Analysis of Q-Learning: Switching System Approach

Donghwan Lee
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
donghwan@illinois.edu

Niao He
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
niaohe@illinois.edu

Abstract

Q-learning is known to be one of the most popular reinforcement learning algorithms to find an optimal policy for an unknown Markov decision process. In this paper, we introduce a new asymptotic convergence analysis of Q-learning based on switching system perspectives and theories. The approach provides a unified viewpoint and greatly simplifies the analysis for a large family of Q-learning algorithms.

1 Introduction

Q-learning, originally introduced by Watkin in [13], is one of the most popular and fundamental reinforcement learning (RL) algorithms for finding the optimal policy of unknown Markov decision processes. There exist few approaches that prove the asymptotic convergence of Q-learning: the original proof [13],the stochastic approximation and contraction mapping-based approach [5, 12], and the stochastic approximation and ODE (ordinary differential equation) approach [2].

The ODE approach analyzes the convergence of general stochastic recursions by examining stability of the associated ODE model [1, 7, 2] and has been used as a convenient analysis tool to prove convergence of many RL algorithms. 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 [11] for policy evaluation is linear, whose asymptotic stability is easier to analyze in general. While [2] 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 [3] by using the max-norm contraction of the Bellman operator. Moreover, the stability analysis does not immediately extend to other Q-learning variants, such as double Q-learning [4] and averaging Q-learning [8].

In this paper, we study a simple and unified framework to analyze Q-learning through switching linear system (SLS) models [9] of the associated ODE. SLSs are an important class of nonlinear hybrid systems, where the system dynamics matrix varies within a finite set of subsystem matrices (or modes) according to a switching signal. The study of SLSs has attracted much attention in the past (see [10] and [9] for comprehensive study and surveys). We show that a nonlinear ODE model associated with Q-learning can be formulated as an SLS, and analyze its asymptotic stability by leveraging particular structure of Q-learning, switching system theories [10, 9], and nonlinear control theories [6]. This switching system approach can be easily extended to other Q-learning variants, such as double Q-learning [4], double Q-learning with regularization terms, averaging Q-learning [8],

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

and Q-learning with linear function approximation. Due to page limits, we only focus on the analysis of the standard Q-learning algorithm here.

2 Preliminaries

2.1 Markov decision problem

In this paper, we consider the infinite-horizon (discounted) Markov decision problem (MDP), where the agent sequentially takes actions to maximize cumulative discounted rewards. In a Markov decision process with the state-space $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ and action-space $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, the decision maker selects an action $a \in \mathcal{A}$ with the current state s, then the state transits to s' with probability $P_a(s, s')$, and the transition incurs a random reward $r_a(s, s')$, where $P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, $a \in \mathcal{A}$, $P_a(s, s')$ is the state transition probability from the current state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ under action $a \in \mathcal{A}$, and $r_a(s, s')$ is the reward random variable conditioned on $a \in \mathcal{A}, s, s' \in \mathcal{S}$ with its expectation $\mathbb{E}[r_a(s, s')|s, a, s'] = R_a(s, s')$. 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 Markov decision problem (MDP) is to find a deterministic optimal policy, π^*, such that the cumulative discounted rewards over infinite time horizons is maximized, i.e.,

$$\pi^* := \mathop{\arg\max}_{\pi\in\Theta} \mathbb{E}\left[\left.\sum_{k=0}^{\infty} \alpha^k r_{a_k}(s_k, s_{k+1})\right| \pi\right],$$

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, and $\mathbb{E}[\cdot|\cdot, \pi]$ is an expectation conditioned on the policy π. The Q-function under policy π is defined as

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

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. 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 [6]. 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 $t \to \infty$. Lastly, we provide the comparison principle below, which will play a central role in the analysis.

Lemma 1 (Comparison principle [6, Lemma 3.4]). *Consider the nonlinear system* (1) *and let v_t be a continuous function whose derivative satisfies the differential inequality*

$$\frac{d}{dt}v_t \le f(v_t), \quad v_0 \le x_0.$$

Then, $v_t \le x_t$ for all $t \in \mathbb{R}_+$.

