on the Borkar and Meyn theorem [4]. We now briefly introduce the Borkar and Meyn's ODE approach for analyzing convergence of the general stochastic recursions

$$\theta_{k+1} = \theta_k + \alpha_k(f(\theta_k) + \varepsilon_{k+1}) \tag{4}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. Basic technical assumptions are given below.

Assumption 1.

1. *The mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous and there exists a function $f_\infty : \mathbb{R}^n \to \mathbb{R}^n$ such that $\lim_{c\to\infty} \frac{f(cx)}{c} = f_\infty(x), \forall x \in \mathbb{R}^n$.*

2. *The origin in $\mathbb{R}^n$ is an asymptotically stable equilibrium for the ODE $\dot{x}_t = f_\infty(x_t)$.*

3. *There exists a unique globally asymptotically stable equilibrium $\theta^e \in \mathbb{R}^n$ for the ODE $\dot{x}_t = f(x_t)$, i.e., $x_t \to \theta^e$ as $t \to \infty$.*

4. *The sequence $\{\varepsilon_k, \mathcal{G}_k, k \geq 1\}$ with $\mathcal{G}_k = \sigma(\theta_i, \varepsilon_i, i \leq k)$ is a martingale difference sequence. In addition, there exists a constant $C_0 < \infty$ such that for any initial $\theta_0 \in \mathbb{R}^n$, we have $\mathbb{E}[\|\varepsilon_{k+1}\|^2|\mathcal{G}_k] \leq C_0(1 + \|\theta_k\|^2), \forall k \geq 0$.*

5. *The step-sizes satisfy $\alpha_k > 0, \sum_{k=0}^\infty \alpha_k = \infty, \sum_{k=0}^\infty \alpha_k^2 < \infty$.*

Lemma 4 ([4, Borkar and Meyn theorem])**.** *Under Assumption 1, for any initial $\theta_0 \in \mathbb{R}^n$, $\sup_{k\geq 0} \|\theta_k\| < \infty$ with probability one. In addition, $\theta_k \to \theta^e$ as $k \to \infty$ with probability one.*

3 Convergence Analysis of Asynchronous Q-learning

We consider the Q-learning updates

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha_k \left\{ r_{a_k}(s_k, s_{k+1}) + \gamma \max_{a\in\mathcal{A}} Q_k(s_{k+1}, a_k) - Q_k(s_k, a_k) \right\}, \tag{5}$$

where $\alpha_k \geq 0$ is the learning rate and (s_k, a_k, s_{k+1}) comes from the trajectory of some behavior policy. For simplicity of presentation, we assume $\{(s_k, a_k)\}_{k=0}^\infty$ is a sequence of i.i.d. random variables from the stationary state-action distribution, $d_a(s)$, such that $d_a(s) > 0$ holds for all $s \in \mathcal{S}, a \in \mathcal{A}$. This assumption is common in the ODE approaches for Q-learning and TD-learning [29] and can potentially be relaxed. Note that different from the original Watkin's Q-learning, we do not require the step-size α_k to depend on the state-action pair.

Before proceeding, we introduce the following compact notations:

$$P := [P_1^T, \cdots, P_{|\mathcal{A}|}^T]^T \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|},\ R := [R_1^T, \cdots, R_{|\mathcal{A}|}^T]^T \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$
$$D_a := \text{diag}[d_a(1), \cdots, d_a(|\mathcal{S}|)] \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|},\ D := \text{diag}[D_1, \cdots, D_{|\mathcal{A}|}] \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}.$$

and $Q := [Q_1^T, \cdots, Q_{|\mathcal{A}|}^T]^T \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, where $Q_a = Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}$, and $R_a(s) := \mathbb{E}[r_a(s, s')|s, a]$. By definition, D is a nonsingular diagonal matrix with strictly positive diagonal elements. In addition, we denote $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ as s-th basis vector (zero except for the s-th component) and a-th basis vector, respectively. For any deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, we define the corresponding distribution vector $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{S}|}$, where $\Delta_{|\mathcal{S}|}$ is the set of all probability distributions over $\mathcal{S}$. Lastly, we denote the matrix

$$\Pi_\pi := [\vec{\pi}(1) \otimes e_1, \vec{\pi}(2) \otimes e_2, \ldots, \vec{\pi}(|S|) \otimes e_{|S|}]^T \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$$

and greedy policy $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} e_s^T Q_a \in \mathcal{A}$. By definition, for any $\pi \in \Theta$, $P\Pi_\pi$ is the state-action pair transition probability matrix under the deterministic policy π.

3.1 Asynchronous Q-learning as affine switching system

Using the notation introduced, the Q-learning update can be rewritten as

$$Q_{k+1} = Q_k + \alpha_k \left(DR + \gamma DP\Pi_{\pi_{Q_k}} Q_k - DQ_k + \varepsilon_{k+1} \right), \tag{6}$$

where

$$\begin{aligned}\varepsilon_{k+1} =& (e_a \otimes e_s)(e_a \otimes e_s)^T R + \gamma (e_a \otimes e_s)(e_{s'})^T \Pi_{\pi_{Q_k}} Q_k \\ & - (e_a \otimes e_s)(e_a \otimes e_s)^T Q_k - (DR + \gamma DP\Pi_{\pi_{Q_k}} Q_k - DQ_k).\end{aligned}$$

It can be easily shown that $\{\varepsilon_{k+1}\}$ is a martingale difference sequence. Using the Bellman equation $(\gamma DP\Pi_{\pi_{Q^*}} - D)Q^* + DR = 0$, (6) can be further rewritten as

$$\begin{aligned}(Q_{k+1} - Q^*) = (Q_k - Q^*) + \alpha_k \big[& (\gamma DP\Pi_{\pi_{Q_k}} - D)(Q_k - Q^*) \\ & + \gamma DP(\Pi_{\pi_{Q_k}} - \Pi_{\pi_{Q^*}})Q^* + \varepsilon_{k+1} \big].\end{aligned} \tag{7}$$

As discussed in Section 2.4, the convergence of (7) can be analyzed by evaluating the stability of the corresponding continuous-time ODE

$$\frac{d}{dt}(Q_t - Q^*) = (\gamma DP\Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^*, Q_0 - Q^* = z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{8}$$

which is an affine switching system. More precisely, if we define a one-to-one map $\psi : \Theta \to \{1, 2, \ldots, |\Theta|\}$, where Θ is the set of all deterministic policies, $x_t := Q_t - Q^*$, and

$$(A_{\psi(\pi)}, b_{\psi(\pi)}) := (\gamma DP\Pi_\pi - D, \gamma DP(\Pi_\pi - \Pi_{\pi_{Q^*}})Q^*)$$

for all $\pi \in \Theta$, then (8) can be represented by the affine switching system

$$\frac{d}{dt} x_t = A_{\sigma(x_t)} x_t + b_{\sigma(x_t)}, \quad x_0 = z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{9}$$

where, $\sigma : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \{1, 2, \ldots, |\Theta|\}$ is a state-feedback switching policy defined by $\sigma(x_t) := \psi(\pi_{Q_t})$, $\pi_{Q_t}(s) = \arg\max_{a \in \mathcal{A}} e_s^T Q_{t,a}$.

Since (9) is a switching system with a state-feedback switching policy, it may cause arbitrary switching behaviors. It is unclear whether its solution exists over all $t \geq 0$ and whether the solution is unique. We establish the existence and uniqueness of its solution, which follows from the global Lipschitz continuity of the affine mapping.

Proposition 1. *The mapping $f(\theta) = (\gamma DP\Pi_{\pi_\theta} - D)\theta$. is globally Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Hence, the solution of the switching system* (9) *exists and is unique for all $t \geq 0$ and $x_0 \in \mathbb{R}^n$.*

3.2 Stability analysis

Note that proving the global asymptotic stability of (9) without the affine term is relatively straightforward based on Lemma 3. However, none of the existing theory supports switching systems with affine terms. To address this issue, we construct two comparison systems by exploiting the special structure of the switching system and the greedy policy and prove their global asymptotic stability. By further building on the vector comparison principle introduced in Lemma 2, we then establish the asymptotic stability of the desired affine switching system.

More specifically, we consider the upper comparison system

$$\frac{d}{dt}(Q_t^u - Q^*) = (\gamma DP\Pi_{\pi_{Q_t^u}} - D)(Q_t^u - Q^*), \quad Q_0^u - Q^* > Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{10}$$

and the lower comparison system

$$\frac{d}{dt}(Q_t^l - Q^*) = (\gamma DP\Pi_{Q^*} - D)(Q_t^l - Q^*), \quad Q_0^l - Q^* < Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}. \tag{11}$$

Observe that, (10) is a linear switching system and (11) is a linear system. We can prove that both systems are asymptotically stable by verifying the strictly negative row dominating diagonal condition required in Lemma 3. By using the vector comparison theorem and the quasi-monotone property, we can prove that the original switching affine system's trajectories are sandwiched by the trajectories of the two systems.

Theorem 1. *Consider the systems (8), (10) and(11). We have*

1. $Q_t^l - Q^* \leq Q_t - Q^* \leq Q_t^u - Q^*$ *for all* $t \geq 0$*;*

2. *The origin is the unique globally asymptotically stable equilibrium point of the affine switching system* (8)*.*

We are now in position to apply the Borkar and Meyn theorem to establishing the convergence of asynchronous Q-learning.

