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

Gen Li* Tsinghua | Yuting Wei† CMU | Yuejie Chi‡ CMU | Yuantao Gu* Tsinghua | Yuxin Chen§ Princeton

June, 2020; Revised: September 2020

**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 $\gamma$-discounted MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, we demonstrate that the $\ell_\infty$-based sample complexity of classical asynchronous Q-learning — namely, the number of samples needed to yield an entrywise $\varepsilon$-accurate estimate of the Q-function — is at most on the order of

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

up to some logarithmic factor, provided that a proper constant learning rate is adopted. Here, $t_{\mathsf{mix}}$ and $\mu_{\mathsf{min}}$ denote respectively the mixing time and the minimum state-action occupancy probability of the sample trajectory. The first term of this bound matches the 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 Qu and Wierman (2020) by a factor of at least $|\mathcal{S}||\mathcal{A}|$. Further, the scaling on the discount complexity can be improved by means of variance reduction.

**Keywords:** model-free reinforcement learning, asynchronous Q-learning, Markovian samples, variance reduction, TD learning, mixing time

# Contents



*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
†Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
§Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA.

# 1 Introduction

Model-free algorithms such as Q-learning (Watkins and Dayan, 1992) play a central role in recent breakthroughs of reinforcement learning (RL) (Mnih et al., 2015). 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, adapt flexibly to changing environments, and are often memory-efficient. Understanding and improving the sample efficiency of model-free algorithms lie at the core of recent research activity (Dulac-Arnold et al., 2019), 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 (or 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

| Paper | Sample complexity | Learning rate |
|---|---|---|
| Even-Dar and Mansour (2003) | $\frac{(t_{\mathsf{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$ | linear: $\frac{1}{t}$ |
| Even-Dar and Mansour (2003) | $\big(\frac{t_{\mathsf{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\big)^{\frac{1}{\omega}} + \big(\frac{t_{\mathsf{cover}}}{1-\gamma}\big)^{\frac{1}{1-\omega}}$ | polynomial: $\frac{1}{t^\omega}$, $\omega \in (\frac{1}{2}, 1)$ |
| Beck and Srikant (2012) | $\frac{t_{\mathsf{cover}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$ | constant: $\frac{(1-\gamma)^4\varepsilon^2}{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\mathsf{cover}}^2}$ |
| Qu and Wierman (2020) | $\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$ | rescaled linear: $\frac{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}}{t+\max\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}, t_{\mathsf{mix}}\}}$ |
| **This work (Theorem 1)** | $\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$ | constant: $\min\big\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\big\}$ |
| **This work (Theorem 2)** | $\frac{t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}$ | constant: $\min\big\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, 1\big\}$ |

Table 1: Sample complexity of asynchronous Q-learning to compute an $\varepsilon$-optimal Q-function in the $\ell_\infty$ norm, where we hide all logarithmic factors. With regards to the Markovian trajectory induced by the behavior policy, we denote by $t_{\mathsf{cover}}$, $t_{\mathsf{mix}}$, and $\mu_{\mathsf{min}}$ the cover time, mixing time, and minimum state-action occupancy probability of the associated stationary distribution, respectively.

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*. We ask a fundamental question:

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

Despite a considerable amount of prior work exploring this algorithm (ranging from the classical work Jaakkola et al. (1994); Tsitsiklis (1994) to the very recent paper Qu and Wierman (2020)), 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 Qu and Wierman (2020) and the one derived for the synchronous counterpart Wainwright (2019a). This raises a natural desire to examine whether there is any bottleneck intrinsic to the asynchronous setting that significantly limits its performance.

**Our 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 $t_{\mathsf{mix}}$, which characterizes how fast the trajectory disentangle itself from the initial state; *(ii)* the smallest state-action occupancy probability $\mu_{\mathsf{min}}$ 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 $\varepsilon$-optimal Q-function estimate — in a strong $\ell_\infty$ 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 (Wainwright, 2019a). In comparison, the second term of (1) — which is unaffected by the accuracy level $\varepsilon$ — is intrinsic to the Markovian nature

[1] Let $\mathcal{X} := \big(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \frac{1}{\varepsilon}\big)$. 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.

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 $t_{\mathsf{cover}}$ 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 $\varepsilon$-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 $\varepsilon \leq \min\big\{1, \frac{1}{(1-\gamma)\sqrt{t_{\mathsf{mix}}}}\big\}$ (Wainwright, 2019b). Moreover, by taking the action space to be a singleton set, the aforementioned results immediately lead to $\ell_\infty$-based sample complexity guarantees for temporal difference (TD) learning (Sutton, 1988) on Markovian samples.

**Comparisons with past work.** A large fraction of the classical literature focused on asymptotic convergence analysis of asynchronous Q-learning (e.g. Jaakkola et al. (1994); Szepesvári (1998); Tsitsiklis (1994)); 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 Qu and Wierman (2020), 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 $1/\mu_{\mathsf{min}} \geq |\mathcal{S}||\mathcal{A}|$, our result (1) improves upon that of Qu and Wierman (2020) 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 addition, we note that several prior work Beck and Srikant (2012); Even-Dar and Mansour (2003) developed sample complexity bounds in terms of the cover time $t_{\mathsf{cover}}$ of the sample trajectory; our result strengthens these bounds by a factor of at least $t_{\mathsf{cover}}^2|\mathcal{S}||\mathcal{A}| \geq |\mathcal{S}|^3|\mathcal{A}|^3$. The interested reader is referred to Table 1 for more precise comparisons, and to Section 5 for discussions of further related work.

**Notation.** Denote by $\Delta(\mathcal{S})$ (resp. $\Delta(\mathcal{A})$) the probability simplex over the set $\mathcal{S}$ (resp. $\mathcal{A}$). For any vector $\boldsymbol{z} = [z_i]_{1\leq i\leq n} \in \mathbb{R}^n$, we overload the notation $\sqrt{\cdot}$ and $|\cdot|$ to denote entry-wise operations, such that $\sqrt{\boldsymbol{z}} := [\sqrt{z_i}]_{1\leq i\leq n}$ and $|\boldsymbol{z}| := [|z_i|]_{1\leq i\leq n}$. For any vectors $\boldsymbol{z} = [a_i]_{1\leq i\leq n}$ and $\boldsymbol{w} = [w_i]_{1\leq i\leq n}$, the notation $\boldsymbol{z} \geq \boldsymbol{w}$ (resp. $\boldsymbol{z} \leq \boldsymbol{w}$) means $z_i \geq w_i$ (resp. $z_i \leq w_i$) for all $1 \leq i \leq n$. 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 $(\varepsilon, \delta)$, and their exact values may change from line to line.

# 2 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. 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 $(s,a) \in \mathcal{S}\times\mathcal{A}$, $P(s' \mid s,a)$ 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, is not yet well understood. See Bertsekas (2017) for an in-depth introduction of this model.

**Q-function and the 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), 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 $\pi$, 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 $\pi$ (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 $\pi$. The action-value function (also Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy $\pi$ is defined by

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a\right],$$

where the actions are taken according to the policy $\pi$ 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

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

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

**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 $\pi_{\mathsf{b}}$ — 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 $\pi_{\mathsf{b}}$, the sample trajectory can be viewed as a sample path of a time-homogeneous Markov chain over all state-action pairs. Throughout this paper, we impose the following assumption regarding uniform ergodicity (Paulin, 2015).

**Assumption 1.** *The Markov chain induced by the stationary behavior policy $\pi_{\mathsf{b}}$ 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

$$\mu_{\mathsf{min}} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mu_{\pi_{\mathsf{b}}}(s,a). \tag{4}$$

Intuitively, $\mu_{\mathsf{min}}$ reflects an information bottleneck — the smaller $\mu_{\mathsf{min}}$ 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

$$t_{\mathsf{mix}} := \min\Big\{t \;\Big|\; \max_{(s_0,a_0) \in \mathcal{S} \times \mathcal{A}} d_{\mathsf{TV}}\big(P^t(\cdot|s_0,a_0), \mu_{\pi_{\mathsf{b}}}\big) \leq \frac{1}{4}\Big\}, \tag{5}$$

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 $d_{\mathsf{TV}}(\mu,\nu)$ stands for the total variation distance between two distributions $\mu$ and $\nu$ (Paulin, 2015). In words, the mixing time $t_{\mathsf{mix}}$ captures how fast the sample trajectory decorrelates from its initial state. Moreover, we define the cover time associated with this Markov chain as follows

$$t_{\mathsf{cover}} := \min\Big\{t \mid \min_{(s_0,a_0) \in \mathcal{S} \times \mathcal{A}} \mathbb{P}\big(\mathcal{B}_t | s_0, a_0\big) \geq \frac{1}{2}\Big\}, \tag{6}$$

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 $(s_0, a_0)$.

**Goal.** Given *a single* sample trajectory $\{s_t, a_t, r_t\}_{t=0}^{\infty}$ generated by the behavior policy $\pi_{\mathsf{b}}$, we aim to compute/approximate the optimal Q-function $Q^\star$ in an $\ell_\infty$ 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 (Qu and Wierman, 2020; Tsitsiklis, 1994) 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.

# 3 Asynchronous Q-learning on a single trajectory

## 3.1 Algorithm

The Q-learning algorithm (Watkins and Dayan, 1992) 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 $\pi_{\mathsf{b}}$, 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

$$\begin{aligned} Q_t(s_{t-1}, a_{t-1}) &= (1-\eta_t) Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \mathcal{T}_t(Q_{t-1})(s_{t-1}, a_{t-1}) \\ Q_t(s,a) &= Q_{t-1}(s,a), \qquad \forall (s,a) \neq (s_{t-1}, a_{t-1}) \end{aligned} \tag{7}$$

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

$$\mathcal{T}_t(Q)(s_{t-1}, a_{t-1}) := r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'). \tag{8}$$

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

$$\forall s \in \mathcal{S}: \qquad V_t(s) := \max_{a \in \mathcal{A}} Q_t(s,a). \tag{9}$$

The complete algorithm is described in Algorithm 1.

**Algorithm 1:** Asynchronous Q-learning

1 **input parameters:** learning rates $\{\eta_t\}$, number of iterations $T$.
2 **initialization:** $Q_0 = 0$.
3 **for** $t = 1, 2, \cdots, T$ **do**
4     Draw action $a_{t-1} \sim \pi_{\mathsf{b}}(s_{t-1})$ and next state $s_t \sim P(\cdot | s_{t-1}, a_{t-1})$.
5     Update $Q_t$ according to (7).

## 3.2 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 $\mu_{\mathsf{min}}$ and $t_{\mathsf{mix}}$ defined respectively in (4) and (5) play a vital role. The proof of this result is provided in Section 6.

**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 $0 < \varepsilon \leq \frac{1}{1-\gamma}$, 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*

$$T \geq \frac{c_0}{\mu_{\mathsf{min}}} \left\{ \frac{1}{(1-\gamma)^5 \varepsilon^2} + \frac{t_{\mathsf{mix}}}{1-\gamma} \right\} \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big) \log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big), \tag{10a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)} \min\left\{ \frac{(1-\gamma)^4\varepsilon^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}} \right\}. \tag{10b}$$

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 $\ell_\infty$-based sample complexity required for Algorithm 1 to attain $\varepsilon$ accuracy is at most

$$\widetilde{O}\Big( \frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)} \Big). \tag{11}$$

A few implications are in order.

**Dependency on the minimum state-action occupancy probability $\mu_{\mathsf{min}}$.** Our sample complexity bound (11) 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 Azar et al. (2013) derived for the setting with independent samples. In comparison, Qu and Wierman (2020, Theorem 7) depends at least quadratically on $1/\mu_{\mathsf{min}}$, which is at least $|\mathcal{S}||\mathcal{A}|$ times larger than our result (11).

