Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction

Gen Li, *Student Member, IEEE*, Yuting Wei, *Member, IEEE*, Yuejie Chi, *Senior Member, IEEE*, Yuantao Gu, *Senior Member, IEEE*, and Yuxin Chen, *Senior Member, IEEE*

***Abstract*— Asynchronous Q-learning aims to learn the optimal action-value function (or Q-function) of a Markov decision process (MDP), based on a single trajectory of Markovian samples induced by a behavior policy. Focusing on a γ-discounted MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, we demonstrate that the ℓ_∞-based sample complexity of classical asynchronous Q-learning — namely, the number of samples needed to yield an entrywise ε-accurate estimate of the Q-function — is at most on the order of $\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}$ up to some logarithmic factor, provided that a proper constant learning rate is adopted. Here, t_{mix} and $\mu_{\min}$ denote respectively the mixing time and the minimum state-action occupancy probability of the sample trajectory. The first term of this bound matches the sample complexity in the synchronous case with independent samples drawn from the stationary distribution of the trajectory. The second term reflects the cost taken for the empirical distribution of the Markovian trajectory to reach a steady state, which is incurred at the very beginning and becomes amortized as the algorithm runs. Encouragingly, the above bound improves upon the state-of-the-art result by a factor of at least $|\mathcal{S}||\mathcal{A}|$ for all scenarios, and by a factor of at least $t_{\mathsf{mix}}|\mathcal{S}||\mathcal{A}|$ for any sufficiently small accuracy level ε. Further, we demonstrate that the scaling on the effective horizon $\frac{1}{1-\gamma}$ can be improved by means of variance reduction.**

Manuscript received September 27, 2020; revised August 3, 2021; accepted October 6, 2021. Date of publication October 14, 2021; date of current version December 23, 2021. The work of Yuting Wei was supported in part by NSF under Grant CCF-2007911, Grant DMS-2015447, and Grant CCF-2106778. The work of Yuejie Chi was supported in part by the Office of Naval Research (ONR) under Grant N00014-18-1-2142 and Grant N00014-19-1-2404; in part by the Army Research Office (ARO) under Grant W911NF-18-1-0303; and in part by NSF under Grant CCF-1806154, Grant CCF-2007911, and Grant CCF-2106778. The work of Yuantao Gu was supported by NSFC under Grant 61971266. The work of Yuxin Chen was supported in part by the Air Force Office of Scientific Research (AFOSR) Young Investigator Program (YIP) Award under Grant FA9550-19-1-0030; in part by ONR under Grant N00014-19-1-2120; in part by the ARO YIP Award under Grant W911NF-20-1-0097 and Grant W911NF-18-1-0303; and in part by NSF under Grant CCF-2106739, Grant CCF-1907661, Grant IIS-1900140, and Grant IIS-2100158. *(Corresponding author: Yuting Wei.)*

Gen Li and Yuxin Chen are with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: genl@princeton.edu; yuxin.chen@princeton.edu).

Yuting Wei is with the Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: ytwei@wharton.upenn.edu).

Yuejie Chi is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Yuantao Gu is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Communicated by R. Venkataramanan, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3120096.

Digital Object Identifier 10.1109/TIT.2021.3120096

***Index Terms*— Model-free reinforcement learning, asynchronous Q-learning, Markovian samples, variance reduction, TD learning, mixing time.**

I. Introduction

MODEL-FREE algorithms such as Q-learning [3] play a central role in recent breakthroughs of reinforcement learning (RL) [4]. In contrast to model-based algorithms that decouple model estimation and planning, model-free algorithms attempt to directly interact with the environment — in the form of a policy that selects actions based on perceived states of the environment — from the collected data samples, without modeling the environment explicitly. Therefore, model-free algorithms are able to process data in an online fashion and are often memory-efficient. Understanding and improving the sample efficiency of model-free algorithms lie at the core of recent research activity [5], whose importance is particularly evident for the class of RL applications in which data collection is costly and time-consuming (such as clinical trials, online advertisements, and so on).

The current paper concentrates on Q-learning, an off-policy model-free algorithm that seeks to learn the optimal action-value function by observing what happens under a behavior policy. The off-policy feature makes it appealing in various RL applications where it is infeasible to change the policy under evaluation on the fly. There are two basic update models in Q-learning. The first one is termed a *synchronous* setting, which hypothesizes on the existence of a simulator (also called a generative model); at each time, the simulator generates an independent sample for every state-action pair, and the estimates are updated simultaneously across all state-action pairs. The second model concerns an *asynchronous* setting, where only a single sample trajectory following a behavior policy is accessible; at each time, the algorithm updates its estimate of a single state-action pair using one state transition from the trajectory. Obviously, understanding the asynchronous setting is considerably more challenging than the synchronous model, due to the Markovian (and hence non-i.i.d.) nature of its sampling process.

Focusing on an infinite-horizon Markov decision process (MDP) with state space $\mathcal{S}$ and action space $\mathcal{A}$, this work investigates asynchronous Q-learning on a single *Markovian trajectory* induced by a behavior policy. We ask a fundamental question:

How many samples are needed for asynchronous Q-learning to learn the optimal Q-function?

Despite a considerable number of prior works analyzing this algorithm (ranging from the classical works [6], [7] to the very recent paper [2]), it remains unclear whether existing sample complexity analysis of asynchronous Q-learning is tight. As we shall elucidate momentarily, there exists a large gap — at least as large as $|\mathcal{S}||\mathcal{A}|$ — between the state-of-the-art sample complexity bound for asynchronous Q-learning [2] and the one derived for the synchronous counterpart [8]. This raises a natural desire to examine whether there is any bottleneck intrinsic to the asynchronous setting that significantly limits its performance.

A. Main Contributions

This paper develops a refined analysis framework that sharpens our understanding about the sample efficiency of classical asynchronous Q-learning on a single sample trajectory. Setting the stage, consider an infinite-horizon MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, and a discount factor $\gamma \in (0,1)$. What we have access to is a sample trajectory of the MDP induced by a stationary behavior policy. In contrast to the synchronous setting with i.i.d. samples, we single out two parameters intrinsic to the Markovian sample trajectory: *(i)* the mixing time t_{mix}, which characterizes how fast the trajectory disentangles itself from the initial state; *(ii)* the smallest state-action occupancy probability μ_{min} of the stationary distribution of the trajectory, which captures how frequent each state-action pair has been at least visited.

With these parameters in place, our findings unveil that: the sample complexity required for asynchronous Q-learning to yield an ε-optimal Q-function estimate — in a strong ℓ_∞ sense — is at most[1]

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}+\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big). \tag{1}$$

The first component of (1) is consistent with the sample complexity derived for the setting with independent samples drawn from the stationary distribution of the trajectory [8]. In comparison, the second term of (1) — which is unaffected by the accuracy level ε — is intrinsic to the Markovian nature of the trajectory; in essence, this term reflects the cost taken for the empirical distribution of the sample trajectory to converge to a steady state, and becomes amortized as the algorithm runs. In other words, the behavior of asynchronous Q-learning would resemble what happens in the setting with independent samples, as long as the algorithm has been run for reasonably long. In addition, our analysis framework readily yields another sample complexity bound

$$\widetilde{O}\Big(\frac{t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}\Big), \tag{2}$$

where t_{cover} stands for the cover time — namely, the time taken for the trajectory to visit all state-action pairs at least once. This facilitates comparisons with several prior results based on the cover time.

Furthermore, we leverage the idea of variance reduction to improve the scaling with the discount complexity $\frac{1}{1-\gamma}$. We demonstrate that a variance-reduced variant of asynchronous Q-learning attains ε-accuracy using at most

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\min\{1,\varepsilon^2\}}+\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big) \tag{3}$$

samples, matching the complexity of its synchronous counterpart if $\varepsilon \le \min\big\{1, \frac{1}{(1-\gamma)\sqrt{t_{\mathsf{mix}}}}\big\}$ [12]. Moreover, by taking the action space to be a singleton set, the aforementioned results immediately lead to ℓ_∞-based sample complexity guarantees for temporal difference (TD) learning [13] on Markovian samples.

Comparisons with past results. A large fraction of the classical literature focused on asymptotic convergence analysis of asynchronous Q-learning (e.g. [6], [7], [14]); these results, however, did not lead to non-asymptotic sample complexity bounds. The state-of-the-art sample complexity analysis was due to the recent work [2], which derived a sample complexity bound $\widetilde{O}\big(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}\big)$. Given the obvious lower bound $1/\mu_{\mathsf{min}} \ge |\mathcal{S}||\mathcal{A}|$, our result (1) improves upon that of [2] by a factor at least on the order of $|\mathcal{S}||\mathcal{A}|\min\big\{t_{\mathsf{mix}}, \frac{1}{(1-\gamma)^4\varepsilon^2}\big\}$. In particular, for sufficiently small accuracy level ε, our improvement exceeds a factor of at least

$$t_{\mathsf{mix}}|\mathcal{S}||\mathcal{A}|.$$

In addition, we note that several prior works [9], [10] developed sample complexity bounds in terms of the cover time t_{cover} of the sample trajectory; our result strengthens these bounds by a factor of at least

$$t_{\mathsf{cover}}^2|\mathcal{S}||\mathcal{A}| \ge |\mathcal{S}|^3|\mathcal{A}|^3.$$

The interested reader is referred to Table I for more precise comparisons, and to Section V for a discussion of further related works.

B. Paper Organization, Notation, and Basic Concept

The remainder of the paper is organized as follows. Section II formulates the problem and introduces some basic quantities and assumptions. Section III presents the asynchronous Q-learning algorithm along with its theoretical guarantees, whereas Section IV accommodates the extension: asynchronous variance-reduced Q-learning. A more detailed account of related works is given in Section V. The analyses of our main theorems are described in Sections VI-IX. We conclude this paper with a summary of our results and a list of future directions in Section X. Several preliminary facts about Markov chains and the proofs of technical lemmas are postponed to the appendix.

Next, we introduce a set of notation that will be used throughout the paper. Denote by $\Delta(\mathcal{S})$ (resp. $\Delta(\mathcal{A})$) the probability simplex over the set $\mathcal{S}$ (resp. $\mathcal{A}$). For any vector $\boldsymbol{z} = [z_i]_{1\le i\le n} \in \mathbb{R}^n$, we overload the notation $\sqrt{\cdot}$ and $|\cdot|$ to denote entry-wise operations, such that $\sqrt{\boldsymbol{z}} := [\sqrt{z_i}]_{1\le i\le n}$ and $|\boldsymbol{z}| := [|z_i|]_{1\le i\le n}$. For any vectors $\boldsymbol{z} = [a_i]_{1\le i\le n}$ and

[1]Let $\mathcal{X} := (|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \frac{1}{\varepsilon})$. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ means there exists a universal constant $C_1 > 0$ such that $f \le C_1\, g$. The notation $\widetilde{O}(\cdot)$ is defined analogously except that it hides any logarithmic factor.

TABLE I

SAMPLE COMPLEXITY OF ASYNCHRONOUS Q-LEARNING AND ITS VARIANTS TO COMPUTE AN ε-OPTIMAL Q-FUNCTION IN THE ℓ_∞ NORM, WHERE WE HIDE ALL LOGARITHMIC FACTORS. WITH REGARDS TO THE MARKOVIAN TRAJECTORY INDUCED BY THE BEHAVIOR POLICY, WE DENOTE BY t_{cover}, t_{mix}, AND $\mu_{\min}$ THE COVER TIME, MIXING TIME, AND MINIMUM STATE-ACTION OCCUPANCY PROBABILITY OF THE ASSOCIATED STATIONARY DISTRIBUTION, RESPECTIVELY

Algorithm	Sample complexity	Learning rate
Asynchronous Q-learning Even-Dar and Mansour, 2003 [9]	$\frac{(t_{\mathsf{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$	linear: $\frac{1}{t}$
Asynchronous Q-learning Even-Dar and Mansour, 2003 [9]	$\left(\frac{t_{\mathsf{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\right)^{\frac{1}{\omega}} + \left(\frac{t_{\mathsf{cover}}}{1-\gamma}\right)^{\frac{1}{1-\omega}}$	polynomial: $\frac{1}{t^\omega}$, $\omega \in (\frac{1}{2}, 1)$
Asynchronous Q-learning Beck and Srikant, 2012 [10]	$\frac{t_{\mathsf{cover}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$	constant: $\frac{(1-\gamma)^4\varepsilon^2}{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\mathsf{cover}}^2}$
Asynchronous Q-learning Qu and Wierman, 2020 [2]	$\frac{t_{\mathsf{mix}}}{\mu_{\min}^2(1-\gamma)^5\varepsilon^2}$	rescaled linear: $\frac{\frac{1}{\mu_{\min}(1-\gamma)}}{t+\max\{\frac{1}{\mu_{\min}(1-\gamma)}, t_{\mathsf{mix}}\}}$
Speedy Q-learning Azar et al., 2011 [11]	$\frac{t_{\mathsf{cover}}}{(1-\gamma)^4\varepsilon^2}$	rescaled linear: $\frac{1}{t+1}$
Asynchronous Q-learning **This work (Theorem 1)**	$\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}$	constant: $\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$
Asynchronous Q-learning **This work (Theorem 2)**	$\frac{t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}$	constant: $\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, 1\right\}$
Asynchronous Q-learning **This work (Theorem 3)**	$\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}$	piecewise constant rescaled linear: (23)
Variance-reduced Q-learning **This work (Theorem 4)**	$\frac{1}{\mu_{\min}(1-\gamma)^3\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}$	constant: $\min\left\{\frac{(1-\gamma)^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$

$\boldsymbol{w} = [w_i]_{1\le i\le n}$, the notation $\boldsymbol{z} \ge \boldsymbol{w}$ (resp. $\boldsymbol{z} \le \boldsymbol{w}$) means $z_i \ge w_i$ (resp. $z_i \le w_i$) for all $1 \le i \le n$. Additionally, we denote by $\mathbf{1}$ the all-one vector, $\boldsymbol{I}$ the identity matrix, and $\mathbb{1}\{\cdot\}$ the indicator function. For any matrix $\boldsymbol{P} = [P_{ij}]$, we denote $\|\boldsymbol{P}\|_1 := \max_i \sum_j |P_{ij}|$. Throughout this paper, we use $c, c_0, c_1, \cdots$ to denote universal constants that do not depend either on the parameters of the MDP or the target levels (ε, δ), and their exact values may change from line to line.

Finally, let us introduce the concept of uniform ergodicity for Markov chains. Consider any Markov chain $(X_0, X_1, X_2, \cdots)$ with transition kernel P, finite state space $\mathcal{X}$ and stationary distribution μ, and denote by $P^t(\cdot \,|\, x)$ the distribution of X_t conditioned on $X_0 = x \in \mathcal{X}$. This Markov chain is said to be *uniformly ergodic* if, for some $\rho < 1$ and $M < \infty$, one has

$$\sup_{x\in\mathcal{X}} d_{\mathsf{TV}}\big(\mu, P^t(\cdot \,|\, x)\big) \le M\rho^t, \tag{4}$$

where $d_{\mathsf{TV}}(\mu, \nu)$ stands for the total variation distance between two distributions μ and ν [15]:

$$d_{\mathsf{TV}}(\mu,\nu) := \frac{1}{2}\sum_{x\in\mathcal{X}} \big|\mu(x) - \nu(x)\big| = \sup_{A\subseteq\mathcal{X}} \big|\mu(A) - \nu(A)\big|. \tag{5}$$

II. MODELS AND BACKGROUND

This paper studies an infinite-horizon MDP with discounted rewards, as represented by a quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote respectively the (finite) state space and action space, whereas $\gamma \in (0, 1)$ indicates the discount factor. Particular emphasis is placed on the scenario with large state/action space and long effective horizon, namely, $|\mathcal{S}|$, $|\mathcal{A}|$ and the effective horizon $\frac{1}{1-\gamma}$ can all be quite large. We use $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ to represent the probability transition kernel of the MDP, where for each state-action pair $(s, a) \in \mathcal{S}\times\mathcal{A}$, $P(s' \,|\, s, a)$ denotes the probability of transiting to state s' from state s when action a is executed. The reward function is represented by $r : \mathcal{S}\times\mathcal{A} \to [0, 1]$, such that $r(s, a)$ denotes the immediate reward from state s when action a is taken; for simplicity, we assume throughout that all rewards lie within $[0, 1]$. We focus on the tabular setting which, despite its basic form, has not yet been well understood. See [16] for an in-depth introduction of this model.

A. Q-Function and Bellman Operator

An action selection rule is termed a *policy* and represented by a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which maps a state to a distribution over the set of actions. A policy is said to be *stationary* if it is time-invariant. We denote by $\{s_t, a_t, r_t\}_{t=0}^{\infty}$ a sample trajectory, where s_t (resp. a_t) denotes the state (resp. the action taken) at time t, and $r_t = r(s_t, a_t)$ denotes the reward received at time t. It is assumed throughout that the rewards are deterministic and depend solely upon the current state-action pair. We denote by $V^\pi : \mathcal{S} \to \mathbb{R}$ the value function of a policy π, namely,

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s\right],$$

which is the expected discounted cumulative reward received when *(i)* the initial state is $s_0 = s$, *(ii)* the actions are taken

based on the policy π (namely, $a_t \sim \pi(s_t)$ for all $t \geq 0$) and the trajectory is generated based on the transition kernel (namely, $s_{t+1} \sim P(\cdot|s_t, a_t)$). It can be easily verified that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$ for any π. The action-value function (also Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π is defined for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ by

$$Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a\right],$$

where the actions are taken according to the policy π except the initial action (i.e. $a_t \sim \pi(s_t)$ for all $t \geq 1$). As is well-known, there exists an optimal policy — denoted by $\pi^\star$ — that simultaneously maximizes $V^\pi(s)$ and $Q^\pi(s, a)$ uniformly over all state-action pairs $(s, a) \in (\mathcal{S} \times \mathcal{A})$. Here and throughout, we shall denote by $V^\star := V^{\pi^\star}$ and $Q^\star := Q^{\pi^\star}$ the optimal value function and the optimal Q-function, respectively.

In addition, the Bellman operator $\mathcal{T}$, which is a mapping from $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to itself, is defined such that the (s, a)-th entry of $\mathcal{T}(Q)$ is given by

$$\mathcal{T}(Q)(s,a) := r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\max_{a' \in \mathcal{A}} Q(s', a')\Big]. \quad (6)$$

It is well known that the optimal Q-function $Q^\star$ is the unique fixed point of the Bellman operator.

B. Sample Trajectory and Behavior Policy

Imagine we have access to a sample trajectory $\{s_t, a_t, r_t\}_{t=0}^\infty$ generated by the MDP $\mathcal{M}$ under a given stationary policy π_{b} — called a *behavior policy*. The behavior policy is deployed to help one learn the "behavior" of the MDP under consideration, which often differs from the optimal policy being sought. Given the stationarity of π_{b}, the sample trajectory can be viewed as a sample path of a time-homogeneous Markov chain over the set of state-action pairs $\{(s, a) \mid s \in \mathcal{S}, a \in \mathcal{A}\}$. Throughout this paper, we impose the following uniform ergodicity assumption [17] (see the definition of uniform ergodicity in Section I-B).

Assumption 1: The Markov chain induced by the stationary behavior policy π_{b} is uniformly ergodic.

There are several properties concerning the behavior policy and its resulting Markov chain that play a crucial role in learning the optimal Q-function. Specifically, denote by $\mu_{\pi_{\mathsf{b}}}$ the stationary distribution (over all state-action pairs) of the aforementioned behavior Markov chain, and define

$$\mu_{\mathsf{min}} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_{\pi_{\mathsf{b}}}(s,a). \quad (7)$$

Intuitively, μ_{min} reflects an information bottleneck; that is, the smaller μ_{min} is, the more samples are needed in order to ensure all state-action pairs are visited sufficiently many times. In addition, we define the associated mixing time of the chain as

$$t_{\mathsf{mix}} := \min\Big\{t \;\Big|\; \max_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} d_{\mathsf{TV}}\big(P^t(\cdot \mid s_0, a_0), \mu_{\pi_{\mathsf{b}}}\big) \leq \frac{1}{4}\Big\}, \quad (8)$$

where $P^t(\cdot|s_0, a_0)$ denotes the distribution of (s_t, a_t) conditional on the initial state-action pair (s_0, a_0), and $d_{\mathsf{TV}}(\mu, \nu)$ is the total variation distance between μ and ν (see (5)). In words, the mixing time t_{mix} captures how fast the sample trajectory decorrelates from its initial state. Moreover, we define the cover time associated with this Markov chain as follows

$$t_{\mathsf{cover}} := \min\Big\{t \mid \min_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\big(\mathcal{B}_t \,|\, s_0, a_0\big) \geq \frac{1}{2}\Big\}, \quad (9)$$

where $\mathcal{B}_t$ denotes the event such that all $(s, a) \in \mathcal{S} \times \mathcal{A}$ have been visited at least once between time 0 and time t, and $\mathbb{P}\big(\mathcal{B}_t \,|\, s_0, a_0\big)$ denotes the probability of $\mathcal{B}_t$ conditional on the initial state (s_0, a_0).

