Finite-Time Analysis of Constant Step-Size Q-Learning: Switching System Approach Revisited
Donghwan Lee

Abstract—This technical note revisits the novel switching system framework in [1] for analyzing the finite-time convergence of Q-learning, where the dynamics of asynchronous Q-learning with a constant step-size is formulated as a discrete-time stochastic switching system model, and a bound on the average iteration is established based on Lyapunov functions. We improve the analysis in the previous paper by replacing the average iteration with the final iteration, which is simpler and more common in the literature. The proposed analysis relies on propagations of the autocorrelation matrix of the state instead of the Lyapunov function analysis. Moreover, we provide comparative analysis of the proposed method and existing approaches, and prove that the proposed approach improves the previous sample complexities in terms of the effective horizon. Besides, the proposed analysis offers additional insights on analysis of Q-learning and reinforcement learning, and complements existing approaches.

Index Terms—Reinforcement learning, Q-learning, switching system, convergence, finite-time analysis

I. INTRODUCTION

Reinforcement learning (RL) addresses the optimal sequential decision making problem for unknown systems through experiences [2]. Recent successes of RL algorithms outperforming humans in several challenging tasks [3]–[10] have triggered a surge of interests in RL algorithms both theoretically and experimentally. Among many others, Q-learning [11] is one of the most fundamental and popular RL algorithms, and its convergence has been extensively studied over the past decades. Classical analysis mostly focuses on asymptotic convergence [12]–[18]. Recently, substantial advances have been made in finite-time convergence analysis [1], [19]–[28], which quantifies how fast the iterations progress toward the solution. Most of the existing results treat the Q-learning dynamics as nonlinear stochastic approximations [29], and use the contraction property of the Bellman equation. Recently, [1] proposed a new perspective of Q-learning based on discrete-time switching system models [30], [31], and established a finite-time analysis based on tools in control theory [32], [33]. The switching system perspective captures unique features of Q-learning dynamics, and allows us to convert the notion of finite-time convergence analysis into the stability analysis of dynamic control systems.

A. Contribution

The main goal of this technical note is to revisit the switching system approach in [1] to finite-time analysis of Q-learning with constant step-sizes for additional insights and complementary analysis. In particular, the analysis in [1] offers finite-time bounds on the average iteration, which requires additional steps to obtain the current estimation. We improve the analysis in [1] by replacing the average iteration with the final iteration, and derive the following bound:

$$E[\|Q_k - Q^*\|_2] \leq \mathcal{O}\left(\frac{d_{\text{min}}^{1/2}|S \times A|^{3/2}}{d_{\text{min}}^{1/2}(1-\gamma)^{3/2}}\right) + \mathcal{O}\left(\frac{|S \times A|^{3/2}}{1-\gamma - \rho}k\right),$$

(1)

which is more commonly addressed in the literature, where $|S \times A|$ is the number of the state-action pairs, $\gamma$ is the discount factor, $d_{\text{min}}$ is the minimum state-action occupation frequency, $\alpha \in (0, 1)$ is the constant step-size, $Q^*$ is a vectorized optimal Q-function, $Q_k$ is the estimation at time $k$, and $\rho := 1 - \alpha d_{\text{min}}(1 - \gamma) \in (0, 1)$. Moreover, the sample complexity, $\tilde{O}\left(\frac{d_{\text{min}}^{1/2}|S \times A|}{(1-\gamma)^{3/2}}\right)$, in [1] for $\varepsilon$-optimal solution can be improved to $\tilde{O}\left(\frac{|S \times A|^3}{\varepsilon^2 d_{\text{min}}(1-\gamma)^2}\right)$ from the proposed approach, where $d_{\text{max}}$ is the maximum state-action occupation frequency, and $\tilde{O}$ ignores the constant and polylogarithmic factors. Compared to other existing approaches in the literature, advantages of the proposed results are summarized as follows:

• The proposed sample complexity improves the existing results in terms of the effective horizon $\frac{1}{1-\gamma}$.

• The proposed finite-time error bound (1) is with respect to the Euclidean norm $\|\cdot\|_2$ instead of the $\infty$-norm, $\|\cdot\|_\infty$, which satisfies $\|\cdot\|_2 \geq \|\cdot\|_\infty$, while most existing bounds are with respect to the $\infty$-norm. Therefore, the proposed bound in (1) is more general.

• The proposed analysis considers constant step-sizes within the interval $(0, 1)$, which is, to the author’s knowledge, one of the most relaxed step-size conditions among the existing results.

The proposed analysis relies on propagations of the autocorrelation matrix, which is new in the literature, instead of the Lyapunov function analysis used in [1]. This approach allows conceptually simple and intuitive analysis with few lines of algebraic manipulations. Besides, the proposed analysis provides additional insights on Q-learning via connections with discrete-time switching systems, and can potentially present a new template for finite-time analysis of more general RL algorithms. Moreover, the new perspective can potentially stimulate synergy between control theory and RL, and open up opportunities to the design of new RL algorithms. We note that this technical note only covers an i.i.d. observation model with a constant step-size for simplicity of the overall analysis. Extensions to more complicated scenarios are not the main purpose of this technical note. Finally, we view our analysis technique as a complement rather than a replacement for existing techniques for Q-learning analysis.

