Is Q-Learning Minimax Optimal? A Tight Sample Complexity Analysis

Gen Li* UPenn | Changxiao Cai† UPenn | Yuxin Chen* UPenn | Yuting Wei* UPenn | Yuejie Chi‡ CMU

February 2021; Revised: November 2021

Abstract

Q-learning, which seeks to learn the optimal Q-function of a Markov decision process (MDP) in a model-free fashion, lies at the heart of reinforcement learning. When it comes to the synchronous setting (such that independent samples for all state-action pairs are drawn from a generative model in each iteration), substantial progress has been made towards understanding the sample efficiency of Q-learning. Consider a γ-discounted infinite-horizon MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$: to yield an entrywise ε-approximation of the optimal Q-function, state-of-the-art theory for Q-learning requires a sample size exceeding the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$, which fails to match existing minimax lower bounds. This gives rise to natural questions: what is the sharp sample complexity of Q-learning? Is Q-learning provably sub-optimal? This paper addresses these questions for the synchronous setting: (1) when $|\mathcal{A}| = 1$ (so that Q-learning reduces to TD learning), we prove that the sample complexity of TD learning is minimax optimal and scales as $\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}$ (up to log factor); (2) when $|\mathcal{A}| \geq 2$, we settle the sample complexity of Q-learning to be on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ (up to log factor). Our theory unveils the strict sub-optimality of Q-learning when $|\mathcal{A}| \geq 2$, and rigorizes the negative impact of over-estimation in Q-learning. Finally, we extend our analysis to accommodate asynchronous Q-learning (i.e., the case with Markovian samples), sharpening the horizon dependency of its sample complexity to be $\frac{1}{(1-\gamma)^4}$.

Keywords: Q-learning, temporal difference learning, effective horizon, sample complexity, minimax optimality, lower bound, over-estimation

Contents

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
†Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA.
‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

1 Introduction

Q-learning is arguably one of the most widely adopted model-free algorithms (Watkins and Dayan, 1992; Watkins, 1989). Characterizing its sample efficiency lies at the core of the statistical foundation of reinforcement learning (RL) (Sutton and Barto, 2018). While classical convergence analyses for Q-learning (Borkar and Meyn, 2000; Jaakkola et al., 1994; Szepesvári, 1998; Tsitsiklis, 1994) focused primarily on the asymptotic regime — in which the number of iterations tends to infinity with other problem parameters held fixed — recent years have witnessed a paradigm shift from asymptotic analyses towards a finite-sample / finite-time framework (Beck and Srikant, 2012; Chen et al., 2020, 2021; Even-Dar and Mansour, 2003; Kearns and Singh, 1999; Lee and He, 2018; Li et al., 2021b; Qu and Wierman, 2020; Wainwright, 2019b; Weng et al., 2020a; Xiong et al., 2020). Drawing insights from high-dimensional statistics (Wainwright, 2019a), a modern non-asymptotic framework unveils more clear and informative impacts of salient problem parameters upon the sample complexity, particularly for those applications with enormous state/action space and long horizon. Motivated by its practical value, a suite of non-asymptotic theory has been recently developed for Q-learning to accommodate multiple sampling mechanisms (Beck and Srikant, 2012; Even-Dar and Mansour, 2003; Jin et al., 2018; Li et al., 2021b; Qu and Wierman, 2020; Wainwright, 2019b).

In this paper, we revisit the sample complexity of Q-learning for tabular Markov decision processes (MDPs). For concreteness, let us consider the synchronous setting, which assumes access to a generative model or a simulator that produces independent samples for all state-action pairs in each iteration (Kakade, 2003; Kearns et al., 2002); this setting is termed "synchronous" as the estimates w.r.t. all state-action pairs are updated at once. We investigate the ℓ_∞-based sample complexity, namely, the number of samples needed for synchronous Q-learning to yield an entrywise ε-accurate estimate of the optimal Q-function. Despite a number of prior works tackling this setting, the dependence of the sample complexity on the effective horizon $\frac{1}{1-\gamma}$ remains unsettled. Take γ-discounted infinite-horizon MDPs for instance: the state-of-the-art sample

paper	learning rates	sample complexity
Even-Dar and Mansour (2003)	linear: $\frac{1}{t}$	$2^{\frac{1}{1-\gamma}}\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
Even-Dar and Mansour (2003)	polynomial: $\frac{1}{t^\omega}$, $\omega\in(1/2,1)$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\Big\{\big(\frac{1}{(1-\gamma)^4\varepsilon^2}\big)^{1/\omega}+\big(\frac{1}{1-\gamma}\big)^{\frac{1}{1-\omega}}\Big\}$
Beck and Srikant (2012)	constant: $\frac{(1-\gamma)^4\varepsilon^2}{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}$	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright (2019b)	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Wainwright (2019b)	polynomial: $\frac{1}{t^\omega}$, $\omega\in(0,1)$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\Big\{\big(\frac{1}{(1-\gamma)^4\varepsilon^2}\big)^{1/\omega}+\big(\frac{1}{1-\gamma}\big)^{\frac{1}{1-\omega}}\Big\}$
Chen et al. (2020)	rescaled linear: $\frac{1}{\frac{1}{(1-\gamma)^2}+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Chen et al. (2020)	constant: $(1-\gamma)^4\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
this work (Q-learning, $\lvert\mathcal{A}\rvert\geq 2$)	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
this work (Q-learning, $\lvert\mathcal{A}\rvert\geq 2$)	constant: $(1-\gamma)^3\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
this work (TD learning, $\lvert\mathcal{A}\rvert=1$)	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^3\varepsilon^2}$
this work (TD learning, $\lvert\mathcal{A}\rvert=1$)	constant: $(1-\gamma)^3\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^3\varepsilon^2}$

Table 1: Comparisons of existing sample complexity upper bounds of *synchronous* Q-learning and TD learning for an infinite-horizon γ-discounted MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, where $0<\varepsilon<1$ is the target accuracy level. Here, sample complexity refers to the total number of samples needed to yield either $\max_{s,a}|\widehat{Q}(s,a)-Q^\star(s,a)|\leq\varepsilon$ with high probability or $\mathbb{E}\big[\max_{s,a}|\widehat{Q}(s,a)-Q^\star(s,a)|\big]\leq\varepsilon$, where $\widehat{Q}$ is the estimate returned by Q-learning. All logarithmic factors are omitted in the table to simplify the expressions.

complexity bounds (Chen et al., 2020; Wainwright, 2019b) scale on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ (up to some log factor), where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and the action space, respectively. However, it is unclear whether this scaling is sharp for Q-learning, and whether it can be further improved via a more refined theory. On the one hand, the minimax lower limit for this setting has been shown to be on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ (up to some log factor) (Azar et al., 2013); this limit is achievable by model-based approaches (Agarwal et al., 2020; Li et al., 2020) and apparently smaller than prior sample complexity bounds for Q-learning. On the other hand, Wainwright (2019c) argued through numerical experiments that "*the usual Q-learning suffers from at least worst-case fourth-order scaling in the discount complexity* $\frac{1}{1-\gamma}$*, as opposed to the third-order scaling* ...", although no rigorous justification was provided therein. Given the gap between the achievability bounds and lower bounds in the status quo, it is natural to seek answers to the following questions:

What is the tight sample complexity characterization of Q-learning?
How does it compare to the minimax sample complexity limit?