Remark 1: It is known that for a finite-state Markov chain, having a finite mixing time t_{mix} implies uniform ergodicity of the chain [17, Page 4]. Thus, our uniform ergodicity assumption is equivalent to the assumption imposed in [2] (which assumes ergodicity in addition to a finite t_{mix}).

C. Goal

Given *a single* sample trajectory $\{s_t, a_t, r_t\}_{t=0}^\infty$ generated by the behavior policy π_{b}, we aim to compute/approximate the optimal Q-function $Q^\star$ in an ℓ_∞ sense. This setting — in which a state-action pair can be updated only when the Markovian trajectory reaches it — is commonly referred to as *asynchronous* Q-learning [2], [6] in tabular RL. The current paper focuses on characterizing, in a non-asymptotic manner, the sample efficiency of classical Q-learning and its variance-reduced variant.

III. Asynchronous Q-Learning on a Single Markovian Trajectory

A. Algorithm

The Q-learning algorithm [3] is arguably one of the most famous off-policy algorithms aimed at learning the optimal Q-function. Given the Markovian trajectory $\{s_t, a_t, r_t\}_{t=0}^\infty$ generated by the behavior policy π_{b}, the asynchronous Q-learning algorithm maintains a Q-function estimate $Q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at each time t and adopts the following iterative update rule

$$\begin{aligned} Q_t(s_{t-1}, a_{t-1}) &= (1-\eta_t) Q_{t-1}(s_{t-1}, a_{t-1}) \\ &\quad + \eta_t \mathcal{T}_t(Q_{t-1})(s_{t-1}, a_{t-1}) \\ Q_t(s,a) &= Q_{t-1}(s,a), \quad \forall (s,a) \neq (s_{t-1}, a_{t-1}) \end{aligned} \quad (10)$$

for any $t \geq 0$, whereas η_t denotes the learning rate or the stepsize. Here, $\mathcal{T}_t$ denotes the empirical Bellman operator w.r.t. the t-th sample, that is,

$$\mathcal{T}_t(Q)\ (s_{t-1}, a_{t-1}) := r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'). \quad (11)$$

It is worth emphasizing that at each time t, only a single entry — the one corresponding to the sampled state-action pair (s_{t-1}, a_{t-1}) — is updated, with all remaining entries unaltered. While the estimate Q_0 can be initialized to arbitrary values, we shall set $Q_0(s, a) = 0$ for all (s, a) unless otherwise noted. The corresponding value function estimate $V_t : \mathcal{S} \to \mathbb{R}$ at time t is thus given by

$$\forall s \in \mathcal{S}: \qquad V_t(s) := \max_{a \in \mathcal{A}} Q_t(s,a). \quad (12)$$

The complete algorithm is described in Algorithm 1.

Algorithm 1: Asynchronous Q-Learning

input parameters: learning rates $\{\eta_t\}$, number of iterations T.
initialization: $Q_0 = 0$.
for $t = 1, 2, \cdots, T$ **do**
  Draw action $a_{t-1} \sim \pi_{\mathsf{b}}(s_{t-1})$, observe reward $r(s_{t-1}, a_{t-1})$, and draw next state $s_t \sim P(\cdot \,|\, s_{t-1}, a_{t-1})$.
  Update Q_t according to (10).

B. Theoretical Guarantees for Asynchronous Q-Learning

We are in a position to present our main theory regarding the non-asymptotic sample complexity of asynchronous Q-learning, for which the key parameters μ_{min} and t_{mix} defined respectively in (7) and (8) play a vital role. The proof of this result is provided in Section VI.

Theorem 1 (Asynchronous Q-Learning): For the asynchronous Q-learning algorithm detailed in Algorithm 1, there exist some universal constants $c_0, c_1 > 0$ such that for any $0 < \delta < 1$ and $0 < \varepsilon \le \frac{1}{1-\gamma}$, one has

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad |Q_T(s,a) - Q^\star(s,a)| \le \varepsilon$$

with probability at least $1-\delta$, provided that the iteration number T and the learning rates $\eta_t \equiv \eta$ obey

$$T \ge \frac{c_0}{\mu_{\mathsf{min}}} \left\{ \frac{1}{(1-\gamma)^5 \varepsilon^2} + \frac{t_{\mathsf{mix}}}{1-\gamma} \right\} \cdot \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big) \log\Big(\frac{1}{(1-\gamma)^2 \varepsilon}\Big), \tag{13a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)} \min\left\{ \frac{(1-\gamma)^4 \varepsilon^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}} \right\}. \tag{13b}$$

Remark 2: The careful reader might immediately remark that the learning rate η studied in Theorem 1 relies on prior knowledge of ε, δ and T. This is more stringent than the learning rates in [2], which do not require pre-determining these parameters. To address this issue, we will explore a more adaptive learning rate schedule shortly in Section III-D, which achieves the same sample complexity without the need of knowing these parameters *a priori*.

Theorem 1 delivers a finite-sample/finite-time analysis of asynchronous Q-learning, given that a fixed learning rate is adopted and chosen appropriately. The ℓ_∞-based sample complexity required for Algorithm 1 to attain ε accuracy is at most

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5 \varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big). \tag{14}$$

A few implications are in order.

1) Dependency on the Minimum State-Action Occupancy Probability μ_{min}: Our sample complexity bound (14) scales linearly in $1/\mu_{\mathsf{min}}$, which is in general unimprovable. Consider, for instance, the ideal scenario where state-action occupancy is nearly uniform across all state-action pairs, in which case $1/\mu_{\mathsf{min}}$ is on the order of $|\mathcal{S}||\mathcal{A}|$. In such a "near-uniform" case, the sample complexity scales linearly with $|\mathcal{S}||\mathcal{A}|$, and this dependency matches the known minimax lower bound [18] derived for the setting with independent samples. In comparison, [2, Theorem 7] depends at least quadratically on $1/\mu_{\mathsf{min}}$, which is at least $|\mathcal{S}||\mathcal{A}|$ times larger than our result (14).

2) Dependency on the Effective Horizon $\frac{1}{1-\gamma}$: The sample size bound (14) scales as $\frac{1}{(1-\gamma)^5\varepsilon^2}$, which coincides with both [8], [19] (for the synchronous setting) and [2], [10] (for the asynchronous setting) with either a rescaled linear learning rate or a constant learning rate. This turns out to be the sharpest scaling known to date for the classical form of Q-learning.

3) Dependency on the Mixing Time t_{mix}: The second additive term of our sample complexity (14) depends linearly on the mixing time t_{mix} and is (almost) independent of the target accuracy ε. The influence of this mixing term is a consequence of the expense taken for the Markovian trajectory to reach a steady state, which is a one-time cost that can be amortized over later iterations if the algorithm is run for reasonably long. Put another way, if the behavior chain mixes not too slowly with respect to ε (in the sense that $t_{\mathsf{mix}} \le \frac{1}{(1-\gamma)^4\varepsilon^2}$), then the algorithm behaves as if the samples were independently drawn from the stationary distribution of the trajectory. In comparison, the influences of t_{mix} and $\frac{1}{(1-\gamma)^5\varepsilon^2}$ in [2] (cf. Table I) are multiplicative regardless of the value of ε, thus resulting in a much higher sample complexity. For instance, if $\varepsilon = O\big(\frac{1}{(1-\gamma)^2\sqrt{t_{\mathsf{mix}}}}\big)$, then the sample complexity result therein is at least

$$\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \ge t_{\mathsf{mix}}|\mathcal{S}||\mathcal{A}|$$

times larger than our result (modulo some log factor).

4) Schedule of Learning Rates: An interesting aspect of our analysis lies in the adoption of a time-invariant learning rate, under which the ℓ_∞ error decays linearly — down to some error floor whose value is dictated by the learning rate. Therefore, a desired statistical accuracy can be achieved by properly setting the learning rate based on the target accuracy level ε and then determining the sample complexity accordingly. In comparison, classical analyses typically adopted a (rescaled) linear or a polynomial learning rule [2], [9]. While the work [10] studied Q-learning with a constant learning rate, their bounds were conservative and fell short of revealing the optimal scaling. Furthermore, we note that adopting time-invariant learning rates is not the only option that enables the advertised sample complexity; as we shall elucidate in Section III-D, one can also adopt carefully designed diminishing learning rates to achieve the same performance guarantees.

5) Mean Estimation Error: The high-probability bound in Theorem 1 readily translates to a mean estimation error guarantee. To see this, let us first make note of the following basic crude bound (see e.g. [10], [20])

$$\big|Q_t(s,a)\big| \le \frac{1}{1-\gamma}, \quad \big|Q_t(s,a) - Q^\star(s,a)\big| \le \frac{1}{1-\gamma} \tag{15}$$

for all $t \geq 0$ and all $(s,a) \in \mathcal{S} \times \mathcal{A}$. By taking $\delta = \varepsilon(1-\gamma)$ in Theorem 1, we immediately reach

$$\mathbb{E}\Big[\max_{s,a}\big|Q_T(s,a) - Q^\star(s,a)\big|\Big] \leq \varepsilon(1-\delta) + \delta\frac{1}{1-\gamma} \leq 2\varepsilon, \tag{16}$$

provided that T obeys (13a). As a result, the sample complexity remains unchanged (up to some logarithmic factor) when the goal is to achieve the mean error bound $\mathbb{E}\big[\max_{s,a}\big|Q_T(s,a) - Q^\star(s,a)\big|\big] \leq 2\varepsilon$.

In addition, our analysis framework immediately leads to another sample complexity guarantee stated in terms of the cover time t_{cover} (cf. (9)), which facilitates comparisons with several past work [9], [10]. The proof follows essentially that of Theorem 1, with a sketch provided in Section VII.

Theorem 2: For the asynchronous Q-learning algorithm detailed in Algorithm 1, there exist some universal constants $c_0, c_1 > 0$ such that for any $0 < \delta < 1$ and $0 < \varepsilon \leq \frac{1}{1-\gamma}$, one has

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad |Q_T(s,a) - Q^\star(s,a)| \leq \varepsilon$$

with probability at least $1-\delta$, provided that the iteration number T and the learning rates $\eta_t \equiv \eta$ obey

$$T \geq \frac{c_0 t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}\log^2\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)\log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big), \tag{17a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2},\ 1\right\}. \tag{17b}$$

Remark 3: The main difference between the cover-time-based analysis and the mixing-time-based analysis lies in the number of visits to each state-action pair (s,a) in every time frame. Owing to the measure concentration of Markov chains, we can see that the number of visits to each (s,a) concentrates around its expected value in each time frame, which in turn ensures that all state-action pairs have been visited at least once as long as the time frame is sufficiently long. This important property allows one to establish an intimate connection between the analysis of Theorem 1 and that of Theorem 2.

In a nutshell, this theorem tells us that the ℓ_∞-based sample complexity of classical asynchronous Q-learning is bounded above by

$$\widetilde{O}\Big(\frac{t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}\Big), \tag{18}$$

which scales linearly with the cover time. This improves upon the prior result [9] (resp. [10]) by an order of at least

$$t_{\mathsf{cover}}^{3.29} \geq |\mathcal{S}|^{3.29}|\mathcal{A}|^{3.29} \qquad (\text{resp. } t_{\mathsf{cover}}^2|\mathcal{S}||\mathcal{A}| \geq |\mathcal{S}|^3|\mathcal{A}|^3).$$

See Table I for detailed comparisons. We shall further make note of some connections between t_{cover} and $t_{\mathsf{mix}}/\mu_{\mathsf{min}}$ to help compare Theorem 1 and Theorem 2: *(i)* in general, $t_{\mathsf{cover}} = \widetilde{O}(t_{\mathsf{mix}}/\mu_{\mathsf{min}})$ for uniformly ergodic chains; *(ii)* one can find some cases where $t_{\mathsf{mix}}/\mu_{\mathsf{min}} = \widetilde{O}(t_{\mathsf{cover}})$. Consequently, while Theorem 1 does not strictly dominate Theorem 2 in all instances, the aforementioned connections reveal that Theorem 1 is tighter for the worst-case scenarios. The interested reader is referred to Section B for details.

C. A Special Case: TD Learning

In the special circumstance that the set of allowable actions $\mathcal{A}$ is a singleton, the corresponding MDP reduces to a Markov reward process (MRP), where the state transition kernel $P: \mathcal{S} \to \Delta(\mathcal{S})$ describes the probability of transitioning between different states, and $r: \mathcal{S} \to [0,1]$ denotes the reward function (so that $r(s)$ is the immediate reward in state s). The goal is to estimate the value function $V: \mathcal{S} \to \mathbb{R}$ from the trajectory $\{s_t, r_t\}_{t=0}^\infty$, which arises commonly in the task of policy evaluation for a given deterministic policy.

The Q-learning procedure in this special setting reduces to the well-known TD learning algorithm, which maintains an estimate $V_t: \mathcal{S} \to \mathbb{R}$ at each time t and proceeds according to the following iterative update[2]

$$\begin{aligned} V_t(s_{t-1}) &= (1-\eta_t)V_{t-1}(s_{t-1}) + \eta_t\left(r(s_{t-1}) + \gamma V_{t-1}(s_t)\right), \\ V_t(s) &= V_{t-1}(s), \qquad \forall s \neq s_{t-1}. \end{aligned} \tag{19}$$

As usual, η_t denotes the learning rate at time t, and V_0 is taken to be 0. Consequently, our analysis for asynchronous Q-learning with a Markovian trajectory immediately leads to non-asymptotic ℓ_∞ guarantees for TD learning, stated below as a corollary of Theorem 1. A similar result can be stated in terms of the cover time as a corollary to Theorem 2, which we omit for brevity.

Corollary 1 (Asynchronous TD learning): Consider the TD learning algorithm (19). There exist some universal constants $c_0, c_1 > 0$ such that for any $0 < \delta < 1$ and $0 < \varepsilon \leq \frac{1}{1-\gamma}$, one has

$$\forall s \in \mathcal{S}: \qquad |V_T(s) - V(s)| \leq \varepsilon$$

with probability at least $1-\delta$, provided that the iteration number T and the learning rates $\eta_t \equiv \eta$ obey

$$\begin{aligned} T \geq \frac{c_0}{\mu_{\mathsf{min}}}&\left\{\frac{1}{(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{1-\gamma}\right\} \\ &\cdot \log\Big(\frac{|\mathcal{S}|T}{\delta}\Big)\log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big), \end{aligned} \tag{20a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}|T}{\delta}\big)}\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}}\right\}. \tag{20b}$$

The above result reveals that the ℓ_∞-sample complexity for TD learning is at most

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big), \tag{21}$$

provided that an appropriate constant learning rate is adopted. We note that prior finite-sample analysis on asynchronous TD learning typically focused on (weighted) ℓ_2 estimation errors with linear function approximation [21], [22], and it is hence difficult to make fair comparisons. The recent paper [23] developed ℓ_∞ guarantees for TD learning, focusing on the synchronous settings with i.i.d. samples rather than Markovian samples.

[2]When $\mathcal{A} = \{a\}$ is a singleton, the Q-learning update rule (10) reduces to the TD update rule (19) by relating $Q(s,a) = V(s)$.

D. Adaptive and Implementable Learning Rates

As alluded to previously, the learning rates recommended in (13b) depend on the mixing time t_{mix}, a parameter that might be either *a priori* unknown or difficult to estimate. Fortunately, it is feasible to adopt a more adaptive learning rate schedule, which does not rely on prior knowledge of t_{mix} while still being capable of achieving the performance advertised in Theorem 1.

1) Learning Rates: In order to describe our new learning rate schedule, we need to keep track of the following quantities for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

- $K_t(s,a)$: the number of times that the sample trajectory visits (s,a) during the first t iterations.

In addition, we maintain an estimate $\widehat{\mu}_{\mathsf{min},t}$ of μ_{min}, computed recursively as follows

$$\widehat{\mu}_{\mathsf{min},t} = \begin{cases} \frac{1}{|\mathcal{S}||\mathcal{A}|}, & \text{if } \min_{s,a} K_t(s,a) = 0; \\ \widehat{\mu}_{\mathsf{min},t-1}, & \text{if } \frac{1}{3} < \frac{\min_{s,a} K_t(s,a)/t}{\widehat{\mu}_{\mathsf{min},t-1}} < 3; \\ \min_{s,a} K_t(s,a)/t, & \text{otherwise.} \end{cases} \tag{22}$$

With the above quantities in place, we propose the following learning rate schedule:

$$\eta_t = \min\Big\{1, c_\eta \exp\Big(\Big\lfloor \log \frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}\Big\rfloor\Big)\Big\}, \tag{23}$$

where $c_\eta > 0$ is some universal constant independent of any MDP parameter[3] and $\lfloor x \rfloor$ denotes the nearest integer less than or equal to x. If $\widehat{\mu}_{\mathsf{min},t}$ forms a reliable estimate of μ_{min}, then one can view (23) as a sort of "piecewise constant approximation" of the rescaled linear stepsizes $\frac{c_\eta \log t}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2 t}$; in fact, this can be viewed as a sort of "doubling trick" — reducing the learning rate by a constant factor every once a while — to approximate rescaled linear learning rates. Theorem 1 can then be readily applied to analyze the performance for each constant segment of this learning rate schedule (23). Noteworthily, such learning rates are fully data-driven and do no rely on any prior knowledge about the Markov chain (like t_{mix} and μ_{min}) or the target accuracy level ε.

2) Performance Guarantees: Encouragingly, our theoretical framework can be readily extended without difficulty to accommodate this adaptive learning rate choice. Specifically, for the Q-function estimates

$$\widehat{Q}_t = \begin{cases} Q_t, & \text{if } \eta_{t+1} \neq \eta_t, \\ \widehat{Q}_{t-1}, & \text{otherwise,} \end{cases} \tag{24}$$

where Q_t is provided by the Q-learning iterations (cf. (10)). We can then establish the following theoretical guarantees, whose proof is deferred to Section VIII.

Theorem 3: Consider asynchronous Q-learning with learning rates (23) and the output (24). There exists some universal constant $C > 0$ such that: for any $0 < \delta < 1$ and $0 < \varepsilon \leq \frac{1}{1-\gamma}$, one has

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \big|\widehat{Q}_T(s,a) - Q^\star(s,a)\big| \leq \varepsilon \tag{25}$$

with probability at least $1-\delta$, provided that

$$T \geq \frac{C}{\gamma^2} \max\Big\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}, \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big\} \cdot \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big) \log\Big(\frac{T}{(1-\gamma)^2\varepsilon}\Big). \tag{26}$$

Remark 4: The interested reader might wonder whether our sample complexity guarantees continue to hold under the linear learning rate $\eta_t = \frac{1}{K_t(s_t,a_t)}$ — a learning rate schedule that has been previously studied in [6], [9]. Nevertheless, as discussed in [8, Section 3.3.1], this linear learning rate can lead to a sample complexity that scales exponentially in the effective horizon $\frac{1}{1-\gamma}$, which is clearly outperformed by a properly rescaled linear learning rate.

IV. Extension: Asynchronous Variance-Reduced Q-Learning

As pointed out in prior literature, the classical form of Q-learning (10) often suffers from sub-optimal dependence on the effective horizon $\frac{1}{1-\gamma}$. For instance, in the synchronous setting, the minimax lower bound is proportional to $\frac{1}{(1-\gamma)^3}$ (see, [18]), while the sharpest known upper bound for vanilla Q-learning scales as $\frac{1}{(1-\gamma)^5}$; see detailed discussions in [8]. To remedy this issue, recent work proposed to leverage the idea of variance reduction to develop accelerated RL algorithms in the synchronous setting [12], [24], as inspired by the seminal SVRG algorithm [25] that originates from the stochastic optimization literature. In this section, we adapt this idea to asynchronous Q-learning and characterize its sample efficiency.

A. Algorithm

In order to accelerate the convergence, it is instrumental to reduce the variability of the empirical Bellman operator $\mathcal{T}_t$ employed in the update rule (10) of classical Q-learning. This can be achieved via the following means. Simply put, assuming we have access to *(i)* a reference Q-function estimate, denoted by $\overline{Q}$, and *(ii)* an estimate of $\mathcal{T}(\overline{Q})$, denoted by $\widetilde{\mathcal{T}}(\overline{Q})$, the variance-reduced Q-learning update rule is given by

$$\begin{aligned} Q_t(s_{t-1},a_{t-1}) &= (1-\eta_t)Q_{t-1}(s_{t-1},a_{t-1}) \\ &\quad + \eta_t\Big(\mathcal{T}_t(Q_{t-1}) - \mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q})\Big)(s_{t-1},a_{t-1}), \\ Q_t(s,a) &= Q_{t-1}(s,a), \qquad \forall (s,a) \neq (s_{t-1},a_{t-1}), \end{aligned} \tag{27}$$

where $\mathcal{T}_t$ denotes the empirical Bellman operator at time t (cf. (11)). The empirical estimate $\widetilde{\mathcal{T}}(\overline{Q})$ can be computed using a set of samples; more specifically, by drawing N consecutive sample transitions $\{(s_i,a_i,s_{i+1})\}_{0\leq i<N}$ from the observed trajectory, we compute

$$\begin{aligned} \widetilde{\mathcal{T}}(\overline{Q})(s,a) &= r(s,a) \\ &\quad + \frac{\gamma \sum_{i=0}^{N-1} \mathbb{1}\{(s_i,a_i)=(s,a)\} \max_{a'} \overline{Q}(s_{i+1},a')}{\sum_{i=0}^{N-1} \mathbb{1}\{(s_i,a_i)=(s,a)\}}. \end{aligned} \tag{28}$$

[3]More precisely, $c_\eta > 0$ can be any universal constant obeying $c_\eta \geq 74 c_0 c_1$ and $c_\eta > 11$, with c_0 and c_1 being the universal constants stated in Theorem 1.