Theorem 2. *Assume that the step-sizes satisfy*

$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{12}$$

Then, $Q_k \to Q^*$ *with probability one.*

The proof is fairly straightforward by invoking Lemma 4, Proposition 1, and Theorem 1. We can see that the convergence of asynchronous Q-learning follows immediately after proving the asymptotic stability of the associated affine switching system. In contrast, the convergence analysis of Q-learning in [4] relies on a nonlinear ODE model, whose asymptotic stability is proved in [5] by using the max-norm contraction of the Bellman operator; yet the analysis only applies to *synchronous Q-learning*, i.e., at each time all entries of the iterate are updated.

Lastly, it is worth mentioning that a number of work has recently established non-asymptotic analysis for asynchronous Q-learning, including [32, 9, 27]. The current best known bound is given in [27] showing a complexity of $\mathcal{O}\left(\frac{(|\mathcal{S}||\mathcal{A}|)^2}{(1-\gamma)^5\epsilon^2}\right)$. However, we stress that unlike this line of work, the purpose of our work is not to provide a tight convergence rate for asynchronous Q-learning, but rather to build an intuitive understanding of the family of Q-learning algorithms through the lens of switching systems. The switching system framework provides a simpler analysis and can be easily extended to deal with many Q-learning variants, as we show in the subsequent sections.

4 Convergence Analysis of Averaging Q-learning

We now consider a variant of the asynchronous Q-learning algorithm, called *averaging Q-learning*, which is newly introduced in [19] and motivated by the success of deep Q-learning [25], in order to improve the stability. The averaging Q-learning maintains two separate estimates, the target estimate and the online estimate: the online estimate is for approximating the state-action value function Q and updated through an online manner, whereas the target estimate is for computing the target values and updated through taking Polyak's averaging. Specifically, the algorithm works as follows:

$$Q_{k+1}^A(s_k, a_k) = Q_k^A(s_k, a_k) + \alpha_k \left\{ r_{a_k}(s_k, s_{k+1}) + \gamma \max_{a \in \mathcal{A}} Q_k^B(s_{k+1}, a_k) - Q_k^A(s_k, a_k) \right\}, \tag{13}$$

$$Q_{k+1}^B(s_k, a_k) = Q_k^B(s_k, a_k) + \alpha_k \delta (Q_k^A(s_k, a_k) - Q_k^B(s_k, a_k)). \tag{14}$$

where $\delta > 0$ is a rescaling constant. Following similar arguments as in the asynchronous Q-learning case, the corresponding ODE model is given by the following switching system:

$$\frac{d}{dt} \begin{bmatrix} Q_t^A - Q^* \\ Q_t^B - Q^* \end{bmatrix} = \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q_t^B}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} Q_t^A - Q^* \\ Q_t^B - Q^* \end{bmatrix} + \begin{bmatrix} \gamma DP(\Pi_{\pi_{Q_t^B}} - \Pi_{\pi_{Q^*}})Q^* \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} Q_0^A - Q^* \\ Q_0^B - Q^* \end{bmatrix} = z \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}, \tag{15}$$

which matches with the general form in (9). We obtain the global asymptotic stability of (15).

Theorem 3. *For any* $\delta > 0$*, the origin is the unique globally asymptotically stable equilibrium point of the affine switching system* (15)*.*

As a result, by invoking Borkar and Meyn's theorem similarly as before, we arrive at

Theorem 4. *For the averaging Q-learning, assuming the step-sizes satisfy* (12)*, then for any* $\delta > 0$*,* $Q_k^A \to Q^*$ *and* $Q_k^B \to Q^*$ *with probability one.*

We remark that this is indeed the first convergence analysis of the averaging Q-learning algorithm. In contrast, previous work [19] only provided the asymptotic convergence of averaging TD-learning. We expect that this analysis would also shed light on the convergence of other target-based Q-learning algorithms, e.g., the double Q-learning [11], periodic Q-learning [20], etc.

5 Convergence Analysis of Q-learning with Linear Function Approximation

When the state-space is large, linear function approximation are commonly used to approximate the optimal Q-function, $Q^* \cong \Phi\theta^*$, where Φ is the feature matrix. In particular, given pre-selected basis (or feature) functions $\phi_1, \ldots, \phi_n : \mathcal{S} \to \mathbb{R}$, the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}|\times n}$ is defined as $\Phi := [\phi(1,1), \phi(2,1), \cdots, \phi(|\mathcal{S}|, |\mathcal{A}|)]^T \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times n}$, where $\phi(s,a)^T := [\phi_1(s,a), \phi_2(s,a), \cdots, \phi_n(s,a)] \in \mathbb{R}^n$. Here $n \ll |\mathcal{S}||\mathcal{A}|$ is a positive integer.

Q-learning with linear function approximation performs the following update:

$$\theta_{k+1} = \theta_k + \alpha_k \phi(s_k, a_k)\big[r_a(s_k, s_{k+1}) + \gamma \max_{a\in\mathcal{A}} (\Phi\theta_k)(s_{k+1}, a) - (\Phi\theta_k)(s_k, a_k)\big], \tag{16}$$

where $\alpha_k \geq 0$ is the learning rate and $\{(s_k, a_k)\}_{k=0}^{\infty}$ are sampled from the stationary state-action distribution $d_a(s)$ under a behavior policy β such that $d_a(s) = \lim_{k\to\infty} \mathbb{P}[(s_k, a_k) = (s, a)|\beta]$. It is well-known that Q-learning with linear function approximation may not converge in general [30]. Under certain conditions, its convergence can be proven. For instance, [24] demonstrates the asymptotic convergence assuming that the following condition holds:

$$\gamma^2 \Phi^T \Pi_\pi^T D^\beta \Pi_\pi \Phi \prec \Phi^T D \Phi, \quad \forall \pi \in \Theta_\Phi, \tag{17}$$

where $\Theta_\Phi := \{\pi \in \Theta : \pi(s) = \arg\max_{a\in\mathcal{A}} (\Phi\theta)(s,a), \forall s \in \mathcal{S}, \theta \in \mathbb{R}^m\}$ and D^β is a diagonal matrix whose diagonal entries correspond to the stationary state distribution of the underlying Markov decision process under the behavior policy β. Recently, [6] considered a slightly stronger condition in order to obtain the convergence rate of Q-learning with linear function approximation.

In this section, we analyze the convergence from the switching system perspective and provide a new sufficient condition that ensures the asymptotic convergence. We start by introducing some basic assumptions.

Assumption 2. $[\Phi]_{ij} \geq 0$ *for all* $i \in \mathcal{S}$ *and* $j \in \{1, 2, \ldots, n\}$.

Assumption 3. *All column vectors of* Φ *are orthogonal.*

Assumption 2 requires all elements of Φ to be nonnegative. This assumption is required in our convergence analysis to obtain lower and upper comparison systems of the affine switching system. In the case that no function approximation is used, Φ is set to be an identity matrix, $\Phi = I$, which automatically satisfies Assumption 2. We emphasize that this assumption is not very restrictive. For instance, if the values of rewards are nonnegative, then it is sufficient to set feature vectors with nonnegative elements when approximating the Q-function. Otherwise, the rewards can always be shifted to nonnegative by adding a large enough constant. Assumption 3 is slightly stricter than the assumption of having full column rank which is usually adopted in the RL literature.This is required in order to guarantee the quasi-monotonicity of the corresponding switching system models.

Following a similar analysis, the associated affine switching system is given by

$$\frac{d}{dt}\theta_t = (\gamma \Phi^T D P \Pi_{\pi_{\Phi\theta_t}} \Phi - \Phi^T D \Phi)\theta_t + \Phi^T D R, \quad \theta_0 \in \mathbb{R}^n,$$

or equivalently,

$$\frac{d}{dt}(\theta_t - \theta^*) = (\gamma \Phi^T D P \Pi_{\pi_{\Phi\theta_t}} \Phi - \Phi^T D \Phi)(\theta_t - \theta^*) + \gamma \Phi^T D P (\Pi_{\pi_{\Phi\theta_t}} - \Pi_{\pi_{\Phi\theta^*}})\Phi\theta^*, \tag{18}$$

where $\theta_0 - \theta^* = z \in \mathbb{R}^n$, $\pi_{\Phi\theta_t}(s) = \arg\max_{a\in\mathcal{A}} (\Phi\theta_t)(s,a)$ and θ^* is the optimal parameter satisfying the projected Bellman equation $\Phi\theta^* = \Gamma(\gamma P \Pi_{\pi_{\Phi\theta^*}} \Phi\theta^* + R)$, and $\Gamma := \Phi(\Phi^T D \Phi)^{-1} \Phi^T D$ is the projection onto the range of Φ.

We first establish the asymptotic stability of the system (18).

Theorem 5. *Suppose that Assumption 2 and Assumption 3 hold. The origin is the unique globally asymptotically stable equilibrium point of the affine switching system* (18) *if the following condition holds:*

$$-\phi_i^T D \phi_i + \phi_i^T \gamma D P \Pi_{\psi(\pi)} \sum_{j\in\{1,2,\ldots,n\}} \phi_j < 0, \quad \pi \in \Theta_\Phi, \tag{19}$$

where ϕ_i^T *is the* i*-th row of the feature matrix* Φ.