B. Related Works

Recently, some progresses have been made in finite-time analysis of Q-learning algorithms [1], [19]–[28]. In particular, [19] provided a finite-time convergence rate with state-action dependent diminishing step-sizes. The authors in [20] analyzed a batch version of synchronous Q-learning, called phased Q-learning, with finite-time bounds. The authors of [21] developed convergence rates for both synchronous and asynchronous Q-learning with polynomial and linear step-sizes. [22] proposed a variant of synchronous Q-learning called speedy Q-learning by adding a momentum term, and obtained an accelerated learning rate. A finite-time analysis of asynchronous Q-learning with constant step-sizes was considered in [23]. Afterwards, many advances have been made recently in finite-time analysis. [28] developed the so-called periodic Q-learning mimicking the stochastic gradient-based training scheme in [3] with periodic target
updates. [24] provided finite-time bounds for general synchronous stochastic approximation, and applied it to a synchronous Q-learning with state-independent diminishing step-sizes. In [25], a finite-time convergence rate of general asynchronous stochastic approximation scheme was derived, and it was applied to asynchronous Q-learning with diminishing step-sizes. Subsequently, [26] obtained sharper bounds under constant step-sizes, [27] provided a Lyapunov method-based analysis for general stochastic approximations and Q-learning with both constant and diminishing step-sizes, and [1] proposed a switching system perspective of Q-learning, and established a finite-time analysis.

As for results with constant step-sizes [23], [26], [27], which are our main focus, [23] gives \( \tilde{O}\left(\frac{1}{\epsilon^2(1-\gamma)}\right) \) sample complexity under properly chosen parameters (detailed later), where \( t_{\text{cover}} \) is the cover time [23], [26] offers \( \tilde{O}\left(\frac{1}{\epsilon^2(1-\gamma)}\right) \) sample complexity, [27] provides \( \tilde{O}\left(\frac{1}{\epsilon^2(1-\gamma)}\right) \) sample complexity, which is most efficient in terms of the effective horizon \( 1/(1-\gamma) \) as pointed out in the previous subsections. Besides, this technical note provides new and simple viewpoints. We view Q-learning update as a discrete-time switching system dynamics with stochastic noises. The proposed error bounds have different features compared to the previous works, and cover different cases detailed throughout this technical note.

II. PRELIMINARIES

A. Markov decision problem

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

A deterministic policy, \( \pi : S \rightarrow A \), maps a state \( s \in S \) to an action \( \pi(s) \in A \). The objective of the Markov decision problem (MDP) is to find a deterministic optimal policy, \( \pi^* \), such that the cumulative discounted rewards over infinite time horizons is maximized, i.e.,

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

where \( \gamma \in [0, 1) \) is the discount factor, \( \Theta \) 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 \( \pi \), and \( \mathbb{E}[\cdot|\pi] \) is an expectation conditioned on the policy \( \pi \). The Q-function under policy \( \pi \) is defined as

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

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

Algorithm 1 Q-Learning with a constant step-size

1: Initialize \( Q_0 \in \mathbb{R}^{|S| \times |A|} \) randomly such that \( \|Q_0\|_{\infty} \leq 1 \).
2: Sample \( s_0 \sim p \).
3: for iteration \( k = 0, 1, \ldots \) do
4: Sample \( a_k \sim \beta(\cdot | s_k \rangle) \) and \( s_k \sim p(\cdot) \).
5: Sample \( s_k' \sim P(\cdot | s_k, a_k, \cdot) \) and \( r_k = r(s_k, a_k, s_k') \).
6: Update \( Q_{k+1}(s_k, a_k) := Q_k(s_k, a_k) + \alpha (r_k + \gamma \max_{a \in A} Q_k(s_k', a) - Q_k(s_k, a_k)) \).
7: end for

B. Switching system

Since the switching system [30], [31] is a special form of nonlinear systems [33], we first consider the nonlinear system

\[
x_{k+1} = f(x_k), \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{1, 2, \ldots\},
\]

where \( x_k \in \mathbb{R}^n \) is the state and \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a nonlinear mapping. An important concept in dealing with the nonlinear system is the equilibrium point. A point \( x = x^* \) in the state-space is said to be an equilibrium point of (2) if it has the property that whenever the state of the system starts at \( x^* \), it will remain at \( x^* \) [33]. For (2), the equilibrium points are the real roots of the equation \( f(x) = 0 \). The equilibrium point \( x^* \) is said to be globally asymptotically stable if for any initial state \( x_0 \in \mathbb{R}^n, x_k \rightarrow x^* \) as \( k \rightarrow \infty \).

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

\[
x_{k+1} = A_0 x_k + b_{x_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},
\]

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

\[
x_{k+1} = A_0 x_k + b_{x_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},
\]

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

C. Assumptions and definitions

In this technical note, we focus on the standard Q-learning algorithm in Algorithm 1 with a constant step-size \( \alpha \in (0, 1) \) under the following setting: \{\( (s_k, a_k) \}_{k=0}^{\infty} \} is an i.i.d. samples under a behavior policy \( \beta \), where the behavior policy is the policy by which the RL agent actually behaves to collect experiences. For simplicity, we assume that the state at each time is sampled from the stationary state distribution \( p \), and in this case, the state-action distribution at each time is identically given by

\[
d(s, a) = p(s) \beta(a|s), \quad (s, a) \in S \times A.\]

Throughout, we make the following standard assumptions.

Assumption 1. \( d(s, a) > 0 \) holds for all \( s \in S, a \in A \).

Assumption 2. The step-size is a constant \( \alpha \in (0, 1) \).

Assumption 3. The reward is bounded as follows:

\[
\max_{(s, a, s') \in S \times A \times S} |r(s, a, s')| =: R_{\text{max}} \leq 1.
\]

Assumption 4. The initial iterate \( Q_0 \) satisfies \( \|Q_0\|_{\infty} \leq 1 \).
Assumption 1 guarantees that every state-action pair is visited infinitely often for sufficient exploration. This assumption is used when the state-action occupation frequency is given. It has been also considered in [26] and [27]. The work in [23] considers another exploration condition, called the cover time condition, which states that there is a certain time period, within which all the state-action pair is expected to be visited at least once. Slightly different cover time conditions have been considered in [26] for convergence rate analysis. Assumption 3 is required to ensure the boundedness of Q-learning iterates, which is applied in almost all RL algorithms. The unit bounds imposed on $R_{max}$ and $Q_{0}$ are just for simplicity of analysis. The constant step-size in Assumption 2 has been also studied in [23] and [27] using different approaches.