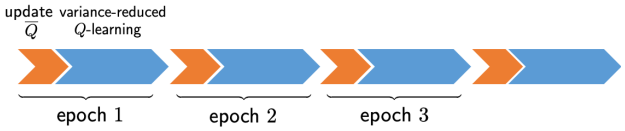

Fig. 1. A pictorial illustration of variance-reduced Q-learning.

Compared with the classical form (10), the original update term $\mathcal{T}_t(Q_{t-1})$ has been replaced by $\mathcal{T}_t(Q_{t-1})-\mathcal{T}_t(\overline{Q})+\widetilde{\mathcal{T}}(\overline{Q})$, in the hope of achieving reduced variance as long as $\overline{Q}$ (which serves as a proxy to $Q^\star$) is chosen properly.

We now take a moment to elucidate the rationale behind the variance-reduced update rule (27). In the vanilla Q-learning update rule (10), the variability in each iteration (conditional on the past) comes primarily from the stochastic term $\mathcal{T}_t(Q_{t-1})$. In order to accelerate convergence, it is advisable to reduce the variability of this term. Suppose now that we have access to a reference point $\overline{Q}$ that is close to Q_{t-1}. By replacing $\mathcal{T}_t(Q_{t-1})$ with

$$\Big\{\mathcal{T}_t(Q_{t-1})-\mathcal{T}_t(\overline{Q})\Big\}+\widetilde{\mathcal{T}}(\overline{Q}),$$

we see that the variability of the first term $\mathcal{T}_t(Q_{t-1})-\mathcal{T}_t(\overline{Q})$ can be small if $Q_{t-1}\approx\overline{Q}$, while the uncertainty of the second term $\widetilde{\mathcal{T}}(\overline{Q})$ can also be well controlled via the use of batch data. Motivated by this simple idea, the variance-reduced Q-learning rule attempts to operate in an epoch-based manner, computing $\widetilde{\mathcal{T}}(\overline{Q})$ once every epoch (so as not to increase the overall sampling burden) and leveraging it to help reduce variability.

For convenience of presentation, we introduce the following notation

$$Q=\textsc{Vr-Q-run-epoch}(\,\overline{Q},N,t_{\mathsf{epoch}}\,) \tag{29}$$

to represent the above-mentioned update rule, which starts with a reference point $\overline{Q}$ and operates upon a total number of $N+t_{\mathsf{epoch}}$ consecutive sample transitions. The first N samples are employed to construct $\widetilde{\mathcal{T}}(\overline{Q})$ via (28), with the remaining samples employed in t_{epoch} iterative updates (27); see Algorithm 3. To achieve the desired acceleration, the proxy $\overline{Q}$ needs to be periodically updated so as to better approximate the truth $Q^\star$ and hence reduce the bias. It is thus natural to run the algorithm in a multi-epoch manner. Specifically, we divide the samples into contiguous subsets called epochs, each containing t_{epoch} iterations and using $N+t_{\mathsf{epoch}}$ samples. We then proceed as follows

$$Q_m^{\mathsf{epoch}}=\textsc{Vr-Q-run-epoch}(\,Q_{m-1}^{\mathsf{epoch}},N,t_{\mathsf{epoch}}\,) \tag{30}$$

for $m=1,\ldots,M$, where M is the total number of epochs, and Q_m^{epoch} denotes the output of the m-th epoch. The whole procedure is summarized in Algorithm 2. Clearly, the total number of samples used in this algorithm is given by $M(N+t_{\mathsf{epoch}})$. We remark that the idea of performing variance reduction in RL is certainly not new, and has been explored in a number of recent works [12], [23], [24], [26]–[28].

B. Theoretical Guarantees for Variance-Reduced Q-Learning

This subsection develops a non-asymptotic sample complexity bound for asynchronous variance-reduced Q-learning on a single trajectory. Before presenting our theoretical guarantees, there are several algorithmic parameters that we shall specify; for given target levels (ε,δ), choose

$$\eta_t\equiv\eta=\frac{c_0}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\big)}\min\left\{\frac{(1-\gamma)^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}}\right\}, \tag{31a}$$

$$N\geq\frac{c_1}{\mu_{\mathsf{min}}}\Big(\frac{1}{(1-\gamma)^3\min\{1,\varepsilon^2\}}+t_{\mathsf{mix}}\Big)\cdot\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big), \tag{31b}$$

$$t_{\mathsf{epoch}}\geq\frac{c_2}{\mu_{\mathsf{min}}}\Big(\frac{1}{(1-\gamma)^3}+\frac{t_{\mathsf{mix}}}{1-\gamma}\Big)\cdot\log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big)\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big), \tag{31c}$$

where $c_0>0$ is some sufficiently small constant, $c_1,c_2>0$ are some sufficiently large constants, and we recall the definitions of μ_{min} and t_{mix} in (7) and (8), respectively. Note that the learning rate (31a) chosen here could be larger than the choice (13b) for the classical form by a factor of $O(\frac{1}{(1-\gamma)^2})$ (which happens if t_{mix} is not too large), allowing the algorithm to progress more aggressively.

Theorem 4 (Asynchronous Variance-Reduced Q-Learning): Let Q_M^{epoch} be the output of Algorithm 2 with parameters chosen according to (31). There exists some constant $c_3>0$ such that for any $0<\delta<1$ and $0<\varepsilon\leq\frac{1}{1-\gamma}$, one has

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\qquad|Q_M^{\mathsf{epoch}}(s,a)-Q^\star(s,a)|\leq\varepsilon$$

with probability at least $1-\delta$, provided that the total number of epochs exceeds

$$M\geq c_3\log\frac{1}{\varepsilon(1-\gamma)^2}. \tag{32}$$

The proof of this result is postponed to Section IX.

In view of Theorem 4, the ℓ_∞-based sample complexity for variance-reduced Q-learning to yield ε accuracy — which is characterized by $M(N+t_{\mathsf{epoch}})$ — can be as low as

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\min\{1,\varepsilon^2\}}+\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big). \tag{33}$$

Except for the second term that depends on the mixing time, the first term matches the result of [12] derived for the synchronous settings with independent samples. In the range $\varepsilon\in(0,\min\{1,\frac{1}{(1-\gamma)\sqrt{t_{\mathsf{mix}}}}\}]$, the sample complexity reduce to $\widetilde{O}(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2})$; the scaling $\frac{1}{(1-\gamma)^3}$ matches the minimax lower bound derived in [18] for the synchronous setting.

Once again, we can immediately deduce guarantees for asynchronous variance-reduced TD learning by reducing the action space to a singleton set (akin to Section III-C), which extends the analysis [23] to Markovian noise. In addition, similar to Section III-D, we can also employ adaptive learning rates in variance-reduced Q-learning — which do not require prior knowledge of t_{mix} and μ_{min} — without compromising

Algorithm 2: Asynchronous Variance-Reduced Q-Learning

input parameters: number of epochs M, epoch length t_{epoch}, recentering length N, learning rate η.
initialization: set $Q_0^{\mathsf{epoch}} \leftarrow 0$.
for each epoch $m = 1, \cdots, M$ **do**
  /* Call Algorithm 3. */
  $Q_m^{\mathsf{epoch}} = \text{VR-Q-RUN-EPOCH}(Q_{m-1}^{\mathsf{epoch}}, N, t_{\mathsf{epoch}})$.

Algorithm 3: function $Q = \text{VR-Q-RUN-EPOCH}(\overline{Q}, N, t_{\mathsf{epoch}})$

Draw N new consecutive samples from the sample trajectory; compute $\widetilde{\mathcal{T}}(\overline{Q})$ according to (28).
Set $s_0 \leftarrow$ current state, and $Q_0 \leftarrow \overline{Q}$.
for $t = 1, 2, \cdots, t_{\mathsf{epoch}}$ **do**
  Draw action $a_{t-1} \sim \pi_{\mathsf{b}}(s_{t-1})$, observe reward $r(s_{t-1}, a_{t-1})$, and draw next state $s_t \sim P(\cdot \,|\, s_{t-1}, a_{t-1})$.
  Update Q_t according to (27).
return: $Q \leftarrow Q_{t_{\mathsf{epoch}}}$.

the sample complexity. For the sake of brevity, we omit these extensions in the current paper.

V. Related Works

In this section, we review several recent lines of works and compare our results with them.

A. The Q-Learning Algorithm and Its Variants

The Q-learning algorithm, originally proposed in [29], has been analyzed in the asymptotic regime by [6], [7], [14], [30] since more than two decades ago. Additionally, finite-time performance of Q-learning and its variants have been analyzed by [2], [8]–[10], [19], [31]–[34] in the tabular setting, by [21], [35]–[43] in the context of function approximations, and by [44] with nonparametric regression. In addition, [11], [12], [24], [45]–[47] studied modified Q-learning algorithms that might potentially improve sample complexities and accelerate convergence. Another line of work studied Q-learning with sophisticated exploration strategies such as UCB exploration (e.g. [48]–[51]), which is beyond the scope of the current work.

B. Finite-Sample ℓ_∞ Guarantees for Q-Learning

We now expand on non-asymptotic ℓ_∞ guarantees available in prior literature, which are the most relevant to the current work. An interesting aspect that we shall highlight is the importance of learning rates. For instance, when a linear learning rate (i.e. $\eta_t = 1/t$) is adopted, the sample complexity results derived in past works [9], [14] exhibit an exponential blow-up in $\frac{1}{1-\gamma}$, which is clearly undesirable. In the synchronous setting, [8]–[10], [19] studied the finite-sample complexity of Q-learning under various learning rate rules; the best sample complexity known to date is $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}\big)$, achieved via either a rescaled linear learning rate [8], [19] or a constant learning rate [19]. When it comes to asynchronous Q-learning (in its classical form), our work provides the first analysis that achieves linear scaling with $1/\mu_{\mathsf{min}}$ or t_{cover}; see Table I for detailed comparisons. Going beyond classical Q-learning, the speedy Q-learning algorithm, which adds a momentum term in the update by using previous Q-function estimates, provably achieves a sample complexity of $\widetilde{O}\big(\frac{t_{\mathsf{cover}}}{(1-\gamma)^4\varepsilon^2}\big)$ [11] in the asynchronous setting, whose update rule takes twice the storage of classical Q-learning. However, the proof idea adopted in the speedy Q-learning paper relies heavily on the specific update rules of speedy Q-learning, which cannot be readily used here to help improve the sample complexity of asynchronous Q-learning in terms of its dependency on $\frac{1}{1-\gamma}$. In comparison, our analysis of the variance-reduced Q-learning algorithm achieves a sample complexity of $\widetilde{O}\big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\big)$ when $\varepsilon < 1$.

C. Finite-Sample Guarantees for Model-Free Algorithms

Convergence properties of several model-free RL algorithms have been studied recently in the presence of Markovian data, including but not limited to TD learning and its variants [21], [22], [28], [52]–[60], Q-learning [35], [36], and SARSA [61]. However, these recent papers typically focused on the (weighted) ℓ_2 error rather than the ℓ_∞ risk, where the latter is often more relevant in the context of RL. In addition, [23] investigated the ℓ_∞ bounds of (variance-reduced) TD learning, although they did not account for Markovian noise.

D. Finite-Sample Guarantees for Model-Based Algorithms

Another contrasting approach for learning the optimal Q-function is the class of model-based algorithms, which has been shown to enjoy minimax-optimal sample complexity in the synchronous setting. More precisely, it is known that by planning over an empirical MDP constructed from $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\big)$ samples, we are guaranteed to find not only an ε-optimal Q-function but also an ε-optimal policy [18], [62], [63]. It is worth emphasizing that the minimax optimality of model-based approach has been shown to hold for the entire ε-range; in comparison, the sample optimality of the model-free approach has only been shown for a smaller range of accuracy level ε in the synchronous setting. We also remark that existing sample complexity analysis for model-based approaches might be generalizable to Markovian data.

VI. Analysis of Asynchronous Q-Learning

This section is devoted to establishing Theorem 1. Before proceeding, we find it convenient to introduce some matrix notation. Let $\mathbf{\Lambda}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be a diagonal matrix obeying

$$\mathbf{\Lambda}_t\big((s,a),(s,a)\big) := \begin{cases} \eta, & \text{if } (s,a) = (s_{t-1}, a_{t-1}), \\ 0, & \text{otherwise,} \end{cases} \tag{34}$$

where $\eta > 0$ is the learning rate. In addition, we use the vector $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ (resp. $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$) to represent our estimate Q_t

(resp. V_t) in the t-th iteration, so that the (s,a)-th (resp. sth) entry of $\boldsymbol{Q}_t$ (resp. $\boldsymbol{V}_t$) is given by $Q_t(s,a)$ (resp. $V_t(s)$). Similarly, let the vectors $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$ represent the optimal Q-function $Q^\star$ and the optimal value function $V^\star$, respectively. We also let the vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ stand for the reward function r, so that the (s,a)-th entry of $\boldsymbol{r}$ is given by $r(s,a)$. In addition, we define the matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ such that

$$\boldsymbol{P}_t\big((s,a),s'\big) := \begin{cases} 1, & \text{if } (s,a,s') = (s_{t-1},a_{t-1},s_t), \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

Clearly, this set of notation allows us to express the Q-learning update rule (10) in the following matrix form

$$\boldsymbol{Q}_t = \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big). \tag{36}$$

A. Error Decay in the Presence of Constant Learning Rates

The main step of the analysis is to establish the following result concerning the dynamics of asynchronous Q-learning. In order to state it formally, we find it convenient to introduce several auxiliary quantities

$$t_{\mathsf{frame}} := \frac{443 t_{\mathsf{mix}}}{\mu_{\min}} \log\Big(\frac{4|\mathcal{S}||\mathcal{A}|T}{\delta}\Big), \tag{37a}$$

$$t_{\mathsf{th}} := \max\left\{\frac{2\log\frac{1}{(1-\gamma)^2\varepsilon}}{\eta\mu_{\min}},\ t_{\mathsf{frame}}\right\}, \tag{37b}$$

$$\mu_{\mathsf{frame}} := \frac{1}{2}\mu_{\min} t_{\mathsf{frame}}, \tag{37c}$$

$$\rho := (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big). \tag{37d}$$

With these quantities in mind, we have the following result.

Theorem 5: Consider the asynchronous Q-learning algorithm in Algorithm 1 with $\eta_t \equiv \eta$. For any $\delta \in (0,1)$ and any $\varepsilon \in (0, \frac{1}{1-\gamma}]$, there exists a universal constant $c > 0$ such that with probability at least $1 - 6\delta$, the following relation holds uniformly for all $t \le T$ (defined in (13a))

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \le (1-\rho)^k \frac{\|\boldsymbol{Q}_0 - \boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty \sqrt{\eta \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} + \varepsilon, \tag{38}$$

provided that $0 < \eta \log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big) < 1$. Here, we define $k := \max\big\{0,\ \big\lfloor \frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}} \big\rfloor\big\}$.

In words, Theorem 5 asserts that the ℓ_∞ estimation error decays linearly — in a blockwise manner — to some error floor that scales with $\sqrt{\eta}$. This result suggests how to set the learning rate based on the target accuracy level, which in turn allows us to pin down the sample complexity under consideration. In what follows, we shall first establish Theorem 5, and then return to prove Theorem 1 using this result.

Before embarking on the proof of Theorem 5, we would like to point out a few key technical ingredients: (i) an epoch-based analysis that focuses on macroscopic dynamics as opposed to per-iteration dynamics, (ii) measure concentration of Markov chains (see Section A) that helps reveal the similarity between epoch-based dynamics and the synchronous counterpart, and (iii) careful analysis of recursive relations. These key ingredients taken collectively lead to a sample complexity bound that improves upon prior analysis in [2].

B. Proof of Theorem 5

We are now positioned to outline the proof of Theorem 5. We remind the reader that for any two vectors $\boldsymbol{z} = [z_i]$ and $\boldsymbol{w} = [w_i]$, the notation $\boldsymbol{z} \le \boldsymbol{w}$ (resp. $\boldsymbol{z} \ge \boldsymbol{w}$) denotes entrywise comparison (cf. Section I), meaning that $z_i \le w_i$ (resp. $z_i \ge w_i$) holds for all i. As a result, for any non-negative matrix $\boldsymbol{A}$, one has $\boldsymbol{A}\boldsymbol{z} \le \boldsymbol{A}\boldsymbol{w}$ as long as $\boldsymbol{z} \le \boldsymbol{w}$.

1) Key Decomposition and a Recursive Formula: The starting point of our proof is the following elementary decomposition

$$\begin{aligned}
\boldsymbol{\Delta}_t &:= \boldsymbol{Q}_t - \boldsymbol{Q}^\star \\
&= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\
&= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\
&= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\
&= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}^\star + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}_t\big(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\big)
\end{aligned} \tag{39}$$

for any $t > 0$, where the first line results from the update rule (36), and the penultimate line follows from the Bellman equation $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$ (see [16]). Applying this relation recursively gives

$$\begin{aligned}
\boldsymbol{\Delta}_t = \underbrace{\gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}^\star}_{=:\boldsymbol{\beta}_{1,t}} \\
+ \underbrace{\gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1} - \boldsymbol{V}^\star\big)}_{=:\boldsymbol{\beta}_{2,t}} \\
+ \underbrace{\prod_{j=1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0}_{=:\boldsymbol{\beta}_{3,t}}.
\end{aligned} \tag{40}$$

Applying the triangle inequality, we obtain

$$|\boldsymbol{\Delta}_t| \le |\boldsymbol{\beta}_{1,t}| + |\boldsymbol{\beta}_{2,t}| + |\boldsymbol{\beta}_{3,t}|, \tag{41}$$

where we recall the notation $|\boldsymbol{z}| := [|z_i|]_{1\le i\le n}$ for any vector $\boldsymbol{z} = [z_i]_{1\le i\le n}$. In what follows, we shall look at these terms separately.

- First of all, given that $\boldsymbol{I} - \boldsymbol{\Lambda}_j$ and $\boldsymbol{\Lambda}_j$ are both non-negative diagonal matrices and that

$$\begin{aligned}
\big\|\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1} - \boldsymbol{V}^\star\big)\big\|_\infty &\le \|\boldsymbol{P}_i\|_1 \|\boldsymbol{V}_{i-1} - \boldsymbol{V}^\star\|_\infty \\
&= \|\boldsymbol{V}_{i-1} - \boldsymbol{V}^\star\|_\infty \\
&\le \|\boldsymbol{Q}_{i-1} - \boldsymbol{Q}^\star\|_\infty = \|\boldsymbol{\Delta}_{i-1}\|_\infty,
\end{aligned}$$

we can easily see that

$$\big|\boldsymbol{\beta}_{2,t}\big| \le \gamma \sum_{i=1}^{t} \|\boldsymbol{\Delta}_{i-1}\|_\infty \prod_{j=i+1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i \boldsymbol{1}. \tag{42}$$

- Next, the term $\boldsymbol{\beta}_{1,t}$ can be controlled by exploiting some sort of statistical independence across different transitions and applying the Bernstein inequality. This is summarized in the following lemma, with the proof deferred to Section C.
Lemma 1: Consider any fixed vector $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$. There exists some universal constant $c > 0$ such that for any $0 < \delta < 1$, one has for all $1 \le t \le T$

$$\left| \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}^\star \right| \le \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} \tag{43}$$

with probability at least $1-\delta$, provided that $0 < \eta \log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big) < 1$. Here, we define

$$\tau_1 := c\gamma\sqrt{\eta \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}. \tag{44}$$

- Additionally, we develop an upper bound on the term $\boldsymbol{\beta}_{3,t}$, which follows directly from the concentration of the empirical distribution of the Markov chain (see Lemma 8). The proof is deferred to Section D.
Lemma 2: For any $\delta > 0$, recall the definition of t_{frame} in (37a). Suppose that $T > t_{\mathsf{frame}}$ and $0 < \eta < 1$. Then with probability exceeding $1-\delta$ one has

$$\left| \prod_{j=1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0 \right| \le (1-\eta)^{\frac{1}{2}t\mu_{\min}} \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1} \tag{45}$$

uniformly over all t obeying $T \ge t \ge t_{\mathsf{frame}}$ and all vector $\boldsymbol{\Delta}_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.
Moreover, in the case where $t < t_{\mathsf{frame}}$, we make note of the straightforward bound

$$\left| \prod_{j=1}^{t} \big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0 \right| \le \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \tag{46}$$

given that $\boldsymbol{I} - \boldsymbol{\Lambda}_j$ is a diagonal non-negative matrix whose entries are bounded by $1 - \eta < 1$.