**Dependency on the discount complexity $\frac{1}{1-\gamma}$.** The sample size bound (11) scales as $\frac{1}{(1-\gamma)^5\varepsilon^2}$, which coincides with both Chen et al. (2020); Wainwright (2019a) (for the synchronous setting) and Beck and Srikant (2012); Qu and Wierman (2020) (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.

**Dependency on the mixing time $t_{\mathsf{mix}}$.** The second additive term of our sample complexity (11) depends linearly on the mixing time $t_{\mathsf{mix}}$ and is (almost) independent of the target accuracy $\varepsilon$. 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 $\varepsilon$ (in the sense that $t_{\mathsf{mix}} \leq \frac{1}{(1-\gamma)^4\varepsilon^2}$), then the algorithm behaves as if the samples were independently drawn from the stationary distribution of the trajectory. In comparison, the influences of $t_{\mathsf{mix}}$ and $\frac{1}{(1-\gamma)^5\varepsilon^2}$ in Qu and Wierman (2020) (cf. Table 1) are multiplicative regardless of the value of $\varepsilon$, 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 $\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \geq t_{\mathsf{mix}}|\mathcal{S}||\mathcal{A}|$ times larger than our result (modulo some log factor).

**Schedule of learning rates.** An interesting aspect of our analysis lies in the adoption of a time-invariant learning rate, under which the $\ell_\infty$ 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 $\varepsilon$ and then determining the sample complexity accordingly. In comparison, classical analyses typically adopted a (rescaled) linear or a polynomial learning rule Even-Dar and Mansour (2003); Qu and Wierman (2020). While the work Beck and Srikant (2012) 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 3.4, one can also adopt carefully designed diminishing learning rates to achieve the same performance guarantees.

In addition, our analysis framework immediately leads to another sample complexity guarantee stated in terms of the cover time $t_{\mathsf{cover}}$ (cf. (6)), which facilitates comparisons with several past work Beck and

Srikant (2012); Even-Dar and Mansour (2003). The proof follows essentially that of Theorem 1, with a sketch provided in Section B.

**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*

$$T \geq \frac{c_0 t_{\mathsf{cover}}}{(1-\gamma)^5 \varepsilon^2} \log^2\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big) \log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big), \tag{12a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)} \min\Big\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2},\ 1\Big\}. \tag{12b}$$

In a nutshell, this theorem tells us that the $\ell_\infty$-based sample complexity of classical asynchronous Q-learning is bounded above by

$$\widetilde{O}\Big(\frac{t_{\mathsf{cover}}}{(1-\gamma)^5\varepsilon^2}\Big), \tag{13}$$

which scales linearly with the cover time. This improves upon the prior result Even-Dar and Mansour (2003) (resp. Beck and Srikant (2012)) by an order of at least $t_{\mathsf{cover}}^{3.29} \geq |\mathcal{S}|^{3.29}|\mathcal{A}|^{3.29}$ (resp. $t_{\mathsf{cover}}^{2}|\mathcal{S}||\mathcal{A}| \geq |\mathcal{S}|^{3}|\mathcal{A}|^{3}$). See Table 1 for detailed comparisons. We shall further make note of some connections between $t_{\mathsf{cover}}$ and $t_{\mathsf{mix}}/\mu_{\mathsf{min}}$ to help compare Theorem 1 and Theorem 2: (1) in general, $t_{\mathsf{cover}} = \widetilde{O}(t_{\mathsf{mix}}/\mu_{\mathsf{min}})$ for uniformly ergodic chains; (2) 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 A.2 for details.

### 3.3 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]

$$\begin{aligned} V_t(s_{t-1}) &= (1-\eta_t)V_{t-1}(s_{t-1}) + \eta_t\left(r(s_{t-1}) + \gamma V_{t-1}(s_t)\right), \\ V_t(s) &= V_{t-1}(s), \qquad \forall s \neq s_{t-1}. \end{aligned} \tag{14}$$

As usual, $\eta_t$ 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 $\ell_\infty$ 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 (14). 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*

$$T \geq \frac{c_0}{\mu_{\mathsf{min}}}\Big\{\frac{1}{(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{1-\gamma}\Big\} \log\Big(\frac{|\mathcal{S}|T}{\delta}\Big) \log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big), \tag{15a}$$

$$\eta = \frac{c_1}{\log\big(\frac{|\mathcal{S}|T}{\delta}\big)} \min\Big\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}}\Big\}. \tag{15b}$$

[2]When $\mathcal{A} = \{a\}$ is a singleton, the Q-learning update rule (7) reduces to the TD update rule (14) by relating $Q(s,a) = V(s)$.

The above result reveals that the $\ell_\infty$-sample complexity for TD learning is at most

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}+\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big), \tag{16}$$

provided that an appropriate constant learning rate is adopted. We note that prior finite-sample analysis on asynchronous TD learning typically focused on (weighted) $\ell_2$ estimation errors with linear function approximation (Bhandari et al., 2018; Srikant and Ying, 2019), and it is hence difficult to make fair comparisons. The recent papers Khamaru et al. (2020); Mou et al. (2020) develop $\ell_\infty$ guarantees for TD learning, with their focus on the synchronous settings with i.i.d. samples rather than Markovian samples.

### 3.4 Adaptive and implementable learning rates

The careful reader might already notice that the learning rates recommended in (10b) depend on the mixing time $t_{\mathsf{mix}}$ — 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 $t_{\mathsf{mix}}$ and which is still capable of achieving the performance advertised in Theorem 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 $\mu_{\mathsf{min}}$, computed recursively as follows

$$\widehat{\mu}_{\mathsf{min},t}=\begin{cases}\frac{1}{|\mathcal{S}||\mathcal{A}|}, & \min_{s,a}K_t(s,a)=0;\\ \widehat{\mu}_{\mathsf{min},t-1}, & \frac{1}{2}<\frac{\min_{s,a}K_t(s,a)/t}{\widehat{\mu}_{\mathsf{min},t-1}}<2;\\ \min_{s,a}K_t(s,a)/t, & \text{otherwise.}\end{cases} \tag{17}$$

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

$$\eta_t=\min\Big\{1,c_\eta\exp\Big(\Big\lfloor\log\frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}\Big\rfloor\Big)\Big\}, \tag{18}$$

where $c_\eta>0$ is some sufficiently large constant, 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 $\mu_{\mathsf{min}}$, then one can view (18) as a sort of "piecewise constant approximation" of the rescaled linear stepsizes $\frac{c_\eta\log t}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2 t}$. Clearly, such learning rates are fully data-driven and do no rely on any prior knowledge about the Markov chain (like $t_{\mathsf{mix}}$ and $\mu_{\mathsf{min}}$) or the target accuracy $\varepsilon$.

**Performance guarantees.** Encouragingly, our theoretical framework can be extended without difficulty to accommodate this adaptive learning rate choice. Specifically, for the Q-function estimates

$$\widehat{Q}_t=\begin{cases}Q_t, & \text{if } \eta_{t+1}\neq\eta_t,\\ \widehat{Q}_{t-1}, & \text{otherwise,}\end{cases} \tag{19}$$

where $Q_t$ is provided by the Q-learning steps (c.f. (7)). We then are ensured of the following theoretical guarantees whose proof is deferred to Appendix C.

**Theorem 3.** *Consider asynchronous Q-learning with learning rates (18). There exists some universal constant $C>0$ such that: for any $0<\delta<1$ and $0<\varepsilon\le\frac{1}{1-\gamma}$, one has*

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\qquad\big|\widehat{Q}_T(s,a)-Q^\star(s,a)\big|\le\varepsilon \tag{20}$$

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

$$T\ge C\max\Big\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2},\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big\}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)\log\Big(\frac{T}{(1-\gamma)^2\varepsilon}\Big). \tag{21}$$

## 4 Extension: asynchronous variance-reduced Q-learning

As pointed out in prior literature, the classical form of Q-learning (7) often suffers from sub-optimal dependence on the discount complexity $\frac{1}{1-\gamma}$. For instance, in the synchronous setting, the minimax lower bound is proportional to $\frac{1}{(1-\gamma)^3}$ (see, Azar et al. (2013)), while the sharpest known upper bound for vanilla Q-learning scales as $\frac{1}{(1-\gamma)^5}$; see detailed discussions in Wainwright (2019a). To remedy this issue, recent work proposed to leverage the idea of variance reduction to develop accelerated RL algorithms in the synchronous setting (Sidford et al., 2018a; Wainwright, 2019b), as inspired by the seminal SVRG algorithm (Johnson and Zhang, 2013) that originates from the stochastic optimization literature. In this section, we adapt this idea to asynchronous Q-learning and characterize its sample efficiency.

### 4.1 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 (7) 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

$$\begin{aligned} Q_t(s_{t-1}, a_{t-1}) &= (1-\eta_t) Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \Big( \mathcal{T}_t(Q_{t-1}) - \mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q}) \Big)(s_{t-1}, a_{t-1}), \\ Q_t(s,a) &= Q_{t-1}(s,a), \qquad \forall (s,a) \neq (s_{t-1}, a_{t-1}), \end{aligned} \tag{22}$$

where $\mathcal{T}_t$ denotes the empirical Bellman operator at time $t$ (cf. (8)). The empirical estimate $\widetilde{\mathcal{T}}(\overline{Q})$ can be computed using a set of samples; more specifically, by drawing $N$ consecutive sample transitions $\{(s_i, a_i, s_{i+1})\}_{0 \le i < N}$ from the observed trajectory, we compute

$$\widetilde{\mathcal{T}}(\overline{Q})(s,a) = r(s,a) + \frac{\gamma \sum_{i=0}^{N-1} \mathbb{1}\{(s_i, a_i) = (s,a)\} \max_{a'} \overline{Q}(s_{i+1}, a')}{\sum_{i=0}^{N-1} \mathbb{1}\{(s_i, a_i) = (s,a)\}}. \tag{23}$$

Compared with the classical form (7), 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.

For convenience of presentation, we introduce the following notation

$$Q = \textsc{Vr-q-run-epoch}(\, \overline{Q}, N, t_{\mathsf{epoch}} \,) \tag{24}$$

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 (23), with the remaining samples employed in $t_{\mathsf{epoch}}$ iterative updates (22); 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 $t_{\mathsf{epoch}}$ iterations and using $N + t_{\mathsf{epoch}}$ samples. We then proceed as follows

$$Q_m^{\mathsf{epoch}} = \textsc{Vr-q-run-epoch}(\, Q_{m-1}^{\mathsf{epoch}}, N, t_{\mathsf{epoch}} \,), \quad m = 1, \dots, M, \tag{25}$$

where $M$ is the total number of epochs, and $Q_m^{\mathsf{epoch}}$ 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 work Du et al. (2017); Khamaru et al. (2020); Sidford et al. (2018a,b); Wainwright (2019b); Xu et al. (2020).

**Algorithm 2:** Asynchronous variance-reduced Q-learning

**input parameters:** number of epochs $M$, epoch length $t_{\mathsf{epoch}}$, recentering length $N$, learning rate $\eta$.