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

**Definition 1.**

1) **Maximum state-action occupation frequency:**

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

2) **Minimum state-action occupation frequency:**

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

3) **Exponential decay rate:**

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



Remark 1. All the assumptions are standard and widely used in the RL literature. All these assumptions will be used throughout this technical note for the convergence proofs. Assumption 1 guarantees that every state-action pair is visited infinitely often for sufficient exploration. This assumption is used when the state-action occupation frequency is given. It has been also considered in [26] and [27]. The work in [23] considers another exploration condition, called the cover time condition, which states that there is a certain time period, within which all the state-action pair is expected to be visited at least once. Slightly different cover time conditions have been considered in [26] for convergence rate analysis. Assumption 3 is required to ensure the boundedness of Q-learning iterates, which is applied in almost all RL algorithms. The unit bounds imposed on $R_{max}$ and $Q_{0}$ are just for simplicity of analysis. The constant step-size in Assumption 2 has been also studied in [23] and [27] using different approaches.

In this section, we study a discrete-time switching system model of Q-learning in Algorithm 1, and establish its finite-time convergence based on the stability analysis of switching system.

**A. Q-learning as a stochastic affine switching system**

Using the notation introduced, the update in Algorithm 1 can be rewritten as

$$Q_{k+1} = Q_k + \alpha \{DR + \gamma DP\Pi Q_k - DQ_k + w_k\}, \quad (6)$$

where

$$w_k = (e_{a_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{s_k}) (e_{s_k}^T\Pi Q_k - (DR + \gamma DP\Pi Q_k - DQ_k)),$$

$$\delta_k = r_k + \gamma (e_{s_k}^T\Pi Q_k - (e_{a_k} \otimes e_s)^T Q_k), \quad (8)$$

which is a linear switching system with an extra affine term, $\gamma DP(\Pi Q_k - \Pi Q^*)Q^*$, and stochastic noise. For any $Q \in [S][A]$, define

$$A_Q := I + \alpha (\gamma DP\Pi Q_k - DQ^*)Q^* + \alpha w_k, \quad (9)$$

where $A_Q$ and $b_Q$ switch among matrices from $\{I + \alpha (\gamma DP\Pi - D) \mid \alpha \in \Theta\}$ and vectors from $\{\gamma DP\Pi - \Pi^*\}Q^* \mid \alpha \in \Theta\}$. Therefore, the convergence of Q-learning is now reduced to analyzing the stability of the above switching system. A main obstacle in proving the stability arises from the presence of the affine and stochastic terms. Without these terms, we can easily establish the exponential stability of the corresponding deterministic switching system:

$$Q_{k+1} - Q^* = A_Q Q_k - Q^* + b_Q + \alpha w_k, \quad (10)$$

where $A_Q$ and $b_Q$ switch among matrices from $\{I + \alpha (\gamma DP\Pi - D) \mid \alpha \in \Theta\}$ and vectors from $\{\gamma DP\Pi - \Pi^*\}Q^* \mid \alpha \in \Theta\}$. Therefore, the convergence of Q-learning is now reduced to analyzing the stability of the above switching system. A main obstacle in proving the stability arises from the presence of the affine and stochastic terms. Without these terms, we can easily establish the exponential stability of the corresponding deterministic switching system.
system, under arbitrary switching policy. Specifically, we have the following result.

Proposition 1 (1). For arbitrary \( H_k \in \mathbb{R}^{[S] \times [A]} \), \( k \geq 0 \), the linear switching system

\[ Q_{k+1}^L - Q^* = A H_k (Q_k - Q^*) , \quad Q_0 - Q^* \in \mathbb{R}^{[S] \times [A]} , \]

is exponentially stable with

\[ \| Q_k - Q^* \|_{\infty} \leq \lambda^k \| Q_0 - Q^* \|_{\infty} , \quad k \geq 0 , \]

where \( \lambda \) is defined in (4).

The above result follows immediately from the key fact that \( \| A Q \|_{\infty} \leq \lambda \), which we formally state in the lemma below.

Lemma 2 (1). For any \( Q \in \mathbb{R}^{[S] \times [A]} \),

\[ \| A Q \|_{\infty} \leq \lambda . \]

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

However, because of the additional affine term and stochastic noises in the original switching system (10), it is not obvious how to directly derive its finite-time convergence. To circumvent the difficulty with the affine term, we will resort to two simpler comparison systems, whose trajectories upper and lower bound that of the original system, and can be more easily analyzed. These systems will be called the upper and lower comparison systems, which capture important behaviors of Q-learning. The upper comparison system, denoted by \( Q^U_k \), upper bounds Q-learning iterate \( Q_k \), while the lower comparison system, denoted by \( Q^L_k \), lower bounds \( Q_k \). The construction of these comparison systems is partly inspired by [17] and exploits the special structure of the Q-learning algorithm. Unlike [17], here we focus on the discrete-time domain and a finite-time analysis. To address the difficulty with the stochastic noise, we introduce a two-phase analysis: the first phase captures the noise effect of the lower comparison system, while the second phase captures the difference between the two comparison systems when noise effect vanishes.

B. Lower comparison system

Consider the stochastic linear system [1]

\[ Q_{k+1}^L - Q^* = A Q^L_k (Q_k^L - Q^*) + \alpha w_k , \quad Q_0^L - Q^* \in \mathbb{R}^{[S] \times [A]} , \]

(11)

where the stochastic noise \( w_k \) is the same as the original system (9). We call it the lower comparison system.