Substituting the preceding bounds into (41), we arrive at

$$|\boldsymbol{\Delta}_t| \le \begin{cases} \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} \\ \quad + \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \big\|\boldsymbol{\Delta}_0\big\|_\infty \mathbf{1}, & t < t_{\mathsf{frame}} \\ \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} \\ \quad + \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} + (1-\eta)^{\frac{1}{2}t\mu_{\min}} \big\|\boldsymbol{\Delta}_0\big\|_\infty \mathbf{1}, & t_{\mathsf{frame}} \le t \le T \end{cases} \tag{47}$$

with probability at least $1-2\delta$, where t_{frame} is defined in (37a). The rest of the proof is thus dedicated to bounding $|\boldsymbol{\Delta}_t|$ based on the above recursive formula (47).

2) Recursive Analysis: We shall start by presenting a crude bound, followed by more refined analysis.

a) A crude bound: We start by observing the following recursive relation

$$|\boldsymbol{\Delta}_t| \le \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \quad 1 \le t \le T, \tag{48}$$

which is a direct consequence of (47). In the sequel, we invoke mathematical induction to establish, for all $1 \le t \le T$, the following crude upper bound

$$\big\|\boldsymbol{\Delta}_t\big\|_\infty \le \frac{\tau_1 \|\boldsymbol{V}^\star\|_\infty + \|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma}, \tag{49}$$

which implies the stability of the asynchronous Q-learning updates.

Towards this, we first observe that (49) holds trivially for the base case (namely, $t = 0$). Now suppose that the inequality (49) holds for all iterations up to $t-1$. In view of (48) and the induction hypotheses,

$$|\boldsymbol{\Delta}_t| \le \frac{\gamma\big(\tau_1 \|\boldsymbol{V}^\star\|_\infty + \big\|\boldsymbol{\Delta}_0\big\|_\infty\big)}{1-\gamma} \sum_{i=1}^{t} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \tag{50}$$

where we invoke the fact that the vector $\prod_{j=i+1}^{t}(\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1}$ is non-negative. Next, define the diagonal matrix $\boldsymbol{M}_i := \prod_{j=i+1}^{t}(\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i$, and denote by $N_i^j(s,a)$ the number of visits to the state-action pair (s,a) between the i-th and the j-th iterations (including i and j). Then the diagonal entries of $\boldsymbol{M}_i$ satisfy

$$\boldsymbol{M}_i((s,a),(s,a)) = \begin{cases} \eta(1-\eta)^{N_{i+1}^t(s,a)}, & \text{if } (s,a) = (s_{i-1}, a_{i-1}), \\ 0, & \text{if } (s,a) \ne (s_{i-1}, a_{i-1}). \end{cases}$$

Letting $\boldsymbol{e}_{(s,a)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a standard basis vector whose only nonzero entry is the (s,a)-th entry, we can easily verify that

$$\begin{aligned} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} &= \boldsymbol{M}_i \mathbf{1} = \boldsymbol{M}_i \boldsymbol{e}_{(s_{i-1}, a_{i-1})} \\ &= \eta(1-\eta)^{N_{i+1}^t(s_{i-1}, a_{i-1})} \boldsymbol{e}_{(s_{i-1}, a_{i-1})} \end{aligned} \tag{51a}$$

and

$$\begin{aligned} &\sum_{i=1}^{t} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} \\ &= \sum_{i=1}^{t} \eta(1-\eta)^{N_{i+1}^t(s_{i-1}, a_{i-1})} \boldsymbol{e}_{(s_{i-1}, a_{i-1})} \\ &= \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \boldsymbol{e}_{(s,a)} \cdot \left\{ \sum_{i=1}^{t} \eta(1-\eta)^{N_{i+1}^t(s,a)} \mathbb{1}\left\{(s_{i-1}, a_{i-1}) = (s,a)\right\} \right\} \\ &\le \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{j=0}^{\infty} \eta(1-\eta)^j \boldsymbol{e}_{(s,a)} = \sum_{j=0}^{\infty} \eta(1-\eta)^j \mathbf{1} = \mathbf{1}. \end{aligned} \tag{51b}$$

Combining the above relations with the inequality (50), one deduces that

$$\begin{aligned}\big\|\boldsymbol{\Delta}_t\big\|_\infty &\le \frac{\gamma(\tau_1\|\boldsymbol{V}^\star\|_\infty + \big\|\boldsymbol{\Delta}_0\big\|_\infty)}{1-\gamma} + \tau_1\|\boldsymbol{V}^\star\|_\infty + \big\|\boldsymbol{\Delta}_0\big\|_\infty \\ &= \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty + \big\|\boldsymbol{\Delta}_0\big\|_\infty}{1-\gamma},\end{aligned}$$

thus establishing (49) for the t-th iteration. This induction analysis thus validates (49) for all $1 \le t \le T$.

b) Refined analysis: Now, we strengthen the bound (49) by means of a recursive argument. To begin with, it is easily seen that the term $(1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\|\boldsymbol{\Delta}_0\|_\infty$ is bounded above by $(1-\gamma)\varepsilon$ for any $t > t_{\mathsf{th}}$, where we remind the reader of the definition of t_{th} in (37b) and the fact that $\|\boldsymbol{\Delta}_0\|_\infty = \|\boldsymbol{Q}^\star\|_\infty \le \frac{1}{1-\gamma}$. It is assumed that $T > t_{\mathsf{th}}$. To facilitate our argument, we introduce a collection of auxiliary quantities u_t as follows

$$u_0 = \frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma}, \tag{52a}$$

$$u_t = \|\boldsymbol{v}_t\|_\infty, \quad \text{with}$$

$$\boldsymbol{v}_t = \begin{cases} \gamma\sum_{i=1}^t \prod_{j=i+1}^t (\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1} + \|\boldsymbol{\Delta}_0\|_\infty\mathbf{1}, \\ \qquad\qquad\qquad\qquad \text{for } 1\le t\le t_{\mathsf{th}}, \\ \gamma\sum_{i=1}^t \prod_{j=i+1}^t (\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}, \\ \qquad\qquad\qquad\qquad \text{for } t > t_{\mathsf{th}}. \end{cases} \tag{52b}$$

These auxiliary quantities are useful as they provide upper bounds on $\|\boldsymbol{\Delta}_t\|_\infty$, as asserted by the following lemma. The proof is deferred to Section E.

Lemma 3: Recall the definition (44) of τ_1 in Lemma 1. With probability at least $1-2\delta$, the quantities $\{u_t\}$ defined in (52) satisfy

$$\|\boldsymbol{\Delta}_t\|_\infty \le \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma} + u_t + \varepsilon. \tag{53}$$

The preceding result motivates us to turn attention to bounding the quantities $\{u_t\}$. Towards this end, we resort to a frame-based analysis by dividing the iterations $[1,t]$ into contiguous frames each comprising t_{frame} (cf. (37a)) iterations. Further, we define another auxiliary sequence:

$$w_k := (1-\rho)^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} = (1-\rho)^k\frac{\|\boldsymbol{Q}_0-\boldsymbol{Q}^\star\|_\infty}{1-\gamma}, \tag{54}$$

where we remind the reader of the definition of ρ in (37d). The connection between $\{w_k\}$ and $\{u_t\}$ is made precise as follows, whose proof is postponed to Section F.

Lemma 4: For any $\delta \in (0, \frac{1}{2})$, with probability at least $1-2\delta$, one has

$$u_t \le w_k, \qquad \text{with} \quad k = \max\left\{0,\ \left\lfloor\frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\right\rfloor\right\}. \tag{55}$$

Combining Lemmas 3-4, we arrive at

$$\begin{aligned}\|\boldsymbol{Q}_t-\boldsymbol{Q}^\star\|_\infty = \|\boldsymbol{\Delta}_t\|_\infty &\le \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma} + w_k + \varepsilon \\ &\le \frac{(1-\rho)^k\|\boldsymbol{Q}_0-\boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma} + \varepsilon,\end{aligned}$$

which finishes the proof of Theorem 5.

C. Proof of Theorem 1

Now we return to complete the proof of Theorem 1. To control $\|\boldsymbol{\Delta}_t\|_\infty$ to the desired level, we first claim that the first term of (38) obeys

$$(1-\rho)^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \le \varepsilon \tag{56}$$

whenever

$$t \ge t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{4}{(1-\gamma)\eta\mu_{\mathsf{min}}}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right), \tag{57}$$

provided that $\eta < 1/\mu_{\mathsf{frame}}$. Furthermore, by taking the learning rate as

$$\eta = \min\left\{\frac{(1-\gamma)^4\varepsilon^2}{c^2\gamma^2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}},\ \frac{1}{\mu_{\mathsf{frame}}}\right\}, \tag{58}$$

one can easily verify that the second term of (38) satisfies

$$\frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} \le \varepsilon, \tag{59}$$

where the last step follows since $\|\boldsymbol{V}^\star\|_\infty \le \frac{1}{1-\gamma}$. Putting the above bounds together ensures $\|\boldsymbol{\Delta}_t\|_\infty \le 3\varepsilon$. By replacing ε with $\varepsilon/3$, we can readily conclude the proof, as long as the claim (56) can be justified.

Proof of the Inequality (56)*:* Observe that

$$(1-\rho)^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \le \exp(-\rho k)\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \le \varepsilon$$

holds true whenever $k \ge \frac{\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right)}{\rho}$, which would hold as long as (according to the definition (55) of k)

$$t \ge t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{t_{\mathsf{frame}}}{\rho}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right). \tag{60}$$

In addition, if $\eta < 1/\mu_{\mathsf{frame}}$, then one has $(1-\eta)^{\mu_{\mathsf{frame}}} \le 1-\eta\mu_{\mathsf{frame}}/2$, thus guaranteeing that

$$\begin{aligned}\rho = (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big) &\ge (1-\gamma)\Big(1-1+\frac{\eta\mu_{\mathsf{frame}}}{2}\Big) \\ &= \frac{1}{2}(1-\gamma)\eta\mu_{\mathsf{frame}}.\end{aligned}$$

This taken collectively with (60) demonstrates that $(1-\rho)^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \le \varepsilon$ holds as long as

$$\begin{aligned}t &\ge t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{2t_{\mathsf{frame}}}{(1-\gamma)\eta\mu_{\mathsf{frame}}}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right) \\ &= t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{4}{(1-\gamma)\eta\mu_{\mathsf{min}}}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right),\end{aligned} \tag{61}$$

where we have made use of the definition of μ_{frame} (cf. (37c)). □

VII. Cover-Time-Based Analysis of Asynchronous Q-Learning

In this section, we sketch the proof of Theorem 2. Before continuing, we recall the definition of t_{cover} in (9), and further introduce a quantity

$$t_{\mathsf{cover,all}} := t_{\mathsf{cover}}\log\frac{T}{\delta}. \tag{62}$$

There are two useful facts regarding $t_{\mathsf{cover,all}}$ that play an important role in the analysis.

Lemma 5: Define the event

$$\mathcal{K}_l := \Big\{\exists (s,a)\in\mathcal{S}\times\mathcal{A} \text{ s.t. it is not visited within iterations } \big(lt_{\mathsf{cover,all}}, (l+1)t_{\mathsf{cover,all}}\big]\Big\},$$

and set $L := \lfloor \frac{T}{t_{\mathsf{cover,all}}} \rfloor$. Then one has $\mathbb{P}\left\{\bigcup_{l=0}^{L}\mathcal{K}_l\right\} \le \delta$.

Proof: See Section H. □

In other words, Lemma 5 tells us that with high probability, all state-action pairs are visited at least once in every time frame $(lt_{\mathsf{cover,all}}, (l+1)t_{\mathsf{cover,all}}]$ with $0 \le l \le \lfloor T/t_{\mathsf{cover,all}} \rfloor$. The next result is a consequence of Lemma 5 as well as the analysis of Lemma 2; the proof can be found in Section D.

Lemma 6: For any $\delta > 0$, recall the definition of $t_{\mathsf{cover,all}}$ in (62). Suppose that $T > t_{\mathsf{cover,all}}$ and $0 < \eta < 1$. Then with probability exceeding $1-\delta$ one has

$$\left|\prod_{j=1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0\right| \le (1-\eta)^{\frac{t}{2t_{\mathsf{cover,all}}}}\|\boldsymbol{\Delta}_0\|_\infty \mathbf{1} \tag{63}$$

uniformly over all t obeying $T \ge t \ge t_{\mathsf{cover,all}}$ and all vector $\boldsymbol{\Delta}_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

With the above two lemmas in mind, we are now positioned to prove Theorem 2. Repeating the analysis of (47) (except that Lemma 2 is replaced by Lemma 6) yields

$$|\boldsymbol{\Delta}_t| \le \begin{cases} \gamma\sum_{i=1}^{t}\big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1} \\ +\tau_1\|\boldsymbol{V}^\star\|_\infty\mathbf{1} + \big\|\boldsymbol{\Delta}_0\big\|_\infty\mathbf{1}, & t < t_{\mathsf{cover,all}} \\ \gamma\sum_{i=1}^{t}\big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1} \\ +\tau_1\|\boldsymbol{V}^\star\|_\infty\mathbf{1} + (1-\eta)^{\frac{t}{2t_{\mathsf{cover,all}}}}\big\|\boldsymbol{\Delta}_0\big\|_\infty\mathbf{1}, & t_{\mathsf{cover,all}} \le t \le T \end{cases}$$

with probability at least $1-2\delta$. This observation resembles (47), except that t_{frame} (resp. μ_{min}) is replaced by $t_{\mathsf{cover,all}}$ (resp. $\frac{1}{t_{\mathsf{cover,all}}}$). As a consequence, we can immediately use the recursive analysis carried out in Section VI-B.2 to establish a convergence guarantee based on the cover time. More specifically, define

$$\widetilde{\rho} := (1-\gamma)\Big(1-(1-\eta)^{\frac{t_{\mathsf{cover,all}}}{2t_{\mathsf{cover,all}}}}\Big) = (1-\gamma)\Big(1-(1-\eta)^{\frac{1}{2}}\Big). \tag{64}$$

Replacing ρ by $\widetilde{\rho}$ in Theorem 5 reveals that with probability at least $1-6\delta$,

$$\|\boldsymbol{Q}_t-\boldsymbol{Q}^\star\|_\infty \le (1-\widetilde{\rho})^k\frac{\|\boldsymbol{Q}_0-\boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} + \varepsilon \tag{65}$$

holds for all $t \le T$, where $k := \max\big\{0,\ \lfloor\frac{t-t_{\mathsf{th,cover}}}{t_{\mathsf{cover,all}}}\rfloor\big\}$ and we abuse notation to define

$$t_{\mathsf{th,cover}} := 2t_{\mathsf{cover,all}}\log\frac{1}{(1-\gamma)^2\varepsilon}.$$

Repeating the proof of the inequality (56) yields

$$(1-\widetilde{\rho})^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \le \varepsilon,$$

whenever $t \ge t_{\mathsf{th,cover}} + t_{\mathsf{cover,all}} + \frac{2t_{\mathsf{cover,all}}}{(1-\gamma)\eta}\log\big(\frac{1}{\varepsilon(1-\gamma)^2}\big)$, with the proviso that $\eta < 1/2$. In addition, setting $\eta = \frac{(1-\gamma)^4}{c^2\gamma^2\varepsilon^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}$ guarantees that

$$\frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} \le \frac{c\gamma}{(1-\gamma)^2}\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} \le \varepsilon.$$

In conclusion, we have $\|\boldsymbol{Q}_t-\boldsymbol{Q}^\star\|_\infty \le 3\varepsilon$ as long as

$$t \ge \frac{c't_{\mathsf{cover,all}}}{(1-\gamma)^5\varepsilon^2}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)\log\Big(\frac{1}{\varepsilon(1-\gamma)^2}\Big),$$

for some sufficiently large constant $c' > 0$. This together with the definition (62) completes the proof.

VIII. Analysis Under Adaptive Learning Rates (Proof of Theorem 3)

A. Useful Preliminary Facts About η_t

To begin with, we make note of several useful properties about η_t.

- Invoking the concentration result in Lemma 8, one can easily show that with probability at least $1-\delta$,

$$\frac{1}{2}\mu_{\mathsf{min}} < \min_{s,a}\frac{K_t(s,a)}{t} < \frac{3}{2}\mu_{\mathsf{min}} \tag{66}$$

holds simultaneously for all t obeying $T \ge t \ge \frac{443t_{\mathsf{mix}}\log(\frac{4|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}}$. In addition, this concentration result taken collectively with the update rule (22) of $\widehat{\mu}_{\mathsf{min},t}$ — in particular, the second case of (22) — implies that $\widehat{\mu}_{\mathsf{min},t}$ "stabilizes" as t grows; to be precise, there exists some quantity $c' \in [1/6, 9/2]$ such that

$$\widehat{\mu}_{\mathsf{min},t} \equiv c'\mu_{\mathsf{min}} \tag{67}$$

holds simultaneously for all t obeying $T \ge t \ge \frac{443t_{\mathsf{mix}}\log(\frac{4|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}}$.

- For any t obeying $t \ge \frac{6c_\eta t_{\mathsf{mix}}\log(\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2}$ (so that $\frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t} \le \frac{1}{c_\eta}$ and $t \ge \frac{443t_{\mathsf{mix}}\log(\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}}$ for $c_\eta \ge 11$), the learning rate (23) simplifies to

$$\eta_t = c_\eta\exp\Big(\Big\lfloor\log\frac{\log t}{c'\mu_{\mathsf{min}}(1-\gamma)\gamma^2 t}\Big\rfloor\Big). \tag{68}$$

Clearly, there exists a sequence of endpoints $t_1 < t_2 < t_3 < \ldots$ with $t_1 \le \frac{6ec_\eta t_{\mathsf{mix}}\log(\frac{2|\mathcal{S}||\mathcal{A}|t_1}{\delta})}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2}$ such that:

$$2t_k < t_{k+1} < 3t_k \qquad \text{and} \tag{69}$$

$$\eta_t = \eta_{(k)} := \frac{\alpha_k\log t_{k+1}}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2 t_{k+1}}, \quad \forall t_k < t \le t_{k+1} \tag{70}$$

for some positive constant $\alpha_k \in \big[\frac{2c_\eta}{9e}, 6c_\eta\big]$; in words, (70) provides a concrete expression/bound for the piecewise constant learning rate, where the t_k's form the change points.

Combining (70) with the definition of $\widehat{Q}_t$ (cf. (22)), one can easily check that for $t > t_1$,

$$\widehat{Q}_t = Q_{t_k}, \qquad \forall t_k < t \le t_{k+1}, \tag{71}$$

meaning that $\widehat{Q}_t$ remains fixed within each time segment $(t_k, t_{k+1}]$. With this property in mind, we only need to analyze Q_{t_k} in the sequel, which can be easily accomplished by invoking Theorem 1.

B. A Crude Bound

Given that $0 < \eta_t \le 1$ and $0 \le r(s,a) \le 1$, the update rule (10) of $\boldsymbol{Q}_t$ implies that

$$\begin{aligned}&\|\boldsymbol{Q}_t\|_\infty \\ &\le \max\big\{(1-\eta_t)\|\boldsymbol{Q}_{t-1}\|_\infty + \eta_t(1+\gamma\|\boldsymbol{Q}_{t-1}\|_\infty), \|\boldsymbol{Q}_{t-1}\|_\infty\big\} \\ &\le \|\boldsymbol{Q}_{t-1}\|_\infty + \gamma,\end{aligned}$$

thus leading to the following crude bound that for any $t > \frac{1}{1-\gamma}$

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \le t + \|\boldsymbol{Q}_0\|_\infty + \|\boldsymbol{Q}^\star\|_\infty \le t + \frac{2}{1-\gamma} \le 3t. \tag{72}$$

Remark 5: As we shall see momentarily, this crude bound allows one to control — in a coarse manner — the error at the beginning of each time interval $[t_{k-1}, t_k]$, which is needed when invoking Theorem 1.

