Finite-Time Analysis of Asynchronous Q-Learning Under Diminishing Step-Size From Control-Theoretic View

HAN-DONG LIM\textsuperscript{1} AND DONGHWAN LEE\textsuperscript{1,2}, (Member, IEEE)

Department of Electrical Engineering, KAIST, Daejeon 34141, South Korea

Corresponding author: Donghwan Lee (donghwan@kaist.ac.kr)

This work was supported in part by the Institute of Information Communications Technology Planning Evaluation (IITP) funded by Korean Government under Grant 2022-0-00469, and in part by KT Corporation as part of the Korea Telecom (KT)-Korea Advanced Institute of Science and Technology (KAIST) Open Research and Development Project.

ABSTRACT Q-learning has long been one of the most popular reinforcement learning algorithms, and theoretical analysis of Q-learning has been an active research topic for decades. Although researches on asymptotic convergence analysis of Q-learning have a long tradition, non-asymptotic convergence has only recently come under active study. The main goal of this paper is to investigate a new finite-time analysis of asynchronous Q-learning under Markovian observation models via a control system viewpoint. In particular, we introduce a discrete-time time-varying switching system model of Q-learning with diminishing step-sizes for our analysis, which significantly improves recent development of the switching system analysis with constant step-sizes, and leads to $O(\sqrt{\log k/k})$ convergence rate that is comparable to or better than most of the state of the art results in the literature. In the meanwhile, we consider the continuous-time Lyapunov equation to avoid the difficulty in the analysis posed by using diminishing step-sizes in discrete-time. The proposed analysis brings in additional insights, covers different scenarios, and provides new simplified templates for analysis to deepen our understanding on Q-learning via its unique connection to discrete-time switching systems.

INDEX TERMS Reinforcement learning, Q-learning, convergence analysis, switching system, control theory.

I. INTRODUCTION

Recently, reinforcement learning has shown remarkable performance in complex and challenging domains. In the seminal work [1], the so-called deep Q-learning has achieved human level performance in challenging tasks such as the Atari video games using deep learning based approaches. Since then, there have been major improvements [2], [3], [4], [5] to solve the Atari benchmark problems as well. On the other hand, significant progress have been made for other various fields such as recommendation system [6], robotics [7], and portfolio management [8] to name just a few.

Developed by Watkins and Dayan in [9], Q-learning is one of the most widely known reinforcement learning algorithms. Theoretical analysis of Q-learning has been an active research topic for decades. In particular, asymptotic convergence of Q-learning has been successfully established through a series of works [10], [11], [12], [13] based on stochastic approximation theory [12]. Although the asymptotic analysis ensures the essential property, the convergence to a solution, the natural question remains unanswered: does the algorithm make a consistent and quantifiable progress toward an optimal solution? To understand this non-asymptotic behavior, meaningful advances have been made recently [14], [15], [16], [17], [18], [19], [20], [21] as well. Among the promising results, our main concern is the control-theoretic framework [16], which provides a
unique switching system [22] perspective of Q-learning. In particular, asynchronous Q-learning with a constant step-size has been naturally formulated as a stochastic discrete-time affine switching system, which has allowed us to transform the convergence analysis into a stability analysis of the switching system model from control theoretic viewpoints. This analysis provides new insights and analysis framework to deepen our understanding on Q-learning via its unique connection to discrete-time switching systems with remarkably simplified proofs. Such viewpoint allows us to follow similar spirit of optimization literature to derive a finite-time bound of the Q-learning algorithm. Nevertheless, this work left key questions unanswered, which are summarized as follows:

1) The study in [16] provides a finite-time analysis for the averaged iterate of Q-learning rather than the final-time iterate. However, the averaged iterate requires additional steps, and it can slow down the convergence in practice.

2) It only considers a constant step-size, which results in biases in the final solution. On the other hand, a diminishing step-size is more widely used in theoretical analysis to guarantee convergence to a solution without the biases.

3) The analysis in [16] adopts i.i.d. observation models. Although this assumption simplifies the overall analysis, it is too idealistic to reflect real world scenarios.

Motivated by the above discussions, the main goal of this paper is to revisit the non-asymptotic analysis of asynchronous Q-learning [16] from discrete-time switching system viewpoints. Contrary to [16], we consider diminishing step-sizes, Markovian observation models, and the final-time convergence. Using a diminishing step-size causes a unique challenge in the analysis due to its time-varying nature. Our main contributions are summarized below.

A. CONTRIBUTIONS

1) We answer the above three questions unresolved in [16]. A new finite-time analysis is proposed for asynchronous Q-learning under Markovian observation models and diminishing step-sizes, which answers the key questions that were left unsolved in [16]. The new analysis offers \( O\left(\sqrt{\frac{\log k}{k}}\right) \) convergence rate in terms of number of iterations, which is comparable to or better than most of the state of the art results [17], [18], [19] under Markovian observation models.

2) We consider the control theoretic analysis in [16] under a diminishing step-size. With the diminishing step-size, it is hard to directly apply the proof techniques in [16] due to time-varying nature of the diminishing step-size, which transfers the underlying linear switching system into a time-varying system, and it is in general much more challenging to analyze. We overcome this difficulty by considering continuous-time Lyapunov equation [23].

3) Finally, we stress that our goal is to provide new insights and analysis framework to deepen our understanding on Q-learning via its unique connection to discrete-time switching systems, rather than improving existing convergence rates. In this respect, we view our analysis technique as a complement rather than a replacement for existing techniques for Q-learning.

B. RELATED WORKS

Since Q-learning was developed by Watkins in [9], several advances have been made in a series of works. Here, we summarize related results in the literature.

1. Asymptotic convergence: Convergence of synchronous Q-learning was analyzed for the first time in the original paper [9]. Subsequently, convergence of asynchronous Q-learning was studied in [10] and [11] using results in stochastic approximation and contraction mapping property. Moreover, an ordinary differential equation (O.D.E) approach was developed in [12] to prove general stochastic approximations, and applied it to prove convergence of synchronous Q-learning. More recently, [13] developed a switching system model of asynchronous Q-learning, and applied the O.D.E method for its convergence.

2. Non-asymptotic convergence with diminishing step-sizes: The early result in [14] considered state-dependent diminishing step-sizes to study non-asymptotic convergence under independent and identically distributed (i.i.d.) observation models. There are a few works focusing on the Markovian sampling setting together with diminishing step-sizes. In particular, [15] obtained sample complexity to achieve \( \epsilon \)-optimal solution with high probabilities by extending the convergence analysis of [24]. More recently, [19] considered a general asynchronous stochastic approximation scheme featuring a weighted infinity-norm contractive operator, and proved a bound on its finite-time convergence rate under the Markovian sampling setting. They used the notion of the so-called sufficient exploration and error decomposition with shifted martingale difference sequences. Another recent work in [18] employed the so-called generalized Moreau envelope and the notion of Lyapunov function. The recent study in [17] provided a probability tail bound through refined analysis based on an adaptive step-size, which depends on the number of visits to each state-action pair. Besides, [25] proposed an optimization-based Q-learning that mimics the recent deep Q-learning in [1], and offered new finite-time convergence results.

3. Non-asymptotic convergence with constant step-sizes: The aforementioned works applied time-varying step-sizes such as diminishing or adaptive step-sizes. On the other hand, there are some works dealing with Q-learning under constant step-sizes. For
instance, [26], [27] employed the concept of the so-called covering time, \( \ell_{\text{cover}} \), to derive mean squared error bounds. Using similar approaches as in the diminishing step-size analysis [17], [18] proposed new convergence analysis under Markovian observation models. More recently, [16] developed a discrete-time switching system model of asynchronous Q-learning to compute a convergence rate of averaged iterated under i.i.d. observation model.

4. Control theoretic viewpoint of RL algorithms: Finally, it is worth summarizing related studies based on control system frameworks. Dynamical system perspectives of reinforcement learning and general stochastic iterative algorithms have a long tradition, and they date back to O.D.E analysis [12], [28], [29], [30]. Recently, [31] studied asymptotic convergence of temporal difference learning (TD-learning) [32] based on a Markovian jump linear system model. They tailored Markovian jump linear system theory developed in the control community to characterize exact behaviors of the first and second order moments of TD-learning algorithms. The paper [13] studied asymptotic convergence of Q-learning [9] through a continuous-time switched linear system model [22]. A finite-time analysis of Q-learning was also investigated in [16] using discrete-time switched linear system models.

The paper is organized as follows: Section II provides preliminary backgrounds on Markov decision process, dynamic system, Q-learning, essential notations and definitions used throughout the paper. In Section III, we introduce a dynamic system model of Q-learning. In particular, Q-learning is represented by a stochastic linear time-varying switching system with an affine input term. Section IV is devoted to analysis of the lower comparison system. We provide essential properties of the lower comparison system, and prove its convergence through a domain transformation under i.i.d. observation model. In Section V, convergence of the upper comparison system is given under i.i.d. observation model. Based on the results in the previous two sections, the convergence of the lower and upper comparison systems, Section VI provides a proof of the convergence of Q-learning under i.i.d. observation model. Section VII extends the analysis in the previous sections with i.i.d. scenario to Markovian observation case. Finally, the overall results are concluded in Section VIII.

II. PRELIMINARIES

A. NOTATIONS

The adopted notation throughout the paper is as follows: \( \mathbb{R} \): set of real numbers; \( \mathbb{R}^n \): \( n \)-dimensional Euclidean space; \( \mathbb{R}^{n \times m} \): set of all \( n \times m \) real matrices; \( \mathbb{N}_0 \): set of natural numbers and zero; \( A^\top \): transpose of matrix \( A \); \( A \succ 0 \) (\( A \prec 0 \), \( A \succeq 0 \), and \( A \preceq 0 \), respectively): symmetric positive definite (negative definite, positive semi-definite, and negative semi-definite, respectively) matrix \( A \); \( x \preceq y \) and \( x \succeq y \) for vectors \( x, y \): element-wise inequalities; \( I \): identity matrix with appropriate dimensions; \( \rho(\cdot) \): spectral radius; for any matrix \( A \), \( A_{ij} \) is the element of \( A \) in \( i \)-th row and \( j \)-th column; \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) for any symmetric matrix \( A \): the minimum and maximum eigenvalues of \( A \), respectively; \( \sigma_{\min}(A) \) and \( \sigma_{\max}(A) \): minimum and maximum singular value of matrix \( A \), respectively; \( |S| \): cardinality of a finite set \( S \).

B. MARKOV DECISION PROCESS

In this paper, we consider an infinite horizon Markov decision process (MDP). It consists of a tuple \( \mathcal{M} = (S, A, P, r, \gamma) \), where \( S \) and \( A \) represent finite state and action spaces, respectively, \( P(s'|s, a) \) denotes the transition probability when action \( a \) is taken at state \( s \), and transition occurs to \( s' \), \( r : S \times A \times S \rightarrow \mathbb{R} \) denotes the reward function, and \( \gamma \in (0, 1) \) is the discount factor. With a stochastic policy \( \pi : S \rightarrow \mathcal{P}(A) \), where \( \mathcal{P}(A) \) is the set of probability distributions over \( A \), agent at the current state \( s \) selects an action \( a \sim \pi(\cdot|s) \), then the agent’s state changes to the next state \( s' \sim P(\cdot|s, a) \), and receives reward \( r := r(s, a, s') \). For a deterministic policy \( \pi : S \rightarrow A \), \( \pi(s) \in A \) is selected without randomness at state \( s \).

The objective of the Markov decision problem (MDP) is to find an optimal policy, denoted by \( \pi^* \), such that the cumulative discounted rewards over infinite time horizons is maximized, i.e.,

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

where \( (s_0, a_0, r_1, s_1, a_1, r_1, \ldots) \) is a state-action trajectory generated by the Markov chain under policy \( \pi \), \( r_k \) denotes the reward received at time \( k \) (will be detailed later), and \( \mathbb{E}[\cdot|\pi] \) is an expectation conditioned on the policy \( \pi \).

The Q-function under policy \( \pi \) is defined as, for \( s, a \in S \times A \)

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

and the optimal Q-function is defined as \( Q^*(s, a) = Q^{\pi^*}(s, a) \) for all \( (s, a) \in S \times A \). Once \( Q^* \) is obtained, then an optimal policy can be retrieved by the greedy action with respect to \( Q^* \), i.e., \( \pi^*(s) = \arg \max_{a \in A} Q^*(s, a) \).

It is known that the optimal Q-function satisfies Bellman equation expressed as follows:

\[
Q^*(s, a) = \mathbb{E} \left[ r_{k+1} + \max_{a_{k+1} \in A} \gamma Q^*(s_{k+1}, a_{k+1}) \mid (s_k, a_k) = (s, a) \right].
\]

Finally, some notations related to MDP are summarized as follows: \( P_a \in \mathbb{R}^{|S| \times |S|}, a \in A \) is the state transition matrix when action \( a \) is taken at state \( s \in S \). Likewise, \( R_a \in \mathbb{R}^{|S|}, a \in A \) is the expected reward vector when action \( a \) is taken at state \( s \in S \), i.e., \( R_a(s) := \mathbb{E}[r(s, a, s')|s, a], s \in S \).
Moreover, the notation $Q_a := Q(\cdot, a) \in \mathbb{R}^{[S]}$, $a \in \mathcal{A}$ will be adopted throughout the paper.

**C. SWITCHING SYSTEM**

In this subsection, we briefly review the switching system, which will play an important role in this paper. The switching system is a special form of nonlinear systems, and hence, we first consider the nonlinear system

$$x_{k+1} = f(x_k), \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\}, \quad (2)$$

where $x_k \in \mathbb{R}^n$ is the state and $f: \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. An important concept in dealing with the nonlinear system is the equilibrium point. A point $x = x^*$ in the state-space 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 $x^*$ [23]. 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 $x_0 \in \mathbb{R}^n$, $x_k \to x^*$ as $k \to \infty$.

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

$$x_{k+1} = A_\sigma x_k, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\}, \quad (3)$$

where $x_k \in \mathbb{R}^n$ is the state, $\sigma \in \mathcal{M} := \{1, 2, \ldots, M\}$ is called the mode, $A_\sigma \in \mathcal{M}$ 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 $A_\sigma = \sigma(x_k)$. A more general class of systems is the affine switching system

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\}, \quad (5)$$

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

Switching system has been studied extensively in the literature. A further study on the theory of switching system can be found in [33], [34], and [35].

**D. LINEAR TIME-VARYING SYSTEM**

In this section, we briefly review the linear time-varying (LTV) system [36]. Let us first consider the continuous-time linear time-invariant (LTI) system

$$\dot{x}_t = Ax_t, \quad t \geq 0, \quad (4)$$

where $x_t \in \mathbb{R}^n$ is the state, $A \in \mathbb{R}^{n \times n}$ is the system matrix. The asymptotic stability of (4) is completely characterized by the notion called Hurwitz matrix [23], [37], whose definition is given in Appendix A. It is well-known that (4) is asymptotically stable if and only if $A$ is Hurwitz. In general, it is hard to check the Hurwitzness without computing the eigenvalues. Nevertheless, a sufficient condition can be obtained in terms of algebraic relations among entries of the corresponding matrix, which is called the strict diagonal dominance whose definition is given in Definition 31 in Appendix A.

Lemma 2: A square matrix $A \in \mathbb{R}^{n \times n}$ is Hurwitz if it is a strictly row diagonally dominant matrix and the diagonal elements are negative.

The proof of Lemma 2 is in Appendix D.

A linear time-varying (LTV) system is a class of linear systems defined as

$$x_{k+1} = A_k x_k, \quad k \in \{0, 1, \ldots\}, \quad (5)$$

where $x_k \in \mathbb{R}^n$ is the state, and $A_k \in \mathbb{R}^{n \times n}$ is the time-varying system matrix. Together with switching systems, LTV system models will be used in the analysis of this paper. Stability analysis of LTV systems is much more challenging than that of a simple linear time-invariant system [37]. For their asymptotic stability, the following sufficient condition will be used in this paper.