2.3 Switching system theory

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

$$\frac{d}{dt}x_t = A_{\sigma_t}x_t, \quad x_0 = z \in \mathbb{R}^n, \quad t \in \mathbb{R}_+, \tag{2}$$

where $x_t \in \mathbb{R}^n$ is the state, $\sigma \in \mathcal{M} := \{1, 2, \ldots, M\}$ is called the mode, and $\sigma_t \in \mathcal{M}$ 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 $\sigma(x_t)$. To prove the global asymptotic stability of the switching system, we will use a fundamental algebraic stability condition of switching systems reported in [10].

Lemma 2 ([10, Theorem 8])**.** *The origin of the linear switching system (2) 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 satisfying*

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

where $[\cdot]_{ij}$ is the (i,j)-element of a matrix $(\cdot)$, such that the matrix relations

$$L A_\sigma = \bar{A}_\sigma L, \quad \forall \sigma \in \mathcal{M},$$

are satisfied.

More comprehensive surveys and study of stability of switching systems can be found in [10] and [9].

2.4 ODE-based stochastic approximation

Due to its simplicity, the convergence analysis of many RL algorithms rely on the ODE (ordinary differential equation) approach [1, 7]. 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 [2]. We now briefly introduce the Borkar and Meyn's ODE approach [2] for analyzing convergence of the general stochastic recursions

$$\theta_{k+1} = \theta_k + \alpha_k (f(\theta_k) + \varepsilon_{k+1}) \tag{3}$$

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*

$$\lim_{c\to\infty} \frac{f(cx)}{c} = f_\infty(x), \quad \forall x \in \mathbb{R}^n.$$

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

The Borkar and Meyn theorem states that under Assumption 1, the stochastic process, $(\theta_k)_{k=0}^\infty$, generated by (3) is bounded and converges to θ^e with probability one.

Lemma 3 (Borkar and Meyn theorem [2])**.** *Suppose that Assumption 1 holds. 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 Revisit Q-learning

In this section, we briefly review the standard Q-learning [13] and introduce an additional assumption adopted in this paper.

The standard Q-learning [13] updates

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

where $0 \leq \alpha_k(s,a) \leq 1$ is called the learning rate associated with the state-action pair (s,a) at iteration k. This value is assumed to be zero if $(s,a) \neq (s_k, a_k)$. 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. Note that the state-action can be visited arbitrarily, which is more general than stochastic visiting rules.

To analyze the convergence based on the switching system model, we consider the stronger assumption that $\{(s_k, a_k)\}_{k=0}^{\infty}$ is a sequence of i.i.d. random variables with a fixed underlying probability distribution, $d_a(s), s \in \mathcal{S}, a \in \mathcal{A}$, of the state and action pair (s,a). This assumption is common in the ODE approaches for Q-learning and TD-learning [11]. Moreover, this assumption can be relaxed by considering a time-varying distribution. However, this direction is not addressed in this paper to simplify the presentation of the proofs.

Throughout the paper, we assume that

Assumption 2. *$d_a(s) > 0$ holds for all $s \in \mathcal{S}, a \in \mathcal{A}$.*

Under this assumption, the modified standard Q-learning is given in Algorithm 1. Compared to the original version, the step-size α_k does not depend on the state-action pair in this version. With a suitable choice on the step-size, Algorithm 1 converges to the optimal Q^* with probability one.

Algorithm 1 Standard Q-Learning

```
Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ randomly.
for iteration $k = 0, 1, \ldots$ do
    Sample $(s,a) \sim d_a(s)$
    Sample $s' \sim P_a(s, \cdot)$ and $r_a(s,s')$
    Update $Q_{k+1}(s,a) = Q_k(s,a) + \alpha_k\{r_a(s,s') + \gamma \max_{a \in \mathcal{A}} Q_k(s',a) - Q_k(s,a)\}$
end for
```

Theorem 1. *Assume that the step-sizes satisfy*

$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{4}$$

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

4 Analysis of Q-learning from Switching System Theory

In this section, we study a switching system-based ODE model of Q-learning and prove the convergence of Q-learning in Theorem 1 based on the switching system analysis.