C. Refined Analysis

Let us define

$$\varepsilon_k := \sqrt{\frac{c_{k,0}\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})\log t_k}{\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2 t_k}}, \tag{73}$$

where the constant $c_{k,0}$ is chosen to be $c_{k,0} = \alpha_{k-1}/c_1 > 0$, with $c_1 > 0$ the universal constant stated in Theorem 1. The property (70) of η_t together with the definition (73) implies that

$$\eta_t = \frac{c_1(1-\gamma)^4\varepsilon_k^2}{\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})} = \frac{c_1}{\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})}\min\Big\{(1-\gamma)^4\varepsilon_k^2, \frac{1}{t_{\mathsf{mix}}}\Big\}$$

for any $t \in (t_{k-1}, t_k]$, as long as $(1-\gamma)^4\varepsilon_k^2 \le 1/t_{\mathsf{mix}}$, or more explicitly, when

$$t_k \ge \frac{c_{k,0}t_{\mathsf{mix}}\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})\log t_k}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2}. \tag{74}$$

In addition, the condition (69) and the definition (73) further tell us that

$$t_k - t_{k-1} > t_{k-1} > \frac{1}{3}t_k = \frac{c_{k,0}\log\big(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta}\big)\log t_k}{3\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2\varepsilon_k^2}.$$

Invoking Theorem 1 with an initialization $\boldsymbol{Q}_{t_{k-1}}$ (which clearly satisfies the crude bound (72)) ensures that

$$\|\boldsymbol{Q}_{t_k} - \boldsymbol{Q}^\star\|_\infty \le \varepsilon_k \tag{75}$$

with probability at least $1-\delta$, with the proviso that

$$\begin{aligned}\frac{1}{3}t_k \ge \frac{c_0}{\mu_{\mathsf{min}}}&\left\{\frac{1}{(1-\gamma)^5\varepsilon_k^2} + \frac{t_{\mathsf{mix}}}{1-\gamma}\right\} \\ &\cdot \log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta}\Big)\log\Big(\frac{t_k}{(1-\gamma)^2\varepsilon_k}\Big)\end{aligned} \tag{76}$$

with $c_0 > 0$ the universal constant stated in Theorem 1. Under the sample size condition (74), this requirement (76) can be guaranteed by adjusting the constant c_η in (23) to satisfy the following inequality:

$$c_{k,0} = \frac{\alpha_{k-1}}{c_1} \ge \frac{2c_\eta}{9ec_1} > 6c_0.$$

Finally, taking $t_{k_{\max}}$ to be the largest change point that does not exceed T, we see from (69) that $\frac{1}{3}T \le t_{k_{\max}} \le T$. Then one has

$$\begin{aligned}\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\|_\infty &= \|\boldsymbol{Q}_{t_{k_{\max}}} - \boldsymbol{Q}^\star\|_\infty \\ &\le \varepsilon_{k_{\max}} = \sqrt{\frac{c_{k,0}\log(\frac{|\mathcal{S}||\mathcal{A}|t_{k_{\max}}}{\delta})\log t_{k_{\max}}}{\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2 t_{k_{\max}}}} \\ &\le \sqrt{\frac{3c_{k,0}\log(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})\log T}{\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2 T}}\end{aligned} \tag{77}$$

These immediately conclude the proof of the theorem under the sample size condition (26), provided that

$$C > \frac{18c_\eta}{c_1} > \frac{3\alpha_{k-1}}{c_1} = 3c_{k,0}.$$

IX. Analysis of Asynchronous Variance-Reduced Q-Learning

This section aims to establish Theorem 4. We carry out an epoch-based analysis, that is, we first quantify the progress made over each epoch, and then demonstrate how many epochs are sufficient to attain the desired accuracy. In what follows, we shall overload the notation by defining

$$t_{\mathsf{frame}} := \frac{443t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big), \tag{78a}$$

$$t_{\mathsf{th}} := \max\left\{\frac{2\log\frac{1}{(1-\gamma)^2\varepsilon}}{\eta\mu_{\mathsf{min}}},\ t_{\mathsf{frame}}\right\}, \tag{78b}$$

$$\rho := (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big), \tag{78c}$$

$$\mu_{\mathsf{frame}} := \frac{1}{2}\mu_{\mathsf{min}}t_{\mathsf{frame}}. \tag{78d}$$

A. Per-Epoch Analysis

We start by analyzing the progress made over each epoch. Before proceeding, we denote by $\widetilde{\boldsymbol{P}} \in [0,1]^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ a matrix corresponding to the empirical probability transition kernel used in (28) from N new sample transitions. Further, we use the vector $\overline{\boldsymbol{Q}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent the reference Q-function, and introduce the vector $\overline{\boldsymbol{V}} \in \mathbb{R}^{|\mathcal{S}|}$ to represent the corresponding value function so that $\overline{V}(s) := \max_a \overline{Q}(s,a)$ for all $s \in \mathcal{S}$.

For convenience, this subsection abuses notation to assume that an epoch starts with an estimate $\boldsymbol{Q}_0 = \overline{\boldsymbol{Q}}$, and consists of the subsequent

$$t_{\mathsf{epoch}} := t_{\mathsf{frame}} + t_{\mathsf{th}} + \frac{8\log\frac{2}{1-\gamma}}{(1-\gamma)\eta\mu_{\mathsf{min}}} \tag{79}$$

iterations of variance-reduced Q-learning updates, where t_{frame} and t_{th} are defined in (78a) and (78b), respectively. In the sequel, we divide all epochs into two phases, depending on the quality of the initial estimate $\overline{\boldsymbol{Q}}$ in each epoch.

1) Phase 1: When $\|\overline{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty > 1/\sqrt{1-\gamma}$: Recalling the matrix notation of $\boldsymbol{\Lambda}_t$ and $\boldsymbol{P}_t$ in (34) and (35), respectively, we can rewrite (27) as follows

$$\boldsymbol{Q}_t = \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}\right). \tag{80}$$

Following similar steps as in the expression (39), we arrive at the following error decomposition

$$\begin{aligned}
\boldsymbol{\Theta}_t &:= \boldsymbol{Q}_t - \boldsymbol{Q}^\star = \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} \\
&\quad + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}\right) - \boldsymbol{Q}^\star \\
&= \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1}-\boldsymbol{Q}^\star\big) \\
&\quad + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} - \boldsymbol{Q}^\star\right) \\
&= \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1}-\boldsymbol{Q}^\star\big) \\
&\quad + \gamma\boldsymbol{\Lambda}_t\left(\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} - \boldsymbol{P}\boldsymbol{V}^\star\right) \\
&= \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\boldsymbol{\Theta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\widetilde{\boldsymbol{P}}-\boldsymbol{P}\big)\overline{\boldsymbol{V}} \\
&\quad + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t-\boldsymbol{P}\big)\big(\boldsymbol{V}^\star-\overline{\boldsymbol{V}}\big) + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}_t\big(\boldsymbol{V}_{t-1}-\boldsymbol{V}^\star\big),
\end{aligned} \tag{81}$$

which once again leads to a recursive relation

$$\begin{aligned}
\boldsymbol{\Theta}_t &= \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\widetilde{\boldsymbol{P}}-\boldsymbol{P}\big)\overline{\boldsymbol{V}}}_{=:\boldsymbol{h}_{0,t}} \\
&\quad + \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i-\boldsymbol{P}\big)\big(\boldsymbol{V}^\star-\overline{\boldsymbol{V}}\big)}_{=:\boldsymbol{h}_{1,t}} \\
&\quad + \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1}-\boldsymbol{V}^\star\big)}_{=:\boldsymbol{h}_{2,t}} \\
&\quad + \underbrace{\prod_{j=1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Theta}_0}_{=:\boldsymbol{h}_{3,t}}.
\end{aligned} \tag{82}$$

This identity takes a very similar form as (40) except for the additional term $\boldsymbol{h}_{0,t}$.

Let us begin by controlling the first term, towards which we have the following lemma. The proof is postponed to Section G.

Lemma 7: Suppose that $\widetilde{\boldsymbol{P}}$ is constructed using N consecutive sample transitions. If $N > t_{\mathsf{frame}}$, then with probability greater than $1-\delta$, one has

$$\begin{aligned}
\|\boldsymbol{h}_{0,t}\|_\infty &\le \gamma\sqrt{\frac{4\log\big(\frac{6N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}}\big\|\overline{\boldsymbol{V}}-\boldsymbol{V}^\star\big\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma}\sqrt{\frac{4\log\big(\frac{6N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}}.
\end{aligned} \tag{83}$$

If $t < t_{\mathsf{frame}}$, then it is straightforwardly seen that

$$|\boldsymbol{h}_{3,t}| \le \|\boldsymbol{\Theta}_0\|_\infty\boldsymbol{1}.$$

Taking this together with the results from Lemma 1 and Lemma 2, we are guaranteed that

$$\begin{aligned}
\big|\boldsymbol{h}_{1,t}\big| &\le \tau_2\|\boldsymbol{V}^\star-\overline{\boldsymbol{V}}\|_\infty\boldsymbol{1} \\
|\boldsymbol{h}_{3,t}| &\le \begin{cases} (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\|\boldsymbol{\Theta}_0\|_\infty\boldsymbol{1}, & \text{if } t_{\mathsf{frame}} \le t \le t_{\mathsf{epoch}} \\ \|\boldsymbol{\Theta}_0\|_\infty\boldsymbol{1}, & \text{if } t < t_{\mathsf{frame}} \end{cases}
\end{aligned}$$

with probability at least $1-2\delta$, where

$$\tau_2 := c'\gamma\sqrt{\eta\log\big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\big)}$$

for some constant $c' > 0$ (similar to (44)). In addition, the term $\boldsymbol{h}_{2,t}$ can be bounded in the same way as $\boldsymbol{\beta}_{2,t}$ in (42). Therefore, repeating the same argument as for Theorem 5 and taking $\xi = \frac{1}{16\sqrt{1-\gamma}}$, we conclude that with probability at least $1-\delta$,

$$\begin{aligned}
\|\boldsymbol{\Theta}_t\|_\infty &\le (1-\rho)^k\frac{\|\boldsymbol{\Theta}_0\|_\infty}{1-\gamma} + \widetilde{\tau} + \xi \\
&= (1-\rho)^k\frac{\|\overline{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \widetilde{\tau} + \xi
\end{aligned} \tag{84}$$

holds simultaneously for all $0 < t \le t_{\mathsf{epoch}}$, where $k = \max\big\{0, \lfloor\frac{t-t_{\mathsf{th},\xi}}{t_{\mathsf{frame}}}\rfloor\big\}$, and

$$\begin{aligned}
\widetilde{\tau} &:= \frac{c\gamma}{1-\gamma}\Bigg\{\sqrt{\frac{\log\frac{N|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^2N\mu_{\mathsf{min}}}} + \|\boldsymbol{V}^\star-\overline{\boldsymbol{V}}\|_\infty \\
&\quad\cdot\left(\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big)} + \sqrt{\frac{4\log\big(\frac{6N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}}\right)\Bigg\}, \\
t_{\mathsf{th},\xi} &:= \max\left\{\frac{2\log\frac{1}{(1-\gamma)^2\xi}}{\eta\mu_{\min}}, t_{\mathsf{frame}}\right\}
\end{aligned}$$

for some constant $c > 0$.

Let $C > 0$ be some sufficient large constant. Setting $\eta_t \equiv \eta = \min\Big\{\frac{(1-\gamma)^2}{C\gamma^2\log\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}}, \frac{1}{\mu_{\mathsf{frame}}}\Big\}$, and ensuring $N \ge \max\{t_{\mathsf{frame}}, C\frac{\log\frac{N|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^3\mu_{\mathsf{min}}}\}$, we can easily demonstrate that

$$\begin{aligned}
\|\boldsymbol{\Theta}_t\|_\infty &\le (1-\rho)^k\frac{\|\overline{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty}{1-\gamma} \\
&\quad + \frac{1}{8\sqrt{1-\gamma}} + \frac{1}{4}\|\boldsymbol{V}^\star-\overline{\boldsymbol{V}}\|_\infty.
\end{aligned}$$

As a consequence, if $t_{\mathsf{epoch}} \geq t_{\mathsf{frame}} + t_{\mathsf{th},\xi} + \frac{8\log\frac{2}{1-\gamma}}{(1-\gamma)\eta\mu_{\mathsf{min}}}$, one has

$$(1-\rho)^k \leq \frac{1}{8}(1-\gamma),$$

which in turn implies that

$$\begin{aligned}\|\boldsymbol{\Theta}_{t_{\mathsf{epoch}}}\|_\infty &\leq \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{1}{8\sqrt{(1-\gamma)}} + \frac{1}{4}\|\boldsymbol{V}^\star - \overline{\boldsymbol{V}}\|_\infty \\ &\leq \frac{1}{2}\max\Big\{\frac{1}{\sqrt{1-\gamma}},\ \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty\Big\}, \end{aligned} \tag{85}$$

where the last step invokes the simple relation $\|\boldsymbol{V}^\star - \overline{\boldsymbol{V}}\|_\infty \leq \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty$. Thus, we conclude that

$$\|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \boldsymbol{Q}^\star\|_\infty \leq \frac{1}{2}\max\Big\{\frac{1}{\sqrt{1-\gamma}},\ \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty\Big\}. \tag{86}$$

2) Phase 2: When $\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \leq 1/\sqrt{1-\gamma}$: The analysis of Phase 2 follows by straightforwardly combining the analysis of Phase 1 and that of the synchronous counterpart in [12]. For the sake of brevity, we only sketch the main steps.

Following the proof idea of [12, Section B.2], we introduce an auxiliary vector $\widehat{\boldsymbol{Q}}$ which is the unique fix point to the following equation, which can be regarded as a population-level Bellman equation with proper reward perturbation, namely,

$$\widehat{\boldsymbol{Q}} = \boldsymbol{r} + \gamma\boldsymbol{P}(\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}. \tag{87}$$

Here, as usual, $\widehat{\boldsymbol{V}} \in \mathbb{R}^{|\mathcal{S}|}$ represents the value function corresponding to $\widehat{\boldsymbol{Q}}$. This can be viewed as a Bellman equation when the reward vector $\boldsymbol{r}$ is replaced by $\widetilde{\boldsymbol{r}} := \boldsymbol{r} + \gamma(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}$. Repeating the arguments in the proof of [12, Lemma 4] (except that we need to apply the measure concentration of $\widetilde{\boldsymbol{P}}$ in the manner performed in the proof of Lemma 7 due to Markovian data), we reach

$$\big\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\big\|_\infty \leq c'\sqrt{\frac{\log\frac{N|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^3 N\mu_{\mathsf{min}}}} \leq \varepsilon \tag{88}$$

with probability at least $1-\delta$ for some constant $c' > 0$, provided that $N \geq (c')^2\frac{\log\frac{N|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^3\varepsilon^2}$ and that $\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \leq 1/\sqrt{1-\gamma}$. It is worth noting that $\widehat{\boldsymbol{Q}}$ only serves as a helper in the proof and is never explicitly constructed in the algorithm, as we don't have access to the probability transition matrix $\boldsymbol{P}$.

In addition, we claim that

$$\big\|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \widehat{\boldsymbol{Q}}\big\|_\infty \leq \frac{\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty}{8} + \frac{\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty}{8} + \varepsilon. \tag{89}$$

Under this claim, the triangle inequality yields

$$\begin{aligned}\|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \boldsymbol{Q}^\star\|_\infty &\leq \|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \widehat{\boldsymbol{Q}}\|_\infty + \|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \\ &\leq \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{9}{8}\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \varepsilon \\ &\leq \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{17}{8}\varepsilon, \end{aligned} \tag{90}$$

where the last inequality follows from (88).

a) Proof of the inequality (88)*:* Suppose that

$$\begin{aligned}\big|\widetilde{\boldsymbol{r}} - \boldsymbol{r}\big| &= \gamma\big|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}\big| \\ &\leq c\left\{\frac{1}{\sqrt{1-\gamma}}\mathbf{1} + \sqrt{\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star)}\right\}\sqrt{\frac{\log\frac{N|\mathcal{S}||\mathcal{A}|}{\delta}}{N\mu_{\mathsf{min}}}}, \end{aligned} \tag{91}$$

holds for some constant $c > 0$. By replacing Lemma 5 in the proof of [12, Lemma 4] with this bound, we can arrive at (88) immediately. In what follows, we demonstrate how to prove the bound (91), which follows a similar argument as in the proof of Lemma 7.

Let us begin with the following triangle inequality:

$$\big|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}\big| \leq \big|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})(\overline{\boldsymbol{V}} - \boldsymbol{V}^\star)\big| + \big|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\boldsymbol{V}^\star\big|, \tag{92}$$

leaving us with two terms to control.

- Similar to (141), by applying the Hoeffding inequality and taking the union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we can control the first term on the right-hand side of (92) as follows:

$$\begin{aligned}&\big\|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})(\overline{\boldsymbol{V}} - \boldsymbol{V}^\star)\big\|_\infty \\ &\quad\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{\frac{2\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{K_N(s,a)}}\big\|\overline{\boldsymbol{V}} - \boldsymbol{V}^\star\big\|_\infty \\ &\quad\leq \sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}(1-\gamma)}} \end{aligned} \tag{93}$$

  with probability at least $1-\delta$. Here, we have made use of the following property of this phase that

$$\big\|\overline{\boldsymbol{V}} - \boldsymbol{V}^\star\big\|_\infty \leq \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \leq 1/\sqrt{1-\gamma}$$

  and $K_N(s,a) \geq N\mu_{\mathsf{min}}/2$ for all (s,a) (see Lemma 8).
- Next, we turn attention to the second term on the right-hand side of (92), towards which we resort to the Bernstein inequality. Note that the (s,a)-th entry of $\big|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\boldsymbol{V}^\star\big|$ is given by

$$\left|\frac{1}{K_N(s,a)}\sum_{i=1}^{K_N(s,a)}\big(\boldsymbol{P}_{t_i+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right|, \tag{94}$$

  where $K_N(s,a)$ denotes the total number of visits to (s,a) during the first N time instances (see also (113)). In addition, let $t_i := t_i(s,a)$ denote the time stamp when the trajectory visits (s,a) for the i-th time (see also (112)). In view of our derivation for (117), the state transitions happening at times $t_1, t_2, \cdots, t_k$ (which are random) are independent for any given integer $k > 0$. It can be calculated that

$$\Big|\big(\boldsymbol{P}_{t_i+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\Big| \leq \frac{1}{1-\gamma}; \tag{95a}$$

$$\begin{aligned}&\mathsf{Var}\left(\frac{1}{k}\sum_{i=1}^{k}\big(\boldsymbol{P}_{t_i+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right) \\ &\qquad= \frac{1}{k}\mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}^\star\big). \end{aligned} \tag{95b}$$

Consequently, invoking the Bernstein inequality implies that with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|}$,

$$\left|\frac{1}{k}\sum_{i=1}^{k}\big(\boldsymbol{P}_{t_i+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right| \le \sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{k}\mathsf{Var}_{\boldsymbol{P}(s,a)}(\boldsymbol{V}^\star)}+\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{3(1-\gamma)k}$$

holds simultaneously for all $1\le k\le N$. Recognizing the bound $\frac{1}{2}N\mu_{\mathsf{min}}\le K_N(s,a)\le N$ and applying the union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$ yield

$$\big|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\boldsymbol{V}^\star\big| \le \sqrt{\frac{2\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star)} + \frac{8\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{3(1-\gamma)N\mu_{\mathsf{min}}}. \tag{96}$$

- Finally, combining (93) and (96) immediately establishes the claim (91).

b) Proof of the Inequality (89)*:* Recalling the variance-reduced update rule (80) and using the Bellman-type equation (87), we obtain

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_t &:= \boldsymbol{Q}_t-\widehat{\boldsymbol{Q}}=\big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)(\boldsymbol{Q}_{t-1}-\widehat{\boldsymbol{Q}})+\boldsymbol{\Lambda}_t\Big(\boldsymbol{r}+\gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}\\ &\quad-\overline{\boldsymbol{V}})+\gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}-\boldsymbol{r}-\gamma\boldsymbol{P}(\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}})-\gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}\Big)\\ &=\big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)(\boldsymbol{Q}_{t-1}-\widehat{\boldsymbol{Q}})\\ &\quad+\boldsymbol{\Lambda}_t\Big(\gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}})-\gamma\boldsymbol{P}(\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}})\Big)\\ &=\big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\widehat{\boldsymbol{\Theta}}_{t-1}\\ &\quad+\gamma\boldsymbol{\Lambda}_t\Big((\boldsymbol{P}_t-\boldsymbol{P})(\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}})+\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\widehat{\boldsymbol{V}})\Big).\end{aligned} \tag{97}$$

Adopting the same expansion as before (see (40)), we arrive at

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_t &= \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i-\boldsymbol{P}\big)(\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}})}_{=:\boldsymbol{\vartheta}_{1,t}}\\ &\quad+\underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1}-\widehat{\boldsymbol{V}}\big)}_{=:\boldsymbol{\vartheta}_{2,t}}\\ &\quad+\underbrace{\prod_{j=1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\widehat{\boldsymbol{\Theta}}_0}_{=:\boldsymbol{\vartheta}_{3,t}}.\end{aligned}$$

Inheriting the results in Lemma 1 and Lemma 2, we can demonstrate that, with probability at least $1-2\delta$,

$$\big|\boldsymbol{\vartheta}_{1,t}\big|\le c\gamma\|\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}}\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big)}\mathbf{1};$$

$$|\boldsymbol{\vartheta}_{3,t}|\le\begin{cases}(1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\|\widehat{\boldsymbol{\Theta}}_0\|_\infty\mathbf{1}, & \text{if } t_{\mathsf{frame}}\le t\le t_{\mathsf{epoch}},\\ \|\widehat{\boldsymbol{\Theta}}_0\|_\infty\mathbf{1}, & \text{if } t<t_{\mathsf{frame}}.\end{cases}$$

Repeating the same argument as for Theorem 5, we reach

$$\|\widehat{\boldsymbol{\Theta}}_t\|_\infty\le(1-\rho)^k\frac{\|\widehat{\boldsymbol{Q}}-\overline{\boldsymbol{Q}}\|_\infty}{1-\gamma}+\frac{c\gamma}{1-\gamma}\|\widehat{\boldsymbol{V}}-\overline{\boldsymbol{V}}\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big)}+\varepsilon$$

for some constant $c>0$, where $k=\max\{0,\lfloor\frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\rfloor\}$ with t_{th} defined in (78b).