1.1 Main contributions

Focusing on γ-discounted infinite-horizon MDPs with state space $\mathcal{S}$ and action space $\mathcal{A}$, this paper settles the ℓ_∞-based sample complexity of synchronous Q-learning. Here and throughout, the standard notation $f(\cdot)=\widetilde{O}(g(\cdot))$ (resp. $f(\cdot)=\widetilde{\Omega}(g(\cdot))$) means that $f(\cdot)$ is orderwise no larger than (resp. no smaller than) $g(\cdot)$ modulo some logarithmic factors. Our main contributions regarding synchronous Q-learning are summarized below.

- When $|\mathcal{A}|=1$, Q-learning coincides with temporal difference (TD) learning in a Markov reward process.

For any $0 < \varepsilon < 1$, we prove that a total sample size of

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) \tag{1}$$

is sufficient for TD learning to guarantee ε-accuracy in an ℓ_∞ sense; see Theorem 1. This is sharp and minimax optimal (up to some log factor).

- Moving on to the case with $|\mathcal{A}| \geq 2$, we demonstrate that a sample size of

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) \tag{2}$$

suffices for Q-learning to yield ε-accuracy in an ℓ_∞ sense for any $0 < \varepsilon < 1$; see Theorem 2. Conversely, we construct a hard MDP instance with 4 states and 2 actions, for which Q-learning provably requires at least

$$\widetilde{\Omega}\Big(\frac{1}{(1-\gamma)^4\varepsilon^2}\Big) \tag{3}$$

iterations to achieve ε-accuracy in an ℓ_∞ sense; see Theorem 3. These two theorems taken collectively lead to the first sharp characterization of the sample complexity of Q-learning, strengthening prior theory (Chen et al., 2020; Wainwright, 2019b) by a factor of $\frac{1}{1-\gamma}$. In addition, the discrepancy between our sharp characterization and the minimax lower bound makes clear that Q-learning is *not* minimax optimal when $|\mathcal{A}| \geq 2$, and is outperformed by, say, the model-based approaches (Agarwal et al., 2020; Li et al., 2020) in terms of the sample efficiency.

Our results cover both rescaled linear and constant learning rates; see Table 1 for more detailed comparisons with previous literature. On the technical side, (i) our analysis for the upper bound relies on a sort of crucial error decompositions and variance control that are previously unexplored, which might shed light on how to pin down the finite-sample efficacy of other variants of Q-learning such as double Q-learning; (ii) the development of our lower bound, which is inspired by Azar et al. (2013); Wainwright (2019c), puts the negative impact of over-estimation on sample efficiency on a rigorous footing.

Finally, we extend our analysis framework to accommodate the asynchronous setting, in which the samples are non-i.i.d. and take the form of a single Markovian trajectory. We show for the first time that the sample complexity of asynchronous Q-learning exhibits a $\frac{1}{(1-\gamma)^4}$ scaling w.r.t. the effective horizon, improving upon the prior state-of-the-art Li et al. (2021b).

1.2 Related works

There is a growing literature dedicated to analyzing the non-asymptotic behavior of value-based model-free RL algorithms in a variety of scenarios. In the discussion below, we subsample the literature and discuss a couple of papers that are the closest to ours.

Finite-sample ℓ_∞-based guarantees for synchronous Q-learning and TD learning. The sample complexities derived in prior literature often rely crucially on the choices of learning rates. Even-Dar and Mansour (2003) studied the sample complexity of Q-learning with linear learning rates $1/t$ or polynomial learning rates $1/t^\omega$, which scales as $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^{2.5}})$ when optimized w.r.t. the effective horizon (attained when $\omega = 4/5$). The resulting sample complexity, however, is sub-optimal in terms of its dependency on not only $\frac{1}{1-\gamma}$ but also the target accuracy level ε. Beck and Srikant (2012) investigated the case of constant learning rates; however, their result suffered from an additional factor of $|\mathcal{S}||\mathcal{A}|$, which could be prohibitively large in practice. More recently, Chen et al. (2020); Wainwright (2019b) further analyzed the sample complexity of Q-learning with either constant learning rates or linearly rescaled learning rates, leading to the state-of-the-art bound $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}\big)$. However, this result remains sub-optimal in terms of its scaling with $\frac{1}{1-\gamma}$. See Table 1 for details. In the special case with $|\mathcal{A}| = 1$, the recent works Khamaru et al. (2020); Mou et al. (2020) developed instance-dependent results for TD learning with Polyak-Ruppert averaging, and studied the local (sub)-optimality of TD learning in a different local minimax framework.

Finite-sample ℓ_∞-based guarantees for asynchronous Q-learning and TD learning. Moving beyond the synchronous model, Beck and Srikant (2012); Chen et al. (2021); Even-Dar and Mansour (2003); Li et al. (2021b); Qu and Wierman (2020); Shah and Xie (2018) developed non-asymptotic convergence guarantees for the asynchronous setting, where the data samples take the form of a single Markovian trajectory (following some behavior policy) and only a single state-action pair is updated in each iteration. A similar scaling of $\widetilde{O}\big(\frac{1}{(1-\gamma)^5}\big)$ also showed up in the state-of-the-art sample complexity bounds for asynchronous Q-learning (Li et al., 2021b), and our theory is the first to sharpen it to $\widetilde{O}\big(\frac{1}{(1-\gamma)^4}\big)$. When it comes to the special case with $|\mathcal{A}| = 1$, the non-asymptotic performance guarantees TD learning with Markovian sample trajectories (assuming that the behavior policy coincides with the target policy) have been recently derived by Bhandari et al. (2021); Mou et al. (2020); Srikant and Ying (2019).

Finite-sample ℓ_∞-based guarantees of other Q-learning variants. With the aim of alleviating the sub-optimal dependency on the effective horizon in vanilla Q-learning and improving sample efficiency, several variants of Q-learning have been proposed and analyzed. Azar et al. (2011) proposed speedy Q-learning, which achieves a sample complexity of $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\big)$ at the expense of doubling the computation and storage complexity. Our result on vanilla Q-learning matches that of speedy Q-learning in an order-wise sense. In addition, Wainwright (2019c) proposed a variance-reduced Q-learning algorithm that is shown to be minimax optimal in the range $\epsilon \in (0, 1)$ with a sample complexity $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\big)$, which was subsequently generalized to the asynchronous setting by Li et al. (2021b). The ℓ_∞ statistical bounds for variance-reduced TD learning have been investigated in Khamaru et al. (2020) for the synchronous setting, and in Li et al. (2021b) for the asynchronous setting. Last but not least, Xiong et al. (2020) established the finite-sample convergence of double Q-learning following the framework of Even-Dar and Mansour (2003); however, it is unclear whether double Q-learning can provably outperform vanilla Q-learning in terms of the sample efficiency.

Others. There are also several other strands of related papers that tackle model-free algorithms but do not pursue ℓ_∞-based non-asymptotic guarantees. For instance, Bhandari et al. (2018); Chen et al. (2019); Doan et al. (2019); Gupta et al. (2019); Lakshminarayanan and Szepesvari (2018); Srikant and Ying (2019); Wu et al. (2020); Xu et al. (2019a,b) developed finite-sample (weighted) ℓ_2 convergence guarantees for several model-free algorithms, which also allow one to accommodate linear function approximation as well as off-policy evaluation. Another line of recent work (Bai et al., 2019; Jin et al., 2018; Li et al., 2021a; Zhang et al., 2020) considered the sample efficiency of Q-learning type algorithms paired with proper exploration strategies (e.g., upper confidence bounds) under the framework of regret analysis. The asymptotic behaviors of some variants of Q-learning, e.g., double Q-learning (Weng et al., 2020b) and relative Q-learning (Devraj and Meyn, 2020) are also studied. The effect of more general function approximation schemes (e.g., certain families of neural network approximations) has been studied in Cai et al. (2019); Fan et al. (2019); Murphy (2005); Wai et al. (2019); Xu and Gu (2020) as well. These are beyond the scope of the present paper.