Lemma 3 [36]: The discrete-time LTV system (5) is asymptotically stable if there exists a sequence of positive definite matrices $\{P_k\}_{k \in \mathbb{N}_0}$, and positive scalars $\{p_k\}_{k \in \mathbb{N}_0}$ such that the following conditions hold for all $k \in \mathbb{N}_0$:

1) $A_k^T P_{k+1} A_k \leq p_k P_k$;
2) $\prod_{k=1}^n p_k \to 0$ as $n \to \infty$;
3) $c \leq \lambda_{\min}(P_k)$ for some positive constant $c > 0$.

The proof is given in Appendix E.

**E. MATRIX REPRESENTATION OF Q-LEARNING**

In this paper, a linear time-varying switching system model of Q-learning will be used for the analysis. To this end, the following matrix notations will be useful to compactly represent Q-learning into a vector and matrix forms:

$$P = \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_1 \\ \vdots \\ Q_{|\mathcal{A}|} \end{bmatrix}. \quad \quad (6)$$

Under policy $\pi$, we have $P^\pi \in \mathbb{R}^{[S]|\mathcal{A}| \times [S]|\mathcal{A}|}$ as the state-action transition matrix, i.e.,

$$(e_s \otimes e_a)^T P^\pi (e_{s'} \otimes e_{a'}) = P[s_{t+1} = s', a_{t+1} = a'|s_t = s, a_t = a, \pi].$$

where $\otimes$ stands for the Kronecker product, and $e_a$ and $e_s$ to represent $a$-th and $s$-th canonical basis vectors in $\mathbb{R}^{|\mathcal{A}|}$ and $\mathbb{R}^{[S]}$, respectively.

In this paper, we represent a policy in a matrix form. In particular, for a given deterministic policy $\pi$, let us define the matrix $\Pi^\pi$ by

$$\Pi^\pi := \begin{bmatrix} (e_{\pi(1)} \otimes e_1)^T \\ \vdots \\ (e_{\pi(|\mathcal{S}|)} \otimes e_{|\mathcal{S}|})^T \end{bmatrix} \in \mathbb{R}^{[S]|\mathcal{S}| \times |\mathcal{A}|}.\quad \quad (7)$$
important properties. Two versions of Q-learning will be used: Algorithm 1, at each iteration we choose an initial state distribution. Under the i.i.d. sampling setting in Algorithm 1, we introduce the following shorthand, arg max
\[ \pi(s, a) = Q_k(s, a) + \alpha_k r_k + \gamma \max_{a' \in A} Q_k(s', a') - Q_k(s, a) \]

Then, we can prove that for any deterministic policy, \( \pi \), we have
\[ \Pi^\pi Q = \begin{bmatrix} Q(1, \pi(1)) \\ Q(2, \pi(2)) \\ \vdots \\ Q(|S|, \pi(|S|)) \end{bmatrix} \]

With this \( \Pi^\pi \), we can prove that for any deterministic policy \( \pi \), \( P^\pi = \Pi^\pi \beta \) is the state-action transition matrix. Using the notations introduced, the Bellman equation can be compactly written as
\[ Q^* = \gamma \Pi P Q^* + R, \]
where \( Q^* \) is the greedy policy defined as
\[ \pi_Q(s) = \arg \max_{a \in A} Q^*(s, a) \]
Moreover, for simplicity of notations, we introduce the following shorthand, \( \Pi_Q := \Pi^\pi \beta \), when the policy is the greedy policy \( \pi_Q(s) = \arg \max_{a \in A} Q(s, a) \) with respect to \( Q \).

**F. REVISIT Q-LEARNING**

This section briefly reviews standard Q-learning and its important properties. Two versions of Q-learning will be investigated in this paper. Algorithm 1 is Q-learning with an i.i.d. observation model, while Algorithm 2 is Q-learning with a Markovian observation model. In Algorithm 1, \( p \) denotes the current state distribution, and \( \beta \) is the so-called behavior policy, which is the policy to collect sample trajectories. On the other hand, in Algorithm 2, \( p_0 \) stands for the initial state distribution. Under the i.i.d. sampling setting in Algorithm 1, at each iteration \( k \), single state-action tuple \((s_k, a_k)\) is sampled from the distribution \( d \) defined by
\[ d(s, a) = \mathbb{P}[s_k = s, a_k = a] = p(s) \beta(a|s), \quad \forall (s, a) \in S \times A. \]

The next state \( s'_k \) is sampled independently at every iteration following the state transition probability \( P(s_k, a_k, \cdot) \). Note that this is similar to the generative setting [20], which assumes that every state-action tuple is generated at once from an oracle. With a slight abuse of notation, \( d \) will be also used to denote the vector \( d \in \mathbb{R}^{[S] \times |A|} \) such that
\[ d^\top (e_i \otimes e_j) := d(s, a), \quad \forall (s, a) \in S \times A. \]

Moreover, we further adopt the following notations:
\[ d_{max} := \max_{(s, a) \in S \times A} d(s, a), \quad d_{min} := \min_{(s, a) \in S \times A} d(s, a). \]

We also define the current reward as
\[ r_k := r(s_k, a_k, s'_k). \]
On the other hand, Q-learning with Markovian observation model in Algorithm 2 uses samples from a single state-action trajectory which follows the underlying Markov chain with the state-action transition matrix \( P^\beta \), where \( \beta \) is the so-called behavior policy. To be more specific, we suppose that starting from an initial state-action tuple \((s_0, a_0) \) under the initial distribution \( \mu_0 \), the underlying MDP generates \((s_0, a_0, r_0, s_1, a_1, r_1, \ldots) \) under the behavior policy \( \beta \) so that the current state-action distribution \( \mu_k \) is defined as
\[ \mu_k(s, a) := \mathbb{P}[s_k = s, a_k = a|\beta], \quad \forall (s, a) \in S \times A. \]

(7)
With the fixed behavior policy \( \beta \), the MDP becomes a Markov chain with the state-action transition probability matrix \( P^\beta := \Pi P^\beta \), i.e.,
\[ \mu^T_0 (P^\beta)^k = \mu^T_k. \]

Throughout the paper, we adopt the following standard assumption on this Markov chain.

**Assumption 4:** Consider the Markov chain of the state-action pair \((s, a) \) \( \in S \times A \) under \( \beta \) and initial state distribution \( p_0 \). We assume that the Markov chain is aperiodic and irreducible [38].

This assumption implies that under \( \beta \), the corresponding Markov chain admits the unique stationary distribution defined as
\[ \mu^T_\infty := \lim_{k \to \infty} \mu^T_k = \lim_{k \to \infty} \mu^T_0 (P^\beta)^k. \]

In the Markovian observation case in Algorithm 2, \( d_{max} \) and \( d_{min} \) are defined as
\[ d_{max} := \max_{(s, a) \in S \times A} \mu_\infty(s, a), \quad d_{min} := \min_{(s, a) \in S \times A} \mu_\infty(s, a). \]

Moreover, the current reward \( r_k \) is defined as
\[ r_k := r(s_k, a_k, s'_k). \]
To analyze the stochastic process generated by Q-learning, we define the notation \( F_k \) as follows:
1) \( F_k \) for i.i.d. observation: \( \sigma \)-field induced by
\[ \{(s_i, a_i, s'_i, r_i, Q_i)\}_{i=0}^k \]
2) \( F_k \) for Markovian observation:
\[ \{(s_i, a_i, s_{i+1}, r_i, Q_i)\}_{i=0}^k \]
In this paper, we will consider the i.i.d. observation model first, and then the corresponding analysis will be extended to the Markovian observation model later. Next, for simplicity of the proof, we assume the following bounds on the initial value, $Q_0$ and the reward vector $R$.

**Assumption 5:** Throughout the paper, we assume that the initial value, $Q_0$, and reward, $R$, is bounded as follows:

$$||Q_0||_\infty \leq 1, \quad ||R||_\infty \leq 1.$$  

With Assumption 5, we can derive bounds on $Q_k$ and $w_k$, which will be useful throughout the paper. First, we establish a bound on $Q_k$. The proof is in Appendix G.

**Lemma 6:** For all $k \geq 0$, we have

$$||Q_k||_\infty \leq \frac{1}{1-\gamma}, \quad \forall k \geq 0.$$

**G. LINEAR STOCHASTIC APPROXIMATION**

Linear stochastic approximation [39] is widely used to find a fixed point $x^*$ of a linear system, $Ax = b$, where $A \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$ are constants, $x \in \mathbb{R}^m$ is an unknown vector. Suppose that direct access to $A$ and $b$ is not available, while their stochastic estimations are accessible. The linear stochastic approximation is the following update:

$$x_{k+1} = x_k + \alpha_k (Ax_k - b + w_k), \quad (9)$$

where $\alpha_k$ is the step-size, and $w_k \in \mathbb{R}^m$ is an i.i.d. noise with zero mean, which is made due to the incomplete information on $(A, b)$. Under some mild assumptions, the stochastic recursion is known to converge to a fixed point, $x^*$, with probability one. When the noises $(w_k)_{k=1}^\infty$ are Markovian, the convergence analysis of (9) is more challenging. To avoid this issue, the notion of uniform ergodicity of Markov chain has been introduced in [38]. Further details on the Markovian sampling will be discussed later in Section VII.

**III. CONTROL SYSTEM APPROACH**

In this section, we reformulate Q-learning update in Algorithm 1 into a state-space representation of a control system [23], which is commonly used in control literature. Using matrix notations, the Q-learning algorithm with i.i.d. observation model in Algorithm 1 can be written as follows:

$$Q_{k+1} = Q_k + \alpha_k (DR + \gamma DQP_k Q_k - DQ_k + w_k), \quad (10)$$

where $D \in \mathbb{R}^{[S]|A| \times [S]|A|}$ is a diagonal matrix with diagonal elements being the current state-action distributions, i.e.,

$$(e_s \otimes e_a)^\top D (e_s \otimes e_a) = d(s, a) = \mathbb{P}[s_k = s, a_k = a],$$

and

$$w_k := (e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^\top R + \gamma (e_{a_k} \otimes e_{s_k})(e_{a_k}^* \otimes e_{s_k})^\top \Pi Q_k Q_k$$

$$- (e_{a_k} \otimes e_{s_k})(e_{a_k} \otimes e_{s_k})^\top Q_k$$

$$- (DR + \gamma DQP_k Q_k - DQ_k)$$

(11)

is an i.i.d. noise term which can be viewed as an error caused by the sampling. Note that the noise term is unbiased conditioned on $Q_k$, i.e., $\mathbb{E}[w_k | Q_k] = 0$. Moreover, using the above bound on $||Q_k||_\infty$, a bound on $w_k$ can be established as follows.

**Lemma 7:** Under i.i.d. observation model, the stochastic noise, $w_k$, defined in (11), is bounded as follows:

$$||w_k||_\infty \leq \frac{4}{1-\gamma}, \quad (12)$$

$$\mathbb{E}[||w_k||^2_F] \leq \frac{9}{(1-\gamma)^2} \quad (13)$$

for all $k \geq 0$, where $F_k$ is a $\sigma$-field induced by $(s_i, a_i, r_i, Q_i)_{i=0}^k$.

The proof is given in Appendix H. Furthermore, using Bellman equation in (6), one can readily prove that the error term $Q_k - Q^*$ evolves according to the following dynamic system:

$$Q_{k+1} - Q^* = A_{Q,k} (Q_k - Q^*) + b_{Q,k} + \alpha_k w_k, \quad (14)$$

where

$$A_{Q,k} := I + \alpha_k T_Q, \quad b_{Q,k} := \alpha_k \gamma D(P_Q - P_{Q^*})Q^*,$$

$$T_Q := \gamma D P Q - D. \quad (15)$$

Note that compared to the linear stochastic approximation, the Q-learning update, has affine term $b_{Q,k}$, which imposes difficulty in the analysis of the algorithm. Before proceeding further, we note that the switching mode depends only on the state, and the state evolves via the Markov decision process. Therefore, once the state is determined, the switching mode is chosen automatically. Moreover, without $w_k$, the system can be viewed as a deterministic state-dependent switched system, which does not include any stochastic components.

Now, we can view $A_{Q,k}$ as a system matrix of a stochastic linear time-varying switching system in the following senses:

1) The system matrix is switched due to the max operator;
2) The system matrix is time-varying due to the diminishing step-size.

In this paper, we will study convergence of Q-learning based on control system perspectives with the above interpretations. In particular, to analyze the convergence, we will use notions in control community such as Lyapunov theory for stability analysis of dynamical systems. First of all, we introduce the following result.

**Lemma 8** [16]: For any $Q \in \mathbb{R}^{[S]|A|}$, and $k \geq 0$.

$$||A_{Q,k}||_\infty \leq 1 - \alpha_k d_{\text{min}} (1 - \gamma).$$

The proof is in Appendix I. Using similar arguments, one can also prove that $T_Q$, which can be thought of as system matrix for continuous dynamics, is Hurwitz matrix, which can be proved using the strict diagonal dominance.

**Lemma 9:** $T_Q$ is Hurwitz for all $Q \in \mathbb{R}^{[S]|A|}$. In addition, we have

$$||T_Q||_\infty \leq 2d_{\text{max}}.$$

The proof is given in Appendix J. Proceeding on, let us consider the stochastic update (14). Ignoring the stochastic term $\alpha_k w_k$, it can be viewed as a discrete-time switched affine system. Unfortunately, analysis of a switched affine system
is more involved than that of a switched linear system. The recent result in [13] overcame this difficulty by introducing the notion of the upper and lower comparison systems corresponding to the original system. In particular, [13] introduced a switching system representation of the continuous-time O.D.E. model corresponding to Q-learning. Then, the stability has been proved based on comparison of the upper and lower comparison systems of the original system in (14). Moreover, [16] extended this framework to a discrete-time setting, and derived finite-time convergence analysis under constant step-sizes. The overall idea is illustrated in Figure 1 and Figure 2 in Appendix B. However, when a time-varying step-size is used, it imposes significant difficulties in the analysis, which will be illustrated in the following sections.

1. Construct \( Q_k^L \) and \( Q_k^U \) such that \( Q_k^L \leq Q_k \leq Q_k^U \).
2. Derive an error bound on \( Q_k^L - Q^* \) and \( Q_k^U - Q^* \), respectively.
3. Combine the error bounds: \( \| Q_k - Q^* \| \leq \max \{ \| Q_k^L - Q^* \|, \| Q_k^U - Q^* \| \} \).

**FIGURE 1.** Flowchart of the overall analysis.

### IV. KEY METHODS: ANALYSIS OF LOWER COMPARISON SYSTEM

This section provides the construction of a lower comparison system and its finite time analysis. The analysis provided in this section will play an important role in further sections. The construction of the lower comparison system follows that of [16]. The convergence analysis follows the spirit of [40], which is a well-known approach to prove convergence rate in optimization literature.

#### A. CONSTRUCTION OF LOWER COMPARISON SYSTEM

By the lower comparison system, we mean a dynamic system whose state at time \( k \), denoted by \( Q_k^L \), bounds the original system’s state \( Q_k \) from below on the orthant, i.e., \( Q_k^L \leq Q_k \).

In particular, the lower comparison system can be expressed as follows:

\[
Q_{k+1}^L = A_{Q^*,k} Q_k^L + w_k.
\]

(16)

Compared to the original system in (14), the main difference is the affine term \( b_{Q_k} \), which does not appear in the lower comparison system. For completeness, we provide the proof of the essential property, \( Q_k^L \leq Q_k \), in the following lemma. The proof is in Appendix K.