**initialization:** set $Q_0^{\mathsf{epoch}} \leftarrow 0$.

**for** each epoch $m = 1, \cdots, M$ **do**

```
/* Call Algorithm 3.                                          */
```

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

## 4.2 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 $(\varepsilon, \delta)$, choose

$$\eta_t \equiv \eta = \frac{c_0}{\log\big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\big)} \min\left\{\frac{(1-\gamma)^2}{\gamma^2},\ \frac{1}{t_{\mathsf{mix}}}\right\}, \tag{26a}$$

$$N \geq \frac{c_1}{\mu_{\mathsf{min}}}\Big(\frac{1}{(1-\gamma)^3\min\{1,\varepsilon^2\}} + t_{\mathsf{mix}}\Big)\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big), \tag{26b}$$

$$t_{\mathsf{epoch}} \geq \frac{c_2}{\mu_{\mathsf{min}}}\Big(\frac{1}{(1-\gamma)^3} + \frac{t_{\mathsf{mix}}}{1-\gamma}\Big)\log\Big(\frac{1}{(1-\gamma)^2\varepsilon}\Big)\log\Big(\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\Big), \tag{26c}$$

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 $\mu_{\mathsf{min}}$ and $t_{\mathsf{mix}}$ in (4) and (5), respectively. Note that the learning rate (26a) chosen here could be larger than the choice (10b) for the classical form by a factor of $O\big(\frac{1}{(1-\gamma)^2}\big)$ (which happens if $t_{\mathsf{mix}}$ is not too large), allowing the algorithm to progress more aggressively.

**Theorem 4** (Asynchronous variance-reduced Q-learning)**.** *Let $Q_M^{\mathsf{epoch}}$ be the output of Algorithm 2 with parameters chosen according to (26). 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*

$$M \geq c_3 \log\frac{1}{\varepsilon(1-\gamma)^2}. \tag{27}$$

The proof of this result is postponed to Section D.

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

$$\widetilde{O}\Big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\min\{1,\varepsilon^2\}} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big). \tag{28}$$

Except for the second term that depends on the mixing time, the first term matches Wainwright (2019b) 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 $\widetilde{O}\big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}\big)$; the scaling $\frac{1}{(1-\gamma)^3}$ matches the minimax lower bound derived in Azar et al. (2013) 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 (similar to Section 3.3), which extends the analysis Khamaru et al. (2020) to Markovian noise. We do not elaborate on this here as it is not the main focus of the current paper.

**Algorithm 3:** function $Q = \textsc{Vr-q-run-epoch}(\overline{Q}, N, t_{\mathsf{epoch}})$

Draw $N$ new consecutive samples from the sample trajectory; compute $\widetilde{\mathcal{T}}(\overline{Q})$ according to (23).
Set $s_0 \leftarrow$ current state, and $Q_0 \leftarrow \overline{Q}$.
**for** $t = 1, 2, \cdots, t_{\mathsf{epoch}}$ **do**
    Draw action $a_{t-1} \sim \pi_{\mathsf{b}}(s_{t-1})$ and next state $s_t \sim P(\cdot|s_{t-1}, a_{t-1})$.
    Update $Q_t$ according to (22).
**return:** $Q \leftarrow Q_{t_{\mathsf{epoch}}}$.

# 5 Related work

**The Q-learning algorithm and its variants.** The Q-learning algorithm, originally proposed in Watkins (1989), has been analyzed in the asymptotic regime by Borkar and Meyn (2000); Jaakkola et al. (1994); Szepesvári (1998); Tsitsiklis (1994) since more than two decades ago. Additionally, finite-time performance of Q-learning and its variants have been analyzed by Beck and Srikant (2012); Chen et al. (2020); Even-Dar and Mansour (2003); Kearns and Singh (1999); Qu and Wierman (2020); Wainwright (2019a) in the tabular setting, by Bhandari et al. (2018); Cai et al. (2019); Chen et al. (2019); Du et al. (2020, 2019); Fan et al. (2019); Weng et al. (2020a,b); Xu and Gu (2020); Yang and Wang (2019) in the context of function approximations, and by Shah and Xie (2018) with nonparametric regression. In addition, Azar et al. (2011); Devraj and Meyn (2020); Ghavamzadeh et al. (2011); Sidford et al. (2018a); Strehl et al. (2006); Wainwright (2019b) 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. Bai et al. (2019); Jin et al. (2018); Wang et al. (2020)), which is beyond the scope of the current work.

**Finite-sample $\ell_\infty$ guarantees for Q-learning.** We now expand on non-asymptotic $\ell_\infty$ 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 work Even-Dar and Mansour (2003); Szepesvári (1998) exhibit an exponential blow-up in $\frac{1}{1-\gamma}$, which is clearly undesirable. In the synchronous setting, Beck and Srikant (2012); Chen et al. (2020); Even-Dar and Mansour (2003); Wainwright (2019a) 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 (Chen et al., 2020; Wainwright, 2019a) or a constant learning rate (Chen et al., 2020). 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 $t_{\mathsf{cover}}$; see Table 1 for detailed comparisons. Going beyond classical Q-learning, the speedy Q-learning algorithm provably achieves a sample complexity of $\widetilde{O}\big(\frac{t_{\mathsf{cover}}}{(1-\gamma)^4\varepsilon^2}\big)$ (Azar et al., 2011) in the asynchronous setting, whose update rule takes twice the storage of classical Q-learning. 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 $\varepsilon < 1$.

**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 (Bhandari et al., 2018; Dalal et al., 2018a,b; Doan et al., 2019; Gupta et al., 2019; Kaledin et al., 2020; Lee and He, 2019; Lin et al., 2020; Srikant and Ying, 2019; Xu et al., 2020, 2019), Q-learning (Chen et al., 2019; Xu and Gu, 2020), and SARSA (Zou et al., 2019). However, these recent papers typically focused on the (weighted) $\ell_2$ error rather than the $\ell_\infty$ risk, where the latter is often more relevant in the context of RL. In addition, Khamaru et al. (2020); Mou et al. (2020) investigated the $\ell_\infty$ bounds of (variance-reduced) TD learning, although they did not account for Markovian noise.

**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 $\varepsilon$-optimal Q-function but also an $\varepsilon$-optimal policy (Agarwal et al., 2019; Azar et al., 2013; Li et al., 2020). It is worth emphasizing that the minimax optimality of model-based approach has been shown to hold for the entire $\varepsilon$-range; in comparison, the sample optimality of the model-free approach has only been shown for a smaller range of accuracy level $\varepsilon$ in the synchronous setting. We also remark that existing sample complexity analysis for model-based approaches might be generalizable to Markovian data.

# 6 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 $\boldsymbol{\Lambda}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be a diagonal matrix obeying

$$\boldsymbol{\Lambda}_t\big((s,a),(s,a)\big) := \begin{cases} \eta, & \text{if } (s,a)=(s_{t-1},a_{t-1}),\\ 0, & \text{otherwise,} \end{cases} \tag{29}$$

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. $s$th) 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

$$\boldsymbol{P}_t\big((s,a),s'\big) := \begin{cases} 1, & \text{if } (s,a,s')=(s_{t-1},a_{t-1},s_t),\\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Clearly, this set of notation allows us to express the Q-learning update rule (7) in the following matrix form

$$\boldsymbol{Q}_t = \big(\boldsymbol{I}-\boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r}+\gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1}\big). \tag{31}$$

## 6.1 Error decay under 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

$$t_{\mathsf{frame}} := \frac{443 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|T}{\delta}\Big), \tag{32a}$$

$$t_{\mathsf{th}} := \max\left\{\frac{2\log\frac{1}{(1-\gamma)^2\varepsilon}}{\eta\mu_{\mathsf{min}}},\ t_{\mathsf{frame}}\right\}, \tag{32b}$$

$$\mu_{\mathsf{frame}} := \frac{1}{2}\mu_{\mathsf{min}} t_{\mathsf{frame}}, \tag{32c}$$

$$\rho := (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big). \tag{32d}$$

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 (10a))*

$$\|\boldsymbol{Q}_t-\boldsymbol{Q}^\star\|_\infty \le (1-\rho)^k\frac{\|\boldsymbol{Q}_0-\boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} + \varepsilon, \tag{33}$$

*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 $\ell_\infty$ 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.

### 6.2 Proof of Theorem 5

#### 6.2.1 Key decomposition and a recursive formula

The starting point of our proof is the following elementary decomposition

$$\begin{aligned}\boldsymbol{\Delta}_t := \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}^\star + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}_t\big(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\big) \end{aligned} \tag{34}$$

for any $t > 0$, where the first line results from the update rule (31), and the penultimate line follows from the Bellman equation $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{V}^\star$ (see Bertsekas (2017)). Applying this relation recursively gives

$$\boldsymbol{\Delta}_t = \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}^\star}_{=:\boldsymbol{\beta}_{1,t}} + \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1} - \boldsymbol{V}^\star\big)}_{=:\boldsymbol{\beta}_{2,t}} + \underbrace{\prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0}_{=:\boldsymbol{\beta}_{3,t}}. \tag{35}$$

Applying the triangle inequality, we obtain

$$|\boldsymbol{\Delta}_t| \le |\boldsymbol{\beta}_{1,t}| + |\boldsymbol{\beta}_{2,t}| + |\boldsymbol{\beta}_{3,t}|, \tag{36}$$

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

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

  we can easily see that

  $$\big|\boldsymbol{\beta}_{2,t}\big| \le \gamma\sum_{i=1}^{t}\|\boldsymbol{\Delta}_{i-1}\|_\infty\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{1}. \tag{37}$$

- 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 E.1.

  **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*

  $$\forall 1 \le t \le T: \quad \left|\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}^\star\right| \le \tau_1\|\boldsymbol{V}^\star\|_\infty\boldsymbol{1} \tag{38}$$

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

  $$\tau_1 := c\gamma\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}. \tag{39}$$

- 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 5). The proof is deferred to Section E.2.

  **Lemma 2.** *For any $\delta > 0$, recall the definition of $t_{\mathsf{frame}}$ in (32a). Suppose that $T > t_{\mathsf{frame}}$ and $0 < \eta < 1$. Then with probability exceeding $1-\delta$ one has*

  $$\left|\prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0\right| \le (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\big|\boldsymbol{\Delta}_0\big| \le (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\|\boldsymbol{\Delta}_0\|_\infty\boldsymbol{1} \tag{40}$$

  *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

$$\left|\prod_{j=1}^{t}\left(\boldsymbol{I}-\boldsymbol{\Lambda}_j\right)\boldsymbol{\Delta}_0\right| \leq \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \tag{41}$$

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 (36), we arrive at

$$|\boldsymbol{\Delta}_t| \leq \begin{cases} \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1\|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \big\|\boldsymbol{\Delta}_0\big\|_\infty \mathbf{1}, & t < t_{\mathsf{frame}} \\ \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1\|\boldsymbol{V}^\star\|_\infty \mathbf{1} + (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\big\|\boldsymbol{\Delta}_0\big\|_\infty \mathbf{1}, & t_{\mathsf{frame}} \leq t \leq T \end{cases} \tag{42}$$

with probability at least $1-2\delta$, where $t_{\mathsf{frame}}$ is defined in (32a). The rest of the proof is thus dedicated to bounding $|\boldsymbol{\Delta}_t|$ based on the above recursive formula (42).