2 Background and algorithms

This paper concentrates on discounted infinite-horizon MDPs (Bertsekas, 2017). We shall start by introducing some basics of tabular MDPs, followed by a description of both Q-learning and TD learning. Throughout this paper, we denote by $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$ and $\mathcal{A} = \{1, \cdots, |\mathcal{A}|\}$ the state space and the action space of the MDP, respectively, and let $\Delta(\mathcal{S})$ represent the probability simplex over the set $\mathcal{S}$.

Basics of discounted infinite-horizon MDPs. Consider an infinite-horizon MDP as represented by a quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\gamma \in (0, 1)$ indicates the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the probability transition kernel (i.e., $P(s' \mid s, a)$ is the probability of transiting to state s' from a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$), and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ stands for the reward function (i.e., $r(s, a)$ is the immediate reward collected in state $s \in \mathcal{S}$ when action $a \in \mathcal{A}$ is taken). Note that the immediate rewards are assumed to lie within $[0, 1]$ throughout this paper. Moreover, we let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ represent a policy, so that $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$ specifies the (possibly randomized) action selection rule in state s. If π is a deterministic policy, then we denote by $\pi(s)$ the action selected by π in state s.

A common objective in RL is to maximize a sort of long-term rewards called value functions or Q-functions. Specifically, given a policy π, the associated value function and Q-function of π are defined respectively by

$$V^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \,\middle|\, s_0 = s\right]$$

for all $s \in \mathcal{S}$, and

$$Q^{\pi}(s,a) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \,\middle|\, s_0 = s, a_0 = a\right]$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here, $\{(s_k, a_k)\}_{k \geq 0}$ is a trajectory of the MDP induced by the policy π (except a_0 when evaluating the Q-function), and the expectations are evaluated with respect to the randomness of the MDP trajectory. Given that the immediate rewards fall within $[0,1]$, it can be straightforwardly verified that $0 \leq V^{\pi}(s) \leq \frac{1}{1-\gamma}$ and $0 \leq Q^{\pi}(s,a) \leq \frac{1}{1-\gamma}$ for any π and any state-action pair (s,a). The optimal value function $V^{\star}$ and optimal Q-function $Q^{\star}$ are defined respectively as

$$V^{\star}(s) \coloneqq \max_{\pi} V^{\pi}(s), \qquad Q^{\star}(s,a) \coloneqq \max_{\pi} Q^{\pi}(s,a)$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. It is well known that there exists a *deterministic* optimal policy, denoted by $\pi^{\star}$, that attains $V^{\star}(s)$ and $Q^{\star}(s,a)$ simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (Sutton and Barto, 2018).

Algorithms: Q-learning and TD learning (the synchronous setting). The synchronous setting assumes access to a generative model (Kearns and Singh, 1999; Sidford et al., 2018) such that: in each iteration t, we collect an independent sample $s_t(s,a) \sim P(\cdot \,|\, s,a)$ for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

With this sampling model in place, the Q-learning algorithm (Watkins and Dayan, 1992) maintains a Q-function estimate $Q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all $t \geq 0$; in each iteration t, the algorithm updates *all* entries of the Q-function estimate at once via the following update rule

$$Q_t = (1 - \eta_t) Q_{t-1} + \eta_t \mathcal{T}_t(Q_{t-1}). \tag{4}$$

Here, $\eta_t \in (0,1]$ denotes the learning rate or the step size in the t-th iteration, and $\mathcal{T}_t$ denotes the empirical Bellman operator constructed by samples collected in the t-th iteration, i.e.,

$$\mathcal{T}_t(Q)(s,a) \coloneqq r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'), \qquad s_t \equiv s_t(s,a) \sim P(\cdot \,|\, s,a) \tag{5}$$

for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Obviously, $\mathcal{T}_t$ is an unbiased estimate of the celebrated Bellman operator $\mathcal{T}$ given by

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

Note that the optimal Q-function $Q^{\star}$ is the unique fixed point of the Bellman operator (Bellman, 1952), that is, $\mathcal{T}(Q^{\star}) = Q^{\star}$. Viewed in this light, synchronous Q-learning can be interpreted as a stochastic approximation scheme (Robbins and Monro, 1951) aimed at solving this fixed-point equation. Throughout this work, we initialize the algorithm in a way that obeys $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$ for every state-action pair (s,a). In addition, the corresponding value function estimate $V_t : \mathcal{S} \to \mathbb{R}$ in the t-th iteration is defined as

$$\forall s \in \mathcal{S}: \qquad V_t(s) \coloneqq \max_{a \in \mathcal{A}} Q_t(s,a). \tag{6}$$

The complete description of Q-learning is summarized in Algorithm 1.

As it turns out, TD learning (Bhandari et al., 2021; Sutton, 1988; Tsitsiklis and Van Roy, 1997) in the synchronous setting can be viewed as a special instance of Q-learning when the action set $\mathcal{A}$ is a singleton (i.e., $|\mathcal{A}| = 1$). In such a case, the MDP reduces to a Markov reward process (MRP) (Bertsekas, 2017), and we shall abuse the notation to use $P : \mathcal{S} \to \Delta(\mathcal{S})$ to describe the probability transition kernel, and employ

Algorithm 1 Synchronous Q-learning for infinite-horizon discounted MDPs.

inputs: learning rates $\{\eta_t\}$, number of iterations T, discount factor γ, initial estimate Q_0.
for $t = 1, 2, \cdots, T$ **do**
  Draw $s_t(s,a) \sim P(\cdot \,|\, s,a)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  Compute Q_t according to (4) and (5).
end for

Algorithm 2 Synchronous TD learning for infinite-horizon discounted MRPs.

inputs: learning rates $\{\eta_t\}$, number of iterations T, discount factor γ, initial estimate V_0.
for $t = 1, 2, \cdots, T$ **do**
  Draw $s_t(s) \sim P(\cdot \,|\, s)$ for each $s \in \mathcal{S}$.
  Compute V_t according to (7).
end for

$r : \mathcal{S} \to [0,1]$ to represent the reward function (with $r(s)$ indicating the immediate reward gained in state s). The TD learning algorithm maintains an estimate $V_t : \mathcal{S} \to \mathbb{R}$ of the value function in each iteration t,[1] and carries out the following iterative update rule

$$V_t(s) = (1-\eta_t)V_{t-1}(s) + \eta_t\big(r(s) + \gamma V_{t-1}(s_t)\big), \qquad s_t \equiv s_t(s) \sim P(\cdot \,|\, s) \tag{7}$$

for each state $s \in \mathcal{S}$. As before, $\eta_t \in (0,1]$ is the learning rate at time t, the initial estimate $V_0(s)$ is taken to be within $\big[0, \frac{1}{1-\gamma}\big]$, and in each iteration, the samples $\{s_t(s) \,|\, s \in \mathcal{S}\}$ are generated independently. The whole algorithm of TD learning is summarized in Algorithm 2.

Finally, while synchronous Q-learning is the main focal point of this paper, we shall also discuss extension to the asynchronous Q-learning, which we will elaborate on in Section 5.