By taking $\eta=c_5\min\big\{\frac{(1-\gamma)^2}{\gamma^2\log\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}},\frac{1}{\mu_{\mathsf{frame}}}\big\}$ for some sufficiently small constant $c_5>0$ and ensuring that

$$t_{\mathsf{epoch}}\ge t_{\mathsf{th}}+t_{\mathsf{frame}}+\frac{c_6}{(1-\gamma)\eta\mu_{\mathsf{min}}}\log\frac{1}{(1-\gamma)^2}$$

for some large constant $c_6>0$, we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}_{t_{\mathsf{epoch}}}\|_\infty&\le\frac{\|\widehat{\boldsymbol{Q}}-\overline{\boldsymbol{Q}}\|_\infty}{8}+\varepsilon\\ &\le\frac{\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty}{8}+\frac{\|\overline{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty}{8}+\varepsilon,\end{aligned}$$

where the last line follows by the triangle inequality.

B. How Many Epochs Are Needed?

We are now ready to pin down how many epochs are needed to achieve ε-accuracy.

- In Phase 1, the contraction result (86) indicates that, if the algorithm is initialized with $\boldsymbol{Q}_0=\mathbf{0}$ at the very beginning, then it takes at most
$$\log_2\left(\frac{\|\boldsymbol{Q}^\star\|_\infty}{\max\{\varepsilon,\frac{1}{\sqrt{1-\gamma}}\}}\right)\le\log_2\Big(\frac{1}{\sqrt{1-\gamma}}\Big)+\log_2\Big(\frac{1}{\varepsilon(1-\gamma)}\Big)$$
epochs to yield $\|\overline{\boldsymbol{Q}}-\boldsymbol{Q}^\star\|_\infty\le\max\{\frac{1}{\sqrt{1-\gamma}},\varepsilon\}$ (so as to enter Phase 2). Clearly, if the target accuracy level $\varepsilon>\frac{1}{\sqrt{1-\gamma}}$, then the algorithm terminates in this phase.
- Suppose now that the target accuracy level $\varepsilon\le\frac{1}{\sqrt{1-\gamma}}$. Once the algorithm enters Phase 2, the dynamics can be characterized by (90). Given that $\overline{\boldsymbol{Q}}$ is also the last iterate of the preceding epoch, the property (90) provides a recursive relation across epochs. Standard recursive analysis thus reveals that: within at most
$$c_7\log\Big(\frac{1}{\varepsilon\sqrt{1-\gamma}}\Big)\le c_7\log\Big(\frac{1}{\varepsilon(1-\gamma)}\Big)$$
epochs (with $c_7>0$ some constant), we are guaranteed to attain an ℓ_∞ estimation error at most 3ε.

To summarize, a total number of $O\big(\log\frac{1}{\varepsilon(1-\gamma)}+\log\frac{1}{1-\gamma}\big)$ epochs are sufficient for our purpose. This concludes the proof.

X. Discussion

This work develops a sharper finite-sample analysis of the classical asynchronous Q-learning algorithm, highlighting and refining its dependency on intrinsic features of the Markovian trajectory induced by the behavior policy. Our sample complexity bound strengthens the state-of-the-art result by an order

of at least $|\mathcal{S}||\mathcal{A}|$. A variance-reduced variant of asynchronous Q-learning is also analyzed, exhibiting improved scaling with the effective horizon $\frac{1}{1-\gamma}$.

Our findings and the analysis framework developed herein suggest a couple of directions for future investigation. For instance, our improved sample complexity of asynchronous Q-learning has a dependence of $\frac{1}{(1-\gamma)^5}$ on the effective horizon, which is inferior to its model-based counterpart. In the synchronous setting, [32], [34] recently demonstrated Q-learning has a dependence of $\frac{1}{(1-\gamma)^4}$, which is tight up to logarithmic factors. In light of this development, it would be important to determine the exact scaling for the asynchronous setting, which is left as future work. In addition, it would be interesting to see whether the techniques developed herein can be exploited towards understanding model-free algorithms with more sophisticated exploration schemes [64]. Finally, asynchronous Q-learning on a single Markovian trajectory is closely related to coordinate descent with coordinates selected according to a Markov chain; one would naturally ask whether our analysis framework can yield improved convergence guarantees for general Markov-chain-based optimization algorithms [65], [66].

APPENDIX

In this section, we gather some basic facts about Markov chains. Before proceeding, we remind the readers of some notation. For any two probability distributions μ and ν, denote by $d_{\mathsf{TV}}(\mu,\nu)$ the total variation distance between μ and ν (cf. (5)). Recall the definition of uniform ergodicity in Section I-B. For any *time-homogeneous* and *uniformly ergodic* Markov chain $(X_0, X_1, X_2, \cdots)$ with transition kernel P, finite state space $\mathcal{X}$ and stationary distribution μ, we let $P^t(\cdot \,|\, x)$ denote the distribution of X_t conditioned on $X_0 = x$. Then the mixing time t_{mix} of this Markov chain is defined by

$$t_{\mathsf{mix}}(\epsilon) := \min\left\{t \;\middle|\; \max_{x\in\mathcal{X}} d_{\mathsf{TV}}\big(P^t(\cdot \,|\, x), \mu\big) \le \epsilon\right\}; \tag{98a}$$

$$t_{\mathsf{mix}} := t_{\mathsf{mix}}(1/4). \tag{98b}$$

A. Concentration of Empirical Distributions of Markov Chains

We first record a result concerning the concentration of measure of the empirical distribution of a uniformly ergodic Markov chain, which makes clear the role of the mixing time.

Lemma 8: Consider the above-mentioned Markov chain. For any $0 < \delta < 1$, if $t \ge \frac{443 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \log \frac{4|\mathcal{X}|}{\delta}$, then for any $y \in \mathcal{X}$, one has

$$\mathbb{P}_{X_1=y}\left\{\exists x \in \mathcal{X} : \left|\sum_{i=1}^{t} \mathbb{1}\{X_i = x\} - t\mu(x)\right| \ge \frac{1}{2} t\mu(x)\right\} \le \delta. \tag{99}$$

Proof: To begin with, consider the scenario when $X_1 \sim \mu$, namely, when X_1 follows the stationary distribution of the chain. Then [17, Theorem 3.4] tells us that: for any given $x \in \mathcal{X}$ and any $\tau \ge 0$,

$$\begin{aligned}
&\mathbb{P}_{X_1\sim\mu}\left\{\left|\sum_{i=1}^{t} \mathbb{1}\{X_i = x\} - t\mu(x)\right| \ge \tau\right\} \\
&\quad\le 2\exp\left(-\frac{\tau^2\gamma_{\mathsf{ps}}}{8(t+1/\gamma_{\mathsf{ps}})\mu(x) + 20\tau}\right) \\
&\quad\le 2\exp\left(-\frac{\tau^2/t_{\mathsf{mix}}}{16(t+2t_{\mathsf{mix}})\mu(x) + 40\tau}\right),
\end{aligned} \tag{100}$$

where γ_{ps} stands for the so-called *pseudo spectral gap* as defined in [17, Section 3.1]. Here, the first inequality relies on the fact $\mathsf{Var}_{X_i\sim\mu}[\mathbb{1}\{X_i = x\}] = \mu(x)(1-\mu(x)) \le \mu(x)$, while the last inequality results from the fact $\gamma_{\mathsf{ps}} \ge 1/(2t_{\mathsf{mix}})$ that holds for uniformly ergodic chains (cf. [17, Proposition 3.4]). Consequently, for any $t \ge t_{\mathsf{mix}}$ and any $\tau \ge 0$, one can continue the bound (100) to obtain

$$\begin{aligned}
(100) &\le 2\exp\left(-\frac{\tau^2}{48t\mu(x)t_{\mathsf{mix}} + 40\tau t_{\mathsf{mix}}}\right) \\
&\le 2\max\left\{\exp\left(-\frac{\tau^2}{96t\mu(x)t_{\mathsf{mix}}}\right), \exp\left(-\frac{\tau}{80t_{\mathsf{mix}}}\right)\right\} \\
&\le \frac{\delta}{|\mathcal{X}|},
\end{aligned}$$

provided that

$$\tau \ge \max\left\{10\sqrt{t\mu(x)t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}},\ 80t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}\right\}.$$

As a result, by taking $\tau = \frac{10}{21}t\mu(x)$ and applying the union bound, we reach

$$\begin{aligned}
&\mathbb{P}_{X_1\sim\mu}\left\{\exists x \in \mathcal{X} : \left|\sum_{i=1}^{t} \mathbb{1}\{X_i = x\} - t\mu(x)\right| \ge \frac{10}{21} t\mu(x)\right\} \\
&\le \sum_{x\in\mathcal{X}} \mathbb{P}_{X_1\sim\mu}\left\{\left|\sum_{i=1}^{t} \mathbb{1}\{X_i = x\} - t\mu(x)\right| \ge \frac{10}{21} t\mu(x)\right\} \le \delta,
\end{aligned} \tag{101}$$

as long as

$$\frac{10}{21}t\mu(x) \ge \max\left\{10\sqrt{t\mu(x)t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}},\ 80t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}\right\}$$

for all $x \in \mathcal{X}$, or equivalently, when

$$t \ge \frac{441 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{2|\mathcal{X}|}{\delta} \qquad \text{with } \mu_{\mathsf{min}} := \min_{x\in\mathcal{X}}\mu(x).$$

Next, we seek to extend the above result to the more general case when X_1 takes an arbitrary state $y \in \mathcal{X}$. From the definition of $t_{\mathsf{mix}}(\cdot)$ (cf. (98a)), we know that

$$d_{\mathsf{TV}}\Big(\sup_{y\in\mathcal{X}} P^{t_{\mathsf{mix}}(\delta)}(\cdot \,|\, y),\ \mu\Big) \le \delta. \tag{102}$$

This taken together with the definition of d_{TV} (cf. (5)) reveals that: for any event $\mathcal{B}$ belonging to the σ-algebra generated by $\{X_\tau\}_{\tau \ge t_{\mathsf{mix}}(\delta)}$, one has

$$\begin{aligned}
&\big|\mathbb{P}\{\mathcal{B} \mid X_1 = y\} - \mathbb{P}\{\mathcal{B} \mid X_1 \sim \mu\}\big| \\
&= \left|\sum_{s\in\mathcal{S}} \mathbb{P}\{\mathcal{B} \mid X_{t_{\mathsf{mix}}(\delta)} = s\}\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)} = s \mid X_1 = y\}\right.
\end{aligned}$$

$$
\begin{aligned}
&\left. -\sum_{s\in\mathcal{S}} \mathbb{P}\{\mathcal{B} \mid X_{t_{\mathsf{mix}}(\delta)}=s\}\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1\sim\mu\}\right| \\
&\le \max\Bigg\{ \sum_{s\in\mathcal{S}_+} \Big[\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1=y\}-\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1\sim\mu\}\Big], \\
&\qquad \sum_{s\in\mathcal{S}_-} \Big[\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1\sim\mu\}-\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1=y\}\Big]\Bigg\} \\
&\le \sup_{A\subseteq\mathcal{S}} \Big|\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}\in A \mid X_1=y\}-\mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}\in A \mid X_1\sim\mu\}\Big| \le \delta,
\end{aligned} \tag{103}
$$

where we define

$$
\begin{aligned}
\mathcal{S}_+ &:= \big\{s\in\mathcal{S}: \\
&\qquad \mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1=y\} > \mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1\sim\mu\}\big\}; \\
\mathcal{S}_- &:= \big\{s\in\mathcal{S}: \\
&\qquad \mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1=y\} < \mathbb{P}\{X_{t_{\mathsf{mix}}(\delta)}=s \mid X_1\sim\mu\}\big\}.
\end{aligned}
$$

Here, the last inequality in (103) follows from the inequality (102) and the definition (5) of the total-variation distance. As a consequence, one obtains

$$
\begin{aligned}
&\sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\left\{\exists x\in\mathcal{X}: \left|\sum_{i=t_{\mathsf{mix}}(\delta)}^{t} \mathbb{1}\{X_i=x\}-\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)\right| \right. \\
&\qquad \left. \ge \frac{10}{21}\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)\right\} \\
&\le \mathbb{P}_{X_1\sim\mu}\left\{\exists x\in\mathcal{X}: \left|\sum_{i=t_{\mathsf{mix}}(\delta)}^{t} \mathbb{1}\{X_i=x\}-\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)\right| \right. \\
&\qquad \left. \ge \frac{10}{21}\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)\right\}+\delta \le 2\delta,
\end{aligned} \tag{104}
$$

with the proviso that $t \ge t_{\mathsf{mix}}(\delta) + \frac{441 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{2|\mathcal{X}|}{\delta}$.

To finish up, we recall from [17, Section 1.1] that $t_{\mathsf{mix}}(\delta) \le 2t_{\mathsf{mix}}\log\frac{2}{\delta}$. Consequently, if $t \ge \frac{443 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{2|\mathcal{X}|}{\delta} \ge t_{\mathsf{mix}}(\delta) + \frac{441 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{2|\mathcal{X}|}{\delta}$, then one has

$$
\begin{aligned}
&\frac{1}{2}\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)-t_{\mathsf{mix}}(\delta) \ge \frac{441 t_{\mathsf{mix}}}{2}\log\frac{2|\mathcal{X}|}{\delta} \ge 100 t_{\mathsf{mix}}(\delta), \\
&\quad\Longrightarrow\quad \frac{1}{2}\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x)-t_{\mathsf{mix}}(\delta) \ge \frac{10}{21}\big(t-t_{\mathsf{mix}}(\delta)\big)\mu(x).
\end{aligned}
$$

These taken together lead to (105), shown at the bottom of the next page. Here, the last inequality of (105) results from (104). Replacing δ with $\delta/2$ thus concludes the proof. □

B. Connection Between the Mixing Time and the Cover Time

Lemma 8 combined with the definition (9) immediately reveals the following upper bound on the cover time:

$$
t_{\mathsf{cover}} = O\Big(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log|\mathcal{X}|\Big). \tag{106}
$$

In addition, while a general matching converse bound (namely, $t_{\mathsf{mix}}/\mu_{\mathsf{min}} = \widetilde{O}(t_{\mathsf{cover}})$) is not available, we can come up with some special examples for which the bound (106) is provably tight.

Example 1: Consider a time-homogeneous Markov chain with state space $\mathcal{X} := \{1,\cdots,|\mathcal{X}|\}$ and probability transition matrix $\boldsymbol{P}\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ as

$$
\boldsymbol{P} = \Big(1-\frac{q(k+1)}{2}\Big)\boldsymbol{I}_{|\mathcal{X}|} + \frac{q}{|\mathcal{X}|}\begin{bmatrix} k\mathbf{1}_{|\mathcal{X}|}\mathbf{1}_{|\mathcal{X}|/2}^\top & \mathbf{1}_{|\mathcal{X}|}\mathbf{1}_{|\mathcal{X}|/2}^\top \end{bmatrix} \tag{107}
$$

for some quantities $q>0$ and $k\ge 1$. Suppose $q(k+1)<2$ and $|\mathcal{X}|\ge 3$. Then this chain obeys

$$
t_{\mathsf{cover}} \ge \frac{t_{\mathsf{mix}}}{\big(8\log 2 + 4\log\frac{1}{\mu_{\mathsf{min}}}\big)\mu_{\mathsf{min}}}. \tag{108}
$$

With the lower bound (108) in place, we conclude that the upper bound (106) is, in general, nearly un-improvable (up to some logarithmic factor).

Remark 6: We shall take a moment to briefly discuss the key design rationale behind Example 1. Let us partition the state space into two halves, denoted respectively by $\mathcal{X}_1$ and $\mathcal{X}_2$. From every state $s\in\mathcal{X}$, it is much easier to transition into the first half $\mathcal{X}_1$ rather than the second half $\mathcal{X}_2$. This leads to two properties: (i) the stationary distribution of any state in $\mathcal{X}_2$ is much lower than that of a state in $\mathcal{X}_1$; (ii) the cover time also increases as the stationary distribution w.r.t. $\mathcal{X}_2$ decreases, given that it becomes more difficult to traverse the second half. As a result, we can guarantee that t_{cover} is proportional to μ_{min} through this type of designs. On the other hand, the example is also constructed in a way such that all states are "lazy", meaning that they are more inclined to stay unchanged rather than moving to a different state. The level of laziness clearly controls how fast the Markov chain mixes, as well as how long it takes to cover all states. This in turn allows one to ensure that t_{cover} is proportional to t_{mix}. More details can be found in the proof below.

Proof: As can be easily verified, this chain is reversible, whose stationary distribution vector $\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{X}|}$ obeys

$$
\boldsymbol{\mu} = \frac{2}{(k+1)|\mathcal{X}|}\begin{bmatrix} k\mathbf{1}_{|\mathcal{X}|/2} \\ \mathbf{1}_{|\mathcal{X}|/2}\end{bmatrix}.
$$

As a result, the minimum state occupancy probability of the stationary distribution is given by

$$
\mu_{\mathsf{min}} := \min_{1\le x\le|\mathcal{X}|}\mu_x = \frac{2}{(k+1)|\mathcal{X}|}. \tag{109}
$$

In addition, the reversibility of this chain implies that the matrix $\boldsymbol{P}^{\mathrm{d}} := \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}^{-\frac{1}{2}}$ with $\boldsymbol{D} := \mathrm{diag}\,[\boldsymbol{\mu}]$ is symmetric and has the same set of eigenvalues as $\boldsymbol{P}$ [67]. A little algebra yields

$$
\begin{aligned}
\boldsymbol{P}^{\mathrm{d}} &= \Big(1-\frac{q(k+1)}{2}\Big)\boldsymbol{I}_{|\mathcal{X}|} \\
&\quad + \frac{q}{|\mathcal{X}|}\begin{bmatrix} k\mathbf{1}_{|\mathcal{X}|/2}\mathbf{1}_{|\mathcal{X}|/2}^\top & \sqrt{k}\mathbf{1}_{|\mathcal{X}|/2}\mathbf{1}_{|\mathcal{X}|/2}^\top \\ \sqrt{k}\mathbf{1}_{|\mathcal{X}|/2}\mathbf{1}_{|\mathcal{X}|/2}^\top & \mathbf{1}_{|\mathcal{X}|/2}\mathbf{1}_{|\mathcal{X}|/2}^\top\end{bmatrix},
\end{aligned}
$$

allowing us to determine the eigenvalues $\{\lambda_i\}_{1\le i\le|\mathcal{X}|}$ as follows

$$
\lambda_1 = 1 \quad \text{and} \quad \lambda_i = 1-\frac{q(k+1)}{2} > 0 \ \ (i\ge 2).
$$

We are now ready to establish the lower bound on the cover time. First of all, the well-known connection between the spectral gap and the mixing time gives [17, Proposition 3.3]

$$t_{\mathsf{mix}} \le \frac{2\log 2 + \log\frac{1}{\mu_{\mathsf{min}}}}{2(1-\lambda_2)} = \frac{2\log 2 + \log\frac{1}{\mu_{\mathsf{min}}}}{q(k+1)}. \tag{110}$$

In addition, let $(x_0, x_1, \cdots)$ be the corresponding Markov chain, and assume that $x_0 \sim \mu$, where μ stands for the stationary distribution. Consider the last state — denoted by $|\mathcal{X}|$, which enjoys the minimum state occupancy probability μ_{min}. For any integer $t > 0$ one has

$$\begin{aligned}
&\mathbb{P}\left\{x_l \neq |\mathcal{X}|,\ \forall 0 \le l \le t\right\} \\
&\overset{\text{(i)}}{=} \mathbb{P}\left\{x_0 \neq |\mathcal{X}|\right\} \prod_{l=1}^{t} \mathbb{P}\Big\{x_l \neq |\mathcal{X}| \,\Big|\, x_0 \neq |\mathcal{X}|, \cdots, x_{l-1} \neq |\mathcal{X}|\Big\} \\
&\overset{\text{(ii)}}{\ge} \mathbb{P}\left\{x_0 \neq |\mathcal{X}|\right\} \prod_{l=1}^{t} \min_{j: j\neq|\mathcal{X}|} \mathbb{P}\big\{x_l \neq |\mathcal{X}| \,\big|\, x_{l-1} = j\big\} \\
&\overset{\text{(iii)}}{=} \Big(1 - \frac{2}{(k+1)|\mathcal{X}|}\Big)\Big(1 - \frac{q}{|\mathcal{X}|}\Big)^t \\
&\overset{\text{(iv)}}{\ge} \Big(1 - \frac{2}{(k+1)|\mathcal{X}|}\Big)\Big(1 - \frac{2qt}{|\mathcal{X}|}\Big),
\end{aligned}$$

where (i) follows from the chain rule, (ii) relies on the Markovian property, (iii) results from the construction (107), and (iv) holds as long as $\frac{q}{|\mathcal{X}|}t < \frac{1}{2}$. Consequently, if $|\mathcal{X}| \ge 3$ and if $t < \frac{|\mathcal{X}|}{8q}$, then one necessarily has

$$\begin{aligned}
&\mathbb{P}\left\{x_l \neq |\mathcal{X}|,\ \forall 0 \le l \le t\right\} \\
&\qquad \ge \Big(1 - \frac{2}{(k+1)|\mathcal{X}|}\Big)\Big(1 - \frac{2qt}{|\mathcal{X}|}\Big) > \frac{1}{2}.
\end{aligned}$$