### 6.2.2 Recursive analysis

**A crude bound.** We start by observing the following recursive relation

$$|\boldsymbol{\Delta}_t| \leq \gamma \sum_{i=1}^{t} \big\|\boldsymbol{\Delta}_{i-1}\big\|_\infty \prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1\|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \qquad 1 \leq t \leq T, \tag{43}$$

which is a direct consequence of (42). In the sequel, we invoke mathematical induction to establish, for all $1 \leq t \leq T$, the following crude upper bound

$$\big\|\boldsymbol{\Delta}_t\big\|_\infty \leq \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty + \|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma}, \tag{44}$$

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

Towards this, we first observe that (44) holds trivially for the base case (namely, $t=0$). Now suppose that the inequality (44) holds for all iterations up to $t-1$. In view of (43) and the induction hypotheses,

$$|\boldsymbol{\Delta}_t| \leq \frac{\gamma\big(\tau_1\|\boldsymbol{V}^\star\|_\infty + \big\|\boldsymbol{\Delta}_0\big\|_\infty\big)}{1-\gamma} \sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} + \tau_1\|\boldsymbol{V}^\star\|_\infty \mathbf{1} + \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, \tag{45}$$

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

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

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

$$\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1} = \boldsymbol{M}_i\mathbf{1} = \boldsymbol{M}_i\boldsymbol{e}_{(s_{i-1},a_{i-1})} = \eta(1-\eta)^{N_{i+1}^t(s_{i-1},a_{i-1})}\boldsymbol{e}_{(s_{i-1},a_{i-1})} \tag{46a}$$

and

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

$$\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{j=0}^{\infty} \eta(1-\eta)^j \boldsymbol{e}_{(s,a)} = \sum_{j=0}^{\infty} \eta(1-\eta)^j \mathbf{1} = \mathbf{1}. \tag{46b}$$

Combining the above relations with the inequality (45), one deduces that

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

thus establishing (44) for the $t$-th iteration. This induction analysis thus validates (44) for all $1 \leq t \leq T$.

**Refined analysis.** Now, we strengthen the bound (44) 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 $t_{\mathsf{th}}$ in (32b) and the fact that $\|\boldsymbol{\Delta}_0\|_\infty = \|\boldsymbol{Q}^\star\|_\infty \leq \frac{1}{1-\gamma}$. It is assumed that $T > t_{\mathsf{th}}$. To facilitate our argument, we introduce a collection of auxiliary quantities $u_t$ as follows

$$u_0 = \frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma}, \tag{47a}$$

$$u_t = \|\boldsymbol{v}_t\|_\infty, \quad \boldsymbol{v}_t = \begin{cases} \gamma \sum_{i=1}^t \prod_{j=i+1}^t (\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} u_{i-1} + \|\boldsymbol{\Delta}_0\|_\infty \mathbf{1}, & \text{for } 1 \leq t \leq t_{\mathsf{th}}, \\ \gamma \sum_{i=1}^t \prod_{j=i+1}^t (\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} u_{i-1}, & \text{for } t > t_{\mathsf{th}}. \end{cases} \tag{47b}$$

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.3.

**Lemma 3.** *Recall the definition* (39) *of* $\tau_1$ *in Lemma 1. With probability at least* $1-2\delta$*, the quantities* $\{u_t\}$ *defined in* (47) *satisfy*

$$\|\boldsymbol{\Delta}_t\|_\infty \leq \frac{\tau_1\|\boldsymbol{V}^\star\|_\infty}{1-\gamma} + u_t + \varepsilon. \tag{48}$$

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 $t_{\mathsf{frame}}$ (cf. (32a)) iterations. Further, we define another auxiliary sequence:

$$w_k := (1-\rho)^k \frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} = (1-\rho)^k \frac{\|\boldsymbol{Q}_0 - \boldsymbol{Q}^\star\|_\infty}{1-\gamma}, \tag{49}$$

where we remind the reader of the definition of $\rho$ in (32d). The connection between $\{w_k\}$ and $\{u_t\}$ is made precise as follows, whose proof is postponed to Section E.4.

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

$$u_t \leq w_k, \qquad \textit{with} \quad k = \max\left\{0, \ \left\lfloor \frac{t - t_{\mathsf{th}}}{t_{\mathsf{frame}}} \right\rfloor\right\}. \tag{50}$$

Combining Lemmas 3-4, we arrive at

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

which finishes the proof of Theorem 5.

## 6.3 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 (33) obeys

$$(1-\rho)^k \frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \leq \varepsilon \tag{51}$$

whenever

$$t \geq t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{2}{(1-\gamma)\eta\mu_{\mathsf{min}}} \log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right), \tag{52}$$

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

$$\eta = \min\left\{\frac{(1-\gamma)^4\varepsilon^2}{c^2\gamma^2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}, \frac{1}{\mu_{\mathsf{frame}}}\right\}, \tag{53}$$

one can easily verify that the second term of (33) satisfies

$$\frac{c\gamma}{1-\gamma}\|\boldsymbol{V}^\star\|_\infty\sqrt{\eta\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)} \leq \varepsilon, \tag{54}$$

where the last step follows since $\|\boldsymbol{V}^\star\|_\infty \leq \frac{1}{1-\gamma}$. Putting the above bounds together ensures $\|\boldsymbol{\Delta}_t\|_\infty \leq 3\varepsilon$. We have thus concluded the proof, as long as the claim (51) can be justified.

*Proof of the inequality* (51). Observe that $(1-\rho)^k\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \leq \exp(-\rho k)\frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} \leq \varepsilon$ holds true whenever $k \geq \frac{\log(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)})}{\rho}$, which would hold as long as (according to the definition (50) of $k$)

$$t \geq t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{t_{\mathsf{frame}}}{\rho}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right). \tag{55}$$

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

$$\rho = (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big) \geq (1-\gamma)\left(1-1+\frac{\eta\mu_{\mathsf{frame}}}{2}\right) = \frac{1}{2}(1-\gamma)\eta\mu_{\mathsf{frame}}.$$

As a consequence, the condition (55) would hold as long as

$$t \geq t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{2t_{\mathsf{frame}}}{(1-\gamma)\eta\mu_{\mathsf{frame}}}\log\left(\frac{1}{\varepsilon(1-\gamma)^2}\right) \geq t_{\mathsf{th}} + t_{\mathsf{frame}} + \frac{2}{(1-\gamma)\eta\mu_{\mathsf{min}}}\log\left(\frac{\|\boldsymbol{\Delta}_0\|_\infty}{\varepsilon(1-\gamma)}\right),$$

where we have made use of the simple bound $\|\boldsymbol{\Delta}_0\|_\infty = \|\boldsymbol{Q}^\star\|_\infty \leq \frac{1}{1-\gamma}$ with $\boldsymbol{Q}_0 = \mathbf{0}$. $\square$

# 7 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 discount complexity $\frac{1}{1-\gamma}$.

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 discount complexity, which is inferior to its model-based counterpart. In the synchronous setting, Wainwright (2019a) demonstrated an empirical lower bound $\frac{1}{(1-\gamma)^4}$ for Q-learning. It would be important to determine the exact scaling in this regard. 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 Dann and Brunskill (2015). 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 Doan et al. (2020); Sun et al. (2020).

## Acknowledgements

Y. Wei is supported in part by the NSF grant CCF-2007911 and DMS-2015447. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1806154 and CCF-2007911. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, IIS-1900140 and DMS-2014279, and the Princeton SEAS Innovation Award. We thank Shicong Cen, Chen Cheng and Cong Ma for numerous discussions about reinforcement learning.


# A Preliminaries on Markov chains

For any two probability distributions $\mu$ and $\nu$, denote by $d_{\mathsf{TV}}(\mu,\nu)$ the total variation distance between $\mu$ and $\nu$ (Brémaud, 2013). 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 $\mu$, we let $P^t(\cdot|x)$ denote the distribution of $X_t$ conditioned on $X_0 = x$. Then the mixing time $t_{\mathsf{mix}}$ of this Markov chain is defined by

$$t_{\mathsf{mix}}(\epsilon) := \min\Big\{t \;\Big|\; \max_{x\in\mathcal{X}} d_{\mathsf{TV}}\big(P^t(\cdot|x),\mu\big) \le \epsilon\Big\}; \tag{56a}$$

$$t_{\mathsf{mix}} := t_{\mathsf{mix}}(1/4). \tag{56b}$$

## A.1 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 5.** *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*

$$\forall y\in\mathcal{X}: \quad \mathbb{P}_{X_1=y}\Bigg\{\exists x\in\mathcal{X}: \sum_{i=1}^{t}\mathbb{1}\{X_i=x\} \le \frac{1}{2}t\mu(x)\Bigg\} \le \delta. \tag{57}$$

*Proof.* To begin with, consider the scenario when $X_1\sim\mu$ (namely, $X_1$ follows the stationary distribution of the chain). Then (Paulin, 2015, Theorem 3.4) tells us that: for any given $x\in\Omega$ and any $\tau\ge 0$,

$$\begin{aligned}\mathbb{P}_{X_1\sim\mu}\left\{\sum_{i=1}^{t}\mathbb{1}\{X_i=x\}\le t\mu(x)-\tau\right\} &\le 2\exp\left(-\frac{\tau^2\gamma_{\mathsf{ps}}}{8(t+1/\gamma_{\mathsf{ps}})\mu(x)+20\tau}\right)\\ &\le 2\exp\left(-\frac{\tau^2/t_{\mathsf{mix}}}{16(t+2t_{\mathsf{mix}})\mu(x)+40\tau}\right),\end{aligned} \tag{58}$$

where $\gamma_{\mathsf{ps}}$ stands for the so-called *pseudo spectral gap* as defined in Paulin (2015, 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. Paulin (2015, Proposition 3.4)). Consequently, for any $t\ge t_{\mathsf{mix}}$ and any $\tau\ge 0$, continue the bound (58) to obtain

$$(58) \le 2\exp\left(-\frac{\tau^2}{48t\mu(x)t_{\mathsf{mix}}+40\tau t_{\mathsf{mix}}}\right) \le 2\max\left\{\exp\left(-\frac{\tau^2}{96t\mu(x)t_{\mathsf{mix}}}\right),\exp\left(-\frac{\tau}{80t_{\mathsf{mix}}}\right)\right\} \le \frac{\delta}{|\mathcal{X}|},$$

provided that $\tau \ge \max\big\{10\sqrt{t\mu(x)t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}},\, 80t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}\big\}$. As a result, by taking $\tau = \frac{10}{21}t\mu(x)$ and applying the union bound, we reach

$$\mathbb{P}_{X_1\sim\mu}\left\{\exists x\in\mathcal{X}:\; \sum_{i=1}^{t}\mathbb{1}\{X_i=x\}\le\frac{11}{21}t\mu(x)\right\} \le \sum_{x\in\mathcal{X}}\mathbb{P}_{X_1\sim\mu}\left\{\sum_{i=1}^{t}\mathbb{1}\{X_i=x\}\le\frac{11}{21}t\mu(x)\right\}\le\delta, \tag{59}$$

as long as $\frac{10}{21}t\mu(x) \ge \max\big\{10\sqrt{t\mu(x)t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}},\, 80t_{\mathsf{mix}}\log\frac{2|\mathcal{X}|}{\delta}\big\}$ for all $x\in\mathcal{X}$, or equivalently, when $t\ge\frac{441t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{2|\mathcal{X}|}{\delta}$ with $\mu_{\mathsf{min}} := \min_{x\in\mathcal{X}}\mu(x)$.