We first introduce the following compact notations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|}, \quad R := \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \quad Q := \begin{bmatrix} Q_1 \\ \vdots \\ Q_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

$$D_a := \begin{bmatrix} d_a(1) & & \\ & \ddots & \\ & & d_a(|\mathcal{S}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}, \quad D := \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|},$$

where $Q_a = Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}, a \in \mathcal{A}$ and $R_a(s) := \mathbb{E}[r_a(s, s')|s, a]$. Note that D is a nonsingular diagonal matrix with strictly positive diagonal elements. Using the notation introduced, the update in Algorithm 1 can be written as

$$Q_{k+1} = Q_k + \alpha_k \{(e_a \otimes e_s)(e_a \otimes e_s)^T R + \gamma (e_a \otimes e_s)(e_{s'})^T \max_{a \in \mathcal{A}} Q(\cdot, a) - (e_a \otimes e_s)(e_a \otimes e_s)^T Q\},$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are s-th basis vector (all components are 0 except for the s-th component which is 1) 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 $\mathcal{S}$, and the matrix

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

Denoting $\pi_Q(s) := \arg\max_{a \in \mathcal{A}} e_s^T Q_a \in \mathcal{A}$, the update can be further simplified as

$$Q_{k+1} = Q_k + \alpha_k \{DR + \gamma DP \Pi_{\pi_{Q_k}} Q_k - DQ_k + \varepsilon_{k+1}\}, \tag{5}$$

where $\varepsilon_{k+1} = (e_a \otimes e_s)(e_a \otimes e_s)^T R + \gamma (e_a \otimes e_s)(e_{s'})^T \Pi_{\pi_{Q_k}} Q_k - (e_a \otimes e_s)(e_a \otimes e_s)^T Q_k - (DR + \gamma DP \Pi_{\pi_{Q_k}} Q_k - DQ_k)$. We note that, for any $\pi \in \Theta$, $P\Pi_\pi$ is the state-action pair transition probability matrix under the deterministic policy π. Using the Bellman equation

$$(\gamma DP \Pi_{\pi_{Q^*}} - D) Q^* + DR = 0,$$

(5) can be rewritten as

$$(Q_{k+1} - Q^*) = (Q_k - Q^*) + \alpha_k \{(\gamma DP \Pi_{\pi_{Q_k}} - D)(Q_t - Q^*) + \gamma DP (\Pi_{\pi_{Q_k}} - \Pi_{\pi_{Q^*}}) Q^* + \varepsilon_{k+1}\}. \tag{6}$$

As discussed in Section 2.4, the convergence of (6) can be analyzed by evaluating the stability of the corresponding continuous-time ODE

$$\frac{d}{dt}(Q_t - Q^*) = (\gamma DP \Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) + \gamma DP (\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}}) Q^*, \; Q_0 - Q^* = z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{7}$$

which is a 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 (7) can be represented by the affine switching system

$$\frac{d}{dt} x_t = A_{\sigma(x_t)} x_t + b_{\sigma(x_t)}, \quad x_0 = z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \tag{8}$$

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

Note that proving the global asymptotic stability of (8) without the affine term is relevantly straightforward based on existing results, e.g., [10, Theorem 8]. However, with the affine term, the proof is no longer trivial with the existing approaches in switching system theories. In what follows, we show that by exploiting the special structure of the switching system and policy associated with the Q-learning update rule, the global asymptotic stability can still be proved.

We first establish the asymptotic stability of the corresponding linear switching system.

Lemma 4. *Consider the affine switching system* (8)*. 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.

The proof follows by applying Lemma 2 with $L = I, \bar{A}_\sigma = A_\sigma$. We defer the proof to the Appendix.

We are now in position to prove the asymptotic stability of (8) associated with Q-learning.

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

Proof. The basic idea of the proof is to find systems whose trajectories lower and upper bound the trajectory of (8) by the comparison principle. Then, by proving asymptotic stability of the two comparison systems, we can prove the asymptotic stability of (8).