3 Main results: sample complexity of synchronous Q-learning

With the above backgrounds in place, we are in a position to state formally our main findings in this section, concentrating on the synchronous setting.

3.1 Minimax optimality of TD learning

We start with the special with $|\mathcal{A}| = 1$ and characterize the ℓ_∞-based sample complexity of synchronous TD learning.

Theorem 1. *Consider any $\delta \in (0,1)$, $\varepsilon \in (0,1]$, and $\gamma \in [1/2, 1)$. Suppose that for any $0 \le t \le T$, the learning rates satisfy*

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}} \tag{8a}$$

for some small enough universal constants $c_1 \ge c_2 > 0$. Assume that the total number of iterations T obeys

$$T \ge \frac{c_3\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{(1-\gamma)^3\varepsilon^2} \tag{8b}$$

for some sufficiently large universal constant $c_3 > 0$. If the initialization obeys $0 \le V_0(s) \le \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, then with probability at least $1-\delta$, Algorithm 2 achieves

$$\max_{s \in \mathcal{S}} \big|V_T(s) - V^\star(s)\big| \le \varepsilon.$$

[1] There is no need to maintain additional Q-estimates, as the Q-function and the value function coincide when $|\mathcal{A}| = 1$.

Remark 1. This high-probability bound immediately translates to a mean estimation error guarantee. Recognizing the crude upper bound $\big|V_T(s) - V^\star(s)\big| \le \frac{1}{1-\gamma}$ (see (93) in Section C.1) and taking $\delta \le \varepsilon(1-\gamma)$, we reach

$$\mathbb{E}\Big[\max_s \big|V_T(s) - V^\star(s)\big|\Big] \le \varepsilon(1-\delta) + \delta\frac{1}{1-\gamma} \le 2\varepsilon, \tag{9}$$

provided that $T \ge \frac{c_3(\log^3 T)\big(\log \frac{|\mathcal{S}|T}{\varepsilon(1-\gamma)}\big)}{(1-\gamma)^3\varepsilon^2}$.

Given that each iteration of synchronous TD learning makes use of $|\mathcal{S}|$ samples, Theorem 1 implies that the sample complexity of TD learning is at most

$$\widetilde{O}\bigg(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\bigg) \tag{10}$$

for any target accuracy level $\varepsilon \in (0,1]$. This non-asymptotic result is valid as long as the learning rates are chosen to be either a proper constant or rescaled linear (see (8a)). Compared to a large number of prior works studying the performance of TD learning (Bhandari et al., 2021; Borkar and Meyn, 2000; Chen et al., 2020; Khamaru et al., 2021; Lakshminarayanan and Szepesvari, 2018; Wainwright, 2019b), Theorem 1 strengthens prior results by uncovering an improved scaling (i.e., $\frac{1}{(1-\gamma)^3}$) in the effective horizon. In fact, prior results on plain TD learning were only able to obtain a scaling as $\frac{1}{(1-\gamma)^5}$ (Wainwright, 2019b).

To assess the tightness of the above result, we take a moment to compare it with the minimax lower bound recently established in the context of value function estimation. Specifically, Pananjady and Wainwright (2020, Theorem 2(b)) asserted that no algorithm whatsoever can obtain an entrywise ε approximation of the value function—in a minimax sense—unless the total sample size exceeds

$$\widetilde{\Omega}\bigg(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\bigg). \tag{11}$$

In turn, this taken together with Theorem 1 unveils the minimax optimality of the sample complexity (modulo some logarithmic factor) of TD learning for the synchronous setting. While prior works have demonstrated how to attain the minimax limit (11) using model-based methods or variance-reduced model-free algorithms (e.g., Azar et al. (2013); Khamaru et al. (2021); Li et al. (2020); Pananjady and Wainwright (2020)), our theory provides the first rigorous evidence that plain TD learning alone is already minimax optimal, without the need of Polyak-Ruppert averaging or variance reduction.

3.2 Tight sample complexity and sub-optimality of Q-learning

Next, we move on to the more general case with $|\mathcal{A}| \ge 2$ and study the performance of Q-learning. As it turns out, Q-learning with $|\mathcal{A}| \ge 2$ is considerably more challenging to analyze than the TD learning case, due to the presence of the nonsmooth max operator. Our ℓ_∞-based sample complexity bound for Q-learning is summarized as follows, strengthening the state-of-the-art results.

Theorem 2. *Consider any $\delta \in (0,1)$, $\varepsilon \in (0,1]$, and $\gamma \in [1/2, 1)$. Suppose that for any $0 \le t \le T$, the learning rates satisfy*

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^3 T}} \le \eta_t \le \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^3 T}} \tag{12a}$$

for some small enough universal constants $c_1 \ge c_2 > 0$. Assume that the total number of iterations T obeys

$$T \ge \frac{c_3\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4\varepsilon^2} \tag{12b}$$

for some sufficiently large universal constant $c_3 > 0$. If the initialization obeys $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, then Algorithm 1 achieves

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

with probability at least $1-\delta$.

Remark 2. Repeating exactly the same argument as in Remark 1, one can readily translate this high-probability bound into the following mean estimation error guarantee:

$$\mathbb{E}\Big[\max_{s,a}\big|Q_T(s,a)-Q^\star(s,a)\big|\Big] \le \varepsilon(1-\delta)+\delta\frac{1}{1-\gamma} \le 2\varepsilon, \tag{13}$$

holds as long as $T \ge \frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\varepsilon(1-\gamma)}\right)}{(1-\gamma)^4\varepsilon^2}$.

In a nutshell, Theorem 2 develops a non-asymptotic bound on the iteration complexity of Q-learning in the presence of the synchronous model. A few remarks and implications are in order.

Sample complexity and sharpened dependency on $\frac{1}{1-\gamma}$. Recognizing that $|\mathcal{S}||\mathcal{A}|$ independent samples are drawn in each iteration, we can see from Theorem 2 the following sample complexity bound

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) \tag{14}$$

in order for Q-learning to attain ε-accuracy ($0<\varepsilon<1$) in an entrywise sense. To the best of our knowledge, this is the first result that breaks the $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ barrier that is present in all state-of-the-art analyses for vanilla Q-learning (Beck and Srikant, 2012; Chen et al., 2020; Li et al., 2021b; Qu and Wierman, 2020; Wainwright, 2019b).

Learning rates. Akin to the TD learning case, our result accommodates two commonly adopted learning rate schemes (cf. (12a)): (i) linearly rescaled learning rates $\frac{1}{1+\frac{c_2(1-\gamma)}{\log^2 T}t}$, and (ii) iteration-invariant learning rates $\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}}$ (which depend on the total number of iterations T but not the iteration number t). In particular, when the sample size is $T=\frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4\varepsilon^2}$, the constant learning rates can be taken to be on the order of

$$\eta_t \equiv \widetilde{O}\big((1-\gamma)^3\varepsilon^2\big), \qquad 0\le t\le T,$$

which depends almost solely on the discount factor γ and the target accuracy ε. Interestingly, both learning rate schedules lead to the same ℓ_∞-based sample complexity bound (in an order-wise sense), making them appealing for practical use.