Next, we move on to the case when $X_1$ takes an arbitrary state $y \in \mathcal{X}$. From the definition of $t_{\mathsf{mix}}(\cdot)$ (cf. (56a)), we know that

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

This together with the definition of $d_{\mathsf{TV}}$ (cf. Paulin (2015, Equation (1.1))) reveals that: for any event $\mathcal{B}$ that can be fully determined by $\{X_\tau\}_{\tau \geq t_{\mathsf{mix}}(\delta)}$, one has

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

and hence

$$\begin{aligned}
&\sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\Bigg\{\exists x \in \mathcal{X} : \sum_{i=t_{\mathsf{mix}}(\delta)}^{t} \mathbb{1}\{X_i = x\} \leq \frac{11}{21}(t - t_{\mathsf{mix}}(\delta))\mu(x)\Bigg\} \\
&\qquad \leq \mathbb{P}_{X_1\sim\mu}\Bigg\{\exists x \in \mathcal{X} : \sum_{i=t_{\mathsf{mix}}(\delta)}^{t} \mathbb{1}\{X_i = x\} \leq \frac{11}{21}(t - t_{\mathsf{mix}}(\delta))\mu(x)\Bigg\} + \delta \leq 2\delta, \qquad (60)
\end{aligned}$$

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

To finish up, we recall from Paulin (2015, Section 1.1) that $t_{\mathsf{mix}}(\delta) \leq 2t_{\mathsf{mix}} \log \frac{2}{\delta}$, which together with the above constraint on $t$ necessarily implies that $\frac{11}{21}(t - t_{\mathsf{mix}}(\delta)) \geq \frac{1}{2}t$. To conclude, if $t \geq \frac{443 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \log \frac{2|\mathcal{X}|}{\delta} \geq t_{\mathsf{mix}}(\delta) + \frac{441 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \log \frac{2|\mathcal{X}|}{\delta}$, one has

$$\sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\Bigg\{\exists x \in \mathcal{X} : \sum_{i=1}^{t} \mathbb{1}\{X_i = x\} \leq \frac{1}{2}t\mu(x)\Bigg\} \leq \sup_{y\in\mathcal{X}} \mathbb{P}_{X_1=y}\Bigg\{\exists x \in \mathcal{X} : \sum_{i=t_{\mathsf{mix}}(\delta)}^{t} \mathbb{1}\{X_i = x\} \leq \frac{1}{2}t\mu(x)\Bigg\} \leq 2\delta.$$

as claimed. □

## A.2 Connection between the mixing time and the cover time

Lemma 5 combined with the definition (6) immediately reveals the following upper bound on the cover time:

$$t_{\mathsf{cover}} = O\Big(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}} \log |\mathcal{X}|\Big). \qquad (61)$$

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 (61) is provably tight.

**Example.** Consider a time-homogeneous Markov chain with state space $\mathcal{X} := \{1, \cdots, |\mathcal{X}|\}$ and probability transition matrix

$$\boldsymbol{P} = \Big(1 - \frac{q(k+1)}{2}\Big)\boldsymbol{I}_{|\mathcal{X}|} + \frac{q}{|\mathcal{X}|}\begin{bmatrix} k\mathbf{1}_{|\mathcal{X}|}\mathbf{1}_{|\mathcal{X}|/2}^\top & \mathbf{1}_{|\mathcal{X}|}\mathbf{1}_{|\mathcal{X}|/2}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \qquad (62)$$

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

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

*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

$$\mu_{\mathsf{min}} := \min_{1 \le x \le |\mathcal{X}|} \mu_x = \frac{2}{(k+1)|\mathcal{X}|}. \tag{63}$$

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 $\boldsymbol{P}$ (Brémaud, 2013). A little algebra yields

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

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

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

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 Paulin (2015, Proposition 3.3)

$$t_{\mathsf{mix}} \le \frac{2\log 2 + \log \frac{1}{\mu_{\mathsf{min}}}}{2(1-\lambda_2)} = \frac{2\log 2 + \log \frac{1}{\mu_{\mathsf{min}}}}{q(k+1)}. \tag{64}$$

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

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

where (i) follows from the chain rule, (ii) relies on the Markovian property, (iii) results from the construction (62), 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

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

This taken collectively with the definition of $t_{\mathsf{cover}}$ (cf. (6)) 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 (63) and (64). $\square$

# B 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 $t_{\mathsf{cover}}$ in (6), and further introduce a quantity

$$t_{\mathsf{cover,all}} := t_{\mathsf{cover}} \log \frac{T}{\delta}. \tag{65}$$

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

**Lemma 6.** *Define the event*

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

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

*Proof.* See Section E.6. □

In other words, Lemma 6 tells us that with high probability, all state-action pairs are visited at least once in every time frame $(l t_{\mathsf{cover,all}}, (l+1) t_{\mathsf{cover,all}}]$ with $0 \leq l \leq \lfloor T / t_{\mathsf{cover,all}} \rfloor$. The next result is an immediate consequence of Lemma 6; the proof can be found in Section E.2.

**Lemma 7.** *For any $\delta > 0$, recall the definition of $t_{\mathsf{cover,all}}$ in (65). Suppose that $T > t_{\mathsf{cover,all}}$ and $0 < \eta < 1$. Then with probability exceeding $1 - \delta$ one has*

$$\left| \prod_{j=1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j) \boldsymbol{\Delta}_0 \right| \leq (1-\eta)^{\frac{t}{2 t_{\mathsf{cover,all}}}} \|\boldsymbol{\Delta}_0\|_\infty \boldsymbol{1} \tag{66}$$

*uniformly over all $t$ obeying $T \geq t \geq t_{\mathsf{cover,all}}$ 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 (42) (except that Lemma 2 is replaced by Lemma 7) yields

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

with probability at least $1 - 2\delta$. This observation resembles (42), except that $t_{\mathsf{frame}}$ (resp. $\mu_{\mathsf{min}}$) 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 6.2.2 to establish a convergence guarantee based on the cover time. More specifically, define

$$\widetilde{\rho} := (1-\gamma)\Big(1 - (1-\eta)^{\frac{t_{\mathsf{cover,all}}}{2 t_{\mathsf{cover,all}}}}\Big) = (1-\gamma)\Big(1 - (1-\eta)^{\frac{1}{2}}\Big). \tag{67}$$

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

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \leq (1-\widetilde{\rho})^k \frac{\|\boldsymbol{Q}_0 - \boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma} \|\boldsymbol{V}^\star\|_\infty \sqrt{\eta \log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} + \varepsilon \tag{68}$$

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

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

Repeating the proof of the inequality (51) yields

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

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

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

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

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

for some sufficiently large constant $c' > 0$. This together with the definition (65) completes the proof.

# C Analysis under adaptive learning rates (proof of Theorem 3)

**Useful preliminary facts.** We first make note of several useful properties about $\eta_t$.

- Invoking the concentration result in Lemma 5, one can easily show that with probability at least $1-\delta$,

  $$\frac{1}{2}\mu_{\mathsf{min}} < \min_{s,a} \frac{K_t(s,a)}{t} < 2\mu_{\mathsf{min}} \tag{69}$$

  holds simultaneously for all $t$ obeying $T \geq t \gtrsim \frac{t_{\mathsf{mix}} \log(\frac{|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}}$. In addition, this concentration result taken collectively with the update rule (17) of $\widehat{\mu}_{\mathsf{min},t}$ (in particular, the second case of (17)) implies that $\widehat{\mu}_{\mathsf{min},t}$ stabilizes as $t$ grows; more precisely, there exists some quantity $c' \in [1/4, 4]$ such that

  $$\widehat{\mu}_{\mathsf{min},t} \equiv c' \mu_{\mathsf{min}} \tag{70}$$

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

- For any $t$ obeying $t \gtrsim \frac{t_{\mathsf{mix}} \log(\frac{|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}(1-\gamma)}$ (so that $\frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}$ is small enough), the learning rate (18) simplifies to

  $$\eta_t = c_\eta \exp\Big(\Big\lfloor \log \frac{\log t}{c' \mu_{\mathsf{min}}(1-\gamma)\gamma^2 t}\Big\rfloor\Big). \tag{71}$$

  Clearly, there exists a sequence of endpoints $t_1 < t_2 < t_3 < \ldots$ as well as a threshold $k_{\mathsf{th}}$ such that: for any $k \geq k_{\mathsf{th}}$ one has

  $$2t_k < t_{k+1} < 3t_k \qquad \text{and} \tag{72}$$

  $$\eta_t = \eta_{(k)} := \frac{\alpha_k \log t_{k+1}}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2 t_{k+1}}, \quad \forall t_k < t \leq t_{k+1} \tag{73}$$

  for some constant $\alpha_k > 0$; in words, (73) provides a concrete expression for the piecewise constant learning rate, where the $t_k$'s form the change points.

Combining (73) with the definition of $\widehat{Q}_t$ (cf. (17)), one can easily check that for $t \gtrsim \frac{t_{\mathsf{mix}} \log(\frac{|\mathcal{S}||\mathcal{A}|t}{\delta})}{\mu_{\mathsf{min}}(1-\gamma)}$,

$$\widehat{Q}_t = Q_{t_k}, \qquad \forall t_k < t \leq t_{k+1}, \tag{74}$$

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

**A crude bound.** Given that $0 < \eta_t \leq 1$ and $0 \leq r(s,a) \leq 1$, the update rule (7) of $\boldsymbol{Q}_t$ implies that

$$\|\boldsymbol{Q}_t\|_\infty \leq \max\left\{(1-\eta_t)\|\boldsymbol{Q}_{t-1}\|_\infty + \eta_t(1+\gamma\|\boldsymbol{Q}_{t-1}\|_\infty),\ \|\boldsymbol{Q}_{t-1}\|_\infty\right\} \leq \|\boldsymbol{Q}_{t-1}\|_\infty + \gamma,$$

thus leading to the following crude bound

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \leq t + \|\boldsymbol{Q}_0\|_\infty + \|\boldsymbol{Q}^\star\|_\infty \leq t + \frac{2}{1-\gamma} \leq 3t, \qquad \text{for any } t > \frac{1}{1-\gamma}. \tag{75}$$

**Refined analysis.** Define

$$\varepsilon_k := \sqrt{\frac{c_{k,0}\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})\log t_k}{\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2 t_k}}, \tag{76}$$

where $c_{k,0} = \alpha_{k-1}/c_1 > 0$ is some constant, and $c_1 > 0$ is the constant stated in Theorem 1. The property (73) of $\eta_t$ together with the definition (76) implies that

$$\eta_t = \frac{c_1(1-\gamma)^4\varepsilon_k^2}{\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})} = \frac{c_1}{\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})}\min\left\{(1-\gamma)^4\varepsilon_k^2, \frac{1}{t_{\mathsf{mix}}}\right\}, \quad \forall t \in (t_{k-1}, t_k],$$

as long as $(1-\gamma)^4\varepsilon_k^2 \leq 1/t_{\mathsf{mix}}$, or more explicitly,

$$t_k \geq \frac{c_{k,0}t_{\mathsf{mix}}\log(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta})\log t_k}{\mu_{\mathsf{min}}(1-\gamma)\gamma^2}. \tag{77}$$

In addition, condition (72) further tells us that

$$\begin{aligned} t_k - t_{k-1} \asymp t_k &= \frac{c_{k,0}\log\left(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta}\right)\log t_k}{\mu_{\mathsf{min}}(1-\gamma)^5\gamma^2\varepsilon_k^2} \\ &\asymp \max\left\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon_k^2}, \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\right\}\log\left(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta}\right)\log t_k \end{aligned}$$

under the sample size condition (77). Now suppose that $c_{k,0}$ is sufficiently large (which can be guaranteed by adjusting the constant $c_\eta$ in (18)). Invoking Theorem 1 with an initialization $\boldsymbol{Q}_{t_{k-1}}$ (which clearly satisfies the crude bound (75)) ensures that

$$\|\boldsymbol{Q}_{t_k} - \boldsymbol{Q}^\star\|_\infty \leq \varepsilon_k \tag{78}$$

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

$$t_k \geq \frac{c_5}{\mu_{\mathsf{min}}}\left\{\frac{1}{(1-\gamma)^5\varepsilon_k^2} + \frac{t_{\mathsf{mix}}}{1-\gamma}\right\}\log\left(\frac{|\mathcal{S}||\mathcal{A}|t_k}{\delta}\right)\log\left(\frac{t_k}{(1-\gamma)^2\varepsilon_k}\right) \tag{79}$$

for some large enough constant $c_5 > 0$.

Finally, taking $t_{k_{\max}}$ to be the largest change point that does not exceed $T$, we see from (72) that $\frac{1}{3}T \leq t_{k_{\max}} \leq T$. By choosing the constant $c_5$ to be sufficiently large, we can ensure that $\varepsilon_{k_{\max}} \leq \varepsilon$. These immediately conclude the proof of the theorem under the sample size condition (21).