**Lemma 10 [16]:** When \( Q_0^L \leq Q_0 \leq Q_0^L \),

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

(17)

To proceed, we will denote \( Q_k^L = Q_k^L - Q^* \) for simplicity. To analyze the stochastic process generated by the lower comparison system, we define the notation \( \mathcal{F}_k^L \) as follows:

1. \( \mathcal{F}_k^L \) for i.i.d. observation: \( Q \)-field induced by

\[
\{(\tau_i, a_i, s_i', r_i, Q_i^L)\}_{i=0}^k.
\]

2. \( \mathcal{F}_k^L \) for Markovian observation: \( Q \)-field induced by

\[
\{(\tau_i, a_i, s_{i+1}, r_i, Q_i^L)\}_{i=0}^k.
\]

#### B. ANALYSIS OF LOWER COMPARISON SYSTEM

This section provides a convergence rate of the lower comparison system. Let us consider the deterministic version of (16) as follows:

\[
x_{k+1} = A_{Q^*,k} x_k, \quad k \in \{0, 1, \ldots\}.
\]

(18)

The above system can be seen as a linear time-varying system. From Lemma 3, the existence of a symmetric positive definite matrix \( G \in \mathbb{R}^{|S| \times |A|} \) and a sequence \( \{p_k\} \) such that

\[
\Pi_{k=0}^\infty p_k \to 0 \quad \text{ensures asymptotic stability of (18), when } G \text{ and } \{p_k\} \text{ satisfies below condition:}
\]

\[
A_{Q^*,k}^T G A_{Q^*,k} \leq p_k G \quad \forall k \geq 0.
\]

(19)

Here, \( p_k \) can be controlled by an appropriate step-size. For the choice of step-size, we will investigate a continuous-time counterpart of (19) because they have a close relation. Roughly speaking, as step-size becomes closer to zero, it will behave as a continuous-time algorithm. Therefore, we first consider a continuous-time Lyapunov equation which will serve as foundational block for our analysis of discrete-time algorithm. Let us choose \( G \) such that

\[
G T_{Q^*}^T + T_{Q^*}^T G = -I.
\]

(20)

If \( T_{Q^*} \) is Hurwitz, then there exists a unique symmetric positive definite matrix \( G \) such that the Lyapunov equation is satisfied, where \( G \) is given by

\[
G = \int_0^\infty (e^{T_{Q^*} t})^T (e^{T_{Q^*} t}) dt.
\]

(21)

Expanding the term \( A_{Q^*,k}^T G A_{Q^*,k} \) in (19), we have

\[
A_{Q^*,k}^T G A_{Q^*,k} - p_k G = (1 - p_k) G + \alpha_k (G T_{Q^*} + T_{Q^*}^T G) + \alpha_k^2 T_{Q^*}^T G T_{Q^*} = (1 - p_k) G - \alpha_k I + \alpha_k^2 T_{Q^*}^T G T_{Q^*}.
\]

(22)

The last equality follows from (20). Let \( p_k = 1 - \nu a_k \), where \( \nu \) is positive scalar whose specific selection will be explained later. Plugging \( p_k \) into (22), we get

\[
A_{Q^*,k}^T G A_{Q^*,k} - (1 - \nu a_k) G = \nu a_k G - \alpha_k I + \alpha_k^2 T_{Q^*}^T G T_{Q^*}.
\]

Choosing \( \nu \) and \( a_k \) small enough, which will be illustrated in the following section, will ensure the following inequality to hold:

\[
A_{Q^*,k}^T G A_{Q^*,k} \leq (1 - \nu a_k) G.
\]

Before proceeding on, we summarize preliminary results on the properties of \( G \). In particular, upper and lower bounds
on the spectral norm of $G$ will play a central role in our analysis.

**Lemma 11:** The following statements hold true:

1) $||G||_2$ satisfies
\[
\frac{1}{4(|S||A|)\gamma d_{\max}} \geq ||G||_2 \leq \frac{|S||A|}{2(1-\gamma)d_{\min}}.
\]

2) $||G^{-1}||_2$ satisfies
\[
\frac{2(1-\gamma)d_{\min}}{|S||A|} \leq ||G^{-1}||_2 \leq 4(|S||A|)\frac{1}{\gamma d_{\max}}.
\]

The proof is given in Appendix L.

1) **CHOICE OF STEP-SIZE**

In this subsection, we elaborate our motivation for choice of the step-size $\alpha_k$. It is known that the diminishing step-size $1/k$ leads to sample complexity which polynomially depends on $1/k$ [15]. In this paper, we consider the following diminishing step-size:

\[
\alpha_k = \frac{\theta}{k+\xi}, \quad (23)
\]

where $\theta > 0$ and $\xi > 0$ are tuning parameters to be determined. There are two conditions to be considered in the choice of the tuning parameters, which are summarized below.

1) The step-size should satisfy the Lyapunov inequality in (19).

2) The choice of step-size should lead to (fast) convergence of both lower and upper comparison system.

In this paper, the two conditions can be satisfied with the following choice of coefficients:

1) First of all, $\xi > 0$ is selected such that
\[
\frac{8|S|^3|A|^3 d_{\max}}{(1-\gamma)^2d_{\min}^2} \leq \xi \leq \frac{16|S|^3|A|^3 d_{\max}}{(1-\gamma)^2d_{\min}^2}, \quad (24)
\]

2) Once $\xi > 0$ is chosen based on the above rule, then $\theta$ can be selected as
\[
\theta = \frac{\xi d_{\min}}{4|S|^3|A|^2d_{\max}}. \quad (25)
\]

Note that the upper bound on $\xi$ is chosen for simplicity of the proof. However, it can be arbitrarily large. With the above setup, we have the following results.

**Lemma 12:** The following statements hold true:

1) Using the step-size in (23), we have
\[
A_{Q^*,k}^T G A_{Q^*,k} \leq (1-\nu\alpha_k)G, \quad (26)
\]

where
\[
\nu = \frac{(1-\gamma)d_{\min}}{|S||A|}. \quad (27)
\]

2) The conditions in (24) and (25) ensure
\[
\nu \theta \geq 2, \quad (28)
\]

where $\nu$ is defined in (27), and
\[
d_{\min}(1-\gamma)\theta \geq 2. \quad (29)
\]

The proof is given in Appendix M. The first item ensures convergence of the deterministic system in (18). The second item shows that the choice of the diminishing step-size in (24) and (25) leads to fast convergence of both lower and upper comparison systems. To be more specific, $\nu$ and $d_{\min}(1-\gamma)$ are important factors in determining the convergence rate of lower and difference of upper and lower comparison systems, respectively. This can be seen as notions that are analogous to the strong convexity in [40]. As can been seen in the further analysis, intuitively, each Lyapunov function of the system can be thought of as decreasing by the factors of $\nu\theta$ and $d_{\min}(1-\gamma)$ at $k$-th iteration. That is, denoting $V_k^L$ and $V_k^U$ the Lyapunov functions of the lower comparison system and difference of upper and lower comparison system at $k$-th iteration, respectively, we will have

\[
V_{k+1}^L \leq (1-\nu\theta)k \nu \alpha_k w_k^L, \quad (29)
\]

\[
V_{k+1}^U \leq (1-d_{\min}(1-\gamma)\theta)k \nu \alpha_k w_k^U, \quad (30)
\]

where $w_k^L$ and $w_k^U$ are bounded noise term for each system. If $\theta$ is small, the decreasing rate would be small, leading to slow convergence. Therefore, the step-size should be chosen to reflect these properties.

2) **PROOF OF CONVERGENCE RATE FOR LOWER COMPARISON SYSTEM**

We now prove the convergence rate for (16). To this end, we will consider the simple quadratic Lyapunov function candidate

\[
V(z) = z^T G z, \quad z \in \mathbb{R}^{|S||A|}. \quad (31)
\]

Convergence of the Lyapunov function candidate is established in the sequel.

**Proposition 13:** Under the step-size condition in (23), we have, for $k \in \mathbb{N}_0,$

\[
\frac{\mathbb{E}}{k+\xi} \left( V(Q_k^L - Q^*) \right) \leq \frac{C_v}{k+\xi},
\]

where $C_v := \max \left[ \frac{16|S|^3|A|^3 d_{\min}}{(1-\gamma)^2d_{\min}^2} \left( \frac{Q_k^L}{\mathbb{E}} \right)^2 \cdot \left( \frac{2|S||A|^3}{1-\gamma} \right) \right].$

**Proof:** Consider the Lyapunov function candidate $V(z) = z^T G z$ for $z \in \mathbb{R}^{|S||A|}$. We first derive an upper bound
\[
\mathbb{E}[V(z_{k+1}) | \mathcal{F}_k^L] \quad \text{from the relations}
\]
\[
\begin{align*}
\mathbb{E}[V(z_{k+1}) | \mathcal{F}_k^L] &= \mathbb{E}[(Q_k^L - Q^*)^\top A_{Q_{\alpha},k}^\top G A_{Q_{\alpha},k} (Q_k^L - Q^*) | \mathcal{F}_k^L] \\
&\quad + \mathbb{E}[(Q_k^L - Q^*)^\top A_{Q_{\alpha},k}^\top G w_k | \mathcal{F}_k^L] + \alpha_k^2 \mathbb{E}[w_k^\top G w_k | \mathcal{F}_k^L] \\
&= \mathbb{E}[(Q_k^L - Q^*)^\top A_{Q_{\alpha},k}^\top G A_{Q_{\alpha},k} (Q_k^L - Q^*) | \mathcal{F}_k^L] \\
&\quad + \alpha_k^2 \mathbb{E}[w_k^\top G w_k | \mathcal{F}_k^L] \\
&\leq (1 - \frac{\theta}{k + \xi}) V(Q_k^L - Q^*) + \frac{1}{(k + \xi)^2} \frac{72(|S||A|)^3}{(1 - \gamma)^5 d_{\min}^3}.
\end{align*}
\]
where \( \mathcal{F}_k^L \) is defined as the \( \sigma \)-field induced by
\[
\{(s_i, a_i, s'_i, r_i, Q_i, Q_i^L)\}_{i=0}^k.
\]
The second equality follows from the relations
\[
\mathbb{E}[2\alpha_k(Q_k^L - Q^*)^\top A_{Q_{\alpha},k}^\top G w_k | \mathcal{F}_k^L] = 0.
\]
The first inequality follows from Lemma 12, and the second last inequality can obtained via bound on \( \mathbb{E}[\|w_k\|_2^2 | \mathcal{F}_k^L] \) in (12) and bound on \( \theta \) in Lemma 33 in the Appendix P. The last inequality follows from inequality in (28). Now, taking the total expectation, one gets
\[
\mathbb{E}[V(z_{k+1})] \leq (1 - \frac{2}{k + \xi}) \mathbb{E}[V(z_k)] + \frac{1}{(k + \xi)^2} \frac{72(|S||A|)^3}{(1 - \gamma)^5 d_{\min}^3}.
\]
To proceed further, an induction argument is applied. For simplicity, let us define the notation \( k' := k + \xi \), and suppose that
\[
\mathbb{E}[V(z_k)] \leq \frac{C_v}{k}
\]
holds where
\[
C_v := \max \left\{ 16|S|^4|A|^4 d_{\max} \frac{\|\tilde{Q}_0\|_2^2}{(1 - \gamma)^4 d_{\min}^4}, \frac{72(|S||A|)^3}{(1 - \gamma)^5 d_{\min}^3} \right\}.
\]
We now prove that \( \mathbb{E}[V(z_{k+1})] \leq \frac{C_v}{k+1} \) holds. It is done by the inequalities
\[
\mathbb{E}[V(z_{k+1})] \leq (1 - \frac{2}{k}) \frac{C_v}{k} + \frac{1}{k^2} \frac{72(|S||A|)^3}{(1 - \gamma)^5 d_{\min}^3} \leq \frac{1}{k+1} \frac{C_v}{k+1} = \frac{C_v}{k + 1},
\]
where the second inequality follows from the fact that
\[
C_v \geq \frac{72(|S||A|)^3}{(1 - \gamma)^4 d_{\min}^4}.
\]
Hence, the inequality at \( k + 1 \) is satisfied. Moreover, when \( k = 0 \), we have \( V(z_0) \leq \frac{C_v}{k+1} \). The induction hypothesis leads to, for all \( k \geq 0 \),
\[
\mathbb{E}[V(z_k)] \leq \frac{C_v}{k + \xi},
\]
which is the desired conclusion.

Proposition 13 provides \( O(\xi) \) convergence rate for
\[
\mathbb{E}[V(Q_k^L - Q^*)] = \mathbb{E}[\|Q_k^L - Q^*\|^2] G(Q_k^L - Q^*).
\]
Using the bound on minimum eigenvalue of \( G \) in Lemma 11, the following theorem establishes its convergence rate.

**Theorem 14:** Under the step-size (23), we have
\[
\mathbb{E}[\|Q_k^L - Q^*\|_{\infty}] \leq \frac{C_L}{\sqrt{k + \xi}},
\]
where
\[
C_L := \max \left\{ \frac{8(|S||A|)^2 d_{\max}}{(1 - \gamma)^3 d_{\min}^3}, \frac{20(|S||A|)^7}{(1 - \gamma)^5 d_{\min}^3} \right\}.
\]
The proof is given in Appendix O. The convergence rate of lower comparison system depends on \((1 - \gamma)^{-\frac{5}{2}} d_{\min}^{-\frac{3}{2}} \) As can be seen in the sequel, the upper comparison system has slower convergence rate then lower comparison system. The bottleneck of convergence rate of Q-learning may be due to the relatively slower convergence of the upper comparison system.

**V. UPPER COMPARISON SYSTEM**

In this section, we focus on the construction of the upper comparison system. Then, its convergence analysis is discussed.

**A. CONSTRUCTION OF UPPER COMPARISON SYSTEM**

As in [16], the upper comparison system in (14) can be derived as follows:
\[
Q_k^U - Q^* = A_{Q_{\alpha},k} (Q_k^U - Q^*) + \alpha_k w_k.
\]
For completeness, we provide the proof that \( Q_k^U \) upper bounds \( Q_k \).

**Lemma 15** [16]: If \( Q_0^U - Q^* \geq Q_0 - Q^* \), then
\[
Q_k^U - Q^* \geq Q_k - Q^*, \quad \forall k \geq 0.
\]
The proof is given in Appendix N. To proceed, we will denote \( Q_0^U = Q_0^L - Q^* \). Since the upper comparison system in (32) is a switching linear system, it is not possible to obtain the corresponding convergence rate with the same techniques as in the lower comparison system. Moreover, we cannot find a common similarity transform matrix as in the lower comparison system case. Instead, we analyze the difference between \( Q_k^U \) and \( Q_k^L \) as in [16]. In particular, we consider the difference system
\[
Q_k^U - Q_k^L = A_{Q_{\alpha},k} (Q_k^U - Q_k^L) + B_{Q_{\alpha},k} (Q_k^L - Q^*)
\]
To prove the last inequality, we need to show
\[ \|Q^U_k - Q^L_k\|_\infty \leq \frac{C_u}{\sqrt{k + \xi}}. \]

where
\[ B_{Q,k} := A_{Q,k} - A_{Q^*,k} = \alpha_k \gamma D P(\Pi Q_k - \Pi Q^*). \]

B. PROOF OF CONVERGENCE RATE FOR DIFFERENCE OF UPPER AND LOWER COMPARISON SYSTEMS

With the above setup, we are now ready to give the convergence rate of \( E[\|Q^U_k - Q^L_k\|_\infty] \). For simplicity of the proof, we assume \( Q^U_0 = Q^L_0 \) without loss of generality.