Proposition 2 (1). Suppose \( Q^U_k \) is \( Q^* \) \( Q_0 \), \( Q^L \), \( Q_k \), \( Q^* \) \( Q_k \), \( Q_k \), \( Q^* \)

for all \( k \geq 0 \).

\[ Q^L_k - Q^* \leq Q_k - Q^* , \]

where \( n := | S \times | A \) and \( w_k \in \mathbb{R}^n \) is a stochastic noise. The noise \( w_k \) has the zero mean, and is bounded. It is formally proved in the following lemma.

Lemma 3. We have

4) \( \mathbb{E}[w_k^T w_k] \leq \frac{9}{(1-\gamma)^2} : = W_{\text{max}} \), for all \( k \geq 0 \).

Proof. For the first statement, we take the conditional expectation on (7) to have \( \mathbb{E}[w_k | x_0] = 0 \). Taking the total expectation again with the law of total expectation leads to the first conclusion. Moreover, the conditional expectation, \( \mathbb{E}[w_k^T w_k | Q_k] \), is bounded as

\[ \mathbb{E}[w_k^T w_k | Q_k] = \mathbb{E}[(e_k \otimes e_k)^T Q_k] \leq \mathbb{E}[\delta_0^2 | Q_k] + (\mathbb{E}[D + \gamma D P Q_k | Q_k])^2 \leq \mathbb{E}[\delta_0^2 | Q_k] + (\mathbb{E}[D + \gamma D P Q_k | Q_k])^2 \leq \mathbb{E}[\delta_0^2 | Q_k] + \mathbb{E}[|w_k|^2 | Q_k]. \]

where \( \delta_k \) is defined in (8), and the last inequality comes from Assumptions 3-4, and Lemma 1. Taking the total expectation, we have the fourth result. Next, taking the square root on both sides of \( \mathbb{E}[\| w_k \|_2^2] \leq W_{\text{max}} \), one gets

\[ \mathbb{E}[\| w_k \|_\infty] \leq \mathbb{E}[\| w_k \|_2] \leq \sqrt{\mathbb{E}[\| w_k \|_2^2]} \leq \sqrt{W_{\text{max}}}, \]

where the first inequality comes from \( \| \cdot \|_\infty \leq \| \cdot \|_2 \). This completes the proof.

To proceed further, let us define the covariance of the noise

\[ \mathbb{E}[w_k w_k^T] = : W_k \leq W_{\text{max}} \geq 0 . \]

An important quantity we use in the main result is the maximum eigenvalue, \( \lambda_{\text{max}}(W_k) \), whose bound can be easily established as follows.

Lemma 4. The maximum eigenvalue of \( W_k \) is bounded as

\[ \lambda_{\text{max}}(W_k) \leq W_{\text{max}} \]

for all \( k \geq 0 \), where \( W_{\text{max}} > 0 \) is given in Lemma 3.

Proof. The proof is completed by noting \( \lambda_{\text{max}}(W_k) \leq \mathbb{E}[w_k w_k^T] = \mathbb{E}[\| w_k \|_2^2] \leq W_{\text{max}} \), where the last inequality comes from Lemma 3, and the second equality uses the fact that the trace is a linear function. This completes the proof.

As a next step, we investigate how the autocorrelation matrix, \( \mathbb{E}[x_k x_k^T] \), propagates over the time. In particular, the autocorrelation matrix is updated through the linear recursion

\[ \mathbb{E}[x_{k+1} x_{k+1}^T] = A \mathbb{E}[x_k x_k^T] + \alpha^2 W_k , \]

where \( \mathbb{E}[w_k w_k^T] = W_k \). Defining \( X_k := \mathbb{E}[x_k x_k^T], k \geq 0 \), it is equivalently written as

\[ X_{k+1} = A X_k A^T + \alpha^2 W_k , \quad \forall k \geq 0 , \]

(13)
Lemma 4

10

and using the bound in Lemma 5 where the first inequality is due to \( k \geq 0 \).

Proof. We first bound \( \lambda_{\text{max}}(X_k) \) as follows:

\[
\lambda_{\text{max}}(X_k) \leq \alpha^2 \sum_{i=0}^{k-1} \lambda_{\text{max}}(A^iW_{k-i-1}(A^T)^i) + \lambda_{\text{max}}(A^kX_0(A^T)^k)
\]

\[
\leq \alpha^2 \sup_{j \geq 0} \lambda_{\text{max}}(W_j) \sum_{i=0}^{k-1} \lambda_{\text{max}}(A^i(A^T)^i)
+ \lambda_{\text{max}}(X_0) \lambda_{\text{max}}(A^k(A^T)^k)
\]

\[
= \alpha^2 \sup_{j \geq 0} \lambda_{\text{max}}(W_j) \sum_{i=0}^{k-1} \|A^i\|_2^2 + \lambda_{\text{max}}(X_0)\|A^k\|_2^2
\]

\[
\leq \alpha^2 W_{\text{max}} n \sum_{i=0}^{k-1} \|A^i\|_2^2 + n \lambda_{\text{max}}(X_0)\|A^k\|_2^2
\]

\[
\leq \alpha^2 W_{\text{max}} n \sum_{i=0}^{k-1} \rho^2i + n \lambda_{\text{max}}(X_0)\rho^{2k}
\]

\[
\leq \alpha^2 W_{\text{max}} n \frac{1}{1 - \rho^2} + n \lambda_{\text{max}}(X_0)\rho^{2k}
\]

where the first inequality is due to \( A^iW_{k-i-1}(A^T)^i \geq 0 \) and \( A^kX_0(A^T)^k \geq 0 \), the third inequality comes from Lemma 4 and \( \|\cdot\|_2 \leq \sqrt{\|\cdot\|_\infty} \), the fourth inequality is due to Lemma 2, the sixth and last inequalities come from \( \rho \in (0, 1) \). On the other hand, since \( X_k \geq 0 \), the diagonal elements are nonnegative. Therefore, we have \( \text{tr}(X_k) \leq n \lambda_{\text{max}}(X_k) \).