A matching lower bound and sub-optimality. The careful reader might remark that there remains a gap between our sample complexity bound for Q-learning and the minimax lower bound (Azar et al., 2013). More specifically, the minimax lower bound scales on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ and is achievable — up to some logarithmic factor — by the model-based approach and variance-reduced methods (Agarwal et al., 2020; Azar et al., 2013; Li et al., 2020; Wainwright, 2019c). This raises natural questions regarding whether our sample complexity bound can be further improved, and whether there is any intrinsic bottleneck that prevents vanilla Q-learning from attaining optimal performance. To answer these questions, we develop the following algorithm-dependent lower bound, which confirms the sharpness of Theorem 2 and reveals the sub-optimality of vanilla Q-learning.

Theorem 3. *Assume that $3/4\le\gamma<1$ and that $T\ge\frac{c_3}{(1-\gamma)^2}$ for some sufficiently large constant $c_3>0$. Suppose that the initialization is $Q_0\equiv 0$, and that the learning rates are taken to be either (i) $\eta_t=\frac{1}{1+c_\eta(1-\gamma)t}$ for all $t\ge 0$, or (ii) $\eta_t\equiv\eta$ for all $t\ge 0$. There exists a γ-discounted MDP with $|\mathcal{S}|=4$ and $|\mathcal{A}|=2$ such that Algorithm 1 — with any $c_\eta>0$ and any $\eta\in(0,1)$ — obeys*

$$\max_s \mathbb{E}\Big[\big|V_T(s)-V^\star(s)\big|^2\Big] \ge \frac{c_{\mathsf{lb}}}{(1-\gamma)^4 T\log^2 T},$$

where $c_{\mathsf{lb}}>0$ is some universal constant.

As asserted by this theorem, it is impossible for Q-learning to attain ε-accuracy (in the sense that $\max_s \mathbb{E}\big[\big|V_T(s) - V^\star(s)\big|^2\big] \le \varepsilon^2$) unless the number of iterations exceeds the order of

$$\frac{1}{(1-\gamma)^4\varepsilon^2}$$

up to some logarithmic factor. Consequently, the upper bound in Theorem 2 is tight in terms of its dependency on the effective horizon $\frac{1}{1-\gamma}$, which is larger than the minimax limit (Azar et al., 2013) by a factor of $\frac{1}{1-\gamma}$. As a consequence, we need to resort to more sophisticated tricks like variance reduction in order to attain minimax optimality when $|\mathcal{A}| \ge 2$ (Li et al., 2021b; Wainwright, 2019c).

4 Key analysis ideas (the synchronous case)

This section outlines the key ideas for the establishment of our main results of Q-learning for the synchronous case, namely Theorem 2 and Theorem 3. The proof for TD learning is deferred to Appendix C. Before delving into the proof details, we first introduce convenient vector and matrix notation that shall be used frequently.

4.1 Vector and matrix notation

To begin with, for any matrix $\boldsymbol{M}$, the notation $\|\boldsymbol{M}\|_1 := \max_i \sum_j |M_{i,j}|$ is defined as the largest row-wise ℓ_1 norm of $\boldsymbol{M}$. For any vector $\boldsymbol{a} = [a_i]_{i=1}^n \in \mathbb{R}^n$, we define $\sqrt{\cdot}$ and $|\cdot|$ in a coordinate-wise manner, i.e. $\sqrt{\boldsymbol{a}} := [\sqrt{a_i}\,]_{i=1}^n \in \mathbb{R}^n$ and $|\boldsymbol{a}| := [|a_i|]_{i=1}^n \in \mathbb{R}^n$. For a set of vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m \in \mathbb{R}^n$ with $\boldsymbol{a}_k = [a_{k,j}]_{j=1}^n$ $(1 \le k \le m)$, we define the max operator in an entrywise fashion such that $\max_{1\le k\le m} \boldsymbol{a}_k := [\max_k a_{k,j}]_{j=1}^n$. For any vectors $\boldsymbol{a} = [a_i]_{i=1}^n \in \mathbb{R}^n$ and $\boldsymbol{b} = [b_i]_{i=1}^n \in \mathbb{R}^n$, the notation $\boldsymbol{a} \le \boldsymbol{b}$ (resp. $\boldsymbol{a} \ge \boldsymbol{b}$) means $a_i \le b_i$ (resp. $a_i \ge b_i$) for all $1 \le i \le n$. We also let $\boldsymbol{a} \circ \boldsymbol{b} = [a_i b_i]_{i=1}^n$ denote the Hadamard product. In addition, we denote by $\boldsymbol{1}$ (resp. $\boldsymbol{e}_i$) the all-one vector (resp. the i-th standard basis vector), and let $\boldsymbol{I}$ be the identity matrix.

We shall also introduce the matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ to represent the probability transition kernel P, whose (s,a)-th row $\boldsymbol{P}_{s,a}$ is a probability vector representing $P(\cdot \,|\, s,a)$. Additionally, we define the *square* probability transition matrix $\boldsymbol{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ (resp. $\boldsymbol{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$) induced by a *deterministic* policy π over the state-action pairs (resp. states) as follows:

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \qquad \text{and} \qquad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}, \tag{15}$$

where $\boldsymbol{\Pi}^\pi \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ is a projection matrix associated with the deterministic policy π:

$$\boldsymbol{\Pi}^\pi = \begin{pmatrix} \boldsymbol{e}_{\pi(1)}^\top & & & \\ & \boldsymbol{e}_{\pi(2)}^\top & & \\ & & \ddots & \\ & & & \boldsymbol{e}_{\pi(|\mathcal{S}|)}^\top \end{pmatrix} \tag{16}$$

with $\boldsymbol{e}_i$ the i-th standard basis vector. Moreover, for any vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, we define $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as follows:

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V}) \circ (\boldsymbol{P}\boldsymbol{V}). \tag{17}$$

In other words, the (s,a)-th entry of $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V})$ corresponds to the variance $\mathsf{Var}_{s'\sim P(\cdot|s,a)}(V(s'))$ w.r.t. the distribution $P(\cdot \,|\, s,a)$.

Moreover, we use the vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent the reward function r, so that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, the (s,a)-th entry of $\boldsymbol{r}$ is given by $r(s,a)$. Analogously, we shall employ the vectors $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent V^π, $V^\star$, V_t, Q^π, $Q^\star$ and Q_t, respectively. Additionally, we define π_t to be the policy associated with Q_t such that for any state-action pair (s,a),

$$\pi_t(s) = \min\Big\{a' \,\Big|\, Q_t(s,a') = \max_{a''} Q_t(s,a'')\Big\}. \tag{18}$$

In other words, for any $s \in \mathcal{S}$, the policy π_t picks out the smallest indexed action that attains the largest Q-value in the estimate $Q_t(s,\cdot)$. As an immediate consequence, one can easily verify

$$Q_t\big(s,\pi_t(s)\big) = V_t(s) \qquad \text{and} \qquad \boldsymbol{P}\boldsymbol{V}_t = \boldsymbol{P}^{\pi_t}\boldsymbol{Q}_t \geq \boldsymbol{P}^{\pi}\boldsymbol{Q}_t \tag{19}$$

for any π, where $\boldsymbol{P}^{\pi}$ is defined in (15). Further, we introduce a matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ such that

$$\boldsymbol{P}_t\big((s,a),s'\big) := \begin{cases} 1, & \text{if } s' = s_t(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

for any (s,a), which is an empirical transition matrix constructed using samples collected in the t-th iteration.