**Theorem 16:** Under the diminishing step-size in (23), we have
\[ E[\|Q^U_k - Q^L_k\|_\infty] \leq \frac{C_u}{\sqrt{k + \xi}}, \]
where
\[ C_u := 32 \max \left\{ \frac{2(|S| |A|)^{13} d_{\max}^2}{(1 - \gamma)^2 d_{\min}^2} \|Q^L_0\|_2, \frac{5(|S| |A|)^{11} d_{\max}^3}{(1 - \gamma)^2 d_{\min}^2} \right\}. \]

**Proof:** The proof follows similar lines as Theorem 14.

We first take an expectation on the difference system (33), and use sub-multiplicativity of the norm as follows:
\[ E[\|Q^U_{k+1} - Q^L_{k+1}\|_\infty] \leq E[\|A_{Q,k}||Q^L_k - Q^L_{k+1}\|_\infty] + E[|B_{Q,k}||Q^L_{k+1} - Q^L_k||_\infty] \leq (1 - (1 - \gamma) d_{\min}) E[\|Q^U_k - Q^L_k\|_\infty] + 2\alpha_k \gamma d_{\max} E[\|Q^L_k - Q^L_k\|_\infty], \]
\[ \leq \left(1 - \frac{2}{k + \xi}\right) E[\|Q^U_k - Q^L_k\|_\infty] + \frac{2\theta \gamma d_{\max} C_L}{k + \xi} \leq \frac{C_u}{(k + \xi) \sqrt{k + \xi}}, \]
where
\[ C_L := \max \left\{ \frac{8(|S| |A|)^3 d_{\max}^3}{(1 - \gamma)^2 d_{\min}^3} \|Q^L_0\|_2, \frac{20(|S| |A|)^2 d_{\max}^5}{(1 - \gamma)^2 d_{\min}^3} \right\}. \]

The second inequality follows from Lemma 8, the third inequality is due to (29), and the last inequality comes from Lemma 33 given in Appendix F. We proceed by induction. Let \( k = k + \xi \), and suppose \( E[\|Q^U_k - Q^L_k\|_\infty] \leq \frac{C_u}{\sqrt{k}} \) for some \( k \geq 0 \). Now, we prove that \( E[\|Q^U_{k+1} - Q^L_{k+1}\|_\infty] \leq \frac{C_u}{\sqrt{k+1}} \) holds as follows:
\[ E[\|Q^U_{k+1} - Q^L_{k+1}\|_\infty] \leq \frac{C_u}{\sqrt{k + 1}} - \frac{2C_u}{k\sqrt{k}} \leq \frac{C_u}{\sqrt{k + 1}}. \]

To prove the last inequality, we need to show
\[ \frac{1}{\sqrt{k + 1}} - \frac{1}{\sqrt{k}} \leq \frac{1}{\sqrt{k + 1}}. \]

Multiplying \( \frac{1}{\sqrt{k + 1}} \) on both sides and rearranging the terms lead to \( (k - 1)\sqrt{k + 1} \leq k\sqrt{k} \). Now, it suffices to prove the last inequality. When \( 0 < \hat{k} < 1 \), the inequality is trivial. When \( \hat{k} \geq 1 \), taking square on both sides, we have
\[ (\hat{k}^2 - 2\hat{k} + 1)(\hat{k} + 1) \leq \hat{k}^3. \]

Rearranging the terms, the inequality becomes
\[ -2\hat{k}^2 + \hat{k} + \hat{k}^2 - 2\hat{k} + 1 = -\hat{k}^2 - \hat{k} + 1 \leq 0, \]
which proves the inequality \( (\hat{k} - 1)\sqrt{k + 1} \leq \hat{k}\sqrt{k} \). This completes the proof.

Compared to the convergence rate of the lower comparison system given in (16), the convergence rate for \( Q^U_k - Q^L_k \) given in (33) depends on \( (1 - \gamma)^{-2} d_{\min}^{-2} \), which leads to slower convergence of the system.

VI. ORIGINAL SYSTEM

With the bounds on upper and lower comparison systems, we are now ready to bound the original system in this section. We use simple triangle inequality \( \|Q_k - Q^L_k\|_\infty \leq \|Q_k - Q^U_k\|_\infty + \|Q^U_k - Q^L_k\|_\infty \) to bound the original iterate \( E[\|Q_k - Q^*\|_\infty] \).

**Theorem 17:** Under the step-size in (23), we have
\[ E[\|Q_k - Q^*\|_\infty] \leq \frac{C_o}{\sqrt{k + \xi}}. \]

where
\[ C_o := \max \left\{ \frac{72(|S| |A|)^{13} d_{\max}^2}{(1 - \gamma)^2 d_{\min}^2} \|Q^L_0\|_2, \frac{180(|S| |A|)^{11} d_{\max}^4}{(1 - \gamma)^2 d_{\min}^2} \right\}. \]

The proof is given in Appendix P. Since \( E[\|Q^U_k - Q^L_k\|_\infty] \) has a worse bound than \( E[\|Q_k - Q^*\|_\infty] \), convergence of \( E[\|Q_k - Q^*\|_\infty] \) is dominated by the bounds on \( E[\|Q^U_k - Q^L_k\|_\infty] \).

VII. FINITE TIME ANALYSIS UNDER MARKOVIAN NOISE

In the previous section, our analysis was based on i.i.d. observation model. However, in real-world scenarios, the state-action tuple follows Markovian observation model. In this section, we assume that the state-action pair follows a Markov chain as summarized in Section II-F. The corresponding Q-learning algorithm with Markovian observation model in Algorithm 2 can be written as follows:
\[ Q_{k+1} = Q_k + \alpha_k (D_{\infty} R + \gamma D_{\infty} P \Pi Q_k - D_{\infty} Q_k + w_k). \]

where
\[ D_{\infty} \in \mathbb{R}^{|S| |A| \times |S| |A|} \]
is a diagonal matrix with diagonal elements being the stationary state-action distributions, i.e.,
\[ (e_s \times e_a)^T D_{\infty} (e_s \times e_a) = \mu_{\infty}(s, a) = \lim_{k \to \infty} \mathbb{P}[s_k = s, a_k = a | \beta], \]
and
\[ w_k := (e_{a_1} \otimes e_{s_1}) (e_{a_1} \otimes e_{s_1})^T R + \gamma (e_{a_1} \otimes e_{s_1}) (e_{s_{k+1}}) \Pi Q_k - (D_{\infty} R + \gamma D_{\infty} P \Pi Q_k - D_{\infty} Q_k) \]

(35)
is a noise term. Using the Bellman equation in (6), one can readily prove that the error term $Q_k - Q^*$ evolves according to the following error dynamic system:

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

where

$$A_{Q,k} := I + \alpha_k T_Q, \quad b_{Q,k} := \alpha_k \gamma D_{\infty} P (\Pi Q - \Pi Q^*) Q^*, \quad T_Q := \gamma D_{\infty} P \Pi Q - D_{\infty}.$$  

(36)

Note that we can view $A_{Q,k}$ as a system matrix of a stochastic linear time-varying switching system as in the previous section. With slight abuse of notation, we use the notation $G$ for the solution of Lyapunov equation (20) with $T_Q$ in (36). Since $G$ only depend on $T_Q$, related lemmas still hold with $d_{\text{min}}$ and $d_{\text{max}}$ defined in Section II-F for the Markovian case. Moreover, due to the correlation of $Q_k$ and $(s_k, a_k, s_{k+1}, r_k)$, the noise term is biased, i.e., $E[w_k | F_k] \neq 0$, where $F_k$ is defined as the $\sigma$-field induced by $(s_i, a_i, s_{i+1}, r_i, Q_i)_{i=0}^t$. Therefore, the so-called crossing term in (30),

$$\xi_k(Q_k^T, Q_k, \alpha_k) := (Q_k^T - Q^*)^T A_{Q,k}^T \alpha_k G w_k.$$

(37)

is not zero, where $\alpha_k := (s_k, a_k, r_{k+1}, s_{k+1})$ is the observation, which is the main difference compared to the analysis in the previous section. To resolve the difficulty, in this paper, we assume that the distribution $\mu_k$ mixes in a geometric rate to the stationary distribution $\mu_{\infty}$, i.e., $\mu_k \rightarrow \mu_{\infty}$ as $k \rightarrow \infty$ exponentially. Note that this assumption is standard and common in the literature [17], [18], [41], [42], [43], which study stochastic optimization algorithms with Markovian observation models. With the help of geometric mixing rate of the underlying Markov chain, our aim is to bound (37) by $O\left(\frac{\tau^{\text{mix}}(\alpha_k)}{k}\right)$. The previous study in [42] has proved that such biased noise in standard TD-learning can be bounded by $O\left(\frac{\tau^{\text{mix}}(\alpha_k)}{k}\right)$ with an additional projection step. In this paper, we prove that with our setups, we can easily follow the spiris of [42], and prove convergence of Q-learning without the additional projection step.

Finally, in this section, we separate the index of the observation $(s_k, a_k, r_{k+1}, s_{k+1})$ and the iterate $(Q_k)_{k=1}^n$ in the noise term, i.e.,

$$w_j(Q_i) := (e_{a_j} \otimes e_{s_j}) \left( (e_{a_j} \otimes e_{s_j})^T R + \gamma e_{s_{j+1}}^T \Pi Q_i Q_i \right) - (e_{a_j} \otimes e_{s_j}) (e_{a_j} \otimes e_{s_j})^T Q_i - (D_{\infty} R + \gamma D_{\infty} P \Pi Q_i Q_i - D_{\infty} Q_i).$$

Moreover, we will use the notation $w_j(e_j, Q_i) := w_j(Q_i)$ interchangeably in order to specify the dependence of the noise on the observation. The subscript of $w_j$ indicates the time index of observation, while the subscript of $Q_i$ is the time index corresponding to Q-iteration. When their indices coincide, that is, having both $k$-th observation of $(s_k, a_k, r_{k+1}, s_{k+1})$ and $k$-th iterate of $(Q_k)_{k=1}^n$, we use $w_k$. Similar to the i.i.d. case, we can bound the noise term as follows:

Lemma 18: Under Markovian observation model in (35), the stochastic noise can be bounded as follows:

$$|w_k| \leq \frac{4}{1-\gamma}, \quad E[|w_k|^2 | F_k] \leq \frac{16 |A| |A|}{(1-\gamma)^2},$$

where $F_k$ is the $\sigma$-field induced by $(s_i, a_i, s_{i+1}, r_i, Q_i)_{i=0}^t$.

The proof is given in Appendix Q.

A. PRELIMINARIES ON MARKOV CHAIN

In this subsection, we introduce some definitions used in information theory, and its relation to the Markov chain under our consideration.

Definition 19 (Total variation distance [38]): The total variation distance between two probability distributions $P$ and $Q$ on the sample space, $\Omega$, is defined as

$$d_{TV}(P, Q) := \max_{A \in \Omega} |P(A) - Q(A)|.$$

Throughout the analysis, we will consider standard assumptions on Markov chain. With Assumption 4, the Markov chain reaches its stationary distribution at a geometric rate. We note that most of the analysis of stochastic optimization on Markov chain [17], [18], [41], [42], [43] relies on geometric convergence to the stationary distribution. Our analysis also adopts this standard scenario.

Lemma 20 [38]: Under Assumption 4, there exist constants $m > 0$ and $\rho \in (0, 1)$ such that

$$d_{TV}(\mu_k, \mu_{\infty}) \leq m \rho^k, \quad \forall k \in \mathbb{N},$$

where $\mu_k$ and $\mu_{\infty}$ are defined in (8).

Now, we can define the mixing-time as follows:

$$\tau^{\text{mix}}(c) := \min\{k \in \mathbb{N} | m \rho^k \leq c\},$$

(38)

which is the minimum time required for the total variation distance between the stationary distribution and the current distribution to be less than a desired level $c > 0$. Moreover, $m$ and $\rho$ are the constants defined in Lemma 20.

Related to the mixing-time, let us introduce the following quantity:

$$K_{\text{mix}} := \min\{k \in \mathbb{N}_0 | \tau \geq 2\tau^{\text{mix}}(\alpha_t), \forall t \geq k\},$$

(39)

where $\alpha_t$ is the step-size defined in (23). Note that $K_{\text{mix}}$ is well-defined, since upper bound of $\tau^{\text{mix}}(\alpha_t)$ is logarithmically proportional to $k$. Intuitively, $K_{\text{mix}}$ roughly means the number of iterations, after which the mixing-time corresponding to the current step-size can be upper bounded by the current time. Therefore, the analysis can be divided into two phases, the first phase $k \leq K_{\text{mix}}$ and the second phase $k > K_{\text{mix}}$. Note that this approach is common in the literature. For example, the recent paper [18] derived a convergence rate after some quantity similar to $K_{\text{mix}}$ iterations, and [17] derived a sample complexity after some time related to the mixing time. For $k \leq K_{\text{mix}}$, we can ensure the iterate remains bounded, and after $k \geq K_{\text{mix}}$, we have $O\left(\frac{\tau^{\text{mix}}(\alpha_k)}{k}\right)$ convergence rate.
B. BOUNDING THE CROSSING TERM $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$

With our assumptions in place, we are now ready to bound the crossing term, $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$, provided that the iterate $Q_k^L - Q^*$ is bounded, which is required because the crossing term is a function of Markovian noise and $Q_k^L - Q^*$. In this subsection, we also need to establish the boundedness of $Q_k^L - Q^*$. As in Assumption 5, we assume that $\|Q_0\|_{\infty} \leq 1$. Note that it is always possible because $\|Q_0\|_{\infty} \leq 1$, and we can always set $Q_0 = Q_0^L$ without loss of generality.

Lemma 21 (Boundedness of $Q_k^L - Q^*$): Suppose that $\|Q_0\|_{\infty} \leq 1$ holds. Then, for all $k \geq 0$, we have:

$$\|Q_k^L - Q^*\|_{\infty} \leq \frac{4}{(1 - \gamma)^2d_{\min}^2}.$$

The proof is given in Appendix R. With the boundedness of $Q_k^L - Q^*$, we can bound the crossing term as follows.

Lemma 22:

1) The crossing term $\xi_k(Q_k^L, Q_k)$ at $k$-th iterate is always bounded as follows:

$$|\xi_k(Q_k^L, Q_k)| \leq \frac{16(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2}.$$

2) $\xi_k(Q_k^L, Q_k)$ is Lipschitz with respect to $(Q_k^L, Q) \in [\mathbb{R}|S||A|] \times [\mathbb{R}|S||A|]$, where $Q_k^L, Q$ is bounded, i.e., $\|Q_k^L - Q^*\|_{\infty} \leq \frac{4}{(1 - \gamma)^2d_{\min}^2}$ and $\|Q\|_{\infty} \leq \frac{1}{1 - \gamma}$:

$$|\xi_k(Q_k^L, Q) - \xi_k(Q_k^L, Q')| \leq \frac{(|S||A|)^2}{2(1 - \gamma)^2d_{\min}^2} \|Q - Q'\|_{\infty} + \frac{2(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} \|Q_k^L - Q'\|_{\infty},$$

where $(Q_k^L, Q) \in [\mathbb{R}|S||A|] \times [\mathbb{R}|S||A|]$, and $Q_k^L, Q'$ is bounded, i.e., $\|Q_k^L - Q^*\|_{\infty} \leq \frac{4}{(1 - \gamma)^2d_{\min}}$ and $\|Q'\|_{\infty} \leq \frac{1}{1 - \gamma}$.

The proof is given in Appendix S. Using the above results, we are now in position to provide a bound on the crossing term, $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$. Before proceeding further, we will first obtain a bound on the delayed term, $\mathbb{E}[\hat{\xi}_k(Q_{k-\tau}^L, Q_{k-\tau})]$, and then derive a bound of the crossing term, $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$.