# D 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

$$t_{\mathsf{frame}} := \frac{443t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\left(\frac{4|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}\right), \tag{80a}$$

$$t_{\mathsf{th}} := \max\left\{\frac{2\log\frac{1}{(1-\gamma)^2\varepsilon}}{\eta\mu_{\mathsf{min}}},\ t_{\mathsf{frame}}\right\}, \tag{80b}$$

$$\rho := (1-\gamma)\big(1-(1-\eta)^{\mu_{\mathsf{frame}}}\big), \tag{80c}$$

$$\mu_{\mathsf{frame}} := \frac{1}{2}\mu_{\mathsf{min}} t_{\mathsf{frame}}. \tag{80d}$$

## D.1 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 (23) 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

$$t_{\mathsf{epoch}} := t_{\mathsf{frame}} + t_{\mathsf{th}} + \frac{8\log\frac{2}{1-\gamma}}{(1-\gamma)\eta\mu_{\mathsf{min}}} \tag{81}$$

iterations of variance-reduced Q-learning updates, where $t_{\mathsf{frame}}$ and $t_{\mathsf{th}}$ are defined in (80a) and (80b), 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.

### D.1.1 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 (29) and (30), respectively, we can rewrite (22) as follows

$$\boldsymbol{Q}_t = (\boldsymbol{I}-\boldsymbol{\Lambda}_t)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}\right). \tag{82}$$

Following similar steps as in the expression (34), we arrive at the following error decomposition

$$\begin{aligned}
\boldsymbol{\Theta}_t := \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= (\boldsymbol{I}-\boldsymbol{\Lambda}_t)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}}\right) - \boldsymbol{Q}^\star \\
&= (\boldsymbol{I}-\boldsymbol{\Lambda}_t)(\boldsymbol{Q}_{t-1}-\boldsymbol{Q}^\star) + \boldsymbol{\Lambda}_t\left(\boldsymbol{r} + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \gamma\widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} - \boldsymbol{Q}^\star\right) \\
&= (\boldsymbol{I}-\boldsymbol{\Lambda}_t)(\boldsymbol{Q}_{t-1}-\boldsymbol{Q}^\star) + \gamma\boldsymbol{\Lambda}_t\left(\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\overline{\boldsymbol{V}}) + \widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} - \boldsymbol{P}\boldsymbol{V}^\star\right) \\
&= (\boldsymbol{I}-\boldsymbol{\Lambda}_t)\boldsymbol{\Theta}_{t-1} + \gamma\boldsymbol{\Lambda}_t(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}} + \gamma\boldsymbol{\Lambda}_t(\boldsymbol{P}_t-\boldsymbol{P})(\boldsymbol{V}^\star-\overline{\boldsymbol{V}}) + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}_t(\boldsymbol{V}_{t-1}-\boldsymbol{V}^\star),
\end{aligned} \tag{83}$$

which once again leads to a recursive relation

$$\begin{aligned}
\boldsymbol{\Theta}_t = &\underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}}_{=:\boldsymbol{h}_{0,t}} + \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i(\boldsymbol{P}_i-\boldsymbol{P})(\boldsymbol{V}^\star-\overline{\boldsymbol{V}})}_{=:\boldsymbol{h}_{1,t}} \\
&+ \underbrace{\gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\boldsymbol{P}_i(\boldsymbol{V}_{i-1}-\boldsymbol{V}^\star)}_{=:\boldsymbol{h}_{2,t}} + \underbrace{\prod_{j=1}^{t}(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Theta}_0}_{=:\boldsymbol{h}_{3,t}}.
\end{aligned} \tag{84}$$

This identity takes a very similar form as (35) 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 E.5.

**Lemma 8.** *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*

$$\|\boldsymbol{h}_{0,t}\|_\infty \le \gamma\sqrt{\frac{4\log\left(\frac{6N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\mathsf{min}}}}\|\overline{\boldsymbol{V}}-\boldsymbol{V}^\star\|_\infty + \frac{\gamma}{1-\gamma}\sqrt{\frac{4\log\left(\frac{6N|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N\mu_{\mathsf{min}}}}. \tag{85}$$

Inheriting 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 (39)). In addition, the term $\boldsymbol{h}_{2,t}$ can be bounded in the same way as $\boldsymbol{\beta}_{2,t}$ in (37). Therefore, repeating the same argument as for Theorem 5, we conclude that with probability at least $1-\delta$,

$$\|\boldsymbol{\Theta}_t\|_\infty \le (1-\rho)^k \frac{\|\boldsymbol{\Theta}_0\|_\infty}{1-\gamma} + \widetilde{\tau} + \xi = (1-\rho)^k \frac{\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty}{1-\gamma} + \widetilde{\tau} + \xi \tag{86}$$

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

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

for some constant $c > 0$.

Let $C > 0$ be some sufficient large constant. Setting $\eta_t \equiv \eta = \min\Big\{\frac{(1-\gamma)^2}{C\gamma^2 \log\frac{|\mathcal{S}||\mathcal{A}|t_{\mathsf{epoch}}}{\delta}}, \frac{1}{\mu_{\mathsf{frame}}}\Big\}$, $\xi = \frac{1}{16\sqrt{1-\gamma}}$ 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

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

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

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

which in turn implies that

$$\|\boldsymbol{\Theta}_{t_{\mathsf{epoch}}}\|_\infty \le \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{1}{8\sqrt{(1-\gamma)}} + \frac{1}{4}\|\boldsymbol{V}^\star - \overline{\boldsymbol{V}}\|_\infty \le \frac{1}{2}\max\Big\{\frac{1}{\sqrt{1-\gamma}},\ \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty\Big\}, \tag{87}$$

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

$$\|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \boldsymbol{Q}^\star\|_\infty \le \frac{1}{2}\max\Big\{\frac{1}{\sqrt{1-\gamma}},\ \|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty\Big\}. \tag{88}$$

#### D.1.2 Phase 2: when $\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \le 1/\sqrt{1-\gamma}$

The analysis of Phase 2 follows by straightforwardly combining the analysis of Phase 1 and that of the synchronous counterpart in Wainwright (2019b). For the sake of brevity, we only sketch the main steps.

Following the proof idea of Wainwright (2019b, 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,

$$\widehat{\boldsymbol{Q}} = \boldsymbol{r} + \gamma \boldsymbol{P}(\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}) + \gamma \widetilde{\boldsymbol{P}} \overline{\boldsymbol{V}}. \tag{89}$$

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 $\boldsymbol{r} + \gamma(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}$. Repeating the arguments in the proof of Wainwright (2019b, Lemma 4) (except that we need to apply the measure concentration of $\widetilde{\boldsymbol{P}}$ in the manner performed in the proof of Lemma 8 due to Markovian data), we reach

$$\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \le c' \sqrt{\frac{\log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^3 N \mu_{\mathsf{min}}}} \le \varepsilon \tag{90}$$

with probability at least $1-\delta$ for some constant $c' > 0$, provided that $N \ge (c')^2 \frac{\log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}}{(1-\gamma)^3 \varepsilon^2}$ and that $\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \le 1/\sqrt{1-\gamma}$. 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

$$\|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \widehat{\boldsymbol{Q}}\|_\infty \le \frac{\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty}{8} + \frac{\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty}{8} + \varepsilon. \tag{91}$$

Under this claim, the triangle inequality yields

$$\begin{aligned} \|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \boldsymbol{Q}^\star\|_\infty &\le \|\boldsymbol{Q}_{t_{\mathsf{epoch}}} - \widehat{\boldsymbol{Q}}\|_\infty + \|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty \le \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{9}{8}\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \varepsilon \\ &\le \frac{1}{8}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}^\star\|_\infty + \frac{17}{8}\varepsilon, \end{aligned} \tag{92}$$

where the last inequality follows from (90).

**Proof of the inequality (91).** Recalling the variance-reduced update rule (82) and using the Bellman-type equation (89), we obtain

$$\begin{aligned} \widehat{\boldsymbol{\Theta}}_t := \boldsymbol{Q}_t - \widehat{\boldsymbol{Q}} &= (\boldsymbol{I} - \boldsymbol{\Lambda}_t)(\boldsymbol{Q}_{t-1} - \widehat{\boldsymbol{Q}}) + \boldsymbol{\Lambda}_t \Big( \boldsymbol{r} + \gamma \boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \overline{\boldsymbol{V}}) + \gamma \widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} - \boldsymbol{r} - \gamma \boldsymbol{P}(\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}) - \gamma \widetilde{\boldsymbol{P}}\overline{\boldsymbol{V}} \Big) \\ &= (\boldsymbol{I} - \boldsymbol{\Lambda}_t)(\boldsymbol{Q}_{t-1} - \widehat{\boldsymbol{Q}}) + \boldsymbol{\Lambda}_t \Big( \gamma \boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \overline{\boldsymbol{V}}) - \gamma \boldsymbol{P}(\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}) \Big) \\ &= (\boldsymbol{I} - \boldsymbol{\Lambda}_t)\widehat{\boldsymbol{\Theta}}_{t-1} + \gamma \boldsymbol{\Lambda}_t \Big( (\boldsymbol{P}_t - \boldsymbol{P})(\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}) + \boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \widehat{\boldsymbol{V}}) \Big). \end{aligned} \tag{93}$$

Adopting the same expansion as before (see (35)), we arrive at

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

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

$$\begin{aligned} |\boldsymbol{\vartheta}_{1,t}| &\le c\gamma \|\widehat{\boldsymbol{V}} - \overline{\boldsymbol{V}}\|_\infty \sqrt{\eta \log\Big(\frac{|\mathcal{S}||\mathcal{A}| t_{\mathsf{epoch}}}{\delta}\Big)}\,\boldsymbol{1}; \\ |\boldsymbol{\vartheta}_{3,t}| &\le \begin{cases} (1-\eta)^{\frac{1}{2} t \mu_{\mathsf{min}}} \|\widehat{\boldsymbol{\Theta}}_0\|_\infty \boldsymbol{1}, & \text{if } t_{\mathsf{frame}} \le t \le t_{\mathsf{epoch}}, \\ \|\widehat{\boldsymbol{\Theta}}_0\|_\infty \boldsymbol{1}, & \text{if } t < t_{\mathsf{frame}}. \end{cases} \end{aligned}$$

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 $t_{\mathsf{th}}$ defined in (80b).

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

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

where the last line follows by the triangle inequality.

### D.2 How many epochs are needed?

We are now ready to pin down how many epochs are needed to achieve $\varepsilon$-accuracy.

- In Phase 1, the contraction result (88) indicates that, if the algorithm is initialized with $\boldsymbol{Q}_0 = \boldsymbol{0}$ at the very beginning, then it takes at most

  $$\log_2\left(\frac{\|\boldsymbol{Q}^\star\|_\infty}{\max\big\{\varepsilon, \frac{1}{\sqrt{1-\gamma}}\big\}}\right) \le \log_2\Big(\frac{1}{\sqrt{1-\gamma}}\Big) + \log_2\Big(\frac{1}{\varepsilon(1-\gamma)}\Big)$$

  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 (92). Given that $\overline{\boldsymbol{Q}}$ is also the last iterate of the preceding epoch, the property (92) 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 $\ell_\infty$ estimation error at most $3\varepsilon$.

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.