Finally, let $\mathcal{X} := \big(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \frac{1}{\varepsilon}\big)$. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ or $f(\mathcal{X}) \lesssim g(\mathcal{X})$ (resp. $f(\mathcal{X}) \gtrsim g(\mathcal{X})$) means that there exists a universal constant $C_0 > 0$ such that $|f(\mathcal{X})| \leq C_0|g(\mathcal{X})|$ (resp. $|f(\mathcal{X})| \geq C_0|g(\mathcal{X})|$). The notation $f(\mathcal{X}) \asymp g(\mathcal{X})$ means $f(\mathcal{X}) \lesssim g(\mathcal{X})$ and $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ hold simultaneously. We define $\widetilde{O}(\cdot)$ in the same way as $O(\cdot)$ except that it hides logarithmic factors.

4.2 Proof outline for Theorem 2

We are now positioned to describe how to establish Theorem 2, towards which we first express the Q-learning update rule (4) and (5) using the above matrix notation. As can be easily verified, Q-learning employs the samples in $\boldsymbol{P}_t$ (cf. (20)) to perform the following update

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \tag{21}$$

in the t-th iteration. In the sequel, we denote by

$$\boldsymbol{\Delta}_t := \boldsymbol{Q}_t - \boldsymbol{Q}^\star \tag{22}$$

the error of the Q-function estimate in the t-th iteration.

4.2.1 Basic decomposition

We start by decomposing the estimation error term $\boldsymbol{\Delta}_t$. In view of the update rule (21), we arrive at the following elementary decomposition:

$$\begin{aligned}
\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\
&= (1-\eta_t)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\},
\end{aligned} \tag{23}$$

where the third line exploits the Bellman equation $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$. Further, the term $\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star)$ can be linked with $\boldsymbol{\Delta}_{t-1}$ using the definition (18) of π_t as follows

$$\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \leq \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1}, \tag{24a}$$

$$\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \geq \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1}, \tag{24b}$$

where we have made use of the relation (19). Substitute (24) into (23) to reach

$$\begin{aligned}
\boldsymbol{\Delta}_t &\leq (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}; \\
\boldsymbol{\Delta}_t &\geq (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}.
\end{aligned} \tag{25}$$

Applying these relations recursively, we obtain

$$\begin{aligned}
\boldsymbol{\Delta}_t &\leq \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\}, \\
\boldsymbol{\Delta}_t &\geq \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\},
\end{aligned} \tag{26}$$

where we define

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^{t}(1-\eta_j), & \text{if } i=0, \\ \eta_i \prod_{j=i+1}^{t}(1-\eta_j), & \text{if } 0<i<t, \\ \eta_t, & \text{if } i=t. \end{cases} \tag{27}$$

Comparisons to prior approaches. We take a moment to discuss how prior analyses handle the above elementary decomposition. Several prior works (e.g., Li et al. (2021b); Wainwright (2019b)) tackled the second term on the right-hand side of the relation (25) via the following crude bounds:

$$\begin{aligned} \boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} &\le \left\|\boldsymbol{P}^{\pi_{i-1}}\right\|_1 \|\boldsymbol{\Delta}_{i-1}\|_\infty \mathbf{1} = \|\boldsymbol{\Delta}_{i-1}\|_\infty \mathbf{1}, \\ \boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{i-1} &\ge -\left\|\boldsymbol{P}^{\pi^\star}\right\|_1 \|\boldsymbol{\Delta}_{i-1}\|_\infty \mathbf{1} = -\|\boldsymbol{\Delta}_{i-1}\|_\infty \mathbf{1}, \end{aligned}$$

which, however, are too loose when characterizing the dependency on $\frac{1}{1-\gamma}$. By contrast, expanding terms recursively without the above type of crude bounding and carefully analyzing the aggregate terms (e.g., $\sum_{i=1}^{t}\eta_i^{(t)}\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}$) play a major role in sharpening the dependence of sample complexity on the effective horizon.

4.2.2 Key intertwined relations underlying $\{\|\boldsymbol{\Delta}_t\|_\infty\}$

By exploiting the crucial relations (26) derived above, we proceed to upper and lower bound $\boldsymbol{\Delta}_t$ separately. To be more specific, defining

$$\beta := \frac{c_4(1-\gamma)}{\log T} \tag{28}$$

for some constant $c_4 > 0$, one can further decompose the upper bound in (26) into several terms:

$$\boldsymbol{\Delta}_t \le \underbrace{\eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{(1-\beta)t}\eta_i^{(t)}\gamma\big(\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big)}_{=:\boldsymbol{\zeta}_t} \tag{29}$$

$$+ \underbrace{\sum_{i=(1-\beta)t+1}^{t}\eta_i^{(t)}\gamma(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}}_{=:\boldsymbol{\xi}_t} + \sum_{i=(1-\beta)t+1}^{t}\eta_i^{(t)}\gamma\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}. \tag{30}$$

Let us briefly remark on the effect of the first two terms:

- Each component in the first term $\boldsymbol{\zeta}_t$ is fairly small, given that $\eta_i^{(t)}$ is sufficiently small for any $i \le (1-\beta)t$ (meaning that each component has undergone contraction — the ones taking the form of $1-\eta_j$ — for sufficiently many times). As a result, the influence of $\boldsymbol{\zeta}_t$ becomes somewhat negligible.
- The second term $\boldsymbol{\xi}_t$, which can be controlled via Freedman's inequality (Freedman, 1975) due to its martingale structure, contributes to the main variance term in the above recursion. Note, however, that the resulting variance term also depends on $\{\boldsymbol{\Delta}_i\}$.

In summary, the right-hand side of the above inequality can be further decomposed into some weighted superposition of $\{\boldsymbol{\Delta}_i\}$ in addition to some negligible effect. This is formalized in the following two lemmas, which make apparent the key intertwined relations underlying $\{\boldsymbol{\Delta}_i\}$.

Lemma 1. *Suppose that $c_1 c_2 \le c_4/8$. With probability at least $1-\delta$,*

$$\boldsymbol{\Delta}_t \le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1 + \max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}$$

holds simultaneously for all $t \ge \frac{T}{c_2\log T}$.

Lemma 2. *Suppose that $c_1 c_2 \le c_4/8$. With probability at least $1-\delta$,*

$$\boldsymbol{\Delta}_t \ge -30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}$$

holds simultaneously for all $t \ge \frac{T}{c_2 \log T}$.

Proof. The proofs of Lemma 1 and Lemma 2 are deferred to Appendices B.2 and B.3, respectively. As a remark, our analysis collects all the error terms accrued through the iterations — instead of bounding them individually — by conducting a high-order nonlinear expansion of the estimation error through recursion, followed by careful control of the main variance term leveraging the structure of the discounted MDP. □

Putting the preceding bounds in Lemmas 1 and 2 together, we arrive at

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)} \tag{31}$$

for all $t \ge \frac{T}{c_2 \log T}$ with probability exceeding $1-2\delta$, which forms the crux of our analysis. Employing elementary analysis tailored to the above recursive relation, one can demonstrate that

$$\|\boldsymbol{\Delta}_T\|_\infty \le O\bigg(\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\bigg) \tag{32}$$

with probability at least $1-2\delta$, which in turn allows us to establish the advertised result under the assumed sample size condition. The details are deferred to Appendix B.4.