Combining the last two inequalities we obtain:

\[
\text{tr}(X_k) \leq n \lambda_{\text{max}}(X_k) \leq \alpha^2 W_{\text{max}} n \frac{1}{1 - \rho^2} + n^2 \lambda_{\text{max}}(X_0)\rho^{2k}.
\]

Moreover, noting the inequality \( \lambda_{\text{max}}(X_0) \leq \text{tr}(X_0) = \text{tr}(x_0x_0^T) = \|x_0\|_2^2 \), and plugging \( \rho = 1 - \alpha d_{\text{min}}(1 - \gamma) \) into \( \rho \) in the last inequality, one gets the desired conclusion.

C. Upper comparison system

Now, consider the stochastic linear switching system [1]

\[
Q_{k+1}^U - Q^* = A_{\sigma_k}(Q_k^U - Q^*) + \alpha w_k, \quad Q_0^U = Q^* \in \mathbb{R}^{[S \times A]},
\]

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

Proposition 3 ([1]). Suppose \( Q_k^U - Q^* \geq Q_0 - Q^* \), where \( \geq \) is used as the element-wise inequality. Then,

\[
Q_k^U - Q^* \geq Q_k - Q^*,
\]

for all \( k \geq 0 \).

According to Proposition 3, the trajectory of the stochastic linear switching system in (16) bounds that of the original system (10) from above. Defining \( x_k := Q_k^U - Q^* \) and \( A_{\sigma_k} := A_{\sigma_k} \), (16) can be concisely represented as the stochastic switching linear system

\[
x_{k+1} = A_{\sigma_k} x_k + \alpha w_k, \quad x_0 = x_0, \quad \forall k \geq 0.
\]

where \( n := |S \times A| \), and \( \sigma_k := \arg \max_{a \in A} Q_k(s_k, a) \). Compared to the lower comparison system (12), which is linear, (17) is a switching system, which is much more complicated due to the dependency of \( Q_k \) and \( Q_k^U \) through the past histories. This is the main reason that [1] failed to find the final iteration convergence. Therefore, the analysis used for the upper comparison system cannot be directly applied, i.e., the autocorrelation matrix \( E[Q_k x_k^T] \) cannot be obtained by using the simple linear recursion given in (15). Still, it is possible to find a bound on the autocorrelation matrix by taking into account the overall trajectories. To proceed further, let us define the history \( F_k = (Q_k, Q_{k-1}, \ldots, Q_0) \), and consider the conditional expectations \( E[|S \times A| F_k] \), \( j \in \{0, 1, \ldots, k-1\} \), whose bounds will play an essential role in our analysis. Unfortunately, since \( w_k \) and \( Q_{k+1} \) are statistically dependent, all the results in Lemma 3 do not hold. Therefore, we need to establish some results for the conditional expectations. The following bounds are in order.

Lemma 6. We have

1) \( \|w_k\|_\infty \leq \frac{1}{1 - \gamma} \) for all \( k \geq 0 \).

2) \( \|w_k w_k^T\|_{\infty} \leq \frac{10(S \times A)^2}{(1 - \gamma)^2} \) for all \( k \geq 0 \).

Proof. Taking the \( \|\cdot\|_\infty \) norm on (7) leads to

\[
\|w_k\|_\infty \leq \|((e_a \otimes e_a) - D) r_k\|_\infty + \gamma \|((e_a \otimes e_a) (e_a)^T - D) r_k\|_\infty + \|((e_a \otimes e_a) (e_a)^T - D)\|_\infty \|Q_k\|_\infty
\]

\[
\leq 2 R_{\text{max}} + \gamma \|((e_a \otimes e_a) (e_a)^T - D)\|_\infty \|Q_k\|_\infty
\]

\[
\leq 2 R_{\text{max}} + \gamma \|Q_k\|_\infty Q_{\text{max}} + 2 Q_{\text{max}}
\]

\[
\leq 2 R_{\text{max}} + 2 \gamma Q_{\text{max}} + 2 Q_{\text{max}}
\]

\[
\leq \frac{4}{1 - \gamma}.
\]
where the last inequality comes from Assumptions 3-4, and Lemma 1, \( e_k \in \mathbb{R}^{|S|} \) is the \( s \)-th basis vector (all components are 0 except for the \( s \)-th component which is 1), and \( e_a \in \mathbb{R}^{|A|} \) is the \( a \)-th basis vector. Using this bound, we can conclude the proof by using

\[
w_k^T w_k = \|w_k\|_2^2 \leq |S \times A| \|w_k\|_2 \leq \frac{16|S \times A|}{(1-\gamma)^2}
\]

This completes the proof. \(\square\)

Note that the bounds in Lemma 6 are deterministic. Therefore, we directly obtain identical bounds for the conditional expectations as follows.

**Corollary 1.** We have

\[
E[w_j^T w_j | F_k] \leq \frac{16|S \times A|}{(1-\gamma)^2} W_{\text{max}}, \quad \forall k \geq 1, \quad 0 \leq j \leq k-1,
\]

where \( F_k = (Q_k, Q_{k-1}, \ldots, Q_0) \).

Based on Corollary 1, we can obtain the following bound as in Lemma 4.

**Lemma 7.** We have

\[
\lambda_{\text{max}}(E[w_j^T w_j | F_k]) \leq W_{\text{max}}, \quad \forall k \geq 1, \quad 0 \leq j \leq k-1,
\]

where \( F_k = (Q_k, Q_{k-1}, \ldots, Q_0) \).

The proof is similar to the proof of Lemma 4, and so omitted here. We are now ready to present the main result of this section.