As a result, this leads to the following convergence:

Theorem 6. *For Q-learning with linear function approximation, under Assumptions 2-3 and the condition specified in (19), we have $\theta_k \to \theta^*$ with probability one.*

We now make some remarks on the sufficient condition (19), which may look abstract at first sight since it purely stems from switching system theory. Similar to Melo's condition, our new condition also suggests that the behavior policy should be close to the optimal policy. In fact, this condition is quite similar to the diagonal dominant condition used in network science fields [37, 7]. Our analysis indicates that this condition is a necessary and sufficient condition for the asymptotic stability of the underlying switching system model of Q-learning, while Melo's condition is only a sufficient condition for the asymptotic stability. Especially, Melo's condition is strong enough to guarantee the existence of a quadratic Lyapunov function for the underlying switching system model, while in general, the switching system does not necessarily admit quadratic Lyapunov functions. This shows the less conservativeness of our new condition.

Proposition 2. *Under the above assumptions, Melo's condition* (17) *implies the condition (19).*

In practice, to derive a computationally tractable sufficient condition, Θ_Φ can be replaced with Θ. A special case where the condition (19) holds is when elements of the feature vectors ϕ_i are binary numbers, $\{0, 1\}$. This clearly holds for the tabular setting.

Proposition 3. *Suppose the elements of the feature matrix Φ are binary numbers, i.e., $\{0, 1\}$, then the condition (19) always holds.*

Lastly, we give a simple MDP example which satisfies the sufficient condition in (19), but violates the Melo's condition (17).

Example 1. *Consider an MDP with $\mathcal{S} = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$, $\gamma = 0.9$, $P_1 = \begin{bmatrix} 1/2 & 1/2 \\ 1 & 0 \end{bmatrix}, P_2 = \begin{bmatrix} 0 & 1 \\ 2/3 & 1/3 \end{bmatrix}$, and a behavior policy β such that $\mathbb{P}[a = 1|s = 1] = 0.2, \mathbb{P}[a = 2|s = 1] = 0.8, \mathbb{P}[a = 1|s = 2] = 0.7, \quad \mathbb{P}[a = 2|s = 2] = 0.3$. The corresponding matrices D^β and D are given by*

$$D^\beta = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \quad D = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.35 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.15 \end{bmatrix}.$$

If the feature matrix is $\Phi^T = \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix}$, then Θ_Φ is given by $\Theta_\Phi = \{\pi_1, \pi_2\}$, where π_1 is a deterministic policy such that $\pi_1(1) = 1$, $\pi_1(2) = 1$ and π_2 such that $\pi_2(1) = 2$, $\pi_2(2) = 2$, which are obtained by considering three cases, $\theta > 0, \theta = 0, \theta < 0$, Here, we assume that whenever $\{1, 2\} = \arg\max_{a \in \mathcal{A}}(\Phi\theta)(s, a)$, we select $a = 1$ in Q-learning. The quantities in (19) are given by -0.885 and -0.03 for all $\pi \in \Theta_\Phi = \{\pi_1, \pi_2\}$, implying convergence of the algorithm. However, the quantity $\gamma^2 \Phi^T \Pi_\pi^T D^\beta \Pi_\pi \Phi - \Phi^T D \Phi$ is computed as -1.2450 and 0.3750 for all $\pi \in \Theta_\Phi = \{\pi_1, \pi_2\}$, respectively. This implies that the condition in (17) *fails to verify the convergence.*

6 Numerical Simulation

Consider an MDP with $\mathcal{S} = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$, $\gamma = 0.9$,

$$P_1 = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}, R_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and a behavior policy β such that

$$\mathbb{P}[a = 1|s = 1] = 0.2, \quad \mathbb{P}[a = 2|s = 1] = 0.8,$$
$$\mathbb{P}[a = 1|s = 2] = 0.7, \quad \mathbb{P}[a = 2|s = 2] = 0.3.$$

Simulated trajectories of the O.D.E. model of Q-learning including the upper and lower comparison systems are depicted in Figure 1. Simulated trajectories of the O.D.E. model of the averaging Q-learning including the upper and lower comparison systems are depicted in Figure 2 for Q_t^A part. The simulation study empirically justifies the bounding principles and asymptotic convergence established in theory.

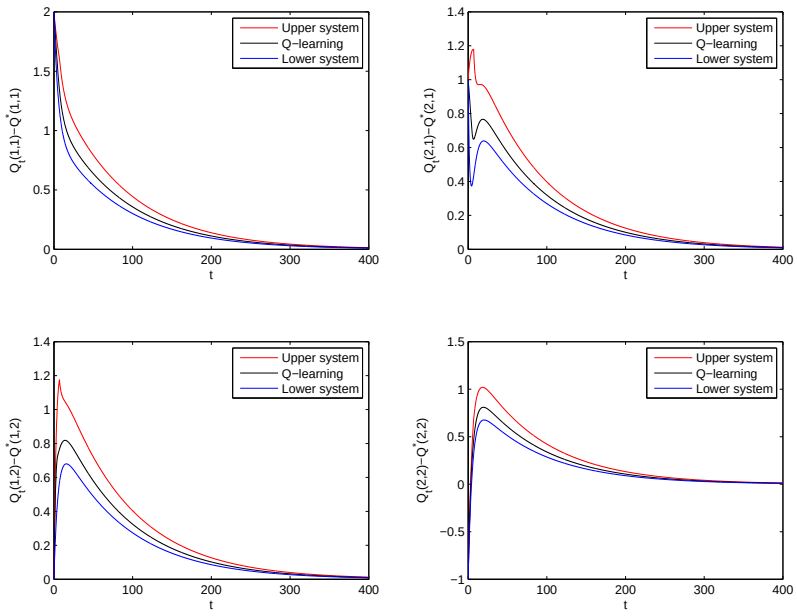

Figure 1: Trajectories of the O.D.E. model of Q-learning and the upper and lower comparison systems

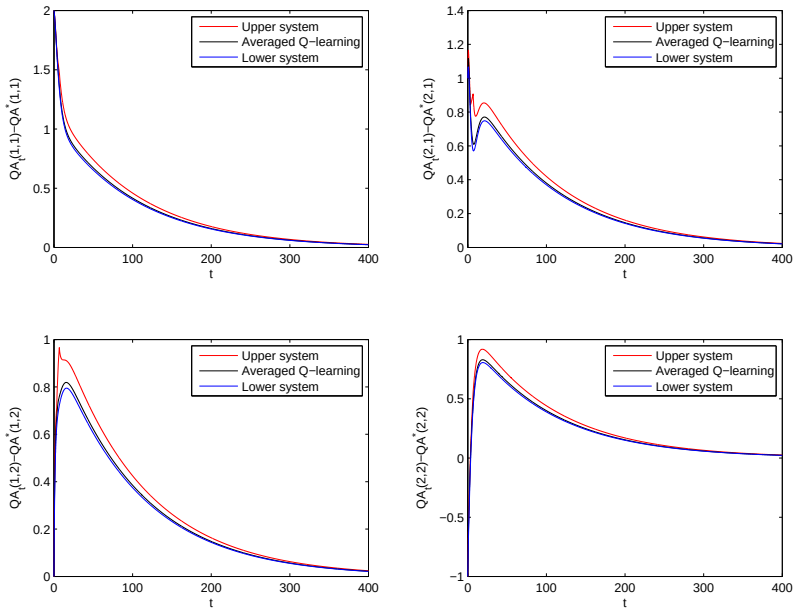

Figure 2: Trajectories of the O.D.E. model of averaging Q-learning and the upper and lower comparison systems (Q_t^A part)

7 Conclusion

In this paper, we offer a unified and convenient convergence analysis of various Q-learning algorithms building on novel connections to switching systems. We establish the first ODE analysis for asynchronous Q-learning and averaging Q-learning, and derive a new sufficient condition to ensure the convergence of Q-learning with linear function approximation. While this work focuses only on the asymptotic convergence of a subset of RL algorithms, we expect that the switching system approach could be leveraged to advance the understanding of many other RL algorithms and extended to non-asymptotic analysis. This may also shed light on the design of more efficient and robust RL algorithms from the control perspective, which we leave for future investigation.

Acknowledgments and Disclosure of Funding

We thank the reviewers and area chair for constructive feedback. We would like to thank Csaba Szepesvari, Bin Hu, and Rohit Gupta for insightful comments. The work was supported by NSF CRII 1755829 and NSF CCF 1934986.

Broader Impact

By bridging Q-learning with switching systems, this work has full potential to promote synergy between two closely related fields/communities: control theory and reinforcement learning, as well as to stimulate further developments in the theory, algorithms and applicability of reinforcement learning. Meanwhile, this paper provides an accessible material on the basics of stochastic approximation, switching system theory, and reinforcement learning theory, which could be beneficial to graduate students, researchers, and even reinforcement learning practitioners. Our analysis could potentially inspire the development of more efficient and robust algorithms that benefit a broad spectrum of data-intensive applications in the realm of reinforcement learning.

References

[1] Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert J Kappen. Speedy Q-learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2411–2419, 2011.

[2] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.