Lemma 23 (Boundedness of $\mathbb{E}[\xi_k(Q_{k-\tau}^L, Q_{k-\tau})]$): For any $\tau$ and $k$ such that $0 \leq \tau \leq k$, there exist $m > 0$ and $0 < \rho < 1$ such that

$$\mathbb{E}[\xi_k(Q_{k-\tau}^L, Q_{k-\tau})] \leq \frac{40(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2}m\rho^\tau.$$

The proof is given in Appendix T. The above lemma states that if $\tau$ and the difference between the new observation $Q_{k-\tau}$ and previous iterate $Q_{k-\tau}$ is large, then the Markovian noise term $w(\alpha_k, Q_{k-\tau})$ inherits the geometric mixing property of the Markovian chain. Now, combining the above result with Lipschitzness of $\xi_k(Q_k^L, Q_k)$, we can bound the crossing term, $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$, as follows. The proof is given in Appendix U.

Lemma 24 (Boundedness of $\mathbb{E}[\xi_k(Q_k^L, Q_k)]$): For $k \geq K_{mix}$, expectation of the cross term can be bounded as follows:

$$\mathbb{E}[\xi_k(Q_k^L, Q_k)] \leq \frac{40(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} + \frac{25(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} \tau_{mix}(\alpha_k) \alpha_k^{\tau_{mix}(\alpha_k)}.$$

C. FINITE TIME ANALYSIS UNDER MARKOVIAN NOISE

Now, we turn to the statement of our main results. The proof logic follows similar arguments as in the i.i.d. setting with few modifications. First, we begin the overall proof by deriving the convergence rate of $Q_k^L$, which corresponds to the lower comparison system with a coordinate transformation. Then, as a next step, we will consider the difference system to obtain a convergence rate of the original system. In the sequel, a generic constant $C > 0$ will be used for unimportant values that change line by line.

Proposition 25 (Convergence Rate of $Q_k^L$): Let $V(z) = z^T G z$ for $z \in \mathbb{R}|S||A|$. Consider the step-size in (23) and Markovian observation model. For any $k \geq K_{mix}$, we have

$$\mathbb{E}[V(Q_k^L)] \leq \frac{C_2}{k + \xi} + \frac{C(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} \tau_{mix}(\alpha_k),$$

where and

$$C_2 := C \max \left\{ \frac{8|S||A|d_1^2}{(1 - \gamma)^2d_{\min}^2} \|Q_{K_{mix}}^L - Q^*\|_2, \frac{(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} \right\}.$$

The proof is given in Appendix V. Using the spectral properties of $G$ in Lemma 11, we can derive the final iterate convergence rate for $\mathbb{E}[\|Q_k^L - Q^*\|_1]$ corresponding to the lower comparison system.

Theorem 26: For any $k \geq K_{mix}$, we have

$$\mathbb{E}[\|Q_k^L - Q^*\|_1] \leq \frac{C_{L}}{\sqrt{k + \xi}} + \frac{C(|S||A|)^{3/2}}{(1 - \gamma)^2d_{\min}^2} \sqrt{\tau_{mix}(\alpha_k)}.$$

where $C_{L} := C \max \left\{ \frac{8|S||A|d_1^2}{(1 - \gamma)^2d_{\min}^2} \|Q_{K_{mix}}^L - Q^*\|_2, \frac{(|S||A|)^2}{(1 - \gamma)^2d_{\min}^2} \right\}$.

The proof is given in Appendix W. Next, we derive a convergence rate for $\mathbb{E}[\|Q_k^L - Q_k^L\|]$. The proof is similar to those of the i.i.d. setting: we use a bound on $\mathbb{E}[\|Q_k^L - Q^*\|_1]$ and then apply an induction logic. The randomness of $Q_k^L - Q_k^L$ mainly relies on $Q_k^L - Q^*$. Therefore the influence of Markovian noise in $Q_k^L - Q^*$ directly affects the difference between the upper and lower comparison systems. That is, the convergence rate of $\mathbb{E}[\|Q_k^L - Q_k^L\|]$ follows that of $\mathbb{E}[\|Q_k^L - Q^*\|_1]$, and sample complexity becomes worse due to slow progress of the upper comparison system as seen in the i.i.d. case.

Theorem 27: For any $k \geq K_{mix}$, the difference between the upper and lower comparison system is bounded as follows:

$$\mathbb{E}[\|Q_k^U - Q_k^L\|] \leq \frac{C_{U}}{\sqrt{k + \xi}} + \frac{C(|S||A|)^{3/2}}{(1 - \gamma)^2d_{\min}^2} \sqrt{\tau_{mix}(\alpha_k)}.$$

where $C_{U} := C \max \left\{ \frac{(|S||A|)^{3/2}}{(1 - \gamma)^2d_{\min}^2} \|Q_{K_{mix}}^L - Q^*\|_2, \frac{(|S||A|)^{3/2}}{(1 - \gamma)^2d_{\min}^2} \right\}$.
The proof is given in Appendix X. Finally, we derive the convergence rate for the original system, which follows from simple triangle inequality.

**Theorem 28:** For \( k \geq K_{\text{mix}} \), we have

\[
\mathbb{E}[\|Q_k - Q^*\|_\infty] \leq C\sqrt{\frac{1}{k+\xi}} + \frac{C}{(1-\gamma)\beta_{\text{max}}^2}\sqrt{t_{\text{mix}}(\alpha_k)},
\]

where \( C := \max \left\{ \frac{(S+1)\beta_{\text{min}}}{(1-\gamma)\beta_{\text{max}}^2} \|Q_{K_{\text{mix}}} - Q^*\|_2 \right\} \).

The proof is given in Appendix Y.

**Remark 29:** Since we can bound the mixing time with logarithmic term by Lemma 40 given in Appendix Z-A, we can achieve \( O\left(\sqrt{\log \frac{1}{k}}\right) \) convergence rate under Markovian observation model. However, from Lemma 42 given in Appendix Z-A, the above result leads to sub-optimal sample complexity. The recent progress [19] achieves \( O\left(\log k\right) \) convergence rate under Markovian observation model, whereas our work uses a diminishing step-size that is independent of the mixing time. Therefore, our analysis covers different scenarios. With the construction of the upper and lower comparison systems, both of which can be derived in this framework, we can achieve the convergence rate of the original system. However, we note that our step-size still depends on the problem parameters including the minimum state visitation distribution \( d_{\text{min}} \).

### D. COMPARATIVE ANALYSIS

The recent paper [17] derives sample complexity in terms of the probability tail bounds using the adaptive step-size

\[
\alpha_k = \min \left\{ 1, c \exp \left( \log \left( \frac{k}{d_{\text{min},k}(1-\gamma)\gamma^2} \right) \right) \right\},
\]

where \( d_{\text{min},k} \) is a number that is updated at every iteration, and \( c \) is some universal constant. Note that the step-size does not depend on \( d_{\text{min}} \) or the mixing time, while it requires an additional step to update \( d_{\text{min},k} \), which brings in additional complexity compared to our diminishing step-size. Moreover, their update requires a state-action dependent process, the number of times that the current state-action visits a certain state-action pair, which requires \( O(|S||A|) \) memory space. With the help of concentration inequality in Markov chain [44] and refined analysis based on number of visits to \((s, a)\), probability tail bounds can be obtained as well.

Another recent paper [18] uses a step-size that satisfies the constraint

\[
\sum_{i=k-1}^{k-i_{\text{mix}}(\alpha_k)} \alpha_i \leq c,
\]

where

\[
t_{\text{mix}}(\alpha_k) := \min \left\{ k \geq 0 \mid \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} d_{\mathcal{T}V}(\mu_k(s, a), \mu) \leq \alpha_k \right\},
\]

and for some constant \( c \) to derive sample complexity in terms of the expectation using analysis based on Moreau envelope [45]. Under ergodic assumption, it derives \( O\left(\sqrt{t_{\text{mix}}(\alpha_k)}\right) \) where \( t_{\text{mix}}(\alpha_k) \) is upper bounded by logarithmic term. The study in [15] showed that extending the analysis of [24] and using the step-size \( \frac{1}{k} \) lead to sample complexity depending polynomially on \( \frac{1}{1-\gamma} \). The paper [19] uses the diminishing step-size \( \alpha_k = \frac{\beta}{k+k_{0}} \) with

\[
k_0 \geq \max \left\{ 4h, \left[ \log \frac{2}{d_{\text{min}}} t_{\text{mix}}(1/4) \right] \right\},
\]

where \( h \geq \frac{4}{d_{\text{min}}(1-\gamma)} \) to derive sample complexity in terms of probability tail bounds. The recent progress [19] proved that analysis on decomposition of error term into shifted Martingale terms leads to probability tail bounds. Most of the previous and recent analysis uses step-sizes with the associated conditions that depend on the mixing time, whereas our work uses a diminishing step-size that is independent of the mixing time. Furthermore, our analysis covers different scenarios. With the construction of the upper and lower comparison systems, both of which can be analyzed with the step-size condition to be independent on the mixing time, we can derive the convergence rate of the original system. However, we note that our step-size still depends on the problem parameters including the minimum state visitation distribution \( d_{\text{min}} \).

### VIII. CONCLUSION

In this paper, we have derived a final iterate convergence rate of asynchronous Q-learning under i.i.d. and Markovian observation model and diminishing step-size. We proved a convergence rate of \( O\left(\sqrt{\log \frac{1}{k}}\right) \), which is the sharpest convergence rate under Markovian observation model. Tightening the sample complexity to the current sharpest bound would be one of the future research directions. Furthermore, extending this framework to Q-learning variants e.g., double Q-learning [46], periodic Q-learning [25], is left as a future work.

### APPENDIX A

**OMITTED DEFINITIONS IN PRELIMINARIES SECTION**

**Definition 30 (Hurwitz Matrix [23], [37]):** A square matrix \( A \in \mathbb{R}^{n \times n} \) is called Hurwitz if its real part of eigenvalues are negative.

**Definition 31 (Strict Diagonal Dominance [47]):** A square matrix \( A \in \mathbb{R}^{n \times n} \) is said to be strictly diagonally dominant if

\[
\sum_{j=1}^{n} |a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \text{for all} \quad i = 1, \ldots, n.
\]
dominant if
\[ |[A]_{ii}| > \sum_{j \neq i,j=1}^n |[A]_{ij}|, \quad \forall i \in \{1, 2, \ldots, n\}. \]

**APPENDIX B**

**OMITTED FIGURES**
See Fig. 2.

**APPENDIX C**

**EXPERIMENTS**
In this section, we conduct a simple experiment on the following MDP:

\[
D := \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & \frac{1}{4} & 0 & 0 \\
0 & 0 & \frac{1}{4} & 0 \\
0 & 0 & 0 & \frac{1}{4}
\end{bmatrix}, \quad P := \begin{bmatrix}
1/2 & 1/2 \\
1/3 & 2/3 \\
1 & 0 \\
0 & 1
\end{bmatrix}
\]
\[ \gamma := 0.99, \quad |S| = 2, \quad |A| = 2. \]

We have initialized \(Q_0\) as the zero vector, and \(Q^U_0\) and \(Q^L_0\) as a vector whose elements are \(-0.1\) smaller or \(0.1\) bigger than \(Q_0\), respectively. Furthermore, we used the step-size, \(\alpha_k := 5/(k+5)\) for \(k \in \mathbb{N}\).

Examining Figure 3a, we observe that the difference between \(Q_k\) and \(Q^*\) converges to zero asymptotically. Additionally, Figure 3b confirms that the inequality \(Q^L_k \leq Q_k \leq Q^U_k\) is maintained throughout the process.

**FIGURE 3. Experimental result.**

(a) Plot for 2000 steps  (b) Plot for 100 steps

**APPENDIX D**

**PROOF OF LEMMA 2**
We first introduce Gerschgorin circle theorem which is essential in proving Lemma 2.

**Lemma 32 (Gerschgorin Circle Theorem [47]):** Let \(A \in \mathbb{R}^{n \times n}\) be a square matrix. Gerschgorin circles are defined as:

\[
\{ z \in \mathbb{C} \mid |z - [A]_{ii}| \leq \sum_{j \neq i,j=1}^n |[A]_{ij}| \}, \quad i = 1, \ldots, n.
\]

The eigenvalues of \(A\) are in the union of Gerschgorin discs,

\[
G(A) = \bigcup_{i=1}^n \left\{ z \in \mathbb{C} \mid |z - [A]_{ii}| \leq \sum_{j \neq i,j=1}^n |[A]_{ij}| \right\}.
\]

**Proof:** Gerschgorin circle theorem states that the eigenvalues of the matrix are within circles, called Gerschgorin discs, which are defined in terms of algebraic relations and entries of the matrix \(A\). Building on Lemma 32, we can prove Lemma 2, a strictly diagonally dominant matrix with negative diagonal terms is Hurwitz matrix.

To this end, first note that \(Re(\lambda_i(A)) - [A]_{ii} \leq |\lambda_i(A) - [A]_{ii}|\), where \(\lambda_i(A)\) denotes the \(i\)-th eigenvalue of \(A\) and \(Re(\lambda_i(A))\) denotes the real part of \(\lambda_i(A)\). Now, applying Lemma 32 yields

\[
Re(\lambda_i(A)) \leq \sum_{j \neq i,j=1}^n |[A]_{ij}| + |A|_{ii} < 0.
\]

Hence, the real part of eigenvalues of \(A\) is negative, which means that \(A\) is Hurwitz matrix. \(\square\)

**APPENDIX E**

**PROOF OF LEMMA 3**

**Proof:** Let \(V_k(x_k) = x_k^T P_k x_k\) be a time-varying Lyapunov function that satisfies the condition in item 1). Recursively applying the inequality in item 1) leads to

\[
V_k(x_k) \leq p_{k-1} V_{k-1}(x_{k-1}) \leq \left( \prod_{i=0}^{k-1} p_i \right) V_0(x_0).
\]

Using item 3), it follows from the last inequality that

\[
c \|x_k\|^2 \leq V_k(x_k) \leq \left( \prod_{i=0}^{k-1} p_i \right) V_0(x_0) \leq \left( \prod_{i=0}^{k-1} p_i \right) \lambda_{\max}(P_0) \|x_0\|^2,
\]

which leads to

\[
\|x_k\|^2 \leq \left( \prod_{i=0}^{k-1} p_i \right) \lambda_{\max}(P_0) \frac{\|x_0\|^2}{c}.
\]

By item 2), we have \(\|x_k\| \to 0\) as \(k \to \infty\). This completes the proof. \(\square\)

**APPENDIX F**

**INEQUALITIES RELATED TO STEP-SIZE**

Now, we introduce a useful inequality used in the proof of Proposition 13.