This taken collectively with the definition of t_{cover} (cf. (9)) reveals that

$$t_{\mathsf{cover}} \ge \frac{|\mathcal{X}|}{8q} \ge \frac{t_{\mathsf{mix}}}{\big(8\log 2 + 4\log\frac{1}{\mu_{\mathsf{min}}}\big)\mu_{\mathsf{min}}},$$

where the last inequality is a direct consequence of (109) and (110). □

C. Proof of Lemma 1

Fix any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, and let us look at $\boldsymbol{\beta}_{1,t}(s,a)$, namely, the (s,a)-th entry of

$$\boldsymbol{\beta}_{1,t} = \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}^\star.$$

For convenience of presentation, we abuse the notation to let $\boldsymbol{\Lambda}_j(s,a)$ denote the (s,a)-th diagonal entry of the diagonal matrix $\boldsymbol{\Lambda}_j$, and $\boldsymbol{P}_t(s,a)$ (resp. $\boldsymbol{P}(s,a)$) the (s,a)-th row of $\boldsymbol{P}_t$ (resp. $\boldsymbol{P}$). In view of the definition (40), we can write

$$\begin{aligned}
\boldsymbol{\beta}_{1,t}(s,a) = \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \Big[&\big(1 - \boldsymbol{\Lambda}_j(s,a)\big)\boldsymbol{\Lambda}_i(s,a) \\
&\cdot \big(\boldsymbol{P}_i(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\Big].
\end{aligned} \tag{111}$$

As it turns out, it is convenient to study this expression by defining

$$\begin{aligned}
t_k(s,a) := \text{the time stamp when the trajectory} \\
\text{visits } (s,a) \text{ for the } k\text{-th time}
\end{aligned} \tag{112}$$

and

$$K_t(s,a) := \max\left\{k \mid t_k(s,a) \le t\right\}, \tag{113}$$

namely, the total number of times — during the first t iterations — that the sample trajectory visits (s,a). With these in place, the special form of $\boldsymbol{\Lambda}_j$ (cf. (34)) allows us to rewrite (111) as

$$\begin{aligned}
\boldsymbol{\beta}_{1,t}(s,a) = \gamma \sum_{k=1}^{K_t(s,a)} \Big[&(1-\eta)^{K_t(s,a)-k}\eta \\
&\cdot \big(\boldsymbol{P}_{t_k+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\Big].
\end{aligned} \tag{114}$$

where we suppress the dependency on (s,a) and write $t_k := t_k(s,a)$ to streamline notation. The main step thus boils down to controlling (114).

Towards this, we claim that: with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{k=1}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right| \\
&\qquad \le \sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\|\boldsymbol{V}^\star\|_\infty
\end{aligned} \tag{115}$$

holds simultaneously for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $1 \le K \le T$, provided that $0 < \eta\log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big) < 1$. Recognizing the trivial bound $K_t(s,a) \le t \le T$ (by construction (113)) and substituting the claimed bound (115) into the expression (114), we arrive at for all $(s,a) \in \mathcal{S}\times\mathcal{A}$

$$\begin{aligned}
|\boldsymbol{\beta}_{1,t}(s,a)| &\le \gamma\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\|\boldsymbol{V}^\star\|_\infty \\
&\le \sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\|\boldsymbol{V}^\star\|_\infty,
\end{aligned} \tag{116}$$

$$\begin{aligned}
&\sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\left\{\exists x \in \mathcal{X} : \left|\sum_{i=1}^{t}\mathbb{1}\{X_i = x\} - t\mu(x)\right| \ge \frac{1}{2}t\mu(x)\right\} \\
&\le \sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\left\{\exists x \in \mathcal{X} : \left|\sum_{i=t_{\mathsf{mix}}(\delta)}^{t}\mathbb{1}\{X_i = x\} - (t - t_{\mathsf{mix}}(\delta))\mu(x)\right| \ge \frac{1}{2}\big(t - t_{\mathsf{mix}}(\delta)\big)\mu(x) - t_{\mathsf{mix}}(\delta)\right\} \\
&\le \sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\left\{\exists x \in \mathcal{X} : \left|\sum_{i=t_{\mathsf{mix}}(\delta)}^{t}\mathbb{1}\{X_i = x\} - (t - t_{\mathsf{mix}}(\delta))\mu(x)\right| \ge \frac{10}{21}\big(t - t_{\mathsf{mix}}(\delta)\big)\mu(x)\right\} \le 2\delta.
\end{aligned} \tag{105}$$

thus concluding the proof of this lemma. It remains to validate the inequality (115).

Proof of the Inequality (115): We first make the observation that: for any *fixed* integer $K > 0$, the following vectors

$$\{\boldsymbol{P}_{t_k+1}(s,a) \mid 1 \le k \le K\}$$

are identically and independently distributed.[4] To justify this observation, let us denote by $\mathbb{P}_{s,a}(\cdot)$ the transition probability from state s when action a is taken. For any $i_1, \cdots, i_K \in \mathcal{S}$, one obtains

$$\begin{aligned}
&\mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K)\right\} \\
&= \mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K-1) \text{ and } s_{t_K+1} = i_K\right\} \\
&= \sum_{m>0} \mathbb{P}\big\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K-1) \text{ and } t_K = m \\
&\qquad \text{and } s_{m+1} = i_K\big\} \\
&\overset{\text{(i)}}{=} \sum_{m>0} \Big[\mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K-1) \text{ and } t_K = m\right\} \\
&\qquad \cdot \mathbb{P}\left\{s_{m+1} = i_K \mid s_m = s, a_m = a\right\}\Big] \\
&= \mathbb{P}_{s,a}(i_K) \\
&\qquad \cdot \sum_{m>0} \mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K-1) \text{ and } t_K = m\right\} \\
&= \mathbb{P}_{s,a}(i_K)\mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K-1)\right\},
\end{aligned}$$

where (i) holds true from the Markov property as well as the fact that t_K is an iteration in which the trajectory visits state s and takes action a. Invoking the above identity recursively, we arrive at

$$\mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1 \le k \le K)\right\} = \prod_{j=1}^{K} \mathbb{P}_{s,a}(i_j), \tag{117}$$

meaning that the state transitions happening at times $\{t_1, \cdots, t_K\}$ are independent, each following the distribution $\mathbb{P}_{s,a}(\cdot)$. This clearly demonstrates the independence of $\{\boldsymbol{P}_{t_k+1}(s,a) \mid 1 \le k \le K\}$.

With the above observation in mind, we resort to the Hoeffding inequality to bound the quantity of interest (which has zero mean). To begin with, notice the facts that for all $k \ge 1$,

$$0 \le \boldsymbol{P}_{t_k+1}(s,a)\boldsymbol{V}^\star \le \|\boldsymbol{V}^\star\|_\infty, \text{ and } 0 \le \boldsymbol{P}(s,a)\boldsymbol{V}^\star \le \|\boldsymbol{V}^\star\|_\infty, \tag{118}$$

which gives

$$\begin{aligned}
&\left|(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right| \\
&\qquad \le (1-\eta)^{K-k}\eta\|\boldsymbol{V}^\star\|_\infty.
\end{aligned}$$

As a consequence, invoking the Hoeffding inequality [68] implies that

$$\begin{aligned}
&\left|\sum_{k=1}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star\right| \\
&\le \sqrt{\frac{1}{2}\sum_{k=1}^{K}\Big((1-\eta)^{K-k}\eta\|\boldsymbol{V}^\star\|_\infty\Big)^2 \log\Big(\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} \\
&\le \sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\,\|\boldsymbol{V}^\star\|_\infty
\end{aligned} \tag{119}$$

with probability exceeding $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$, where the last line holds since

$$\sum_{k=1}^{K}\Big((1-\eta)^{K-k}\eta\Big)^2 \le \eta^2\sum_{j=0}^{\infty}(1-\eta)^j = \frac{\eta^2}{1-(1-\eta)} = \eta.$$

Taking the union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $1 \le K \le T$ then reveals that: with probability at least $1-\delta$, the inequality (119) holds simultaneously over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $1 \le K \le T$. This concludes the proof. □

D. Proof of Lemma 2 and Lemma 6

Proof of Lemma 2: Let $\boldsymbol{\beta}_{3,t} = \prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0$. Denote by $\boldsymbol{\beta}_{3,t}(s,a)$ (resp. $\boldsymbol{\Delta}_0(s,a)$) the (s,a)-th entry of $\boldsymbol{\beta}_{3,t}$ (resp. $\boldsymbol{\Delta}_0$). From the definition of $\boldsymbol{\beta}_{3,t}$, it is easily seen that

$$\big|\boldsymbol{\beta}_{3,t}(s,a)\big| = (1-\eta)^{K_t(s,a)}\big|\boldsymbol{\Delta}_0(s,a)\big|, \tag{120}$$

where $K_t(s,a)$ denotes the number of times the sample trajectory visits (s,a) during the iterations $[1,t]$ (cf. (113)). By virtue of Lemma 8 and the union bound, one has, with probability at least $1-\delta$, that

$$K_t(s,a) \ge t\mu_{\mathsf{min}}/2 \tag{121}$$

simultaneously over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all t obeying $\frac{443\tau_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{4|\mathcal{S}||\mathcal{A}|T}{\delta} \le t \le T$. Substitution into the relation (120) establishes that, with probability greater than $1-\delta$,

$$|\boldsymbol{\beta}_3(s,a)| \le (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\big|\boldsymbol{\Delta}_0(s,a)\big|. \tag{122}$$

holds uniformly over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all t obeying $\frac{443\tau_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{4|\mathcal{S}||\mathcal{A}|T}{\delta} \le t \le T$, as claimed.

Proof of Lemma 6: The proof of this lemma is essentially the same as that of Lemma 2, except that we use instead the following lower bound on $K_t(s,a)$ (which is an immediate consequence of Lemma 5)

$$K_t(s,a) \ge \Big\lfloor\frac{t}{t_{\mathsf{cover,all}}}\Big\rfloor \ge \frac{t}{2t_{\mathsf{cover,all}}} \tag{123}$$

for all $t > t_{\mathsf{cover,all}}$. Therefore, replacing $t\mu_{\mathsf{min}}$ with $t/t_{\mathsf{cover,all}}$ in the above analysis, we establish Lemma 6.

E. Proof of Lemma 3

We prove this fact via an inductive argument. The base case with $t = 0$ is a consequence of the crude bound (49). Now, assume that the claim holds for all iterations up to $t-1$, and we would like to justify it for the t-th iteration as well. Towards this, define

$$h(t) := \begin{cases} \|\boldsymbol{\Delta}_0\|_\infty, & \text{if } t \le t_{\mathsf{th}}, \\ (1-\gamma)\varepsilon, & \text{if } t > t_{\mathsf{th}}. \end{cases} \tag{124}$$

[4]The Markov chain w.r.t. the sample trajectory should be viewed as being infinitely long, although we only get to observe its first T samples. The random variables $\{t_k\}$ are, in truth, independent of the choice of T.

Recall that $(1-\eta)^{\frac{1}{2}t\mu_{\min}} \le (1-\gamma)\varepsilon$ for any $t \ge t_{\mathsf{th}}$. Therefore, combining the inequality (47) with the induction hypotheses indicates that

$$\begin{aligned}|\boldsymbol{\Delta}_t| &\le \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}\cdot\left(\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}+u_{i-1}+\varepsilon\right)\\ &\quad+\tau_1\|\boldsymbol{V}^\star\|_\infty\mathbf{1}+h(t)\mathbf{1}\\ &=\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}+\tau_1\|\boldsymbol{V}^\star\|_\infty\mathbf{1}+h(t)\mathbf{1}\\ &\quad+\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}\left(\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}+\varepsilon\right).\end{aligned}$$

Taking this together with the inequality (51b) and rearranging terms, we obtain

$$\begin{aligned}|\boldsymbol{\Delta}_t| &\le \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}+\frac{\gamma\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}\mathbf{1}\\ &\quad+\gamma\varepsilon\mathbf{1}+\tau_1\|\boldsymbol{V}^\star\|_\infty\mathbf{1}+h(t)\mathbf{1}\\ &=\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}\mathbf{1}+\gamma\varepsilon\mathbf{1}+\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}\\ &\quad+h(t)\mathbf{1}\\ &=\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}\mathbf{1}+\gamma\varepsilon\mathbf{1}+\boldsymbol{v}_t+(1-\gamma)\varepsilon\,\mathbb{1}\{t>t_{\mathsf{th}}\}\mathbf{1}\\ &\le\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}\mathbf{1}+\varepsilon\mathbf{1}+\boldsymbol{v}_t,\end{aligned}\tag{125}$$

where we have used the definition of $\boldsymbol{v}_t$ in (52). This taken collectively with the definition $u_t=\|\boldsymbol{v}_t\|_\infty$ establishes that

$$\|\boldsymbol{\Delta}_t\|_\infty\le\frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma}+\varepsilon+u_t$$

as claimed. This concludes the proof.

F. Proof of Lemma 4

We shall prove this result by induction over the index k. To start with, consider the base case where $k=0$ and $t<t_{\mathsf{th}}+t_{\mathsf{frame}}$. By definition, it is straightforward to see that $u_0\le\|\boldsymbol{\Delta}_0\|_\infty/(1-\gamma)=w_0$. In fact, repeating our argument for the crude bound (see Section VI-B.2) immediately reveals that

$$\forall t\ge 0:\qquad u_t\le\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma}=w_0,\tag{126}$$

thus indicating that the inequality (55) holds for the base case. In what follows, we assume that the inequality (55) holds up to $k-1$, and would like to extend it to the case with all t obeying $\left\lfloor\frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\right\rfloor=k$.

Consider any $0\le j<t_{\mathsf{frame}}$. In view of the definition of $\boldsymbol{v}_t$ (cf. (52)) as well as our induction hypotheses, one can arrange terms to derive

$$\begin{aligned}\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}&=\gamma\sum_{i=1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}\ \prod_{n=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(\boldsymbol{I}-\boldsymbol{\Lambda}_n)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}\\ &=\gamma\sum_{s=0}^{k-1}\left\{\sum_{i:\max\left\{\lfloor\frac{i-j-1-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\rfloor,0\right\}=s}\ \prod_{n=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(\boldsymbol{I}-\boldsymbol{\Lambda}_n)\boldsymbol{\Lambda}_i\mathbf{1}u_{i-1}\right\}\\ &\le\gamma\sum_{s=0}^{k-1}\left\{\sum_{i:\max\left\{\lfloor\frac{i-j-1-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\rfloor,0\right\}=s}\ \prod_{n=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(\boldsymbol{I}-\boldsymbol{\Lambda}_n)\boldsymbol{\Lambda}_i\mathbf{1}\right\}w_s,\end{aligned}\tag{127}$$

where the last inequality follows from our induction hypotheses, the non-negativity of $(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}$, and the fact that w_s is non-increasing.

Given any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, let us look at the (s,a)-th entry of $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}$ — denoted by $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)$, towards which it is convenient to pause and introduce some notation. Recall that $N_i^n(s,a)$ has been used to denote the number of visits to the state-action pair (s,a) between iteration i and iteration n (including i and n). To help study the behavior in each timeframe, we introduce the following quantities

$$L_h^{k-1}:=N_i^n(s,a)\tag{128}$$

with $i=t_{\mathsf{th}}+ht_{\mathsf{frame}}+j+1,\ n=t_{\mathsf{th}}+kt_{\mathsf{frame}}+j$ for every $h\le k-1$. Lemma 8 tells us that, with probability at least $1-2\delta$,

$$L_h^{k-1}\ge(k-h)\mu_{\mathsf{frame}}\qquad\text{with }\mu_{\mathsf{frame}}=\frac{1}{2}\mu_{\min}t_{\mathsf{frame}},\tag{129}$$

which holds uniformly over all state-action pairs (s,a). Armed with this set of notation, it is straightforward to use the expression (127) to verify that

$$\begin{aligned}\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)&\le\gamma\sum_{h=0}^{k-1}\eta\Big\{(1-\eta)^{L_h^{k-1}-1}+(1-\eta)^{L_h^{k-1}-2}\\ &\qquad+\cdots+(1-\eta)^{L_{h+1}^{k-1}}\Big\}w_h\\ &=\gamma\sum_{h=0}^{k-1}\left((1-\eta)^{L_{h+1}^{k-1}}-(1-\eta)^{L_h^{k-1}}\right)w_h\\ &=:\gamma\sum_{h=0}^{k-1}(\alpha_{h+1}-\alpha_h)\,w_h,\end{aligned}\tag{130}$$

where we denote $\alpha_h:=(1-\eta)^{L_h^{k-1}}$ for any $h\le k-1$ and $\alpha_k:=1$.

A little algebra further leads to

$$\begin{aligned}\gamma\sum_{h=0}^{k-1}(\alpha_{h+1}-\alpha_h)\,w_h&=\gamma(\alpha_k w_{k-1}-\alpha_0 w_0)\\ &\quad+\gamma\sum_{h=1}^{k-1}\alpha_h\,(w_{h-1}-w_h).\end{aligned}\tag{131}$$

Thus, in order to control the quantity $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)$, it suffices to control the right-hand side of (131), for which we start by bounding the last term. Plugging in the definitions of w_h and α_h yields

$$\begin{aligned}\frac{1-\gamma}{\|\boldsymbol{\Delta}_0\|_\infty}\sum_{h=1}^{k-1}\alpha_h\,(w_{h-1}-w_h)&=\sum_{h=1}^{k-1}(1-\eta)^{L_h^{k-1}}(1-\rho)^{h-1}\rho\\ &\le\rho\sum_{h=1}^{k-1}(1-\eta)^{(k-h)\mu_{\mathsf{frame}}}(1-\rho)^{h-1},\end{aligned}$$

where the last inequality results from the fact (129). Additionally, direct calculation yields

$$\begin{aligned}
&\rho\sum_{h=1}^{k-1}(1-\eta)^{(k-h)\mu_{\mathsf{frame}}}(1-\rho)^{h-1}\\
&\quad=\rho(1-\eta)^{(k-1)\mu_{\mathsf{frame}}}\sum_{h=1}^{k-1}\Big(\frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}}\Big)^{h-1}\\
&\quad=\rho(1-\eta)^{(k-1)\mu_{\mathsf{frame}}}\frac{1-\big(\frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}}\big)^{k-1}}{1-\frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}}}\\
&\quad=\rho(1-\eta)^{\mu_{\mathsf{frame}}}\frac{(1-\rho)^{k-1}-(1-\eta)^{(k-1)\mu_{\mathsf{frame}}}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}}\\
&\quad\le\rho(1-\eta)^{\mu_{\mathsf{frame}}}\frac{(1-\rho)^{k-1}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}},
\end{aligned}\tag{132}$$

where the last inequality makes use of the fact that

$$\begin{aligned}
&(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}\\
&\quad=1-(1-\gamma)(1-(1-\eta)^{\mu_{\mathsf{frame}}})-(1-\eta)^{\mu_{\mathsf{frame}}}\\
&\quad=\gamma\left\{1-(1-\eta)^{\mu_{\mathsf{frame}}}\right\}=\frac{\gamma}{1-\gamma}\rho\ge 0.
\end{aligned}\tag{133}$$

Combining the inequalities (130), (131) and (132) and using the fact $\alpha_0 w_0\ge 0$ give

$$\begin{aligned}
&\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)\le\gamma\sum_{h=1}^{k-1}\alpha_h\left(w_{h-1}-w_h\right)+\gamma\alpha_k w_{k-1}\\
&\le\frac{\big\|\boldsymbol{\Delta}_0\big\|_\infty}{1-\gamma}\Bigg\{\gamma\rho(1-\eta)^{\mu_{\mathsf{frame}}}\frac{(1-\rho)^{k-1}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}}\\
&\qquad+\gamma(1-\rho)^{k-1}\Bigg\}.
\end{aligned}\tag{134}$$

We are now ready to justify that $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)\le w_k$. Note that the observation (133) implies

$$\begin{aligned}
\gamma\frac{\rho(1-\eta)^{\mu_{\mathsf{frame}}}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}}&=\gamma\frac{\rho(1-\eta)^{\mu_{\mathsf{frame}}}}{\frac{\gamma}{1-\gamma}\rho}\\
&=(1-\gamma)(1-\eta)^{\mu_{\mathsf{frame}}}.
\end{aligned}$$

This combined with the bound (134) yields

$$\begin{aligned}
&\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}(s,a)\\
&\le\frac{\big\|\boldsymbol{\Delta}_0\big\|_\infty}{1-\gamma}\big\{(1-\gamma)(1-\eta)^{\mu_{\mathsf{frame}}}(1-\rho)^{k-1}+\gamma(1-\rho)^{k-1}\big\}\\
&\le\frac{\big\|\boldsymbol{\Delta}_0\big\|_\infty}{1-\gamma}\big(\gamma+(1-\gamma)(1-\eta)^{\mu_{\mathsf{frame}}}\big)(1-\rho)^{k-1}\\
&=(1-\rho)^k\frac{\big\|\boldsymbol{\Delta}_0\big\|_\infty}{1-\gamma}=w_k,
\end{aligned}\tag{135}$$

where the last line follows from the definition of ρ (cf. (37d)). Since the above inequality holds for all state-action pair (s,a), we conclude that

$$u_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}=\big\|\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}+j}\big\|_\infty\le w_k.\tag{136}$$

As a consequence, we have established the inequality (55) for all t obeying $\big\lfloor\frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\big\rfloor=k$, which together with the induction argument completes the proof of this lemma.