[3] Shalabh Bhatnagar, H. L. Prasad, and L. A. Prashanth. *Stochastic recursive algorithms for optimization: simultaneous perturbation methods*, volume 434. Springer, 2012.

[4] Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[5] Vivek S Borkar and K Soumyanatha. An analog scheme for fixed point computation. i. theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(4): 351–355, 1997.

[6] Zaiwei Chen, Sheng Zhang, Thinh T. Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of Q-learning with linear function approximation: Stability and finite-time analysis, 2019.

[7] Lj Cvetković and Vladimir Kostić. Application of generalized diagonal dominance in wireless sensor network optimization problems. *Applied Mathematics and Computation*, 218(9):4798–4805, 2012.

[8] Adithya M Devraj and Sean Meyn. Zap Q-learning. In *Advances in Neural Information Processing Systems*, pages 2235–2244, 2017.

[9] Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

[10] Harsh Gupta, R Srikant, and Lei Ying. Adaptive learning rate selection for temporal difference learning. *ICML workshop on Real-world Sequential Decision Making*, 2019.

[11] Hado V Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.

[12] Morris W Hirsch and Hal Smith. Monotone dynamical systems. In *Handbook of differential equations: ordinary differential equations*, volume 2, pages 239–357. Elsevier, 2006.

[13] Bin Hu and Usman Syed. Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. In *Advances in Neural Information Processing Systems*, pages 8477–8488, 2019.

[14] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

[15] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[16] Hassan K Khalil. Nonlinear systems. *Upper Saddle River*, 2002.

[17] Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[18] Donghwan Lee and Niao He. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In *2019 American Control Conference (ACC)*, pages 4897–4902. IEEE, 2019.

[19] Donghwan Lee and Niao He. Target-based temporal-difference learning. In *International Conference on Machine Learning*, pages 3713–3722, 2019.

[20] Donghwan Lee and Niao He. Periodic Q-learning. *arXiv preprint arXiv:2002.09795*, 2020.

[21] Daniel Liberzon. *Switching in systems and control*. Springer Science & Business Media, 2003.

[22] Hai Lin and Panos J Antsaklis. Stability and stabilizability of switched linear systems: a survey of recent results. *IEEE Transactions on Automatic control*, 54(2):308–322, 2009.

[23] Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *International Conference on Machine Learning*, pages 719–726, 2010.

[24] Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[26] André Platzer. Vector barrier certificates and comparison systems. In *Formal Methods: 22nd International Symposium, FM 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 15-17, 2018, Proceedings*, volume 10951, page 418, 2018.

[27] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*, 2020.

[28] R Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830, 2019.

[29] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[30] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

[31] Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000. ACM, 2009.

[32] Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070, 1998.

[33] John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

[34] Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[35] Wolfgang Walter. *Ordinary differential equations (graduate texts in mathematics)*. Springer, 1998.

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

[37] Lin Xiao and Stephen Boyd. Optimal scaling of a gradient method for distributed resource allocation. *Journal of optimization theory and applications*, 129(3):469–488, 2006.

[38] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675, 2019.

A Notations

Some notations used in the paper are summarized below:

1. State space: $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$
2. Action space: $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$
3. Transition probability: $P_a(s, s')$
4. Random reward: $r_a(s, s')$
5. $P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, a \in \mathcal{A}$
6. $P_a(s, s')$: the state transition probability from the current state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ under action $a \in \mathcal{A}$
7. $r_a(s, s')$: the reward random variable conditioned on $a \in \mathcal{A}, s, s' \in \mathcal{S}$
8. $R_a(s, s') := \mathbb{E}[r_a(s, s')|s, a, s']$
9. $$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \quad R := \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \quad Q := \begin{bmatrix} Q_1 \\ \vdots \\ Q_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$
   where $Q_a = Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}, a \in \mathcal{A}$ and $R_a(s) := \mathbb{E}[r_a(s, s')|s, a]$.
10. $$D_a := \begin{bmatrix} d_a(1) & & \\ & \ddots & \\ & & d_a(|\mathcal{S}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \quad D := \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$$
11. $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$: s-th basis vector (zero except for the s-th component) and a-th basis vector, respectively
12. $\Delta_{|\mathcal{S}|}$: the set of all probability distributions over $\mathcal{S}$
13. $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{S}|}$
14. $$\Pi_\pi := \begin{bmatrix} \vec{\pi}(1)^T \otimes e_1^T \\ \vec{\pi}(2)^T \otimes e_2^T \\ \vdots \\ \vec{\pi}(|S|)^T \otimes e_{|S|}^T \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}.$$
15. Feature vector: $\phi(s, a)^T := [\phi_1(s, a), \phi_2(s, a), \cdots, \phi_n(s, a)] \in \mathbb{R}^n$.
16. Feature matrix:
$$\Phi := \begin{bmatrix} \phi(1, 1)^T \\ \phi(2, 1)^T \\ \vdots \\ \phi(|\mathcal{S}|, |\mathcal{A}|)^T \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times n},$$

B Proof of Lemma 2

Proof. We simplify and summarize the ideas of the proofs in the literature, [35, page 112],[12, Theorem 3.2.], in the following proof. Instead of (2), first consider

$$\frac{d}{dt}v_\varepsilon(t) = \underline{f}(v_\varepsilon(t)) - \varepsilon\mathbf{1}_n, \quad v_\varepsilon(0) < x(0), \quad \forall t \geq 0$$

where $\varepsilon > 0$ is a sufficiently small real number and $\mathbf{1}_n$ is a vector where all elements are ones, where we use a different notation for the time index for convenience. Suppose that the statement is not true, and let

$$t^* := \inf\{t \geq 0 : \exists i \text{ such that } v_{\varepsilon,i}(t) > x_i(t)\} < \infty,$$

and let i be such index. By the definition of t^*, we have that $v_{\varepsilon,i}(t^*) = x_i(t^*)$ and $v_{\varepsilon,j}(t^*) \leq x_j(t^*)$ for any $j \neq i$. Then, since $\overline{f}$ is quasi-monotone increasing, we have

$$\overline{f}_i(v_\varepsilon(t^*)) \leq \overline{f}_i(x(t^*)). \tag{20}$$

On the other hand, by the definition of t^*, there exists a small $\delta > 0$ such that

$$v_{\varepsilon,i}(t^* + \Delta t) > x_i(t^* + \Delta t)$$

for all $0 < \Delta t < \delta$. Dividing both sides by Δt and taking the limit $\Delta t \to 0$, we have

$$\dot{v}_{\varepsilon,i}(t^*) \geq \dot{x}_i(t^*) = \overline{f}_i(x(t^*)). \tag{21}$$

By the hypothesis, it holds that

$$\frac{d}{dt} v_\varepsilon(t) = \underline{f}(v_\varepsilon(t)) - \varepsilon \mathbf{1}_n < \underline{f}(v_\varepsilon(t)) \leq \overline{f}(v_\varepsilon(t))$$

holds for all $t \geq 0$. The inequality implies $\dot{v}_{\varepsilon,i}(t) < \overline{f}_i(v_\varepsilon(t))$, which in combination with (21) leads to $\overline{f}_i(v_\varepsilon(t^*)) > \overline{f}_i(x(t^*))$. However, it contradicts with (20). Therefore, $v_\varepsilon(t) \leq x(t)$ holds for all $t \geq 0$. Since the solution $v_\varepsilon(t)$ continuously depends on $\varepsilon > 0$ [35, Chap. 13], taking the limit $\varepsilon \to 0$, we conclude $v_0(t) \leq x(t)$ holds for all $t \geq 0$. This completes the proof.

□

C Proof of Proposition 1

Proof. The proof is completed by the inequalities

$$\begin{aligned}
\|f(x) - f(y)\|_\infty =& \|(\gamma DP\Pi_{\pi_x} - D)x - (\gamma DP\Pi_{\pi_y} - D)y\|_\infty \\
\leq& \|\gamma DP\|_\infty \|\Pi_{\pi_x} x - \Pi_{\pi_y} y\|_\infty + \|D\|_\infty \|x - y\|_\infty \\
=& \|\gamma DP\|_\infty \max_{s \in \mathcal{S}} |\max_{a \in \mathcal{A}} x_a(s) - \max_{a \in \mathcal{A}} y_a(s)| + \|D\|_\infty \|x - y\|_\infty \\
\leq& \|\gamma DP\|_\infty \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} |x_a(s) - y_a(s)| + \|D\|_\infty \|x - y\|_\infty \\
=& \|\gamma DP\|_\infty \|x - y\|_\infty + \|D\|_\infty \|x - y\|_\infty \\
\leq& \left(\|\gamma DP\|_\infty + \|D\|_\infty\right) \|x - y\|_\infty,
\end{aligned}$$

indicating that f is globally Lipschitz continuous with respect to the $\|\cdot\|_\infty$ norm. □

D Proof of Theorem 1

Lemma 5. *Consider the affine switching system* (9)*. The origin of the associated linear switching system*

$$\frac{d}{dt} x_t = A_{\sigma_t} x_t,$$

is the unique globally asymptotically stable equilibrium point under arbitrary switchings, σ_t.