**Lemma 33:** For \(\theta\) and \(\xi\) defined in (25), we have the following bounds:

\[
\theta \leq \frac{4|S||A|}{(1 - \gamma)d_{\min}}.
\]

**Proof:** The inequality directly follows from the definition of \(\xi\) in (24):

\[
\theta \leq \frac{(1 - \gamma)d_{\min}}{4|S|^3|A|^2 d_{\max}} = \frac{4|S||A|}{(1 - \gamma)^2 d_{\min}}.
\]

This completes the proof. \(\square\)
APPENDIX G
PROOF OF LEMMA 6

Proof: Assume \(|Q_k(s_k, a_k)| \leq \frac{1}{1 - \gamma}\). Then, using triangle inequality to Q-learning update in Algorithm 1 and the assumption that \(|R|_\infty \leq 1\) leads to
\[
|Q_{k+1}(s_k, a_k)| \leq (1 - \alpha_k) \frac{1}{1 - \gamma} + \alpha_k \left(1 + \gamma \frac{1}{1 - \gamma}\right)
\]
\[
\leq \frac{1}{1 - \gamma}.
\]
Therefore, using induction, we have the desired inequality. □

APPENDIX I
PROOF OF LEMMA 8

Proof: The proof of Lemma 8 is as follows: The column sum for each row becomes
\[
\sum_{j=1}^{\left|S\right|} |A_{Q_k}|_{ij} = \sum_{j=1}^{\left|S\right|} |I - \alpha_k D + \alpha_k \gamma D \Pi Q|_{ij}
\]
\[
= |I - \alpha_k D|_{ii} + \sum_{j=1}^{\left|S\right|} |\alpha_k \gamma D \Pi Q|_{ij}
\]
\[
= 1 - \alpha_k |D|_{ii} + \alpha_k \gamma |D|_{ii} \sum_{j=1}^{\left|S\right|} |P \Pi Q|_{ij}
\]
\[
= 1 - \alpha_k |D|_{ii} + \alpha_k \gamma |D|_{ii},
\]
where the second equality is due to the fact that each element of \(A_{Q_k}\) is positive. Taking maximum over the row index, we get
\[
||A_{Q_k}||_\infty = \max_{i \in \{1, 2, \ldots, \left|S\right|\}} \left(1 + \alpha_k |D|_{ii} (\gamma - 1)\right)
\]
\[
= 1 - \alpha_k \min_{(s, a) \in S \times A} d(s, a) (1 - \gamma).\]

This completes the proof. □

APPENDIX J
PROOF OF LEMMA 9

Proof: The proof of Lemma 9 is as follows: From Lemma 2, it suffices to prove that \(T_Q\) is strictly row diagonally dominant which is defined in Definition 31. Each component can be calculated as follows:
\[
|T_Q|_{ii} = |D|_{ii} (1 + \gamma e_i^T \Pi Q e_i),
\]
\[
\sum_{j \neq i \in \left|S\right|} |T_Q|_{ij} = \sum_{j \neq i \in \left|S\right|} \gamma |D|_{ii} e_i^T \Pi Q e_j
\]
\[
= \gamma |D|_{ii} (1 - e_i^T \Pi Q e_i).
\]

Now, we can prove that the diagonal part dominates the off-diagonal part, i.e.,
\[
||T_Q||_1 = \sum_{j \neq i \in \left|S\right|} ||T_Q|_{ij}|,
\]
\[
\leq \sum_{j \neq i \in \left|S\right|} |T_Q|_{ii} = ||T_Q||_1,
\]
from the following relation:
\[
||T_Q||_1 - \sum_{j \neq i \in \left|S\right|} ||T_Q|_{ij}| = |D|_{ii} (1 - \gamma e_i^T \Pi Q e_i) - \gamma |D|_{ii} (1 - e_i^T \Pi Q e_i)
\]
\[
= (1 - \gamma) |D|_{ii} > 0,
\]
since the diagonal term is negative, \(T_Q\) is Hurwitz by Lemma 2. Moreover, collecting each elements of \(T_Q\) in (41)
and (42), we have

\[ ||T_Q||_\infty = \max_{i \in \{1, 2, \ldots, |S|\}} \sum_{j=1}^{|A|} ||T_Q||_{i,j} \]

\[ \leq \max_{i \in \{1, 2, \ldots, |S|\}} \left( \left( 1 + \sum_{j=1}^{|A|} [T_Q]_{i,j} + O(s^2) \right) \right) \]

This completes the proof. \[ \square \]

**APPENDIX K**

**PROOF OF LEMMA 10**

*Proof:* The proof is done by induction. Assume for some \( k \geq 0 \), equation 17 satisfies. Now, we prove that the inequality is also satisfied for \((k + 1)\)-th iterate.

\[
Q_{k+1}^\star - Q^\star = A_{Q,k}^\star (Q_k^\star - Q^\star) + \alpha_k w_k \\
\leq A_{Q,k}^\star (Q_k^\star - Q^\star) + \alpha_k w_k \\
= A_{Q,k}^\star (Q_k^\star - Q^\star) + \alpha_k (Q_k^\star - Q^\star) + \alpha_k w_k \\
= A_{Q,k}^\star (Q_k^\star - Q^\star) + \alpha_k (Q_k^\star - Q^\star) \\
\leq A_{Q,k}^\star (Q_k^\star - Q^\star) + \alpha_k (Q_k^\star - Q^\star), \quad (44)
\]

where the first inequality is due to the hypothesis and the fact that \( A_{Q,k}^\star \) is positive matrix for any \( Q \). Moreover, the second inequality can be driven using the fact that \( \Pi_{Q,k} Q_k \leq \Pi_{Q,k} Q_k \) and \( \Pi_{Q,k} Q^\star \leq \Pi_{Q,k} Q^\star \).

**APPENDIX L**

**PROOF OF LEMMA 11**

Before proceeding on, we introduce the following lemma which is required for the main proof.

*Lemma 34 ([48], Chapter C.1):* For \( 1 \leq k \leq n \), we have

\[
\binom{n}{k} \leq \left( \frac{en}{k} \right)^k,
\]

where \( \binom{n}{k} := \frac{n!}{k!(n-k)!} \).

Now, we prove the first item in Lemma 11.

*Proof:* For any \( s > 0 \), we can bound the matrix exponential term in (21) as follows:

\[
\left\| e^{T_Q s^\star} \right\|_{\infty} = \left\| I + T_Q^{s^\star} + \frac{(T_Q^{s^\star})^2}{2!} + \frac{(T_Q^{s^\star})^3}{3!} + \cdots \right\|_{\infty} \\
= \max_{i \in \{1, 2, \ldots, |S|\}} \left( \left( 1 + \sum_{j=1}^{|A|} [T_Q^{s^\star}]_{i,j} + O(s^2) \right) \right) \\
\leq \max_{i \in \{1, 2, \ldots, |S|\}} \left( \left( 1 + \sum_{j=1}^{|A|} [T_Q^{s^\star}]_{i,j} + O(s^2) \right) \right) \\
= e^{(1 - \gamma) d_{\min} s} \quad (45)
\]

where the first equality expands matrix exponential in terms of matrix polynomials, and the second equality follows from the definition of infinity norm. Bounding the first term of (45) with the diagonally dominant property in Lemma 9, we have

\[
\left\| e^{T_Q s^\star} \right\|_{\infty} \leq 1 - (1 - \gamma) d_{\min} s + O(s^2). \quad (46)
\]

Using the last inequality, we can obtain the following relation:

\[
\left\| e^{T_Q s^\star} \right\|_{\infty} = \left\| (e^{T_Q s^\star})^n \right\|_{\infty} \\
\leq \left\| e^{T_Q s^\star} \right\|_{\infty}^n \\
= \left\{ 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} + O \left( \frac{s^2}{n^2} \right) \right\}^n \quad (47)
\]

where the first inequality follows from the sub-multiplicativity of matrix norm, and the second inequality follows from directly applying (46).

Taking the limit on both sides and using the relation, \( \lim_{n \to \infty} \left( 1 - \frac{a}{n} \right)^n = e^{-a} \) for any \( a > 0 \), we can derive the following bound:

\[
\left\| e^{T_Q s^\star} \right\|_{\infty} \leq \lim_{n \to \infty} \left( 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} + O \left( \frac{s^2}{n^2} \right) \right)^n \\
= \lim_{n \to \infty} \left( 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} \right)^n \left( O \left( \frac{s^2}{n^2} \right) \right)^k \\
= \lim_{n \to \infty} \left( 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} \right)^n \left( O \left( \frac{s^2}{n^2} \right) \right)^k \\
\leq \lim_{n \to \infty} \left( 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} + \lim_{n \to \infty} \sum_{k=1}^n O \left( \frac{s^2}{n^2} \right)^k \right) \\
= \lim_{n \to \infty} \left( 1 - (1 - \gamma) d_{\min} s^\star \frac{1}{n} + \lim_{n \to \infty} \sum_{k=1}^n O \left( \frac{s^2}{n^2} \right)^k \right) \\
= e^{(1 - \gamma) d_{\min} s} \\
\]

where \( \binom{n}{k} := \frac{n!}{k!(n-k)!} \). The first equality follows from the binomial theorem, and the second inequality follows from Lemma 34, which bounds the binomial coefficient. The second last equality follows from summation of geometric series. Therefore, we have

\[
\left\| e^{T_Q s^\star} \right\|_{\infty} \leq e^{-(1 - \gamma) d_{\min} s}. \quad (48)
\]
Next, we bound the spectral norm of $G$ as follows:

$$\|G\|_2 = \int_0^\infty \left\| \left( e^{tG^t} \right) \right\|_2^2 dt \leq |S| |A| \int_0^\infty \left\| \left( e^{tG^t} \right) \right\|_\infty^2 dt \leq \int_0^\infty e^{-2(1-\gamma)d_{\text{max}}t} dt \leq \frac{|S| |A|}{2(1-\gamma)d_{\text{min}}},$$

where the first equality comes from the fact that $\left\| e^{tG^t} \right\|_2 = \left\| e^{tG} \right\|_2$, and the second inequality follows from (48). The lower bound of $\|G\|_2$ follows from applying simple triangle inequality to (20).

$$\|G\|_2 \geq \frac{1}{2\|TQ^r\|_2} \geq \frac{1}{2\sqrt{|S| |A| \|TQ^r\|_\infty}} \geq \frac{1}{4\sqrt{|S| |A| d_{\text{max}}}},$$

where the last inequality follows from Lemma 9. Now, we prove the second item, bound on $\|G^{-1}\|_2$. Multiplying $G^{-1}$ on the right and left of (20), we have

$$TQ^r G^{-1} + G^{-1} TQ^r = -G^{-2}. \quad (49)$$

Now taking the spectral norm on (49), and using the triangle inequality, we have

$$\|G^{-2}\|_2 \leq 2\|TQ^r\|_2 \|G^{-1}\|_2. \quad (50)$$

Since $G^{-1}$ is symmetric positive definite, it follows that $\|G^{-2}\|_2 = \|G^{-1}\|_2^2$. Therefore, we have

$$\|G^{-1}\|_2 \leq 2(|S| |A|)^{1/2} \|TQ^r\|_\infty \leq 4(|S| |A|)^{1/2} d_{\text{max}},$$

where the first inequality follows from (50) and the fact that $\|G^{-1}\|_2 > 0$, and the last inequality comes from Lemma 9. Next, the lower bound of $\|G^{-1}\|_2$ can be obtained as follows:

$$\|G^{-1}\|_2 = \frac{1}{\sigma_{\text{min}}(G)} \geq \frac{2(1-\gamma)d_{\text{min}}}{|S| |A|},$$

where the first equality is due to the fact that $G$ is positive symmetric definite, and the last inequality follows from the upper bound on $\|G\|_2$. This completes the proof.

**APPENDIX N**

**PROOF OF LEMMA 15**

**Proof:** The proof follows from induction. To this end, suppose that $Q^{k}_U - Q^* \geq Q_k - Q^*$ holds for some $k \geq 0$. Then, we have

$$Q^{k+1}_U - Q^* = A_{Q_k} (Q^{k}_U - Q^*) + \alpha_k w_k \geq A_{Q_k} (Q^{k}_U - Q^*) + b_{Q_k} \alpha_k w_k \geq A_{Q_k} (Q_k - Q^*) + b_{Q_k} \alpha_k w_k = Q_{k+1} - Q^*,$$

where the first inequality follows from the fact that $b_{Q_k} \leq 0$. The second inequality is due to positiveness of $A_{Q_k}$. This completes the proof.

**APPENDIX O**

**PROOF OF THEOREM 14**

**Proof:** We have

$$\mathbb{E}[\|Q^*_k - Q^*\|_2^2] = \frac{1}{\lambda_{\text{min}}(G)} \mathbb{E} \left[ (Q^*_k - Q^*)^T (Q^*_k - Q^*) \right] \leq \frac{C_i^2}{k + \frac{\gamma}{\xi}},$$

where the second inequality follows from Lemma 11 and Proposition 13. Applying Jensen’s inequality to the above result yields

$$\mathbb{E}[\|Q^*_k - Q^*\|_\infty] \leq \sqrt{\mathbb{E}[\|Q^*_k - Q^*\|_2^2]} \leq \sqrt{\mathbb{E}[\|Q^*_k - Q^*\|_2]} \leq \frac{C_i}{\sqrt{k + \frac{\gamma}{\xi}}},$$

which is the desired conclusion.
APPENDIX P

PROOF OF THEOREM 17

Proof: The inequality follows from applying triangle inequality of the norm as follows:

\[
\mathbb{E}[||Q_k - Q^*||_\infty] \\
\leq \mathbb{E}[||Q^* - Q^*_k||_\infty] + \mathbb{E}[||Q_k - Q^*_k||_\infty] \\
\leq \mathbb{E}[||Q^* - Q^*_k||_\infty] + \mathbb{E}[||Q^*_k - Q_k||_\infty]
\]

where the first inequality follows from triangle inequality. The second inequality is due to the fact that \( ||Q_k - Q^*_k||_\infty \leq ||Q^*_k - Q_k||_\infty \). This completes the proof.

\[\square\]

APPENDIX Q

PROOF OF LEMMA 18

Proof: The bound on \( ||w_k||_\infty \) can be proved with the same logic as in Lemma 7. For \( \mathbb{E}[||w_k||_\infty^2 | \mathcal{F}_k] \), we have

\[
\mathbb{E}[||w_k||_\infty^2 | \mathcal{F}_k] \leq |S||A|||w_k||_\infty^2 | \mathcal{F}_k| \leq \frac{16|S||A|}{(1 - \gamma)^2},
\]

where \( \mathcal{F}_k \) is a \( \sigma \)-field induced by \((s_i, a_i, s_{i+1}, r_i, Q_i)\) for all \( i \), the first inequality follows from \( ||w_k||_\infty \leq |S||A| ||w_k||_\infty \), and the last inequality is due to the bound on \( ||w_k||_\infty \). This completes the proof.

\[\square\]

APPENDIX R

PROOF OF LEMMA 21

Proof: We first prove the boundedness of \( Q^*_k - Q^* \) by an induction argument. For \( k = 0 \), we have

\[
||Q^*_0 - Q^*||_\infty \leq \frac{4}{(1 - \gamma)^2d_{\min}},
\]

because

\[
||Q^*_0 - Q^*||_\infty \leq ||Q^*_0||_\infty + ||Q^*||_\infty
\]

\[
\leq 1 + \frac{1}{1 - \gamma} = \frac{2 - \gamma}{1 - \gamma} \leq \frac{4}{(1 - \gamma)^2d_{\min}},
\]

where the last inequality follows from Lemma 6. Next, suppose that \( ||Q^*_k - Q^*||_\infty \leq \frac{4}{(1 - \gamma)^2d_{\min}} \) holds for some \( k \geq 0 \). Then, we have

\[
||Q^*_k - Q^*||_\infty
\leq ||A_{Q^*_k}||_\infty ||Q^*_k - Q^*||_\infty + \alpha_k ||w_k||_\infty
\]

\[
\leq (1 - d_{\min}(1 - \gamma))\alpha_k ||Q^*_k - Q^*||_\infty + \alpha_k \frac{4}{1 - \gamma}
\]

\[
\leq \frac{4}{(1 - \gamma)^2d_{\min}},
\]

where the first inequality follows from the update of lower comparison system (16), and the second inequality comes from the bound on \( w_k \) in Lemma 18 and \( A_{Q^*_k} \) in Lemma 8. The proof is completed by induction.

\[\square\]

APPENDIX S

PROOF OF LEMMA 22

Before proceeding, an important lemma is first established.

Lemma 35: \( w_k(Q) \) is Lipschitz with respect to \( Q \in \mathbb{R}^{|S||A|} \):

\[
||w_k(Q) - w_k(Q')||_\infty \leq 4||Q - Q'||_\infty,
\]

where \( Q' \in \mathbb{R}^{|S||A|} \).