# E Proofs of technical lemmas

## E.1 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 (35), we can write

$$\boldsymbol{\beta}_{1,t}(s,a) = \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \big(1 - \boldsymbol{\Lambda}_j(s,a)\big)\boldsymbol{\Lambda}_i(s,a)\big(\boldsymbol{P}_i(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^\star. \tag{94}$$

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

$$t_k(s,a) := \text{the time stamp when the trajectory visits } (s,a) \text{ for the } k\text{-th time} \tag{95}$$

and

$$K_t(s,a) := \max\left\{k \mid t_k(s,a) \le t\right\}, \tag{96}$$

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. (29)) allows us to rewrite (94) as

$$\boldsymbol{\beta}_{1,t}(s,a) = \gamma \sum_{k=1}^{K_t(s,a)} (1-\eta)^{K_t(s,a)-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}. \tag{97}$$

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 (97).

Towards this, we claim that: there exists some constant $c>0$ such that with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a) - \boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}\right| \le c\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\|\boldsymbol{V}^{\star}\|_{\infty} \tag{98}$$

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 (96)) and substituting the bound (98) into the expression (97), we arrive at

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad |\boldsymbol{\beta}_{1,t}(s,a)| \le c\sqrt{\eta\log\Big(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)}\|\boldsymbol{V}^{\star}\|_{\infty}, \tag{99}$$

thus concluding the proof of this lemma. It remains to validate the inequality (98).

*Proof of the inequality (98).* Before proceeding, we introduce some additional notation. Let $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V})^{\star}\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a vector whose $(s,a)$-th entry is given by the variance of $\boldsymbol{V}^{\star}$ w.r.t. the transition probability $P_{s,a}(\cdot)$ from state $s$ when action $a$ is taken, namely,

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}, \qquad \big[\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^{\star})\big]_{(s,a)} := \sum_{s'\in\mathcal{S}} P_{s,a}(s')\big(V^{\star}(s')\big)^2 - \Big(\sum_{s'\in\mathcal{S}} P_{s,a}(s')V^{\star}(s')\Big)^2. \tag{100}$$

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. 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

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

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

$$\mathbb{P}\left\{s_{t_k+1} = i_k \ (\forall 1\le k\le K)\right\} = \prod_{j=1}^{K}\mathbb{P}_{s,a}(i_j), \tag{101}$$

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 Bernstein inequality to bound the quantity of interest (which has zero mean). To begin with, the variance parameter can be characterized by

$$\begin{aligned}
\mathsf{Var}\left[\sum_{k=1}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}\right] &= \sum_{k=1}^{K}(1-\eta)^{2K-2k}\eta^2\mathsf{Var}\left[\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}\right] \\
&= \eta^2\sum_{k=1}^{K}(1-\eta)^{2K-2k}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left[\boldsymbol{V}^{\star}\right] \\
&\le \eta^2\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left[\boldsymbol{V}^{\star}\right]\sum_{j=0}^{\infty}(1-\eta)^j = \frac{\eta^2}{1-(1-\eta)}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left[\boldsymbol{V}^{\star}\right] \\
&= \eta\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left[\boldsymbol{V}^{\star}\right] =: \sigma_K^2.
\end{aligned}$$

In addition, each term in the summation clearly satisfies

$$\left|(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}\right| \le 2\eta\|\boldsymbol{V}^{\star}\|_{\infty} =: D, \qquad 1 \le k \le K.$$

As a consequence, invoking the Bernstein inequality implies that

$$\begin{aligned}
\left|\sum_{k=1}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}^{\star}\right| &\le \tilde{c}\left(\sqrt{\sigma_K^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)} + D\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)\right) \\
&\le \tilde{c}\left(\sqrt{\eta\|\boldsymbol{V}^{\star}\|_{\infty}^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)} + 2\eta\|\boldsymbol{V}^{\star}\|_{\infty}\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)\right) \\
&\le 3\tilde{c}\sqrt{\eta\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}\|\boldsymbol{V}^{\star}\|_{\infty} \qquad (102)
\end{aligned}$$

with probability exceeding $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$, where the second line relies on the simple bound $\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left[\boldsymbol{V}^{\star}\right] \le \|\boldsymbol{V}^{\star}\|_{\infty}^2$, and the last line holds if $0 < \eta\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right) < 1$. 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 (102) holds simultaneously over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $1 \le K \le T$. This concludes the proof. $\square$

## E.2 Proof of Lemma 2 and Lemma 7

**Proof of Lemma 2.** Let $\boldsymbol{\beta}_{3,t} = \prod_{j=1}^{t}\left(\boldsymbol{I}-\boldsymbol{\Lambda}_j\right)\boldsymbol{\Delta}_0$. 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

$$|\boldsymbol{\beta}_{3,t}(s,a)| = (1-\eta)^{K_t(s,a)}\left|\boldsymbol{\Delta}_0(s,a)\right|, \qquad (103)$$

where $K_t(s,a)$ denotes the number of times the sample trajectory visits $(s,a)$ during the iterations $[1,t]$ (cf. (96)). By virtue of Lemma 5 and the union bound, one has, with probability at least $1-\delta$, that

$$K_t(s,a) \ge t\mu_{\mathsf{min}}/2 \qquad (104)$$

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 (103) establishes that, with probability greater than $1-\delta$,

$$|\boldsymbol{\beta}_3(s,a)| \le (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}}\left|\boldsymbol{\Delta}_0(s,a)\right|. \qquad (105)$$

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 7.** 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 6)

$$K_t(s,a) \geq \left\lfloor \frac{t}{t_{\mathsf{cover,all}}} \right\rfloor \geq \frac{t}{2t_{\mathsf{cover,all}}} \tag{106}$$

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 7.

## E.3 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 (44). 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

$$h(t) := \begin{cases} \|\boldsymbol{\Delta}_0\|_\infty, & \text{if } t \leq t_{\mathsf{th}}, \\ (1-\gamma)\varepsilon, & \text{if } t > t_{\mathsf{th}}. \end{cases} \tag{107}$$

Recall that $(1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}} \leq (1-\gamma)\varepsilon$ for any $t \geq t_{\mathsf{th}}$. Therefore, combining the inequality (42) with the induction hypotheses indicates that

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

Taking this together with the inequality (46b) and rearranging terms, we obtain

$$\begin{aligned} |\boldsymbol{\Delta}_t| &\leq \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} u_{i-1} + \frac{\gamma \tau_1 \|\boldsymbol{V}^\star\|_\infty}{1-\gamma} \mathbf{1} + \gamma\varepsilon \mathbf{1} + \tau_1 \|\boldsymbol{V}^\star\|_\infty \mathbf{1} + h(t)\mathbf{1} \\ &= \frac{\tau_1 \|\boldsymbol{V}^\star\|_\infty}{1-\gamma} \mathbf{1} + \gamma\varepsilon \mathbf{1} + \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \mathbf{1} u_{i-1} + h(t)\mathbf{1} \\ &= \frac{\tau_1 \|\boldsymbol{V}^\star\|_\infty}{1-\gamma} \mathbf{1} + \gamma\varepsilon \mathbf{1} + \boldsymbol{v}_t + (1-\gamma)\varepsilon \, \mathbb{1}\{t > t_{\mathsf{th}}\}\mathbf{1} \\ &\leq \frac{\tau_1 \|\boldsymbol{V}^\star\|_\infty}{1-\gamma} \mathbf{1} + \varepsilon \mathbf{1} + \boldsymbol{v}_t, \end{aligned} \tag{108}$$

where we have used the definition of $\boldsymbol{v}_t$ in (47). This taken collectively with the definition $u_t = \|\boldsymbol{v}_t\|_\infty$ establishes that

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

as claimed. This concludes the proof.

## E.4 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 $u_0 \leq \|\boldsymbol{\Delta}_0\|_\infty / (1-\gamma) = w_0$. In fact, repeating our argument for the crude bound (see Section 6.2.2) immediately reveals that

$$\forall t \geq 0: \qquad u_t \leq \frac{\|\boldsymbol{\Delta}_0\|_\infty}{1-\gamma} = w_0, \tag{109}$$

thus indicating that the inequality (50) holds for the base case. In what follows, we assume that the inequality (50) holds up to $k-1$, and would like to extend it to the case with all $t$ obeying $\left\lfloor \frac{t - t_{\mathsf{th}}}{t_{\mathsf{frame}}} \right\rfloor = k$.

Let us focus on the case when $t = t_{\mathsf{th}} + kt_{\mathsf{frame}}$; the case with $t = t_{\mathsf{th}} + kt_{\mathsf{frame}} + j$ $(1 \le j < t_{\mathsf{frame}})$ follows from an analogous argument and is omitted for brevity. In view of the definition of $\boldsymbol{v}_t$ (cf. (47)) as well as our induction hypotheses, one can arrange terms to derive

$$\begin{aligned}
\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}} &= \gamma \sum_{i=1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}} \prod_{j=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \boldsymbol{1} u_{i-1} \\
&= \gamma \sum_{s=0}^{k-1} \left\{ \sum_{i:\, \max\left\{\lfloor \frac{i-1-t_{\mathsf{th}}}{t_{\mathsf{frame}}} \rfloor, 0\right\} = s} \prod_{j=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \boldsymbol{1} u_{i-1} \right\} \\
&\le \gamma \sum_{s=0}^{k-1} \left\{ \sum_{i:\, \max\left\{\lfloor \frac{i-1-t_{\mathsf{th}}}{t_{\mathsf{frame}}} \rfloor, 0\right\} = s} \prod_{j=i+1}^{t_{\mathsf{th}}+kt_{\mathsf{frame}}} (\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i \boldsymbol{1} \right\} w_s,
\end{aligned} \tag{110}$$

where the last inequality follows from our induction hypotheses and the non-negativity of $(\boldsymbol{I} - \boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\boldsymbol{1}$.

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

$$N_s^{k-1} := N_i^j(s,a) \qquad \text{with } i = t_{\mathsf{th}} + st_{\mathsf{frame}} + 1,\ j = t_{\mathsf{th}} + kt_{\mathsf{frame}} \tag{111}$$

for every $s \le k-1$; in words, $N_s^{k-1}$ stands for the total number of visits to $(s,a)$ between the $s$-th frame and the $(k-1)$-th frame. Lemma 5 tells us that, with probability at least $1-2\delta$,

$$N_s^{k-1} \ge (k-s)\mu_{\mathsf{frame}} \qquad \text{with } \mu_{\mathsf{frame}} = \frac{1}{2}\mu_{\mathsf{min}} t_{\mathsf{frame}}, \tag{112}$$

which actually holds uniformly over all state-action pairs $(s,a)$. Armed with this set of notation, it is straightforward to use the expression (110) to verify that

$$\begin{aligned}
\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}(s,a) &\le \gamma \sum_{s=0}^{k-1} \eta \left\{ (1-\eta)^{N_s^{k-1}-1} + (1-\eta)^{N_s^{k-1}-2} + \cdots + (1-\eta)^{N_{s+1}^{k-1}} \right\} w_s \\
&= \gamma \sum_{s=0}^{k-1} \left( (1-\eta)^{N_{s+1}^{k-1}} - (1-\eta)^{N_s^{k-1}} \right) w_s =: \gamma \sum_{s=0}^{k-1} (\alpha_{s+1} - \alpha_s)\, w_s,
\end{aligned} \tag{113}$$

where we denote $\alpha_s := (1-\eta)^{N_s^{k-1}}$ for any $s \le k-1$ and $\alpha_k := 1$.