G. Proof of Lemma 7

Recalling that $\mathbf{0}\le\sum_{i=1}^t\prod_{j=i+1}^t(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1}\le\mathbf{1}$ (cf. (51b)), we obtain

$$\begin{aligned}
\|\boldsymbol{h}_{0,t}\|_\infty&\le\gamma\Big\|\sum_{i=1}^t\prod_{j=i+1}^t\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\Big\|_1\big\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty\\
&\le\gamma\big\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty.
\end{aligned}\tag{137}$$

As a result, it remains to upper bound $\big\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty$.

Suppose that $\widetilde{\boldsymbol{P}}$ is constructed using N consecutive sample transitions. Without loss of generality, assume that these N sample transitions are the transitions between the following $N+1$ samples

$$(s_0,a_0),(s_1,a_1),(s_2,a_2),\cdots,(s_N,a_N).$$

Then the (s,a)-th row of $\widetilde{\boldsymbol{P}}$ — denoted by $\widetilde{\boldsymbol{P}}(s,a)$ — is given by

$$\begin{aligned}
\widetilde{\boldsymbol{P}}(s,a)&=\frac{1}{K_N(s,a)}\sum_{i=0}^{N-1}\boldsymbol{P}_{i+1}(s,a)\overline{\boldsymbol{V}}\,\mathbb{1}\{(s_i,a_i)=(s,a)\}\\
&=\frac{1}{K_N(s,a)}\sum_{i=1}^{K_N(s,a)}\boldsymbol{P}_{t_i+1}(s,a)\overline{\boldsymbol{V}},
\end{aligned}\tag{138}$$

where $\boldsymbol{P}_i$ is defined in (35), and $\boldsymbol{P}_i(s,a)$ denotes its (s,a)-th row. Here, $K_N(s,a)$ denotes the total number of visits to (s,a) during the first N time instances (cf. (113)), and $t_k:=t_k(s,a)$ denotes the time stamp when the trajectory visits (s,a) for the k-th time (cf. (112)).

In view of our derivation for (117), the state transitions happening at time $t_1,t_2,\cdots,t_k$ are independent for any given integer $k>0$. This together with the Hoeffding inequality implies that

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{k}\left|\sum_{i=1}^k\big(\boldsymbol{P}_{t_i+1}(s,a)-\boldsymbol{P}(s,a)\big)\overline{\boldsymbol{V}}\right|\ge\tau\right\}\\
&\qquad\le 2\exp\left\{-\frac{k\tau^2}{2\|\overline{\boldsymbol{V}}\|_\infty^2}\right\}.
\end{aligned}\tag{139}$$

Consequently, with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|}$ one has

$$\begin{aligned}
&\left|\frac{1}{k}\sum_{i=1}^k\big(\boldsymbol{P}_{t_i+1}(s,a)-\boldsymbol{P}(s,a)\big)\overline{\boldsymbol{V}}\right|\\
&\qquad\le\sqrt{\frac{2\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{k}}\big\|\overline{\boldsymbol{V}}\big\|_\infty,\qquad 1\le k\le N.
\end{aligned}$$

Recognizing the simple bound $K_N(s,a)\le N$, the above inequality holds for each state-action pair (s,a) when k is replaced by $K_N(s,a)$. Then, applying the union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we obtain

$$\big\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty\le\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{\frac{2\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{K_N(s,a)}}\big\|\overline{\boldsymbol{V}}\big\|_\infty\tag{140}$$

with probability at least $1-\delta$.

In addition, for any $N\ge t_{\mathsf{frame}}$, Lemma 8 guarantees that with probability $1-2\delta$, each state-action pair (s,a) is visited

at least $N\mu_{\min}/2$ times, namely, $K_N(s,a) \geq \frac{1}{2}N\mu_{\min}$ for all (s,a). This combined with (141) yields

$$\begin{aligned}
&\left\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\right\|_\infty \\
&\quad\leq \sqrt{\frac{4\log\left(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\min}}}\left\|\overline{\boldsymbol{V}}\right\|_\infty \\
&\quad\leq \sqrt{\frac{4\log\left(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\min}}}\left(\left\|\overline{\boldsymbol{V}}-\boldsymbol{V}^\star\right\|_\infty+\left\|\boldsymbol{V}^\star\right\|_\infty\right) \\
&\quad\leq \sqrt{\frac{4\log\left(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\min}}}\left\|\overline{\boldsymbol{V}}-\boldsymbol{V}^\star\right\|_\infty \\
&\qquad+\frac{1}{1-\gamma}\sqrt{\frac{4\log\left(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\min}}} \qquad (141)
\end{aligned}$$

with probability at least $1-3\delta$, where the second inequality follows from the triangle inequality, and the last inequality follows from $\left\|\boldsymbol{V}^\star\right\|_\infty \leq \frac{1}{1-\gamma}$. Putting this together with (137) concludes the proof.

H. Proof of Lemma 5

For notational convenience, set $t_l := t_{\mathsf{cover}}l$, and define

$$\mathcal{H}_l := \left\{\exists(s,a)\in\mathcal{S}\times\mathcal{A} \text{ that is not visited within } \left(t_l, t_{l+1}\right]\right\}$$

for any integer $l \geq 0$. In view of the definition of t_{cover}, we see that for any given $(s',a') \in \mathcal{S}\times\mathcal{A}$,

$$\mathbb{P}\left\{\mathcal{H}_l \mid (s_{t_l}, a_{t_l}) = (s',a')\right\} \leq \frac{1}{2}. \qquad (142)$$

Consequently, for any integer $L > 0$, one can invoke the Markovian property to obtain

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_L\right\} = \mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\} \\
&\quad\cdot\mathbb{P}\left\{\mathcal{H}_L \mid \mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\} \\
&= \mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\}\sum_{s',a'}\Big[\mathbb{P}\left\{\mathcal{H}_L \mid (s_{t_l},a_{t_l})=(s',a')\right\} \\
&\quad\cdot\mathbb{P}\left\{(s_{t_l},a_{t_l})=(s',a') \mid \mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\}\Big] \\
&\leq \frac{1}{2}\mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\} \\
&\quad\cdot\sum_{s',a'}\mathbb{P}\left\{(s_{t_l},a_{t_l})=(s',a') \mid \mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\} \\
&= \frac{1}{2}\mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_{L-1}\right\},
\end{aligned}$$

where the inequality follows from (142). Repeating this derivation recursively, we deduce that

$$\mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_L\right\} \leq \frac{1}{2^L}.$$

This tells us that

$$\begin{aligned}
&\mathbb{P}\left\{\exists(s,a)\in\mathcal{S}\times\mathcal{A} \text{ that is not visited between } (0, t_{\mathsf{cover,all}}]\right\} \\
&\quad\leq \mathbb{P}\left\{\mathcal{H}_1\cap\cdots\cap\mathcal{H}_{\log_2\frac{T}{\delta}}\right\} \leq \frac{1}{2^{\log_2\frac{T}{\delta}}} = \frac{\delta}{T},
\end{aligned}$$

which in turn establishes the advertised result by applying the union bound.

ACKNOWLEDGMENT

The authors thank Shicong Cen, Chen Cheng, and Cong Ma for numerous discussions about reinforcement learning.

REFERENCES

[1] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7031–7043.

[2] G. Qu and A. Wierman, "Finite-time analysis of asynchronous stochastic approximation and Q-learning," in *Proc. Conf. Learn. Theory*, 2020, pp. 3185–3205.

[3] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[4] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[5] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," 2019, *arXiv:1904.12901*. [Online]. Available: http://arxiv.org/abs/1904.12901

[6] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learn.*, vol. 16, no. 16, pp. 185–202, Sep. 1994.

[7] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 703–710.

[8] M. J. Wainwright, "Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning," 2019, *arXiv:1905.06265*. [Online]. Available: https://arxiv.org/abs/1905.06265

[9] E. Even-Dar and Y. Mansour, "Learning rates for Q-learning," *J. Mach. Learn. Res.*, vol. 5, pp. 1–25, Dec. 2004.

[10] C. L. Beck and R. Srikant, "Error bounds for constant step-size Q-learning," *Syst. Control Lett.*, vol. 61, no. 12, pp. 1203–1208, 2012.

[11] M. G. Azar, R. Munos, M. Ghavamzadeh, and H. Kappen, "Reinforcement learning with a near optimal rate of convergence," INRIA, Tech. Rep. INRIA-00636615v2, 2011.

[12] M. J. Wainwright, "Variance-reduced Q-learning is minimax optimal," 2019, *arXiv:1906.04697*. [Online]. Available: https://arxiv.org/abs/1906.04697

[13] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.

[14] C. Szepesvári, "The asymptotic convergence-rate of Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 1998, pp. 1064–1070.

[15] A. B. Tsybakov and V. Zaiats, *Introduction to Nonparametric Estimation*, vol. 11. Springer, 2009.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. Athena Scientific, 2017.

[17] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electron. J. Probab.*, vol. 20, pp. 1–32, Jan. 2015.

[18] M. Gheshlaghi Azar, R. Munos, and H. J. Kappen, "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model," *Mach. Learn.*, vol. 91, no. 3, pp. 325–349, Jun. 2013.

[19] Z. Chen, S. Theja Maguluri, S. Shakkottai, and K. Shanmugam, "Finite-sample analysis of stochastic approximation using smooth convex envelopes," 2020, *arXiv:2002.00874*. [Online]. Available: http://arxiv.org/abs/2002.00874

[20] A. Gosavi, "Boundedness of iterates in Q-learning," *Syst. Control Lett.*, vol. 55, no. 4, pp. 347–349, 2006.

[21] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Proc. Conf. Learn. Theory*, 2018, pp. 1691–1692.

[22] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and TD learning," in *Proc. Conf. Learn. Theory*, 2019, pp. 2803–2830.

[23] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan, "Is temporal difference learning optimal? An instance-dependent analysis," 2020, *arXiv:2003.07337*. [Online]. Available: http://arxiv.org/abs/2003.07337

[24] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, "Near-optimal time and sample complexities for solving Markov decision processes with a generative model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5186–5196.

[25] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[26] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou, "Stochastic variance reduction methods for policy evaluation," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1049–1058.
[27] A. Sidford, M. Wang, X. Wu, and Y. Ye, "Variance reduced value iteration and faster algorithms for solving Markov decision processes," in *Proc. 29th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2018, pp. 770–787.
[28] T. Xu, Z. Wang, Y. Zhou, and Y. Liang, "Reanalysis of variance reduced temporal difference learning," 2020, *arXiv:2001.01898*. [Online]. Available: http://arxiv.org/abs/2001.01898
[29] C. J. C. H. Watkins, "Learning from delayed rewards," King's College, Cambridge, U.K., Tech. Rep., 1989.
[30] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000.
[31] M. J. Kearns and S. P. Singh, "Finite-sample convergence rates for Q-learning and indirect algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 996–1002.
[32] G. Li, C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, "Tightening the dependence on horizon in the sample complexity of Q-learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6296–6306.
[33] H. Xiong, L. Zhao, Y. Liang, and W. Zhang, "Finite-time analysis for double Q-learning," 2020, *arXiv:2009.14257*. [Online]. Available: http://arxiv.org/abs/2009.14257
[34] G. Li, C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, "Is Q-learning minimax optimal? A tight sample complexity analysis," 2021, *arXiv:2102.06548*. [Online]. Available: http://arxiv.org/abs/2102.06548
[35] Z. Chen, S. Zhang, T. T. Doan, J.-P. Clarke, and S. Theja Maguluri, "Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning," 2019, *arXiv:1905.11425*. [Online]. Available: http://arxiv.org/abs/1905.11425
[36] P. Xu and Q. Gu, "A finite-time analysis of Q-learning with neural network function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10555–10565.
[37] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2019, *arXiv:1901.00137*. [Online]. Available: http://arxiv.org/abs/1901.00137
[38] S. S. Du, Y. Luo, R. Wang, and H. Zhang, "Provably efficient Q-learning with function approximation via distribution shift error checking Oracle," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8058–8068.
[39] S. S. Du, J. D. Lee, G. Mahajan, and R. Wang, "Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity," 2020, *arXiv:2002.07125*. [Online]. Available: http://arxiv.org/abs/2002.07125
[40] Q. Cai, Z. Yang, J. D. Lee, and Z. Wang, "Neural temporal-difference and Q-learning converges to global optima," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11312–11322.
[41] L. Yang and M. Wang, "Sample-optimal parametric Q-learning using linearly additive features," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6995–7004.
[42] B. Weng, H. Xiong, L. Zhao, Y. Liang, and W. Zhang, "Momentum Q-learning with finite-sample convergence guarantee," 2020, *arXiv:2007.15418*. [Online]. Available: http://arxiv.org/abs/2007.15418
[43] W. Weng, H. Gupta, N. He, L. Ying, and R. Srikant, "Provably-efficient double Q-learning," 2020, *arXiv:2007.05034*. [Online]. Available: http://arxiv.org/abs/2007.05034
[44] D. Shah and Q. Xie, "Q-learning with nearest neighbors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3111–3121.
[45] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman, "PAC model-free reinforcement learning," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 881–888.
[46] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos, "Speedy Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2411–2419.
[47] A. M. Devraj and S. P. Meyn, "Q-learning with uniformly bounded variance: Large discounting is not a barrier to fast learning," 2020, *arXiv:2002.10301*. [Online]. Available: http://arxiv.org/abs/2002.10301
[48] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4863–4873.
[49] Y. Wang, K. Dong, X. Chen, and L. Wang, "Q-learning with UCB exploration is sample efficient for infinite-horizon MDP," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[50] Y. Bai, T. Xie, N. Jiang, and Y.-X. Wang, "Provably efficient Q-learning with low switching cost," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8002–8011.
[51] G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi, "Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–60.
[52] T. Xu, S. Zou, and Y. Liang, "Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10633–10643.
[53] H. Gupta, R. Srikant, and L. Ying, "Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4706–4715.
[54] T. Doan, S. Maguluri, and J. Romberg, "Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1626–1635.
[55] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with Markovian noise," 2020, *arXiv:2002.01268*. [Online]. Available: http://arxiv.org/abs/2002.01268
[56] G. Dalal, G. Thoppe, B. Szörényi, and S. Mannor, "Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning," in *Proc. Conf. Learn. Theory*, 2018, pp. 1199–1233.
[57] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite sample analyses for TD(0) with function approximation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6144–6160.
[58] D. Lee and N. He, "Target-based temporal difference learning," 2019, *arXiv:1904.10945*. [Online]. Available: http://arxiv.org/abs/1904.10945
[59] Y. Lin, G. Qu, L. Huang, and A. Wierman, "Multi-agent reinforcement learning in time-varying networked systems," 2020, *arXiv:2006.06555*. [Online]. Available: http://arxiv.org/abs/2006.06555
[60] W. Mou, C. Junchi Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, "On linear stochastic approximation: Fine-grained Polyak-ruppert and non-asymptotic concentration," 2020, *arXiv:2004.04719*. [Online]. Available: http://arxiv.org/abs/2004.04719
[61] S. Zou, T. Xu, and Y. Liang, "Finite-sample analysis for SARSA with linear function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8665–8675.
[62] A. Agarwal, S. Kakade, and L. F. Yang, "Model-based reinforcement learning with a generative model is minimax optimal," 2019, *arXiv:1906.03804*. [Online]. Available: http://arxiv.org/abs/1906.03804
[63] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Breaking the sample size barrier in model-based reinforcement learning with a generative model," 2020, *arXiv:2005.12900*. [Online]. Available: http://arxiv.org/abs/2005.12900
[64] C. Dann and E. Brunskill, "Sample complexity of episodic fixed-horizon reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2818–2826.
[65] T. Sun, Y. Sun, Y. Xu, and W. Yin, "Markov chain block coordinate descent," *Comput. Optim. Appl.*, vol. 75, pp. 35–61, Oct. 2020.
[66] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Convergence rates of accelerated Markov gradient descent with applications in reinforcement learning," 2020, *arXiv:2002.02873*. [Online]. Available: http://arxiv.org/abs/2002.02873
[67] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, vol. 31. Springer, 2013.
[68] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

Gen Li (Student Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University in 2016 and 2021, respectively, under the advice of Yuantao Gu. He is currently a Post-Doctoral Scholar advised by Yuxin Chen with the Department of Electrical and Computer Engineering, Princeton University. His recent research interests include machine learning, reinforcement learning, high-dimensional statistics, and nonconvex optimization.

Yuting Wei (Member, IEEE) received the Ph.D. degree in statistics from the University of California, Berkeley, advised by Martin Wainwright and Aditya Guntuboyina. She spent two years at Carnegie Mellon University as an Assistant Professor of statistics and one year at Stanford University as a Stein Fellow. She is currently an Assistant Professor with the Department of Statistics and Data Science, Wharton School, University of Pennsylvania. Her research interests include high-dimensional and non-parametric statistics, statistical machine learning, and reinforcement learning. She was a recipient of the 2018 Erich L. Lehmann Citation from the Berkeley Statistics Department for her Ph.D. dissertation in theoretical statistics.

Yuejie Chi (Senior Member, IEEE) received the B.E. (Hons.) degree in electrical engineering from Tsinghua University, Beijing, China, in 2007, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University in 2009 and 2012, respectively. She was with The Ohio State University from 2012 to 2017. Since 2018, she has been with the Department of Electrical and Computer Engineering, Carnegie Mellon University, where she is currently a Professor and held the Inaugural Robert E. Doherty Early Career Development Professorship from 2018 to 2020. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, and machine learning and inverse problems, with applications in sensing and societal systems, broadly defined. Among others, she was a recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE), the Inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing, and named a Goldsmith Lecturer by the IEEE Information Theory Society. She serves as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Yuantao Gu (Senior Member, IEEE) received the B.E. degree in electronic engineering from Xi'an Jiaotong University in 1998 and the Ph.D. degree (Hons.) in electronic engineering from Tsinghua University in 2003.

He joined the Faculty of Tsinghua University in 2003 where he is currently a Full Professor with the Department of Electronic Engineering. He was a Visiting Scientist at Microsoft Research Asia from 2005 to 2006, the Research Laboratory of Electronics, Massachusetts Institute of Technology, from 2012 to 2013, and the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 2015. His research interests include machine learning, decentralized optimization, graph signal processing, and compressed sensing.

Prof. Gu has been an Elected Member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee since 2019 and the IEEE Signal Processing Theory and Methods (SPTM) Technical Committee since 2017. He received the Best Paper Award of IEEE Global Conference on Signal and Information Processing (GlobalSIP) in 2015, the Award for Best Presentation of Journal Paper of IEEE International Conference on Signal and Information Processing (ChinaSIP) in 2015, and the Zhang Si-Ying (CCDC) Outstanding Youth Paper Award (with his student) in 2017. He was one of the most Outstanding Reviewers for IEEE ICASSP in 2019 and received the Outstanding Editorial Board Member Award for the IEEE TRANSACTIONS ON SIGNAL PROCESSING Editorial Board in 2021. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2015 to 2019 and a Handling Editor for Elsevier *Digital Signal Processing* from 2015 to 2017. He has been a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2019.

Yuxin Chen (Senior Member, IEEE) received the B.E. degree (Hons.) from Tsinghua University in 2008, the M.S. degree in electrical and computer engineering from The University of Texas at Austin in 2010, and the Ph.D. degree in electrical engineering and the M.S. degree in statistics from Stanford University in January 2015. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA. Prior to joining Princeton, he was a Post-Doctoral Scholar with the Department of Statistics, Stanford University. His research interests include high-dimensional data analysis, convex and nonconvex optimization, reinforcement learning, statistical learning, statistical signal processing, and information theory. He has received the 2020 Princeton Graduate Mentoring Award, the 2019 AFOSR Young Investigator Award, the 2020 ARO Young Investigator Award, the ICCM Best Paper Award (Gold Medal), and the 2021 Princeton SEAS Junior Faculty Award, and was selected as the Finalist for the Best Paper Prize for Young Researchers in Continuous Optimization 2019.