Proof: The proof follows from simple algebraic inequalities:

\[
||w_k(Q) - w_k(Q')||_\infty
\leq ||[(\gamma (e_{a_k} \otimes e_{a_k} (e_{s_k+1})^T - \gamma D_{\infty} P) (\Pi Q - \Pi Q')) ||_\infty

\leq (1 + \gamma)(1 + d_{\min})||Q - Q'||_\infty
\leq 4||Q - Q'||_\infty,
\]

This completes the proof.

Now, we prove Lemma 22.

Proof: The proof simply uses boundedness of \( Q^*_k - Q^* \) as follows:

\[
||\xi_k(Q^*_k, Q_k)|| \leq ||A_{Q^*_k}(Q^*_k - Q^*)||_2 ||Gw_k||_2
\]

\[
\leq 4(|S||A|)\frac{1}{2} \frac{|S||A|}{(1 - \gamma)^2d_{\min}} \frac{4}{(1 - \gamma)^2d_{\min}}
\]

\[
\leq \frac{16(|S||A|)^2}{(1 - \gamma)}d_{\min},
\]

where the first inequality follows from Hölder’s inequality, the second inequality is due to the fact that \( ||A_{Q^*_k}||_2 \leq \sqrt{|S||A|} ||A_{Q^*_k}||_\infty \), bound on \( G \) in Lemma 11, and bound on \( w_k \) in Lemma 18.

Moreover, the Lipchitzness of \( \xi_k(Q^*_k, Q) \) can be also proved using boundedness of \( Q^*_k \) and \( w_k \) as follows:

\[
||\xi_k(Q^*_k, Q) - \xi_k(Q^*_k, Q')||
\leq ||(Q^*_k - Q^*)^T A_{Q^*_k} Gw_k(Q_k) - (Q^*_k - Q^*)^T A_{Q^*_k} Gw_k(Q')||
\leq ||(Q^*_k - Q^*)^T A_{Q^*_k} Gw_k(Q_k) - (Q^*_k - Q^*)^T A_{Q^*_k} Gw_k(Q')||
\]

\[
\leq \left( ||G A_{Q^*_k}(Q^*_k - Q^*)||_2 ||w_k(Q_k) - w_k(Q')||_2
\]

\[
\leq \left( ||G A_{Q^*_k}(Q^*_k - Q^*)||_2 \right)^2 \frac{16(|S||A|)^2}{(1 - \gamma)^2d_{\min}}
\]

\[
\leq \frac{16(|S||A|)^2}{(1 - \gamma)^2d_{\min}},
\]

\[\square\]
where the first inequality is due to the fact that $a_1^T b_1 - a_2^T b_2 = (a_1 - a_2)^T (b_1 + a_2^T (b_1 - b_2))$. The second inequality follows from Hölder’s inequality; the third inequality is due to the Lipschitzness of $w_Q$ in Lemma 35, and the last inequality comes from collecting bounds on each terms. This completes the proof.

□

APPENDIX T

PROOF OF LEMMA 23

Before we proceed on we introduce definitions and lemmas required to prove Lemma 23.

Definition 36 (Conditional independence [49]): For random variables $X, Y, Z$,

$$X \rightarrow Y \rightarrow Z$$

denotes that $X$ and $Z$ are conditionally independent given $Y$, that is, $P(Z | X, Y) = P(Z | Y)$.

Lemma 37 (Control of coupling [42]): For some $k \geq 0$, and $\tau > 0$, consider two random variables $X$ and $Y$ such that

$$X \rightarrow (s_k, a_k) \rightarrow (s_{k+\tau}, a_{k+\tau}) \rightarrow Y.$$  

Let the underlying Markov chain, $\{(s_k, a_k)\}_{k=0}^\infty$, satisfies Assumption 4. Moreover, let us construct independent random variables $X'$ and $Y'$ drawn from the marginal distributions of $X$ and $Y$, respectively, i.e., $X' \sim P(X = \cdot)$ and $Y' \sim P(Y = \cdot)$ so that

$$P(X' = \cdot, Y' = \cdot) = P(X' = \cdot)P(Y' = \cdot) = P(X = \cdot)P(Y = \cdot).$$

Then, for any function $f : X \times Y \rightarrow \mathbb{R}$ such that $\sup_{(x,y)\in X\times Y} |f(x,y)| < \infty$, we have

$$|E[f(X,Y)] - E[f(X',Y')]| \leq 2 \sup_{(x,y)\in X\times Y} |f(x,y)| (mp^\tau).$$

Proof: Now, we state the proof of Lemma 23. Using some simple decomposition, we can get

$$|E[\xi_k(Q^L_{k-r},Q_{k-r})]| = |E[\xi_k(Q^L_{k-r},Q_{k-r})] - E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$$

$$+ |E[\xi(Q^L_{k-r},Q_{k-r},o_k')]| \leq |E[\xi(Q^L_{k-r},Q_{k-r},o_k)] - E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$$

$$+ |E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|,$$

where the first inequality is due to triangle inequality. The first term in the above inequality can be bounded with the coupling lemma in Lemma 37, and the second term can be bounded with the help of independent coupling construction and ergodic assumption in Assumption 4.

First, we will bound the term in (51):

$$|E[\xi(Q^L_{k-r},Q_{k-r},o_k)] - E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$$

To apply Lemma 37, first note the relations

$$(Q^L_{k-r},Q_{k-r}) \rightarrow (s_{k-r},a_{k-r}) \rightarrow (s_k,a_k) \rightarrow o_k,$$

where the arrow notation is defined in Definition 36, and $o_k := (s_k, o_k, r_k+1, s_{k+1})$. Now, let us construct the coupling distribution of $(Q^L_{k-r},Q_{k-r})$ and $o_k'$ drawn independently from some marginal distributions of $(Q^L_{k-r},Q_{k-r})$ and $o_k$, respectively, i.e., $(Q^L_{k-r},Q_{k-r}) \sim \mathbb{P}((Q^L_{k-r},Q_{k-r}) = \cdot)$ and $o_k' \sim \mathbb{P}(o_k = \cdot)$. From the fact that $\xi_k$ is bounded, which is given in the first statement of Lemma 22, we can now apply Lemma 37, leading to

$$|E[\xi(Q^L_{k-r},Q_{k-r},o_k)] - E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$$

$$\leq \frac{32(|S||A|)^2}{(1-\gamma^2)d_{\text{min}}^2} m \rho^k.$$

Next, we will bound $|E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$ in (51) from the construction of coupling that $(Q^L_{k-r},Q_{k-r})$ and $o_k'$ are independent, and ergodic assumption from Assumption 4.

Using the tower property of expectation, i.e., $E[X] = E[E[X|Y]]$, convexity of $|\cdot|$, and Jensen’s inequality, we have

$$|E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$$

$$\leq E[|E[\xi(Q^L_{k-r},Q_{k-r},o_k)Q^L_{k-r},Q_{k-r})]|.$$

Now, our aim is to bound $|E[\xi(Q^L_{k-r},Q_{k-r},o_k')]|$ $Q^L_{k-r},Q_{k-r})$ in the last inequality. Denoting the noise term in (35) with the observation $o_k'$ as $w_q(Q^L_{k-r})$, we first bound the noise term. Let us define $D_{\text{uk}}$ to denote the state-action distribution at $k$-th time step as follows:

$$(e_x \times e_a)^T D_{\text{uk}}(e_x \times e_a) = \mu_k(s,a) = P(s_k = s, a_k = a|s)$$

where $\mu_k$ is defined in (7). The noise term $w_q(Q^L_{k-r})$ defined in (35) can be bounded as follows:

$$||E[w_q(Q^L_{k-r})Q^L_{k-r},Q_{k-r})]|$$

$$\leq ||D_{\text{uk}} - D_{\text{uk}}||_{\infty} ||R + \gamma \Pi_{Q_{k-r}} Q^L_{k-r},Q_{k-r})||_{\infty}$$

$$\leq \frac{2}{1-\gamma} ||D_{\text{uk}} - D_{\text{uk}}||_{\infty}$$

$$= \frac{2}{1-\gamma} \max_{(s,a) \in S \times A} |\mu_k(s,a) - \mu_{\infty}(s,a)|$$

$$= \frac{2}{1-\gamma} d_{TV}(\mu_k, \mu_{\infty})$$

where the first inequality follows from the fact that $o_k'$ is independent with $Q^L_{k-r}$, and the second inequality follows from Assumption 5 and Lemma 6. In particular, the first equality comes from the definition of the infinity norm, and the second equality follows from definition of the total variation distance in Definition 19. The last inequality follows from the ergodicity assumption of the Markov chain in Assumption 20. With the above result, we can now bound the quantity

$$|E[\xi(Q^L_{k-r},Q_{k-r},o_k')Q^L_{k-r},Q_{k-r})]|$$

149934

VOLUME 12, 2024

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
as follows:

\[
\| \mathbb{E}[\xi(Q_{k^t-t}, Q_{k^t-r}, o_{k^t})Q_{k^t-t}^L, Q_{k^t-r}] \| \\
\leq \|4[S]^2|A|^2 \frac{1}{(1-\gamma^2)} \| \mathbb{E}[w_k'(Q_{k^t-r})Q_{k^t-t}^L, Q_{k^t-r}] \| \| A \| \| A \|
\leq \frac{8[S]^2|A|^2}{(1-\gamma^4)} \|Q_{k^t-r}, Q_{k^t-r}, o_k'\| \| A \| \| A \|
\leq \frac{8[S]^2|A|^2}{(1-\gamma^4)} \| \mathbb{E}[\xi(Q_{k^t-r}, Q_{k^t-r}, o_k')] \| \| A \| \| A \| \| A \|
\]

where the first equality is due to the definition of the crossing term in (37). The first inequality is due to Hölder’s inequality, and the last inequality follows from Lemma 11 and Lemma 21 and applying (55). Combining the above result with (53) leads to

\[
\| \mathbb{E}[\xi(Q_{k^t-r}, Q_{k^t-r}, o_k')] \| \leq \frac{8[S]^2|A|^2}{(1-\gamma^4)} \| A \| \| A \| \| A \|
\]

Now, we are ready to finish the proof. Bounding each term in (51) with (52) and (56), we have

\[
\| \mathbb{E}[\xi(Q_{k^t-r}, Q_{k^t-r})] \| \leq \frac{40([S]|A|^2)^2}{(1-\gamma^4)} \| A \| \| A \| \| A \|
\]

This completes the proof. □

APPENDIX U
PROOF OF LEMMA 24

Proof: From Lipschitzness of \( \xi_k(Q_{k^t}, Q_k) \) in Lemma 22, we have

\[
\| \xi_k(Q_{k^t}, Q_k) - \xi_k(Q_{k^t-r}, Q_{k^t-r}) \| \leq \| \xi_k(Q_{k^t}, Q_k) - \xi_k(Q_{k^t-r}, Q_{k^t-r}) \| \\
\leq \| (|S| |A|^2)^2 \| Q_k - Q_{k^t-r} \| \| Q_k - Q_{k^t-r} \| \| A \| \| A \| \| A \| \\
\leq \| (|S| |A|^2)^2 \| Q_k - Q_{k^t-r} \| \| A \| \| A \| \| A \|
\]

In the sequel, \( \| Q_k - Q_{k^t-r} \| \| Q_{k^t-r} - Q_{k^t-r} \| \) will be bounded in terms of step-size. From the update of \( Q_{k^t} \) in (17), for \( i \geq 0 \), we have

\[
\| Q_{k^t+1} - Q_{k^t} \| \leq \| Q_i^t \| + \alpha_i T_0 Q_{k^t}^i + \alpha_i w_i \| \| A \| \| A \| \| A \| \| A \|
\]

Similarly, for \( Q_{i} \), we have

\[
\| Q_{i+1} - Q_{i} \| \leq \| Q_{i+1} \| + \alpha_i (e_i \otimes e_i) \| \| A \| \| A \| \| A \| \| A \|
\]

For any \( k \geq 1 \) and \( \tau \) such that \( 1 \leq \tau \leq k \), summing up (58) and (59) from \( i = k - \tau \) to \( k - 1 \), respectively, we get

\[
\| Q_{k^t} - Q_{k^t-r} \| \leq \sum_{i=k-\tau}^{k-1} \alpha_i \| Q_{i+1} - Q_{i} \| \| A \| \| A \| \\
\leq \frac{12}{1-\gamma} \| A \| \| A \| \| A \| \| A \|
\]

The first inequality follows from triangle inequality. The second inequality follows from (58). The last inequality follows from the fact that \( \alpha_k \) is decreasing sequence. Similarly, using (59), we have

\[
\| Q_k - Q_{k^t-r} \| \leq \sum_{i=k-\tau}^{k-1} \alpha_i \| Q_{i+1} - Q_{i} \| \| A \| \| A \| \\
\leq \frac{2}{1-\gamma} \| A \| \| A \| \| A \| \| A \|
\]

Therefore, combining the above results with (57) leads to

\[
\| \xi_k(Q_{k^t}, Q_k) - \xi_k(Q_{k^t-r}, Q_{k^t-r}) \| \leq \frac{25[S]^2|A|^2}{(1-\gamma^4)} \| Q_k - Q_{k^t-r} \| \| A \| \| A \| \| A \|
\]

Next, we will bound \( \mathbb{E}[\xi_k(Q_{k^t}, Q_k)] \). In particular, taking total expectations on (60) yields

\[
\mathbb{E}[\xi_k(Q_{k^t}, Q_k)] \leq \mathbb{E}[\xi_k(Q_{k^t-r}, Q_{k^t-r})] + \frac{25[S]^2|A|^2}{(1-\gamma^4)} \| A \| \| A \| \| A \|
\]

Since the above inequality is satisfied for any \( 1 \leq \tau \leq k \), for \( k \geq K_{mix} \), we can set \( \tau = \tau_{mix}(\alpha_k) \) from the definition.
of \( \tau^{\text{mix}}(\alpha_k) \). Now, we have

\[
\mathbb{E}[\xi_k(Q^k_\ell, Q_k)] \\
\leq \frac{40(|S||A|)^2}{(1 - \gamma)^4d_{\text{min}}^2}
\begin{align*}
\lambda^2 d_{\text{min}}^2
+ 25|S|^2|A|^2
\tau^{\text{mix}}(\alpha_k)\alpha_k - \tau^{\text{mix}}(\alpha_k) \\
\leq \left( \frac{40(|S||A|)^2}{(1 - \gamma)^4d_{\text{min}}^2}
\right)
\alpha_k
+ 25|S|^2|A|^2
\tau^{\text{mix}}(\alpha_k)\alpha_k - \tau^{\text{mix}}(\alpha_k) \\
\leq \left( \frac{40(|S||A|)^2}{(1 - \gamma)^4d_{\text{min}}^2}
\right)
\alpha_k
+ 25|S|^2|A|^2
\tau^{\text{mix}}(\alpha_k)\alpha_k - \tau^{\text{mix}}(\alpha_k),
\end{align*}
\]

where the first inequality follows from Lemma 23, the second inequality comes from the definition of mixing time in (38), and the last inequality follows from the fact that \( \alpha_k \) is a decreasing sequence.

\[
\text{Proof of Proposition 25}
\]

Proof: We will follow similar lines as in the proof of Proposition 13. For simplicity of the induction proof, let

\[
C_1 := \max \left\{ \frac{8|S|^4|A|^4}{(1 - \gamma)^3d_{\text{min}}^2}, \frac{256|S|^2|A|^2}{(1 - \gamma)^2d_{\text{min}}^2} \right\}.
\]