Proof. The proof follows by applying Lemma 3 with $L = I, \bar{A}_\sigma = A_\sigma$. In this case, the condition, $LA_\sigma = \bar{A}_\sigma L$ holds. It remains to prove the strictly negative row dominating diagonal property. For notational convenience, we definte Π_σ, $\sigma \in \mathcal{M}$ as $\Pi_{\pi_{Q_t^B}}$ such that $\sigma = \psi(\pi_{Q_t^B})$. Then,

$$\begin{aligned}
[A_\sigma]_{ii} + \sum_{j \in \{1,2,\ldots,n\} \setminus \{i\}} |[A_\sigma]_{ij}| =& [D]_{ii}[\gamma P\Pi_\sigma - I]_{ii} + \sum_{j \in \{1,2,\ldots,n\} \setminus \{i\}} [D]_{ii} |[\gamma P\Pi_\sigma - I]_{ij}| \\
\leq& [\gamma P\Pi_\sigma - I]_{ii} + \sum_{j \in \{1,2,\ldots,n\} \setminus \{i\}} |[\gamma P\Pi_\sigma - I]_{ij}| \\
=& [\gamma P\Pi_\sigma]_{ii} - 1 + \sum_{j \in \{1,2,\ldots,n\} \setminus \{i\}} |[\gamma P\Pi_\sigma]_{ij}| \\
=& \gamma - 1 < 0, \quad \forall \sigma \in \mathcal{M},
\end{aligned}$$

which proves the global asymptotic stability. □

Proof of Theorem 1. The basic idea of the proof is to find systems whose trajectories lower and upper bounds the trajectory of (9) by the vector comparison principle. Then, by proving the asymptotic stability of the two comparison systems, we can prove the asymptotic stability of (9).

Since each element of $\Pi_{\pi_{Q^*}} Q^*$ takes the maximum value across a, it is clear that $(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^* \leq 0$ holds, where the inequality is element-wise. Moreover, since γDP has nonnegative elements, $\gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^* \leq 0$ holds. Therefore, we have $(\gamma D_\beta P \Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^* \leq (\gamma DP\Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) \leq (\gamma DP\Pi_{\pi_{Q_t - Q^*}} - D)(Q_t - Q^*)$ for all $t \in \mathbb{R}_+$. To proceed, define the vector functions

$$\begin{aligned}\overline{f}(y) =& (\gamma DP\Pi_{\pi_y} - D)y,\\ \underline{f}(z) =& (\gamma DP\Pi_{\pi_{z+Q^*}} - D)z + \gamma DP(\Pi_{\pi_{z+Q^*}} - \Pi_{\pi_{Q^*}})Q^*,\end{aligned}$$

and consider the systems

$$\begin{aligned}\frac{d}{dt}y_t &= \overline{f}(y_t), \quad y_0 > Q_0 - Q^*,\\ \frac{d}{dt}z_t &= \underline{f}(z_t), \quad z_0 = Q_0 - Q^*,\end{aligned}$$

for all $t \geq 0$. To apply Lemma 2, we will prove that $\overline{f}$ is quasi-monotone increasing. For any $z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, consider a nonnegative vector $p \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that its ith element is zero. Then, for any $i \in \mathcal{S}$, we have

$$\begin{aligned}e_i^T \overline{f}(z+p) =& e_i^T(\gamma DP\Pi_{z+p} - D)(z+p)\\ =& \gamma e_i^T DP\Pi_{z+p}(z+p) - e_i^T Dz - e_i^T Dp\\ =& \gamma e_i^T DP\Pi_{z+p}(z+p) - e_i^T Dz\\ =& \gamma e_i^T DP \begin{bmatrix} \max_a(z_a(1)+p_a(1)) \\ \max_a(z_a(2)+p_a(2)) \\ \vdots \\ \max_a(z_a(|\mathcal{S}|)+p_a(|\mathcal{S}|)) \end{bmatrix} - e_i^T Dz\\ \geq& \gamma e_i^T DP \begin{bmatrix} \max_a z_a(1) \\ \max_a z_a(2) \\ \vdots \\ \max_a z_a(|S|) \end{bmatrix} - e_i^T Dz\\ =& e_i^T \overline{f}(z),\end{aligned}$$

which proves the quasi-monotone increasing property, where the second line is due to $e_i^T Dp = 0$. Moreover, following similar lines of the proof of Proposition 1, one can prove that $\overline{f}$ is Lipschitz continuous. Using $\underline{f}(z) = (\gamma DP\Pi_{\pi_{(z+Q^*)}} - D)(z + Q^*) + DR$ and following similar lines of the proof of Proposition 1, we conclude that $\underline{f}$ is Lipschitz continuous as well. Now, by Lemma 2, $Q_t - Q^* \leq Q_t^u - Q^*$ holds for every $t \in \mathbb{R}_+$, where $Q_t^u - Q^*$ is the solution of the switching system, which we refer to as an upper comparison system

$$\frac{d}{dt}(Q_t^u - Q^*) = (\gamma DP\Pi_{\pi_{Q_t^u}} - D)(Q_t^u - Q^*), \quad Q_0^u - Q^* > Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

By Lemma 5, the origin of the above switching system is globally asymptotically stable even under arbitrary switchings. Therefore, $Q_t - Q^*$ is asymptotically upper bounded by the zero vector as $t \to \infty$.

On the other hand, we have

$$\begin{aligned}(\gamma DP\Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^* =& (\gamma DP\Pi_{\pi_{Q_t}} - D)Q_t + DR\\ \geq& (\gamma DP\Pi_{\pi_{Q^*}} - D)Q_t + DR = (\gamma DP\Pi_{\pi_{Q^*}} - D)(Q_t - Q^*),\end{aligned}$$

where the first inequality is due to $\gamma DP\Pi_{\pi_{Q_t}} Q_t \geq \gamma DP\Pi_{\pi_{Q^*}} Q_t$, and the second equality uses $DQ^* = \gamma DP\Pi_{\pi_{Q^*}} Q^* + DR$. Again, define the vector functions for lower comparison parts

$$\begin{aligned}\overline{f}(y) =& (\gamma DP\Pi_{\pi_y} - D)y + DR,\\ \underline{f}(z) =& (\gamma DP\Pi_{\pi_{Q^*}} - D)z + DR\end{aligned}$$

and consider the systems

$$\frac{d}{dt}y_t = \overline{f}(y_t), \quad y_0 = Q_0,$$
$$\frac{d}{dt}z_t = \underline{f}(z_t), \quad z_0 < Q_0,$$

for all $t \geq 0$. To apply Lemma 2, we can prove that $\overline{f}$ is quasi-monotone increasing following the same lines as above. $\overline{f}$ is Lipschitz continuous by Proposition 1 and $\underline{f}$ is Lipschitz continuous as it is linear. Therefore, we can invoke Lemma 2, to prove the inequality $Q_t^l - Q^* \leq Q_t - Q^*$ for all $t \geq 0$, where $Q_t^l - Q^*$ is the solution of the following linear system called the lower comparison system:

$$\frac{d}{dt}(Q_t^l - Q^*) = (\gamma DP\Pi_{Q^*} - D)(Q_t^l - Q^*), \quad Q_0^l - Q^* < Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

The origin of the above linear system is globally asymptotically stable equilibrium point by Lemma 5. Therefore, $Q_t - Q^*$ is asymptotically lower bounded by the zero vector as $t \to \infty$. Combining the bounds, we conclude that $Q_t - Q^* \to 0$ as $t \to \infty$. This completes the proof of Theorem 1. □

E Proof of Theorem 2

Proof of Theorem 2. First of all, note that the affine switching system model in (9) corresponds to the ODE model, $\frac{d}{dt}x_t = f(x_t)$, that appears in Assumption 1. The proof is completed by examining all the statements in Assumption 1:

1. Q-learning in (7) can be expressed as the stochastic recursion in (4) with
$$f(\theta) = (\gamma DP\Pi_{\pi_\theta} - D)\theta + \gamma DP(\Pi_{\pi_\theta} - \Pi_{\pi_{Q^*}})Q^*.$$
To prove the first statement of Assumption 1, we note that
$$\frac{f(c\theta)}{c} = (\gamma DP\Pi_{\pi_\theta} - D)\theta + \frac{\gamma DP(\Pi_{\pi_\theta} - \Pi_{\pi_{Q^*}})Q^*}{c},$$
where the last equality is due to the homogeneity of the policy, $\pi_{c\theta}(s) = \arg\max_{a\in\mathcal{A}} e_s^T c\theta_a = \arg\max_{a\in\mathcal{A}} e_s^T\theta_a$. By taking the limit, we have
$$\begin{aligned}\lim_{c\to\infty}\frac{f(c\theta)}{c} &= (\gamma DP\Pi_{\pi_\theta} - D)\theta + \lim_{c\to\infty}\frac{\gamma DP(\Pi_{\pi_\theta} - \Pi_{\pi_{Q^*}})Q^*}{c} \\ &= (\gamma DP\Pi_{\pi_\theta} - D)\theta = f_\infty(\theta).\end{aligned}$$
Moreover, f is globally Lipschitz continuous according to Proposition 1. Therefore, the proof is completed.