**Theorem 2.** For any \( k \geq 0 \), we have

\[
E[\|Q_k^U - Q^*\|_2] \leq \frac{4A^{1/2}|S \times A|^{3/2}}{\gamma_{\text{min}}^2 (1-\gamma)^{3/2}} + |S \times A|\|Q_0^U - Q^*\|_2 \rho^k.
\]

(18)

**Proof.** From the recursion (17), we have

\[
x_k = \prod_{i=0}^{k-1} A_{x_i} x_0 + \alpha \sum_{j=0}^{k-1} \prod_{i=j+2}^{k-1} A_{x_i} w_j,
\]

where

\[
\prod_{i=0}^{k-1} A_{x_i} = A_{x_{k-1}} A_{x_{k-2}} \cdots A_{x_1} A_{x_0}.
\]

Using this expression, the matrix \( x_k^T x_k \) can be directly calculated as

\[
x_k^T x_k = \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T + \alpha \sum_{j=0}^{k-1} \left\{ \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) x_0 w_j^T \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right\} + \alpha \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) w_j x_0^T \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T + \alpha^2 \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) w_j w_j^T \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T.
\]

Defining the past history, \( F_k := (Q_{k-1}, Q_{k-2}, \ldots, Q_0, Q_k^U, Q_{k-1}^U, Q_{k-2}^U, \ldots, Q_0^U) \), the conditional expectation is

\[
E[x_k^T x_k | F_k] = E\left[ \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right| F_k]
\]

\[
+ \alpha^2 E\left[ \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) w_j w_j^T \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right| F_k] = \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T + \alpha^2 \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) E[w_j w_j^T | F_k] \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T.
\]

Taking the total expectation leads to

\[
E[x_k^T x_k] = E\left[ \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right] + \alpha^2 E\left[ \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) E[w_j w_j^T | F_k] \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right]
\]

Taking the trace on both sides yields the following chain of inequalities:

\[
\text{tr}(E[x_k^T x_k]) = E[\text{tr}(x_k^T x_k)] = E\left[ \text{tr}\left( \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right) \right]
\]

\[
+ \alpha^2 E\left[ \text{tr}\left( \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) E[w_j w_j^T | F_k] \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right) \right] \leq E \left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) E[w_j w_j^T | F_k] \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right] \leq E \left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \text{tr}(E[w_j w_j^T | F_k]) \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \lambda_{\text{max}} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \lambda_{\text{max}} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right] \leq E \left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) x_0 x_0^T \left( \prod_{i=0}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \lambda_{\text{max}} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right]
\]

\[
+ \alpha^2 E\left[ n \lambda_{\text{max}} \left( \prod_{i=0}^{k-1} A_{x_i} \right) \sum_{j=0}^{k-1} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \lambda_{\text{max}} \left( \prod_{i=j+2}^{k-1} A_{x_i} \right) \left( \prod_{i=j+2}^{k-1} A_{x_i} \right)^T \right].
\]
Theorem 3. For any $k \geq 0$, we have
\[
\mathbb{E}[\|Q_k - Q^*\|_2] \leq \frac{10\alpha^2 n |S \times A|^3/2}{d_{\min}^2 (1 - \gamma)^{3/2}} + \frac{4|S \times A|^{3/2}}{1 - \gamma} \rho^k.
\]
Proof. Setting $Q_0 = Q_k = Q_{\ell_0}$, noting (19), and using the triangle inequality, we have
\[
\mathbb{E}[\|Q_k - Q^*\|_2] = \mathbb{E}[\|Q_k - Q_k^\ell + Q_k^\ell - Q^*\|_2] \\
\leq \mathbb{E}[\|Q_k - Q_k^\ell\|_2] + \mathbb{E}[\|Q_k^\ell - Q^*\|_2] \\
\leq \mathbb{E}[\|Q_k^\ell - Q_k\|_2] + \mathbb{E}[\|Q_k^\ell - Q^*\|_2] + \mathbb{E}[\|Q_k^\ell - Q^*\|_2] \\
\leq 2\mathbb{E}[\|Q_k^\ell - Q^*\|_2] + \mathbb{E}[\|Q_k^\ell - Q_k\|_2] \\
\leq 6|S \times A|^{1/2} d_{\min}^{-1/2} (1 - \gamma)^3/2 + |S \times A| |Q_0 - Q^*|_2 |S \times A| \rho^k \\
+ \frac{4|S \times A|^{3/2}}{d_{\min}^2 (1 - \gamma)^{3/2}} + |S \times A| |Q_0 - Q^*|_2 \rho^k \\
\leq \frac{10\alpha^2 n |S \times A|^3/2}{d_{\min}^2 (1 - \gamma)^{3/2}} + 2|S \times A| |Q_0 - Q^*|_2 \rho^k,
\]
where the second inequality is due to $0 \leq Q_k - Q_k^\ell \leq Q_k^\ell - Q^\ell_k$, and the last inequality is due to Theorem 1 and Theorem 2. Next, one can derive
\[
|Q_0 - Q^*|_2 \leq n^{1/2} |Q_0 - Q^*|_\infty
\]
where the first inequality comes from $\|\cdot\|_2 \leq \sqrt{n} \|\cdot\|_\infty$, the second inequality is due to the triangle inequality, and the last inequality applies Lemma 1 and Assumption 4. Finally, combining the last two inequalities leads to the desired conclusion.

Note that in Theorem 3, the second term vanishes exponentially as $k \to \infty$, while the first term represents a bias due to the constant step-size. As $\alpha \to 0$, the first term also vanishes. Therefore, by appropriately choosing the step-size $\alpha \in (0 , 1)$, the error can be made arbitrarily small with a sufficiently large $k$ and a sufficiently small $\alpha \in (0 , 1)$. Using Markov’s inequality, a concentration bound can be easily obtained as well.