4.3 Proof outline for Theorem 3

Construction of a hard instance and its property Let us construct an MDP $\mathcal{M}_{\mathsf{hard}}$ with state space $\mathcal{S}=\{0,1,2,3\}$ (see a pictorial illustration in Figure 4.3). We shall denote by $\mathcal{A}_s$ the action space associated with state s. The probability transition kernel and reward function of $\mathcal{M}_{\mathsf{hard}}$ are specified as follows

$$\mathcal{A}_0=\{1\}, \qquad P(0\,|\,0,1)=1, \qquad r(0,1)=0, \tag{33a}$$
$$\mathcal{A}_1=\{1,2\}, \qquad P(1\,|\,1,1)=p, \quad P(0\,|\,1,1)=1-p, \quad r(1,1)=1, \tag{33b}$$
$$P(1\,|\,1,2)=p, \quad P(0\,|\,1,2)=1-p, \quad r(1,2)=1, \tag{33c}$$
$$\mathcal{A}_2=\{1\}, \qquad P(2\,|\,2,1)=p, \quad P(0\,|\,2,1)=1-p, \quad r(2,1)=1, \tag{33d}$$
$$\mathcal{A}_3=\{1\}, \qquad P(3\,|\,3,1)=1, \qquad r(3,1)=1, \tag{33e}$$

where the parameter p is taken to be

$$p=\frac{4\gamma-1}{3\gamma}. \tag{34}$$

Before moving forward to analyze the behavior of Q-learning, we first characterize the optimal value function and Q-function of this MDP; the proof is postponed to Section D.4.

Lemma 3. *Consider the MDP $\mathcal{M}_{\mathsf{hard}}$ constructed in (33). One has*

$$V^\star(0)=Q^\star(0,1)=0; \tag{35a}$$
$$V^\star(1)=Q^\star(1,1)=Q^\star(1,2)=V^\star(2)=Q^\star(2,1)=\frac{1}{1-\gamma p}=\frac{3}{4(1-\gamma)}; \tag{35b}$$
$$V^\star(3)=Q^\star(3,1)=\frac{1}{1-\gamma}. \tag{35c}$$

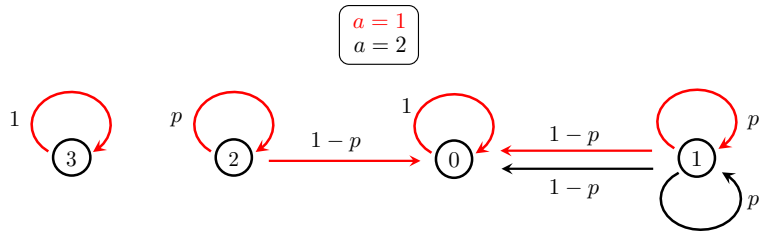

Figure 1: The constructed hard MDP instance used in the analysis of Theorem 3, where $p = \frac{4\gamma-1}{3\gamma}$ and the specifications are described in (33).

Recognizing the elementary decomposition

$$\mathbb{E}\left[\left(V^\star(s) - V_T(s)\right)^2\right] = \left(\mathbb{E}\left[V^\star(s) - V_T(s)\right]\right)^2 + \mathsf{Var}\big(V_T(s)\big) \tag{36}$$

for any state s, our proof consists of lower bounding either the squared bias term $\big(\mathbb{E}[V^\star(s) - V_T(s)]\big)^2$ or the variance term $\mathsf{Var}\big(V_T(s)\big)$. In short, we shall primarily analyze the dynamics w.r.t. state 2 to handle the case when the learning rates are either too small or too large, and analyze the dynamics w.r.t. state 1 to cope with the case with medium learning rates (with state 3 serving as a helper state to simplify the analysis). The latter case — corresponding to the learning rates adopted in establishing the upper bounds — is the most challenging: critically, from state 1 the agent can take one of two identical actions, whose value tends to be estimated with a high positive bias due to maximizing over the empirical state-action values, highlighting the well-recognized "over-estimation" issue of Q-learning in practice (Hasselt, 2010). The complete proof is deferred to Appendix D.

5 Extension: sample complexity of asynchronous Q-learning

Moving beyond the synchronous setting, another scenario of practical importance is the case where the acquired samples take the form of a single Markovian trajectory (Tsitsiklis, 1994). In this section, we extend our analysis framework for synchronous Q-learning to accommodate Markovian non-i.i.d. samples.

5.1 Markovian samples and asynchronous Q-learning

Markovian sample trajectory. Suppose that we obtain a Markovian sample trajectory $\{(s_t, a_t, r_t)\}_{t=0}^{\infty}$, which is generated by the MDP of interest when a stationary behavior policy π_{b} is employed; in other words,

$$a_t \sim \pi_{\mathsf{b}}(\cdot \,|\, s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot \,|\, s_t, a_t), \qquad t \geq 0. \tag{37}$$

When π_{b} is stationary, the trajectory $\{(s_t, a_t)\}_{t=0}^{\infty}$ can be viewed as a sample path of a time-homogeneous Markov chain; in what follows, we shall denote by $\mu_{\pi_{\mathsf{b}}}$ the stationary distribution of this Markov chain. Note that the behavior policy π_{b} can often be quite different from the target optimal policy $\pi^\star$.

Asynchronous Q-learning. In the presence of a single Markovian sample trajectory, the Q-learning algorithm implements the following iterative update rule

$$Q_t(s_{t-1}, a_{t-1}) = (1 - \eta_t) Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \Big\{ r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t, a') \Big\}, \tag{38a}$$

$$Q_t(s, a) = Q_{t-1}(s, a) \qquad \text{for all } (s, a) \neq (s_{t-1}, a_{t-1}) \tag{38b}$$

for all $t \geq 1$, where $0 < \eta_t \leq 1$ stands for the learning rate at time t. It is often referred to as *asynchronous Q-learning*, as only a single state-action pair is updated in each iteration (in contrast, synchronous Q-learning updates all state-action pairs simultaneously in each iteration). This also leads to the following estimate for the value function at time t:

$$V_t(s) := \max_{a \in \mathcal{A}} Q_t(s, a) \qquad \text{for all } s \in \mathcal{S}. \tag{39}$$

As can be expected, the presence of Markovian non-i.i.d. data considerably complicates the analysis for asynchronous Q-learning.

Assumptions. In order to ensure sufficient coverage of the sample trajectory over the state/action space, we make the following assumption throughout this section, which is also commonly imposed in prior literature.

Assumption 1. *The Markov chain induced by the behavior policy π_{b} is uniformly ergodic.*[2]

In addition, there are two crucial quantities concerning the sample trajectory that dictate the performance of asynchronous Q-learning. The first one is the minimum state-action occupancy probability of the sample trajectory, defined formally as

$$\mu_{\mathsf{min}} \coloneqq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_{\pi_{\mathsf{b}}}(s,a). \tag{40}$$

This metric captures the information bottleneck incurred by the least visited state-action pair. The second key quantity is the mixing time associated with the sample trajectory, denoted by

$$t_{\mathsf{mix}} \coloneqq \min\left\{t \,\middle|\, \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mathsf{TV}}\big(P^t(\cdot\,|\,s,a), \mu_{\pi_{\mathsf{b}}}\big) \le \frac{1}{4}\right\}. \tag{41}$$

Here, $d_{\mathsf{TV}}(\mu,\nu) \coloneqq \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|$ indicates the total variation distance between two measures μ and ν over $\mathcal{X}$ (Tsybakov and Zaiats, 2009), whereas $P^t(\cdot\,|\,s,a)$ stands for the distribution of (s_t,a_t) when the sample trajectory is initialized at $(s_0,a_0)=(s,a)$. In words, the mixing time reflects the time required for the Markov chain to become nearly independent of the initial states. See Li et al. (2021b, Section 2) for a more detailed account of these quantities and assumptions.