Expanding the term \( Q^k_{\ell + 1} \), we have the following relation:

\[
\mathbb{E}[V(z_{k+1})| F^k_{\ell}]
\leq \left( 1 - \frac{2}{k + \xi} \right) \mathbb{E}[V(Q^l)]
+ \frac{C_1}{(k + \xi)^2}
+ 2\alpha_k \mathbb{E}[\zeta_k(Q^l, Q_k)| F^k_{\ell}],
\]

\[
\mathbb{E}[\xi_k(Q^l, Q_k)]
\leq \frac{40(|S||A|)^2}{(1 - \gamma)^4d_{\text{min}}^2} \left( \lambda \rho^{\tau^{\text{mix}}(\alpha_k)} \right)
+ 25|S|^2|A|^2 \tau^{\text{mix}}(\alpha_k)\alpha_k - \tau^{\text{mix}}(\alpha_k),
\]

where \( F^k_{\ell} \) is a \( \sigma \)-field induced by \( (s_i, a_i, s_{i+1}, r_i, Q^l, Q_k) \) for \( i = 0 \), the first inequality follows from Lemma 12, and positive definiteness of \( G \), the second inequality follows from the step-size condition in Lemma 12, and collecting the bound on \( \lambda(G) \) in Lemma 11, \( w_k \) in Lemma 18, and \( \theta \) in Lemma 33. Taking total expectations, it follows that

\[
\mathbb{E}[V(Q^l_{\ell + 1})] \leq \left( 1 - \frac{2}{k + \xi} \right) \mathbb{E}[V(Q^l)]
+ \frac{C_1}{(k + \xi)^2}
+ 2\alpha_k \mathbb{E}[\zeta_k(Q^l, Q_k)].
\]

The main difference compared to (30) in Proposition 13 is that we have the non-zero term \( \mathbb{E}[\zeta_k(Q^l, Q_k)] \). Now, bounding

\[
\mathbb{E}[\xi_k(Q^l, Q_k)]
\leq \left( 1 - \frac{2}{k + \xi} \right) \mathbb{E}[V(Q^l)]
+ \frac{C_1}{(k + \xi)^2}
+ 2\alpha_k \mathbb{E}[\zeta_k(Q^l, Q_k)].
\]

\[
\mathbb{E}[V(Q^l_{\ell + 1})]
\leq \left( 1 - \frac{2}{k + \xi} \right) \mathbb{E}[V(Q^l)] + \frac{C_1}{(k + \xi)^2}
+ 2\theta \left( \frac{2}{k + \xi} \right) \mathbb{E}[\xi_k(Q^l, Q_k)]
\leq \left( 1 - \frac{2}{k + \xi} \right) \mathbb{E}[V(Q^l)] + \frac{C_1}{(k + \xi)^2}
+ 2\theta \left( \frac{2}{k + \xi} \right) \mathbb{E}[\xi_k(Q^l, Q_k)].
\]

\[
\text{APPENDIX W}
\]

\text{Proof of Theorem 26}

Proof: Using the spectral properties of \( G \) in Lemma 11, and from Proposition 25, we have

\[
\mathbb{E}[\|Q^l - Q^*\|^2] \leq \frac{1}{\lambda_{\text{min}}(G)} \left( \frac{C_2}{k + \xi} + \frac{C(|S||A|)^4}{(1 - \gamma)^3d_{\text{min}}^2} \frac{\tau^{\text{mix}}(\alpha_k)}{k + \xi} \right) \]

\[
\leq \left( \frac{2}{k + \xi} \right) \mathbb{E}[\xi_k(Q^l, Q_k)]
\leq \left( \frac{2}{k + \xi} \right) \mathbb{E}[\xi_k(Q^l, Q_k)].
\]

The main difference compared to (30) in Proposition 13 is that we have the non-zero term \( \mathbb{E}[\zeta_k(Q^l, Q_k)] \). Now, bounding
APPENDIX X
PROOF OF THEOREM 27

Proof: As in Theorem 16, we take expectation on (33), and use sub-multiplicativity of the norm.

\[ \mathbb{E}[||Q^U_{k+1} - Q^L_{k+1}||_\infty] \]

\[ \leq \left(1 - \frac{2}{k + \xi}\right) \mathbb{E}[||Q^U_k - Q^L_k||_\infty] \\
+ \frac{\theta}{k + \xi} \gamma d_{max}\mathbb{E}[||Q^L_k - Q^*||_\infty] \\
\leq \left(1 - \frac{2}{k + \xi}\right) \mathbb{E}[||Q^U_k - Q^L_k||_\infty] \\
+ \frac{2}{k + \xi} \gamma d_{max}\left(\frac{C^U}{\sqrt{k + \xi}} + \frac{C(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^4 d_{min}^3} \sqrt{\tau_{max}(\alpha_k)}\right). \]

where the first inequality is due to triangle inequality, and the last inequality follows from upper bound on \( \theta \) in Lemma 33 given in Appendix F. Applying the same induction in Theorem 16, we have

\[ \mathbb{E}[||Q^U_k - Q^L_k||_\infty] \leq \frac{C^U}{\sqrt{k + \xi}} + \frac{C(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^4 d_{min}^3} \sqrt{\tau_{max}(\alpha_k)}. \]

This completes the proof. \( \square \)

APPENDIX Y
PROOF OF THEOREM 28

Proof: Using triangle inequality, we can derive the upper bound when \( k \geq K_{mix} \).

\[ \mathbb{E}[||Q_k - Q^*||_\infty] \leq \mathbb{E}[||Q^U_k||_\infty] + \mathbb{E}[||Q^L_k - Q^*||_\infty] + \mathbb{E}[||Q^L_k - Q^L_k||_\infty] \]

\[ \leq \frac{C^*}{\sqrt{k + \xi}} + \frac{C(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^4 d_{min}^3} \sqrt{\tau_{max}(\alpha_k)}, \]

where the first inequality is due to triangle inequality, and the second inequality follows from combining the bound on the lower comparison system, Theorem 26 and the difference of the upper and lower comparison systems, Theorem 27.

APPENDIX Z
SAMPLE COMPLEXITY UNDER I.I.D. OBSERVATION MODELS

Lemma 38 (Sample Complexity With Respect to Expectation): To achieve \( \mathbb{E}[||Q_k - Q^*||_\infty] \leq \epsilon \), we need at most following number of samples:

\[ O\left(\frac{(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} \frac{1}{\epsilon^2}\right) \]

Proof: To achieve \( \mathbb{E}[||Q_k - Q^*||_\infty] \leq \epsilon \), one sufficient condition using the bound in Theorem 17 is, for some universal constant \( C \), let Then,

\[ \mathbb{E}[||Q_k - Q^*||_\infty] \leq \frac{C^*}{\sqrt{k + \xi}} \leq \epsilon, \]

which leads to

\[ k \geq \frac{1}{\epsilon^2} \max \left\{ \frac{C(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} ||Q^L_0||^2_2, \frac{C(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} \right\}. \]

This completes the proof.

Now, the convergence in expectation can lead to the following probability tail bound.

Lemma 39 (Sample Complexity With Respect to Probability Tail Bound): With probability at least \( 1 - \delta \), we can achieve \( ||Q_k - Q^*||_\infty \leq \epsilon \) with at most following number of samples:

\[ O\left(\frac{(|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} \frac{1}{\delta^2} \right). \]

Proof: For simplicity let

\[ C_t := C \max \left\{ \frac{|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} ||Q^L_0||^2_2, \frac{|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} \right\}. \]

for some universal constant \( C \). Applying Markov inequality to Theorem 17 yields

\[ \mathbb{P}(||Q_k - Q^*||_\infty \geq \epsilon) \leq \frac{\mathbb{E}[||Q_k - Q^*||_\infty]}{\epsilon} \leq \frac{1}{\epsilon} \frac{C_t}{\sqrt{k + \xi}}. \]

The corresponding complement of the event satisfies

\[ \mathbb{P}(||Q_k - Q^*||_\infty \leq \epsilon) = 1 - \mathbb{P}(||Q_k - Q^*||_\infty \geq \epsilon) \]

\[ \geq 1 - \frac{1}{\epsilon} \frac{C_t}{\sqrt{k + \xi}}, \]

where the last inequality follows from (61). Now, to achieve \( \mathbb{P}(||Q_k - Q^*||_\infty \leq \epsilon) \geq 1 - \delta \), we lower bound the above inequality with \( 1 - \delta \), which leads to

\[ \frac{1}{\epsilon} \frac{C_t}{\sqrt{k + \xi}} \leq \delta. \]

The above inequality results in

\[ k \geq \frac{C}{\epsilon \delta^2} \max \left\{ \frac{|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} ||Q^L_0||^2_2, \frac{|S||A|)^{\frac{13}{2}}}{(1 - \gamma)^2 d_{min}^3} \right\}. \]

This completes the proof. \( \square \)

A. SAMPLE COMPLEXITY UNDER MARKOVIAN NOISE

In this subsection, we derive sample complexity under Markovian noise. As preliminary results, some lemmas are introduced first.

Lemma 40: The mixing time, \( \tau_{mix}(\alpha_k) \), has the following bound:

\[ \tau_{mix}(\alpha_k) \leq \frac{1}{\log \frac{1}{\theta}} \log \left( \frac{m(k + \xi)}{\theta} \right) + 1. \]

VOLUME 12, 2024
Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Proof: By the definition of the mixing time in (38), \( \tau^{\text{mix}}(\alpha_k) \) satisfies
\[
m \rho^{\tau^{\text{mix}}(\alpha_k)} \leq \alpha_k.
\] (62)
Hence, simply rearranging the terms, we have the following inequalities:
\[
\tau^{\text{mix}}(\alpha_k) \log \rho + \log m \leq \log \alpha_k
\]
\[
1 - \frac{1}{\log \rho} \log \frac{m}{\alpha_k} \leq \tau^{\text{mix}}(\alpha_k)
\]
\[
1 - \frac{1}{\log \rho} \log \frac{m}{\alpha_k} = \frac{1}{\log \rho} \log \frac{m(k + \xi)}{\theta} \leq \tau^{\text{mix}}(\alpha_k),
\] (63)
where the first inequality follows by taking logarithm on both sides of (62), the second inequality is obtained by re-arranging the first inequality, and The last inequality is due to the definition of step-size \( \alpha_k \) in (23). From the definition of mixing time in (38), \( \tau^{\text{mix}}(\alpha_k) \) is the minimum integer achieving the bound (63). Therefore we have \( \tau^{\text{mix}}(\alpha_k) \leq \frac{1}{\log \rho} \log \frac{m(k + \xi)}{\theta} + 1 \), which concludes the proof. □

Without loss of generality, to keep simplicity, we assume
\[
\tau^{\text{mix}}(\alpha_k) \leq \frac{1}{\log \rho} \log \frac{m(k + \xi)}{\theta}.
\]

Lemma 41: To achieve \( \mathbb{E}[||Q_k - Q^*||_\infty] \leq \epsilon \), we need at most following number of samples:
\[
\tilde{O}\left(\left(\frac{||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}(\log \frac{1}{\rho})^2} \right).
\]
where \( \tilde{O}(\cdot) \) hides logarithmic factors.

Proof: Using the bound in Theorem 28, we need to achieve
\[
\mathbb{E}[||Q_k - Q^*||_\infty] \leq \frac{C^*}{\sqrt{k + \xi}} + \frac{C(||S|||A|\right)^{13}d_{\text{max}} \sqrt{\tau^{\text{mix}}(\alpha_k)}}{(1 - \gamma)^{16d_{\text{min}}^2}(\sqrt{k + \xi})} \leq \epsilon.
\]
A sufficient condition to achieve the bound is to make each term in Theorem 28 smaller than \( \frac{\epsilon}{2} \). Dosing so leads to the following two inequalities:
\[
\frac{C^*}{\sqrt{k + \xi}} \leq \frac{\epsilon}{2},
\]
\[
\frac{C(||S|||A|\right)^{13}d_{\text{max}} \sqrt{\tau^{\text{mix}}(\alpha_k)}}{(1 - \gamma)^{16d_{\text{min}}^2}(\sqrt{k + \xi})} \leq \epsilon.
\]
The first inequality leads to
\[
k \geq \frac{C}{\epsilon^2} \text{max} \left( \left(\frac{||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}}, \frac{C(||S|||A|\right)^{13}d_{\text{max}}}{(1 - \gamma)^{16d_{\text{min}}^2}} \right).
\] (64)
For the second inequality, we first bound the mixing time with a logarithmic term using Lemma 40, and use the fact that \( \log x \leq 2\sqrt{x} \). The second inequality can be achieved by bounding the upper bound
\[
\frac{C(||S|||A|\right)^{13}d_{\text{max}} \sqrt{\tau^{\text{mix}}(\alpha_k)}}{(1 - \gamma)^{16d_{\text{min}}^2}} \leq \frac{C(||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}} \frac{\log \frac{m}{\rho}}{2} \frac{1}{\sqrt{k + \xi}}
\]
\[
+ \frac{2C(||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}} \frac{1}{\sqrt{\log \frac{1}{\rho}}} \frac{1}{\sqrt{k + \xi}}.
\]
Next, bounding each term on the right side of the above inequality with \( \frac{\epsilon}{2} \) leads to
\[
k \geq \frac{C}{\epsilon^2} \text{max} \left( \left(\frac{||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}}, \frac{\log \frac{m}{\rho}}{2}, \frac{1}{\sqrt{\log \frac{1}{\rho}}} \frac{1}{\sqrt{k + \xi}} \right).
\] (65)
and
\[
k \geq \frac{C(||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}} \frac{1}{\epsilon^2}.
\] (66)
Collecting the largest coefficients for \( \frac{1}{\tau}, ||S||, |A|, \) and \( d_{\text{min}} \) in (64), (65), and (66), we get
\[
\tilde{O}\left(\left(\frac{||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}(\log \frac{1}{\rho})^2} \right).
\]
This completes the proof. □

Now, we derive sample complexity in terms of probability tail bounds using Markov inequality. The proof is similar to the i.i.d. case, and hence, it is omitted here.

Lemma 42: With probability at least \( 1 - \delta \), we can achieve \( ||Q^k - Q^*||_\infty \leq \epsilon \) with at most following number of samples:
\[
\tilde{O}\left(\left(\frac{||S|||A|\right)^{13}}{(1 - \gamma)^{16d_{\text{min}}^2}(\log \frac{1}{\rho})^2} \right).
\]

REFERENCES
[1] V. Mnih, “Human-level control through deep reinforcement learning,” Nature, vol. 518, pp. 529–533, Feb. 2015.
[2] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, “Recurrent experience replay in distributed reinforcement learning,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–9.
[3] J. Schrittwieser, I. Antongoulou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, “Mastering Atari, go, chess and shogi by planning with a learned model,” Nature, vol. 588, no. 7839, pp. 604–609, Dec. 2020.
[4] A. P. Badia, P. Sprechmann, A. Vitrivskiy, D. Guo, B. Piot, S. Kapturowski, O. Tieleman, M. Arjovsky, A. Pritzel, A. Bolt, and C. Blundell, “Never give up: Learning directed exploration strategies,” 2020, arXiv:2002.06038.
[5] A. P. Badia, “Agent57: Outperforming the Atari human benchmark,” in Proc. Int. Conf. Mach. Learn., Vienna, Austria, Jul. 2020, pp. 507–517.
[6] M. M. Afsar, T. Crump, and B. Far, “Reinforcement learning based recommender systems: A survey,” 2021, arXiv:2101.06286.
[7] J. Koher, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” Int. J. Robot. Res., vol. 32, no. 11, pp. 1238–1274, Sep. 2013.
[8] Z. Jiang and J. Liang, “Cryptocurrency portfolio management with deep reinforcement learning,” in Proc. Intell. Syst. Conf. (IntelliSys), Sep. 2017, pp. 905–913.
[9] C. J. C. H. Watkins and P. Dayan, “Q-learning,” Mach. Learn., vol. 8, pp. 279–292, May 1992.