2. The second statement of Assumption 1 follows from Lemma 5..

3. The third statement of Assumption 1 follows from Theorem 1.

4. Next, we prove the remaining parts. Recall that the Q-learning update is expressed as
$$Q_{k+1} = Q_k + \alpha_k(f(Q_k) + \varepsilon_{k+1})$$
with the stochastic error
$$\begin{aligned}\varepsilon_{k+1} =&(e_a \otimes e_s)(e_a \otimes e_s)^T R + \gamma(e_a \otimes e_s)(e_{s'})^T\Pi_{\pi_{Q_k}}Q_k \\ &- (e_a \otimes e_s)(e_a \otimes e_s)^T Q_k - (DR + \gamma DP\Pi_{\pi_{Q_k}}Q_k - DQ_k)\end{aligned}$$
and
$$f(Q) = (\gamma DP\Pi_{\pi_Q} - D)Q + \gamma DP(\Pi_{\pi_Q} - \Pi_{\pi_{Q^*}})Q^*.$$
Define the history $\mathcal{G}_k := (\varepsilon_k, \varepsilon_{k-1}, \ldots, \varepsilon_1, Q_k, Q_{k-1}, \ldots, Q_0)$, and the process $(M_k)_{k=0}^\infty$ with $M_k := \sum_{i=1}^k \varepsilon_i$. Then, we can prove that $(M_k)_{k=0}^\infty$ is Martingale. To do so, we first

prove $\mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k]=0$ by

$$\begin{aligned}\mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k] =&\mathbb{E}[(e_a\otimes e_s)(e_a\otimes e_s)^T R|\mathcal{G}_k]+\mathbb{E}[\gamma(e_a\otimes e_s)(e_{s'})^T\Pi_{\pi_{Q_k}}Q_k|\mathcal{G}_k]\\ &-\mathbb{E}[(e_a\otimes e_s)(e_a\otimes e_s)^T Q_k|\mathcal{G}_k]-\mathbb{E}[DR+\gamma DP\Pi_{\pi_{Q_k}}Q_k-DQ_k|\mathcal{G}_k]\\ =&\mathbb{E}[DR+\gamma DP\Pi_{\pi_{Q_k}}Q_k-DQ_k|\mathcal{G}_k]-\mathbb{E}[DR+\gamma DP\Pi_{\pi_{Q_k}}Q_k-DQ_k|\mathcal{G}_k]\\ =&0,\end{aligned}$$

where the second equality is due to the i.i.d. assumption of samples. Using this identity, we have

$$\begin{aligned}\mathbb{E}[M_{k+1}|\mathcal{G}_k] =&\mathbb{E}\left[\sum_{i=1}^{k+1}\varepsilon_i\middle|\mathcal{G}_k\right]=\mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k]+\mathbb{E}\left[\sum_{i=1}^{k}\varepsilon_i\middle|\mathcal{G}_k\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{k}\varepsilon_i\middle|\mathcal{G}_k\right]=\sum_{i=1}^{k}\varepsilon_i=M_k.\end{aligned}$$

Therefore, $(M_k)_{k=0}^{\infty}$ is a Martingale sequence, and $\varepsilon_{k+1}=M_{k+1}-M_k$ is a Martingale difference. Moreover, it can be easily proved that the fourth condition of Assumption 1 is satisfied by algebraic calculations. Therefore, the fourth condition is met.

□

F Proof of Theorem 3

Proof. Using $\gamma DP(\Pi_{\pi_{Q_t^B}}-\Pi_{\pi_{Q^*}})Q^*\le 0$, we obtain

$$\begin{aligned}\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{Q_t^B}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}Q_t^A-Q^*\\ Q_t^B-Q^*\end{bmatrix}+\begin{bmatrix}\gamma DP(\Pi_{\pi_{Q_t^B}}-\Pi_{\pi_{Q^*}})Q^*\\ 0\end{bmatrix}&\le\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{Q_t^B}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}Q_t^A-Q^*\\ Q_t^B-Q^*\end{bmatrix}\\ &\le\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{Q_t^B-Q^*}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}Q_t^A-Q^*\\ Q_t^B-Q^*\end{bmatrix}.\end{aligned}$$

Consider the upper comparison system

$$\frac{d}{dt}\begin{bmatrix}Q_t^{A,u}-Q^*\\ Q_t^{B,u}-Q^*\end{bmatrix}=\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{Q_t^{B,u}-Q^*}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}Q_t^{A,u}-Q^*\\ Q_t^{B,u}-Q^*\end{bmatrix},\ \begin{bmatrix}Q_0^{A,u}-Q^*\\ Q_0^{B,u}-Q^*\end{bmatrix}>\begin{bmatrix}Q_0^{A}-Q^*\\ Q_0^{B}-Q^*\end{bmatrix}\in\mathbb{R}^{2|\mathcal{S}||\mathcal{A}|},$$

and define the vector functions

$$\begin{aligned}\overline{f}(y_1,y_2):=\begin{bmatrix}\overline{f}_1(y_1,y_2)\\ \overline{f}_2(y_1,y_2)\end{bmatrix}&:=\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{y_2}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}\\ \underline{f}(y_1,y_2):=\begin{bmatrix}\underline{f}_1(z_1,z_2)\\ \underline{f}_2(z_1,z_2)\end{bmatrix}&:=\begin{bmatrix}-D & \gamma DP\Pi_{\pi_{z_2+Q^*}}\\ \delta I & -\delta I\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix}+\begin{bmatrix}\gamma DP(\Pi_{\pi_{z_2+Q^*}}-\Pi_{\pi_{Q^*}})Q^*\\ 0\end{bmatrix},\end{aligned}$$

and consider the systems

$$\begin{aligned}\frac{d}{dt}\begin{bmatrix}y_{t,1}\\ y_{t,2}\end{bmatrix}&=\begin{bmatrix}\overline{f}_1(y_{t,1},y_{t,2})\\ \overline{f}_2(y_{t,1},y_{t,2})\end{bmatrix},\quad y_0>\begin{bmatrix}Q_0^A-Q^*\\ Q_0^B-Q^*\end{bmatrix},\\ \frac{d}{dt}\begin{bmatrix}z_{t,1}\\ z_{t,2}\end{bmatrix}&=\begin{bmatrix}\underline{f}_1(z_{t,1},z_{t,2})\\ \underline{f}_2(z_{t,1},z_{t,2})\end{bmatrix},\quad z_0=\begin{bmatrix}Q_0^A-Q^*\\ Q_0^B-Q^*\end{bmatrix},\end{aligned}$$

for all $t\ge 0$. We first prove that $\overline{f}$ is quasi-monotone increasing. We will check the condition of the quasi-monotone increasing function for $\overline{f}_1$ and $\overline{f}_2$, separately. Assume that $p_1\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $p_2\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are nonnegative vectors, and an $i$the element of p_1 is zero. For $\overline{f}_1$, we have

$$\begin{aligned}e_i^T\overline{f}_1(y_1+p_1,y_2+p_2) =&-e_i^T D(y_1+p_1)+\gamma e_i^T DP\Pi_{\pi_{(y_2+p_2)}}(y_2+p_2)\\ =&-e_i^T Dy_1+\gamma e_i^T DP\Pi_{\pi_{(y_2+p_2)}}(y_2+p_2)\\ \ge&-e_i^T Dy_1+\gamma e_i^T DP\Pi_{\pi_{y_2}}y_2\\ =&e_i^T\overline{f}_1(y_1,y_2),\end{aligned}$$

where the second line is due to $-e_i^T D p_1 = 0$. Similarly, assuming that $p_1 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $p_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are nonnegative vectors, and an $i$the element of p_2 is zero, we get

$$\begin{aligned} e_i^T \overline{f}_2(y_1 + p_1, y_2 + p_2) =& \delta e_i^T (y_1 + p_1) - \gamma \delta e_i^T (y_2 + p_2) \\ =& \delta e_i^T (y_1 + p_1) - \gamma \delta e_i^T y_2 \\ \geq& \delta e_i^T y_1 - \gamma \delta e_i^T y_2 \\ =& e_i^T \overline{f}_2(y_1, y_2), \end{aligned}$$

where the second line is due to $e_i^T p_2 = 0$. Therefore, $\overline{f}$ is quasi-monotone increasing. The Lipschitz continuity of $\overline{f}$ and $\underline{f}$ can be easily proved. Therefore, by Lemma 2, $\begin{bmatrix} Q_t^A - Q^* \\ Q_t^B - Q^* \end{bmatrix} \leq \begin{bmatrix} Q_t^{A,u} - Q^* \\ Q_t^{B,u} - Q^* \end{bmatrix}$ holds for all $t \geq 0$, where $\begin{bmatrix} Q_t^{A,u} - Q^* \\ Q_t^{B,u} - Q^* \end{bmatrix}$ is the solution of the upper comparison system.