A litter algebra further leads to

$$\gamma \sum_{s=0}^{k-1} (\alpha_{s+1} - \alpha_s)\, w_s = \gamma(\alpha_k w_{k-1} - \alpha_0 w_0) + \gamma \sum_{s=1}^{k-1} \alpha_s\, (w_{s-1} - w_s)\,. \tag{114}$$

Thus, in order to control the quantity $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}(s,a)$, it suffices to control the right-hand side of (114), for which we start by bounding the last term. Plugging in the definitions of $w_s$ and $\alpha_s$ yields

$$\frac{1-\gamma}{\|\boldsymbol{\Delta}_0\|_\infty} \sum_{s=1}^{k-1} \alpha_s\, (w_{s-1} - w_s) = \sum_{s=1}^{k-1} (1-\eta)^{N_s^{k-1}} (1-\rho)^{s-1} \rho \le \rho \sum_{s=1}^{k-1} (1-\eta)^{(k-s)\mu_{\mathsf{frame}}} (1-\rho)^{s-1},$$

where the last inequality resuls from the fact (112). Additionally, direct calculation yields

$$\rho \sum_{s=1}^{k-1} (1-\eta)^{(k-s)\mu_{\mathsf{frame}}} (1-\rho)^{s-1} = \rho (1-\eta)^{(k-1)\mu_{\mathsf{frame}}} \sum_{s=1}^{k-1} \left( \frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}} \right)^{s-1}$$

$$
\begin{aligned}
&= \rho(1-\eta)^{(k-1)\mu_{\mathsf{frame}}} \frac{1-\left(\frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}}\right)^{k-1}}{1-\frac{1-\rho}{(1-\eta)^{\mu_{\mathsf{frame}}}}} \\
&= \rho(1-\eta)^{\mu_{\mathsf{frame}}} \frac{(1-\rho)^{k-1}-(1-\eta)^{(k-1)\mu_{\mathsf{frame}}}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}} \\
&\leq \rho(1-\eta)^{\mu_{\mathsf{frame}}} \frac{(1-\rho)^{k-1}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}},
\end{aligned} \tag{115}
$$

where the last inequality makes use of the fact that

$$
\begin{aligned}
(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}} &= 1-(1-\gamma)(1-(1-\eta)^{\mu_{\mathsf{frame}}})-(1-\eta)^{\mu_{\mathsf{frame}}} \\
&= \gamma\left\{1-(1-\eta)^{\mu_{\mathsf{frame}}}\right\} = \frac{\gamma}{1-\gamma}\rho \geq 0.
\end{aligned} \tag{116}
$$

Combining the inequalities (113), (114) and (115) and using the fact $\alpha_0 w_0 \geq 0$ give

$$
\begin{aligned}
\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}(s,a) &\leq \gamma \sum_{s=1}^{k-1} \alpha_s\left(w_{s-1}-w_s\right)+\gamma\alpha_k w_{k-1} \\
&\leq \frac{\left\|\boldsymbol{\Delta}_0\right\|_\infty}{1-\gamma}\left\{\gamma\rho(1-\eta)^{\mu_{\mathsf{frame}}}\frac{(1-\rho)^{k-1}}{(1-\rho)-(1-\eta)^{\mu_{\mathsf{frame}}}}+\gamma(1-\rho)^{k-1}\right\}.
\end{aligned} \tag{117}
$$

We are now ready to justify $\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}(s,a) \leq w_k$. Note that the observation (116) implies

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

This combined with the bound (117) yields

$$
\begin{aligned}
\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}(s,a) &\leq \frac{\left\|\boldsymbol{\Delta}_0\right\|_\infty}{1-\gamma}\left\{(1-\gamma)(1-\eta)^{\mu_{\mathsf{frame}}}(1-\rho)^{k-1}+\gamma(1-\rho)^{k-1}\right\} \\
&\leq \frac{\left\|\boldsymbol{\Delta}_0\right\|_\infty}{1-\gamma}\left(\gamma+(1-\gamma)(1-\eta)^{\mu_{\mathsf{frame}}}\right)(1-\rho)^{k-1} \\
&= (1-\rho)^k\frac{\left\|\boldsymbol{\Delta}_0\right\|_\infty}{1-\gamma} = w_k,
\end{aligned} \tag{118}
$$

where the last line follows from the definition of $\rho$ (cf. (32d)). Since the above inequality holds for all state-action pair $(s,a)$, we conclude that

$$
u_{t_{\mathsf{th}}+kt_{\mathsf{frame}}} = \left\|\boldsymbol{v}_{t_{\mathsf{th}}+kt_{\mathsf{frame}}}\right\|_\infty \leq w_k. \tag{119}
$$

We have thus finished the proof for the case when $t = t_{\mathsf{th}}+kt_{\mathsf{frame}}$. As mentioned before, the case with $t = t_{\mathsf{th}}+kt_{\mathsf{frame}}+j$ $(j=1,\ldots,t_{\mathsf{frame}}-1)$ can be justified using the same argument. As a consequence, we have established the inequality (50) for all $t$ obeying $\left\lfloor\frac{t-t_{\mathsf{th}}}{t_{\mathsf{frame}}}\right\rfloor = k$, which together with the induction argument completes the proof of this lemma.

## E.5 Proof of Lemma 8

Recalling that $\mathbf{0} \leq \sum_{i=1}^t \prod_{j=i+1}^t (\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\mathbf{1} \leq \mathbf{1}$ (cf. (46b)), we obtain

$$
\|\boldsymbol{h}_{0,t}\|_\infty \leq \gamma\left\|\sum_{i=1}^t\prod_{j=i+1}^t(\boldsymbol{I}-\boldsymbol{\Lambda}_j)\boldsymbol{\Lambda}_i\right\|_1\left\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\right\|_\infty \leq \gamma\left\|(\widetilde{\boldsymbol{P}}-\boldsymbol{P})\overline{\boldsymbol{V}}\right\|_\infty. \tag{120}
$$

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

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

$$\widetilde{\boldsymbol{P}}(s,a) = \frac{1}{K_N(s,a)} \sum_{i=0}^{N-1} \boldsymbol{P}_{i+1}(s,a)\overline{\boldsymbol{V}}\, \mathbb{1}\{(s_i, a_i) = (s,a)\} = \frac{1}{K_N(s,a)} \sum_{i=1}^{K_N(s,a)} \boldsymbol{P}_{t_i+1}(s,a)\overline{\boldsymbol{V}}, \tag{121}$$

where $\boldsymbol{P}_i$ is defined in (30), 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. (96)), and $t_k := t_k(s,a)$ denotes the time stamp when the trajectory visits $(s,a)$ for the $k$-th time (cf (95)).

In view of our derivation for (101), 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

$$\mathbb{P}\left\{ \frac{1}{k} \left| \sum_{i=1}^{k} \big(\boldsymbol{P}_{t_i+1}(s,a) - \boldsymbol{P}(s,a)\big)\overline{\boldsymbol{V}} \right| \geq \tau \right\} \leq 2\exp\left\{ -\frac{k\tau^2}{2\|\overline{\boldsymbol{V}}\|_\infty^2} \right\}. \tag{122}$$

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

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

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

$$\big\|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty \leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{\frac{2\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{K_N(s,a)}} \big\|\overline{\boldsymbol{V}}\big\|_\infty \tag{123}$$

with probability at least $1 - \delta$.

In addition, for any $N \geq t_{\mathsf{frame}}$, Lemma 5 guarantees that with probability $1 - 2\delta$, each state-action pair $(s,a)$ is visited at least $N\mu_{\mathsf{min}}/2$ times, namely, $K_N(s,a) \geq \frac{1}{2}N\mu_{\mathsf{min}}$ for all $(s,a)$. This combined with (124) yields

$$\begin{aligned}
\big\|(\widetilde{\boldsymbol{P}} - \boldsymbol{P})\overline{\boldsymbol{V}}\big\|_\infty &\leq \sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}} \big\|\overline{\boldsymbol{V}}\big\|_\infty \\
&\leq \sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}} \left(\big\|\overline{\boldsymbol{V}} - \boldsymbol{V}^\star\big\|_\infty + \big\|\boldsymbol{V}^\star\big\|_\infty\right) \\
&\leq \sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}} \big\|\overline{\boldsymbol{V}} - \boldsymbol{V}^\star\big\|_\infty + \frac{1}{1-\gamma}\sqrt{\frac{4\log\big(\frac{2N|\mathcal{S}||\mathcal{A}|}{\delta}\big)}{N\mu_{\mathsf{min}}}}
\end{aligned} \tag{124}$$

with probability at least $1 - 3\delta$, where the second inequality follows from the triangle inequality, and the last inequality follows from $\big\|\boldsymbol{V}^\star\big\|_\infty \leq \frac{1}{1-\gamma}$. Putting this together with (120) concludes the proof.

## E.6 Proof of Lemma 6

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

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

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

$$\mathbb{P}\left\{\mathcal{H}_l \mid (s_{t_l}, a_{t_l}) = (s', a')\right\} \leq \frac{1}{2}. \tag{125}$$

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

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

where the inequality follows from (125). 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

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

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

# References


Agarwal, A., Kakade, S., and Yang, L. F. (2019). Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804*.

Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. (2011). Reinforcement learning with a near optimal rate of convergence. Technical report, INRIA.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient $q$-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8002–8011.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208.

Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

Brémaud, P. (2013). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. (2019). Neural temporal-difference and q-learning converges to global optima. In *Advances in Neural Information Processing Systems*, pages 11312–11322.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. (2019). Performance of Q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018a). Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dalal, G., Thoppe, G., Szörényi, B., and Mannor, S. (2018b). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

Devraj, A. M. and Meyn, S. P. (2020). Q-learning with uniformly bounded variance: Large discounting is not a barrier to fast learning. *arXiv preprint arXiv:2002.10301*.

Doan, T., Maguluri, S., and Romberg, J. (2019). Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1626–1635.

Doan, T. T., Nguyen, L. M., Pham, N. H., and Romberg, J. (2020). Convergence rates of accelerated markov gradient descent with applications in reinforcement learning. *arXiv preprint arXiv:2002.02873*.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1049–1058. JMLR. org.

Du, S. S., Lee, J. D., Mahajan, G., and Wang, R. (2020). Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. *arXiv preprint arXiv:2002.07125*.

Du, S. S., Luo, Y., Wang, R., and Zhang, H. (2019). Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8058–8068.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2019). A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*.

Ghavamzadeh, M., Kappen, H. J., Azar, M. G., and Munos, R. (2011). Speedy Q-learning. In *Advances in neural information processing systems*, pages 2411–2419.

Gupta, H., Srikant, R., and Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Kaledin, M., Moulines, E., Naumov, A., Tadic, V., and Wai, H.-T. (2020). Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. *arXiv preprint arXiv:2002.01268*.

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2020). Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*.

Lee, D. and He, N. (2019). Target-based temporal difference learning. *arXiv preprint arXiv:1904.10945*.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900, accepted to Neural Information Processing Systems*.

Lin, Y., Qu, G., Huang, L., and Wierman, A. (2020). Distributed reinforcement learning in multi-agent networked systems. *arXiv preprint arXiv:2006.06555*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. (2020). On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *arXiv preprint arXiv:2004.04719*.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conference on Learning Theory*.

Shah, D. and Xie, Q. (2018). Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM.

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

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. (2006). PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888.

Sun, T., Sun, Y., Xu, Y., and Yin, W. (2020). Markov chain block coordinate descent. *Computational Optimization and Applications*, pages 1–27.

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

Szepesvári, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp $\ell_\infty$-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

Wang, Y., Dong, K., Chen, X., and Wang, L. (2020). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Weng, B., Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020a). Momentum Q-learning with finite-sample convergence guarantee. *arXiv preprint arXiv:2007.15418*.

Weng, W., Gupta, H., He, N., Ying, L., and Srikant, R. (2020b). Provably-efficient double Q-learning. *arXiv preprint arXiv:2007.05034*.

Xu, P. and Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. *accepted to International Conference on Machine Learning*.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2020). Reanalysis of variance reduced temporal difference learning. *ICLR, arXiv preprint arXiv:2001.01898*.

Xu, T., Zou, S., and Liang, Y. (2019). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643.

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675.