Proposition 4. For any $k \geq 0$ and $\varepsilon > 0$, we have
\[
\mathbb{P}[\|Q_k - Q^*\|_2 < \varepsilon] \leq 1 - \frac{10\alpha^2 n |S \times A|^{3/2}}{d_{\min}^2 (1 - \gamma)^{3/2}} \rho^k.
\]
Proof. Using the Markov’s inequality leads to
\[
\mathbb{P}[\|Q_k - Q^*\|_2 \geq \varepsilon] \leq \frac{10\alpha^2 n |S \times A|^{3/2}}{d_{\min}^2 (1 - \gamma)^{3/2}} \rho^k.
\]
The probability corresponding to the complement event is the desired conclusion.

Based on the expected error bound and the concentration bound, the sample complexity (or iteration complexity) can be obtained as follows.

Proposition 5 (Sample complexity). To achieve $\|Q_k - Q^*\|_2 < \varepsilon$ with probability at least $1 - \delta$, we need the number of samples at most
\[
O \left( \frac{|S \times A|^{3/2} \ln \left( \frac{|S \times A|^{3/2}}{(1 - \gamma)^{2\delta}} \right)}{\delta^2 \varepsilon^2 d_{\min}^2 (1 - \gamma)^4} \right).
\]
Moreover, to achieve $\mathbb{E}[\|Q_k - Q^*\|_2] < \varepsilon$ we need the number of samples at most
\[
O \left( \frac{|S \times A|^{3/2} \ln \left( \frac{|S \times A|^{3/2}}{(1 - \gamma)^{2\delta}} \right)}{\varepsilon^2 d_{\min}^2 (1 - \gamma)^4} \right).
\]

The proof can be found in Appendix A. Based on the finite-time error bound on the Q-learning iterates in Proposition 5, we can derive an upper bound on the sample or iteration complexity of Q-learning: To achieve $\|Q_k - Q^*\|_2 < \varepsilon$ with probability at least $1 - \delta$, we need the number of samples at most
\[
O \left( \frac{|S \times A|^3}{\delta^2 \varepsilon^2 d_{\min}^2 (1 - \gamma)^4} \right),
\]
where $\tilde{O}$ ignores the polylogarithmic factors. If the state-action pair is sampled uniformly from $S \times A$, then $d(s, a) = 1/|S|$, $\forall (s, a) \in S \times A$ and $d_{\min} = d_{\max} = 1/|S||A|$. In this case, the sample complexity becomes
\[
O \left( \frac{|S \times A|^5}{\delta^2 \varepsilon^2 (1 - \gamma)^4} \right).
\]
For the expected error bound $\mathbb{E}[\|Q_k - Q^*\|_2] < \varepsilon$, the sample complexity is
\[
O \left( \frac{|S \times A|^5}{\varepsilon^2 (1 - \gamma)^4} \right).
\]
In the next subsection, we compare sample complexities of the proposed and other approaches, and discuss advantages and disadvantages of the proposed analysis.

E. Comparative analysis

The finite-time analysis of asynchronous Q-learning with constant step-sizes was first considered in [23], has been recently studied in [26], and the concurrent work [27]. Comparisons of several approaches are summarized below.

- Based on the cover time assumption, [23] provides the bound

\[
\mathbb{E}[\|Q_k - Q^\star\|_\infty] \\
\leq (1 - \alpha(1 - \gamma))^{k/t_{\text{cover}}} Q_{\text{max}} + \frac{\gamma t_{\text{cover}}}{(1 - \gamma)} \sqrt{\frac{\alpha}{2 - \alpha}} W_{\text{max}}|S \times A| \\
+ \frac{t_{\text{cover}}(t_{\text{cover}} - 1)}{2(1 - \gamma)} \alpha(Q_{\text{max}} + \|Q^\star\|_\infty + W_{\text{max}} + \|Q^\star\|_\infty) \\
+ \sqrt{\frac{\alpha}{2 - \alpha}} W_{\text{max}}^2|S \times A|
\]

in terms of \(\infty\)-norm under the assumption \(\alpha < \frac{1}{t_{\text{cover}}}\), where \(t_{\text{cover}}\) is the cover time defined in [23, Assumpt. 1]. For a reasonably fair comparisons, letting \(R_{\text{max}} = 1, Q_{\text{max}} = \frac{1}{\gamma}, W_{\text{max}} = \frac{1}{\alpha(1 - \gamma)^2}, \|Q^\star\|_\infty \leq \frac{1}{\gamma}\) and simplifying the bounds, we have

\[
\mathbb{E}[\|Q_k - Q^\star\|_\infty] \leq \frac{(1 - \alpha(1 - \gamma))^{k/t_{\text{cover}}}}{1 - \gamma} + \frac{30t_{\text{cover}}\alpha^{1/2}}{(1 - \gamma)^3},
\]

where \([\cdot]\) denotes the floor function

Following similar lines as in the proof of Proposition 5, we get the sample complexity

\[
\tilde{O}\left(\frac{t_{\text{cover}}^3}{\varepsilon^2(1 - \gamma)^3}\right)
\]

Noting that \(t_{\text{cover}} \geq |S \times A|\), the best achievable complexity is

\[
\tilde{O}\left(\frac{|S \times A|^3}{\varepsilon^2(1 - \gamma)^3}\right)
\]

We can observe the following facts:

1. The proposed bound in (21) improves that of [23] by the factor of \((1 - \gamma)^3\).
2. The bound in (21) is with respect to the Euclidean norm \(\|\cdot\|_2\), which satisfies \(\|\cdot\|_2 \geq \|\cdot\|_\infty\), while the bound in [23] is with respect to the \(\infty\)-norm. Therefore, the proposed bound is more general.
3. The bound in (21) assumes \(\alpha \leq \frac{1}{t_{\text{cover}}^2}\), while the proposed analysis covers more general case \(\alpha \in (0, 1)\).