Moreover, using the inequality $\gamma DP\Pi_{\pi_{Q_t^B}} Q_t^B \geq \gamma DP\Pi_{\pi_{Q^*}} Q_t^B$, we obtain

$$\frac{d}{dt} \begin{bmatrix} Q_t^A \\ Q_t^B \end{bmatrix} = \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q_t^B}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} Q_t^A \\ Q_t^B \end{bmatrix} + \begin{bmatrix} DR \\ 0 \end{bmatrix} \geq \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q^*}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} Q_t^A \\ Q_t^B \end{bmatrix} + \begin{bmatrix} DR \\ 0 \end{bmatrix}.$$

Using this relation, consider the lower comparison system

$$\frac{d}{dt} \begin{bmatrix} Q_t^{A,l} - Q^* \\ Q_t^{B,l} - Q^* \end{bmatrix} = \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q^*}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} Q_t^{A,l} - Q^* \\ Q_t^{B,l} - Q^* \end{bmatrix}, \quad \begin{bmatrix} Q_0^{A,l} - Q^* \\ Q_0^{B,l} - Q^* \end{bmatrix} < \begin{bmatrix} Q_0^A - Q^* \\ Q_0^B - Q^* \end{bmatrix} \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|},$$

or equivalently,

$$\frac{d}{dt} \begin{bmatrix} Q_t^{A,l} \\ Q_t^{B,l} \end{bmatrix} = \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q^*}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} Q_t^{A,l} \\ Q_t^{B,l} \end{bmatrix} + \begin{bmatrix} DR \\ 0 \end{bmatrix}.$$

To proceed, define the vector functions

$$\begin{aligned} \overline{f}(y_1, y_2) &:= \begin{bmatrix} \overline{f}_1(y_1, y_2) \\ \overline{f}_2(y_1, y_2) \end{bmatrix} = \begin{bmatrix} \gamma DP\Pi_{\pi_{y_2}} & -D \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} DR \\ 0 \end{bmatrix} \\ \underline{f}(z_1, z_2) &:= \begin{bmatrix} \underline{f}_1(z_1, z_2) \\ \underline{f}_2(z_1, z_2) \end{bmatrix} = \begin{bmatrix} -D & \gamma DP\Pi_{\pi_{Q^*}} \\ \delta I & -\delta I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} DR \\ 0 \end{bmatrix}, \end{aligned}$$

and consider the systems

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} y_{t,1} \\ y_{t,2} \end{bmatrix} &= \begin{bmatrix} \overline{f}_1(y_{t,1}, y_{t,2}) \\ \overline{f}_2(y_{t,1}, y_{t,2}) \end{bmatrix}, \quad y_0 = \begin{bmatrix} Q_0^A - Q^* \\ Q_0^B - Q^* \end{bmatrix}, \\ \frac{d}{dt} \begin{bmatrix} z_{t,1} \\ z_{t,2} \end{bmatrix} &= \begin{bmatrix} \underline{f}_1(z_{t,1}, z_{t,2}) \\ \underline{f}_2(z_{t,1}, z_{t,2}) \end{bmatrix}, \quad z_0 < \begin{bmatrix} Q_0^A - Q^* \\ Q_0^B - Q^* \end{bmatrix}, \end{aligned}$$

for all $t \geq 0$. Similar to the upper comparison systems, we can easily prove that $\overline{f}$ is quasi-monotone increasing, $\overline{f}$ and $\underline{f}$ are Lipschitz continuous. Therefore, applying similar steps as before and using Lemma 2, we have that $\begin{bmatrix} Q_t^A - Q^* \\ Q_t^B - Q^* \end{bmatrix} \geq \begin{bmatrix} Q_t^{A,l} - Q^* \\ Q_t^{B,l} - Q^* \end{bmatrix}$ holds for all $t \geq 0$, where $\begin{bmatrix} Q_t^{A,l} - Q^* \\ Q_t^{B,l} - Q^* \end{bmatrix}$ is the solution of the linear system

Now, it remains to prove the asymptotic convergence of the comparison systems. For notational convenience, we define $\Pi_\sigma, \sigma \in \mathcal{M}$ as $\Pi_{\pi_{Q_t^B}}$ such that $\sigma = \psi(\pi_{Q_t^B})$. Then, for the upper comparison switching system, we apply Lemma 3 with $A_\sigma = \begin{bmatrix} -D & \gamma DP\Pi_\sigma \\ \delta I & -\delta I \end{bmatrix}$ and $L = \begin{bmatrix} I & 0 \\ 0 & \gamma^{1/2} I \end{bmatrix}$, which satisfies $LA_\sigma = \bar{A}_\sigma L$ with $\bar{A}_\sigma = \begin{bmatrix} -D & \gamma^{1/2} DP\Pi_\sigma \\ \gamma^{1/2}\delta I & -\delta I \end{bmatrix}$. To check the strictly negative row

dominating diagonal condition, for $i \in \{1, 2, \ldots, |\mathcal{S}||\mathcal{A}|\}$, we have

$$\begin{aligned}[\bar{A}_\sigma]_{ii} + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[\bar{A}_\sigma]_{ij}| =& [-D]_{ii} + \gamma^{1/2}[-D]_{ii} \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[P\Pi_\sigma]_{ij}| \\ \leq& [-D]_{ii} + \gamma^{1/2}[-D]_{ii} \\ \leq& -1 + \gamma^{1/2} < 0.\end{aligned}$$

For $i \in \{|\mathcal{S}||\mathcal{A}| + 1, |\mathcal{S}||\mathcal{A}| + 2, \ldots, 2|\mathcal{S}||\mathcal{A}|\}$, we also have

$$[\bar{A}_\sigma]_{ii} + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[\bar{A}_\sigma]_{ij}| = -\delta + \delta\gamma^{1/2} = \delta(-1 + \gamma^{1/2}) < 0$$

for any $\delta > 0$. Therefore, the strictly negative row dominating diagonal condition is satisfied. By Lemma 3, the origin of the switching system (15) is globally asymptotically stable. The lower comparison system's stability can be proved in an equivalent way. Since the switching system's solution is upper and lower bounded by the corresponding comparison systems, it asymptotically converges to the origin. This completes the proof. □

G Proof of Theorem 5

Proof. By Assumption 2, it holds that $\gamma\Phi^T DP(\Pi_{\pi_{\Phi\theta_t^l}} - \Pi_{\pi_{\Phi\theta^*}})\Phi\theta^* \leq 0$, $\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta_t^u}} - D)\Phi(\theta_t^u - \theta^*) \leq \Phi^T(\gamma DP\Pi_{\pi_{\Phi(\theta_t^u - \theta^*)}} - D)\Phi(\theta_t^u - \theta^*)$, and we obtain the upper comparison system

$$\begin{aligned}\frac{d}{dt}(\theta_t^u - \theta^*) =& \Phi^T(\gamma DP\Pi_{\pi_{\Phi(\theta_t^u - \theta^*)}} - D)\Phi(\theta_t^u - \theta^*), \\ \theta_0^u - \theta^* >& \theta_0 - \theta^* \in \mathbb{R}^n.\end{aligned} \tag{22}$$

To proceed, define the vector functions

$$\begin{aligned}\overline{f}(y) =& \Phi^T(\gamma DP\Pi_{\pi_{\Phi y}} - D)\Phi y \\ \underline{f}(z) =& (\gamma\Phi^T DP\Pi_{\pi_{\Phi(z+\theta^*)}}\Phi - \Phi^T D\Phi)z + \gamma\Phi^T DP(\Pi_{\pi_{\Phi(z+\theta^*)}} - \Pi_{\pi_{\Phi\theta^*}})\Phi\theta^*, ,\end{aligned}$$

and consider the systems

$$\begin{aligned}\frac{d}{dt}y_t &= \overline{f}(y_t), \quad y_0 > \theta_0 - \theta^*, \\ \frac{d}{dt}z_t &= \underline{f}(z_t), \quad z_0 = \theta_0 - \theta^*,\end{aligned}$$

for all $t \geq 0$. To apply Lemma 3, we first check the quasi-monotonicity of $\overline{f}$. For any nonnegative vector p such that its ith element is zero, we have

$$\begin{aligned}e_i^T\overline{f}(y+p) =& e_i^T(\gamma\Phi^T DP\Pi_{\Phi(y+p)}\Phi - \Phi^T D\Phi)(y+p) \\ =& \gamma e_i^T\Phi^T DP\Pi_{\Phi(y+p)}\Phi(y+p) - e_i^T\Phi^T D\Phi p - e_i^T\Phi^T D\Phi y \\ =& \gamma e_i^T\Phi^T DP\Pi_{\Phi(y+p)}\Phi(y+p) - e_i^T\Phi^T D\Phi y \\ =& \gamma e_i^T\Phi^T DP\begin{bmatrix}\max_a(\Phi(y+p))_a(1) \\ \max_a(\Phi(y+p))_a(2) \\ \vdots \\ \max_a(\Phi(y+p))_a(|\mathcal{S}|)\end{bmatrix} - e_i^T\Phi^T D\Phi y \\ \geq& \gamma e_i^T\Phi^T DP\begin{bmatrix}\max_a(\Phi(y))_a(1) \\ \max_a(\Phi(y))_a(2) \\ \vdots \\ \max_a(\Phi(y))_a(|\mathcal{S}|)\end{bmatrix} - e_i^T\Phi^T D\Phi y \\ =& \gamma e_i^T\Phi^T DP\Pi_{\Phi(y)}\Phi(y) - e_i^T\Phi^T D\Phi y \\ =& e_i^T\overline{f}(y),\end{aligned}$$

where the third line is due to Assumption 3 and the fact that $\Phi^T D\Phi$ is a diagonal matrix. Therefore, $\overline{f}$ is quasi-monotone increasing. The Lipschitz continuity of $\overline{f}$ and $\underline{f}$ can be provided following similar

lines of the proof of Proposition 1, where we can use the fact that $f(z) = (\gamma DP\Pi_{\pi_{(z+Q^*)}} - D)(z + Q^*) + DR$. Therefore, the vector comparison principle, Lemma 2, leads to $\theta_t \leq \theta_t^u$ as $t \to \infty$.