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 $(\gamma D_\beta P\Pi_{\pi_{Q_t}} - D)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi_{Q^*}})Q^* \leq (\gamma DP\Pi_{\pi_{Q_t}} - D)(Q_t - Q^*)$ for all $t \in \mathbb{R}_+$. By the comparison principle, Lemma 1, $Q_t - Q^* \leq Q_t^u - Q^*$ 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

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

where we consider the initial state identical to the initial state of the original SLS (8). By Lemma 4, 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, we invoke the comparison principle, Lemma 1, 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:

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

where the initial state is identical to the initial state of the original SLS (8). The origin of the above linear system is globally asymptotically stable equilibrium point by Lemma 4. 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 2. □

Based on the results, we can now apply the Borkar and Meyn theorem, Lemma 3, to prove Theorem 1. The proof follows typical routines of the ODE approaches [1], thus omitted here due to the space limit and deferred to the Appendix.

5 Conclusion

In this paper, we studied the standard Q-learning algorithm through the switching system perspective, and provided a simple proof for the asymptotic convergence of Q-learning by leveraging existing theory on the stability of linear switching systems and comparison principles. The switching system approach also provides a convenient tool for analysis of other Q-learning variants, and shed light on the underlying dynamics of RL algorithms. For future work, we would like to investigate the non-asymptotic convergence of Q-learning algorithms based on discrete-time stochastic switching system models.

References

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

[2] 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.

[3] 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.

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

[5] 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.

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

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

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

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

[10] 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.

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

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

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

Appendix

Proof of Lemma 4: We apply Lemma 2 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 $\sigma = \psi(\pi_{Q_t^B})$. Letting $n = |\mathcal{S}||\mathcal{A}|$, the property is proved by

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

$$
\begin{aligned}
=&\gamma - 1\\
<&0, \quad \forall \sigma \in \mathcal{M},
\end{aligned}
$$

which proves the global asymptotic stability. □

Proof of Theorem 1: First of all, note that the affine switching system model in (8) 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. In the following, we itemize the proofs of the statements in Assumption 1 in the same order.

1. Q-learning in (6) can be expressed as the stochastic recursion in (3) 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

$$
\begin{aligned}
\frac{f(c\theta)}{c} =& \frac{(\gamma DP\Pi_{\pi_{c\theta}} - D)c\theta + \gamma DP(\Pi_{\pi_{c\theta}} - \Pi_{\pi_{Q^*}})Q^*}{c}\\
=&(\gamma DP\Pi_{\pi_\theta} - D)\theta + \frac{\gamma DP(\Pi_{\pi_\theta} - \Pi_{\pi_{Q^*}})Q^*}{c},
\end{aligned}
$$

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 the inequalities

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

implying that f is globally Lipschitz continuous with the parameter $\|\gamma DP\|_\infty \max_{\pi\in\Theta} \|\Pi_\pi\|_\infty + \|D\|_\infty$. Therefore, the proof is completed.

2. The second statement of Assumption 1 is directly proved by **Claim** in the proof of Theorem 2.

3. The third statement of Assumption 1 is directly proved by Theorem 2.

4. Next, we prove the remaining parts. If we define $M_k := \sum_{i=0}^k \varepsilon_i$, then M_k is Martingale as

$$
\begin{aligned}
&\mathbb{E}[M_{k+1}|\mathcal{G}_k]\\
=&\mathbb{E}\left[\sum_{i=0}^{k+1} \varepsilon_i \middle| (\varepsilon_i, \theta_i)_{i=1}^k\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}[\varepsilon_{k+1}|(\varepsilon_i,\theta_i)_{i=1}^k] + \mathbb{E}\left[\sum_{i=0}^k \varepsilon_i \middle| (\varepsilon_i,\theta_i)_{i=1}^k\right] \\
=&\mathbb{E}\left[\sum_{i=0}^k \varepsilon_i \middle| (\varepsilon_i,\theta_i)_{i=1}^k\right] = \sum_{i=0}^k \varepsilon_i = M_k
\end{aligned}
$$

and ε_k is a Martingale difference sequence. Moreover, it can be easily proved that the fourth condition of Assumption 1 is satisfied. Therefore, the fourth condition is met. □