- The results in [26] provide \(\tilde{O}\left(\frac{1}{\min(1 - \gamma)^3 \varepsilon^2}\right)\) to achieve \(\mathbb{E}[\|Q_k - Q^\star\|_\infty] < \varepsilon\). If the state-action pair is sampled uniformly from \(S \times A\), then \(d_{\min} = \frac{1}{|S \times A|}\), and the complexity is \(\tilde{O}\left(\frac{|S \times A|^3}{(1 - \gamma)^3 \varepsilon^2}\right)\). The proposed bound in (21) improves that of [23] by the factor of \((1 - \gamma)^{-1}\), while the dependency in terms of \(|S \times A|\) is worse. Moreover, the bound in (21) is with respect to the Euclidean norm \(\|\cdot\|_2\), which satisfies \(\|\cdot\|_2 \geq \|\cdot\|_\infty\), while the bound in [26] is with respect to the \(\infty\)-norm. Therefore, the proposed bound is more general.

- The complexity \(\tilde{O}\left(\frac{1}{\min(1 - \gamma)^3 \varepsilon^2}\right)\) is given in [27] to achieve \(\mathbb{E}[\|Q_k - Q^\star\|_\infty] < \varepsilon\). If the state-action pair is sampled uniformly from \(S \times A\), then \(d_{\min} = \frac{1}{|S \times A|}\), and the complexity is

\[
\tilde{O}\left(\frac{|S \times A|^3}{(1 - \gamma)^3 \varepsilon^2}\right)
\]

The proposed bound in (21) improves that of [27] by the factor of \((1 - \gamma)^{-1}\), while the dependency in terms of \(|S \times A|\) is worse. Moreover, the bound in (21) is with respect to the Euclidean norm \(\|\cdot\|_2\), which satisfies \(\|\cdot\|_2 \geq \|\cdot\|_\infty\), while the bound in [26] is with respect to the \(\infty\)-norm. Therefore, the proposed bound is more general.

IV. Conclusion

In this technical note, we have revisited the switching system framework in [1] to analyze the finite-time convergence of Q-learning. We have improved the analysis in the previous paper by replacing the average iteration with the final iteration, which is simpler and more common in the literature. The proposed analysis relies on propagations of the autocorrelation matrix of the state, which is new, and conceptually simple and intuitive with few lines of algebraic manipulations. Moreover, we have provided comparative analysis of the proposed method and existing approaches, and proved that the proposed approach improves the previous sample complexities in terms of the effective horizon. The proposed finite-time error bounds are more general than most existing bounds for the constant step-size Q-learning in terms of the norm used for the error, and the allowable range of step-sizes. Besides, the proposed analysis potentially offers additional insights on analysis of Q-learning and reinforcement learning, and complements existing approaches. Potential future topics include finite-time analysis of SARSA, double Q-learning, and actor-critic using similar dynamic system viewpoints.

References

[1] D. Lee, J. Hu, and N. He, “A discrete-time switching system analysis of Q-learning,” arXiv preprint arXiv:2102.08583, 2021.
[2] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT Press, 1998.
[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, p. 529, 2015.
[4] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, “Dueling network architectures for deep reinforcement learning,” in International conference on machine learning, 2016, pp. 1995–2003.
[5] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in International conference on learning representations, 2016.
[6] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver, “Memory-based control with recurrent neural networks,” arXiv preprint arXiv:1512.04455, 2015.
[7] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in Proceedings of the AAAI conference on artificial intelligence, vol. 30, no. 1, 2016.
[8] M. G. Bellemare, W. Dabney, and R. Munos, “A distributional perspective on reinforcement learning,” in International Conference on Machine Learning, 2017, pp. 449–458.
[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in International conference on machine learning, 2015, pp. 1889–1897.
[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, 2017.
[11] C. J. C. H. Watkins and P. Dayan, “Q-learning,” Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.
[12] J. N. Tsitsiklis, “Asynchronous stochastic approximation and q-learning,” Machine learning, vol. 16, no. 3, pp. 185–202, 1994.
[13] T. Jaakkola, M. I. Jordan, and S. P. Singh, “Convergence of stochastic iterative dynamic programming algorithms,” in Advances in neural information processing systems, 1994, pp. 703–710.
Using the inequality $1 - \frac{1}{x} \leq \ln x \leq x - 1$ for all $x > 0$, the right-hand side of (23) is upper bounded by $\frac{\ln (\frac{4|S \times A|^{3/2}}{d^2 \min(1 - \gamma)^3})}{\alpha d \min(1 - \gamma)}$.

Therefore, (23) holds if

$$k \geq \frac{\ln (\frac{4|S \times A|^{3/2}}{d^2 \min(1 - \gamma)^3})}{\alpha d \min(1 - \gamma)}.$$  \hfill (24)

Now, consider the special case $\alpha = \frac{\delta^2 \varepsilon^2 d \min(1 - \gamma)^3}{400|S \times A|^3}$ in (22), and plug it into $\alpha$ in (24), we have

$$k \geq \frac{400|S \times A|^3 \ln \left(\frac{4|S \times A|^{3/2}}{d^2 \min(1 - \gamma)^3}\right)}{\delta^2 \varepsilon^2 d \min(1 - \gamma)^3},$$

Therefore, with the number of samples at most $O\left(\frac{|S \times A|^3 \ln \left(\frac{4|S \times A|^{3/2}}{d^2 \min(1 - \gamma)^3}\right)}{\delta^2 \varepsilon^2 d \min(1 - \gamma)^3}\right)$, we achieve $\|Q_k - Q^*\|_2 < \varepsilon$ with probability at least $1 - \delta$. The second statement can be proved similarly. This completes the proof.