On the other hand, by Assumption 2, it holds that $\gamma\Phi^T DP\Pi_{\pi_{\Phi\theta_t}}\Phi\theta_t \geq \gamma\Phi^T DP\Pi_{\pi_{\Phi\theta^*}}\Phi\theta_t$, and we obtain the lower comparison system

$$\begin{aligned}\frac{d}{dt}(\theta_t^l - \theta^*) =&\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta^*}} - D)\Phi(\theta_t^l - \theta^*),\\ &\theta_0^l - \theta^* <\theta_0 - \theta^* \in \mathbb{R}^n,\end{aligned}$$

or equivalently,

$$\begin{aligned}\frac{d}{dt}\theta_t^l =&\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta^*}} - D)\Phi\theta_t^l - \Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta^*}} - D)\Phi\theta^*,\\ \theta_0^l <&\theta_0 \in \mathbb{R}^n.\end{aligned}$$

To proceed, define the vector functions

$$\begin{aligned}\overline{f}(y) =&\Phi^T(\gamma DP\Pi_{\pi_{\Phi y}} - D)\Phi y + \Phi^T DR\\ \underline{f}(z) =&\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta^*}} - D)\Phi z - \Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta^*}} - D)\Phi\theta^*,\end{aligned}$$

and consider the systems

$$\begin{aligned}\frac{d}{dt}y_t = \overline{f}(y_t), \quad y_0 = \theta_0,\\ \frac{d}{dt}z_t = \underline{f}(z_t), \quad z_0 < \theta_0,\end{aligned}$$

for all $t \geq 0$. To apply Lemma 3, we check the quasi-monotonicity of $\overline{f}$, which can be easily proved following the steps for the upper comparison system. The Lipschitz continuity of $\overline{f}$ and $\underline{f}$ can be also proved following similar lines of the proof of Proposition 1. Therefore, Lemma 2 leads to $\theta_t \geq \theta_t^l$ as $t \to \infty$.

To prove the asymptotic stability of the original system (18), it is sufficient to prove that the upper and lower comparison systems are globally asymptotically stable. In this respect, we can apply Lemma 3 to obtain a sufficient condition for the stability. In particular, both the upper and lower comparison systems are globally asymptotically stable if the switching system is globally asymptotically stable

$$\frac{d}{dt}\theta_t = A_{\sigma_t}\theta_t,$$

under arbitrary switchings, σ_t, where $A_{\psi(\pi)} = \Phi^T(\gamma DP\Pi_{\psi(\pi)} - D)\Phi$ for all $\pi \in \Theta_\Phi$. By Lemma 3, it is true if and only if

$$\begin{aligned}&[A_{\psi(\pi)}]_{ii} + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[A_{\psi(\pi)}]_{ij}|\\ =&[\Phi^T(\gamma DP\Pi_{\psi(\pi)} - D)\Phi]_{ii} + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} |[\Phi^T(\gamma DP\Pi_{\psi(\pi)} - D)\Phi]_{ij}|\\ =&\phi_i^T(\gamma DP\Pi_{\psi(\pi)})\phi_i - \phi_i^T D\phi_i + \sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} \phi_i^T(\gamma DP\Pi_{\psi(\pi)} - D)\phi_j\\ =&-\phi_i^T D\phi_i + \sum_{j\in\{1,2,\ldots,n\}} \phi_i^T\gamma DP\Pi_{\psi(\pi)}\phi_j\\ =&-\phi_i^T D\phi_i + \phi_i^T\gamma DP\Pi_{\psi(\pi)}\sum_{j\in\{1,2,\ldots,n\}} \phi_j\\ <&0\end{aligned}$$

for all $i \in \{1, 2, \ldots, n\}, \pi \in \Theta_\Phi$, where the second line is due to Assumption 2, and the fourth line is due to Assumption 3 and the fact that $\phi_i^T D\phi_j = 0$ for $j \neq i$. This completes the proof. □

H Proof of Proposition 3

Proof. If the elements of the feature matrix Φ are binary numbers, then since the columns of Φ consist of sums of distinct basis vectors, $e_i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and it follows that

$$\sum_{j\in\{1,2,\dots,n\}} \phi_j \leq \mathbf{1}_{|\mathcal{S}||\mathcal{A}|}, \tag{23}$$

where $\mathbf{1}_{|\mathcal{S}||\mathcal{A}|}$ is the vector with all elements being ones. The right-hand side of the condition in Theorem 6 is bounded as

$$\begin{aligned}
[A_{\psi(\pi)}]_{ii} + \sum_{j\in\{1,2,\dots,n\}\setminus\{i\}} |[A_{\psi(\pi)}]_{ij}| =& -\phi_i^T D\phi_i + \phi_i^T \gamma DP\Pi_{\psi(\pi)} \sum_{j\in\{1,2,\dots,n\}} \phi_j \\
\leq& -\phi_i^T D\phi_i + \phi_i^T \gamma DP\Pi_{\psi(\pi)} \mathbf{1}_{|\mathcal{S}||\mathcal{A}|} \\
=& -\phi_i^T D\phi_i + \gamma\phi_i^T D\mathbf{1}_{|\mathcal{S}||\mathcal{A}|} \\
=& -\phi_i^T D\phi_i + \gamma\phi_i^T D\phi_i \\
=& (\gamma-1)\phi_i^T D\phi_i \\
<& 0,
\end{aligned}$$

where the first line comes from Theorem 6, the second line is due to (23), and the third line is due to the fact that $P\Pi_{\psi(\pi)}$ is a stochastic matrix, i.e., its low sums are one. This completes the proof. □

I Proof of Proposition 2

Proof. The basic idea of the proof relies on the fact that Melo's sufficient condition ensures the existence of a quadratic Lyapunov function for the upper comparison system (22) following the results in [24]. Since the new sufficient condition, Proposition 2, is a necessary and sufficient condition for the global asymptotic stability of the upper comparison system, the Melo's condition implies the proposed new condition. Suppose that Melo's sufficient condition holds, and consider the quadratic Lyapunov function candidate: $V(\theta_t - \theta^*) := \frac{1}{2}(\theta_t - \theta^*)^T(\theta_t - \theta^*)$.

Its time derivative along the state trajectories of the upper comparison system (22) is given by

$$\begin{aligned}
&\frac{d}{dt}V(\theta_t - \theta^*) = (\theta_t - \theta^*)^T\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta_t}} - D)\Phi(\theta_t - \theta^*) \\
=&\gamma(\theta_t - \theta^*)^T\Phi^T DP\Pi_{\pi_{\Phi\theta_t}}\Phi(\theta_t - \theta^*) - (\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) \\
=& -(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) + \gamma\mathbb{E}[(\theta_t - \theta^*)^T\Phi^T(e_a \otimes e_s)(e_{s'})^T\Pi_{\pi_{\Phi\theta_t}}\Phi(\theta_t - \theta^*)\},
\end{aligned}$$

where (s, a) is sampled from the stationary state-action distribution and $s' \sim P_a(s, \cdot)$. Similar to the ideas in [24], using Holder's inequality leads to

$$\begin{aligned}
&\frac{d}{dt}V(\theta_t - \theta^*) \\
=&(\theta_t - \theta^*)^T\Phi^T(\gamma DP\Pi_{\pi_{\Phi\theta_t}} - D)\Phi(\theta_t - \theta^*) \\
\leq& -(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) + \gamma\sqrt{\mathbb{E}[(\theta_t - \theta^*)^T\Phi^T(e_a \otimes e_s)(e_a \otimes e_s)^T\Phi(\theta_t - \theta^*)]} \\
&\times\sqrt{\mathbb{E}[(\theta_t - \theta^*)^T\Phi^T\Pi^T_{\pi_{\Phi\theta_t}}(e_{s'})(e_{s'})^T\Pi_{\pi_{\Phi\theta_t}}\Phi(\theta_t - \theta^*)]} \\
=& -(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) + \gamma\sqrt{(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*)} \times\sqrt{(\theta_t - \theta^*)^T\Phi^T\Pi^T_{\pi_{\Phi\theta_t}}D^\beta\Pi_{\pi_{\Phi\theta_t}}\Phi(\theta_t - \theta^*)},
\end{aligned}$$

where the last equality uses the fact that the distribution of s' is identical to the distribution of s. Now, we apply the Melo's condition to have

$$\begin{aligned}
&\frac{d}{dt}V(\theta_t - \theta^*) \\
<& -(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) + \gamma\sqrt{(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*)}\sqrt{\frac{1}{\gamma^2}(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*)} \\
=& -(\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) + (\theta_t - \theta^*)^T\Phi^T D\Phi(\theta_t - \theta^*) \\
=&0, \quad \forall\theta_t - \theta^* \neq 0.
\end{aligned}$$

This implies that V is a Lyapunov function. By the standard Lyapunov theorem, the origin of the upper comparison system (22) is globally asymptotically stable. The proof holds even if the upper comparison system is arbitrarily switching. Since the new sufficient condition in Proposition 2 is a necessary and sufficient condition for the global asymptotic stability of the upper comparison system (22) under arbitrary switching, this implies that the proposed new condition holds. □