5.2 Sample complexity of asynchronous Q-learning

While a number of previous works have been dedicated to understanding the performance of asynchronous Q-learning, its sample complexity bound remains loose when it comes to the dependency on the effective horizon $\frac{1}{1-\gamma}$. Encouragingly, the analysis framework laid out in this paper allows us to tighten the dependency on $\frac{1}{1-\gamma}$, as stated below.

Theorem 4. *Consider any $\delta\in(0,1)$, $\varepsilon\in(0,1]$, and $\gamma\in[1/2,1)$. Suppose that for any $0\le t\le T$, the learning rates satisfy*

$$\eta_t \equiv \eta = \frac{c_1\log^3 T}{(1-\gamma)T\mu_{\mathsf{min}}} \tag{42a}$$

for some universal constants $0<c_1\le 1$. Assume that the total number of iterations T obeys

$$T \ge \frac{c_2\log^2\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\mu_{\mathsf{min}}}\max\left\{\frac{\log^3 T}{(1-\gamma)^4\varepsilon^2}, \frac{t_{\mathsf{mix}}}{1-\gamma}\right\} \tag{42b}$$

for some sufficiently large universal constant $c_2>0$. If the initialization obeys $0\le Q_0(s,a)\le\frac{1}{1-\gamma}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, then asynchronous Q-learning (cf. (38)) satisfies

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

with probability at least $1-\delta$.

Remark 3. Similar to Remark 1 and Remark 2, one can immediately translate the above high-probability result into the following mean estimation error bound:

$$\mathbb{E}\Big[\max_{s,a}\big|Q_T(s,a)-Q^\star(s,a)\big|\Big] \le \varepsilon(1-\delta)+\delta\frac{1}{1-\gamma} \le 2\varepsilon, \tag{43}$$

which holds as long as $T\ge \frac{c_2\log^2\frac{|\mathcal{S}||\mathcal{A}|T}{\varepsilon(1-\gamma)}}{\mu_{\mathsf{min}}}\max\left\{\frac{\log^3 T}{(1-\gamma)^4\varepsilon^2}, \frac{t_{\mathsf{mix}}}{1-\gamma}\right\}$ for some large enough constant $c_2>0$.

[2] See Paulin (2015, Section 1.2) for the definition of uniform ergodicity.

This theorem demonstrates that with high probability, the total sample size needed for asynchronous Q-learning to yield entrywise ε accuracy is

$$\widetilde{O}\bigg(\frac{1}{\mu_{\min}(1-\gamma)^4\varepsilon^2}+\frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}\bigg), \tag{44}$$

provided that the learning rates are taken to be some proper constant (see (42a)). The first term in (44) resembles our sample complexity characterization of synchronous Q-learning (cf. (14)), except that we replace the number $|\mathcal{S}||\mathcal{A}|$ of state-action pairs in (14) with $1/\mu_{\min}$ in order to account for non-uniformity across state-action pairs. The second term in (44) is nearly independent of the target accuracy (except for some logarithmic scaling), and can be viewed as the burn-in time taken for asynchronous Q-learning to mimic synchronous Q-learning despite Markovian data.

We now pause to compare Theorem 4 with prior non-asymptotic theory for asynchronous Q-learning. As far as we know, all existing sample complexity bounds (Beck and Srikant, 2012; Chen et al., 2021; Even-Dar and Mansour, 2003; Li et al., 2021b; Qu and Wierman, 2020) scale at least as $\frac{1}{(1-\gamma)^5}$ in terms of the dependency on the effective horizon, with Theorem 4 being the first result to sharpen this dependency to $\frac{1}{(1-\gamma)^4}$. In particular, our sample complexity bound strengthens the state-of-the-art result Li et al. (2021b) by a factor up to $\frac{1}{1-\gamma}$, while improving upon Qu and Wierman (2020) by a factor of at least $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}\min\big\{t_{\mathsf{mix}}, \frac{1}{(1-\gamma)^3\varepsilon^2}\big\}$.[3]

6 Concluding remarks

In this paper, we have settled the sample complexity of synchronous Q-learning in γ-discounted infinite-horizon MDPs, which is shown to be on the order of $\widetilde{O}\big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\big)$ when $|\mathcal{A}|=1$ and $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\big)$ when $|\mathcal{A}|\geq 2$. A matching lower bound has been developed when $|\mathcal{A}|\geq 2$ through studying the dynamics of Q-learning on a hard MDP instance, which unveils the negative impact of an inevitable over-estimation issue. Our theory has been further extended to accommodate asynchronous Q-learning, resulting in tight dependency of the sample complexity on the effective horizon. The analysis framework developed herein—which exploits novel error decompositions and variance control that differ substantially from prior approaches—might suggest a plausible path towards sharpening the sample complexity of, as well as understanding the algorithmic bottlenecks for, other model-free algorithms (e.g., double Q-learning (Hasselt, 2010)).

Acknowledgements

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773. Y. Wei is supported in part by the the NSF grants CCF-2106778, DMS-2147546/2015447 and CAREER award DMS-2143215. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, the NSF grants CCF-1806154 and CCF-2007911. The authors are grateful to Laixi Shi for helpful discussions about the lower bound, and thank Shaocong Ma for pointing out some errors in an early version of this work. Part of this work was done while G. Li, Y. Chen and Y. Wei were visiting the Simons Institute for the Theory of Computing.

A Freedman's inequality

The analysis of this work relies heavily on Freedman's inequality (Freedman, 1975), which is an extension of the Bernstein inequality and allows one to establish concentration results for martingales. For ease of presentation, we include a user-friendly version of Freedman's inequality as follows.

[3] The sample complexity of Li et al. (2021b) scales as $\widetilde{O}\big(\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2}+\frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}\big)$, while the sample complexity of Qu and Wierman (2020) scales as $\widetilde{O}\big(\frac{t_{\mathsf{mix}}}{\mu_{\min}^2(1-\gamma)^5\varepsilon^2}\big)$. It is worth noting that $1/\mu_{\min}\geq|\mathcal{S}||\mathcal{A}|$ and is therefore a large factor.

Theorem 5. *Suppose that* $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, *where* $\{X_k\}$ *is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}\left[X_k \mid \{X_j\}_{j:j<k}\right] = 0 \qquad \text{for all } k \ge 1.$$

Define

$$W_n := \sum_{k=1}^n \mathbb{E}_{k-1}\left[X_k^2\right],$$

where we write $\mathbb{E}_{k-1}$ *for the expectation conditional on* $\{X_j\}_{j:j<k}$. *Then for any given* $\sigma^2 \ge 0$, *one has*

$$\mathbb{P}\left\{|Y_n| \ge \tau \text{ and } W_n \le \sigma^2\right\} \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{45}$$

In addition, suppose that $W_n \le \sigma^2$ *holds deterministically. For any positive integer* $K \ge 1$, *with probability at least* $1-\delta$ *one has*

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}. \tag{46}$$

Proof. See Freedman (1975); Tropp (2011) for the proof of (45). As an immediate consequence of (45), one has

$$\mathbb{P}\left\{|Y_n| \ge \sqrt{4\sigma^2\log\frac{2}{\delta}} + \frac{4}{3}R\log\frac{2}{\delta} \text{ and } W_n \le \sigma^2\right\} \le \delta. \tag{47}$$

Next, we turn attention to (46). Consider any positive integer K. As can be easily seen, the event

$$\mathcal{H}_K := \left\{|Y_n| \ge \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}\right\}$$

is contained within the union of the following K events

$$\mathcal{H}_K \subseteq \bigcup_{0\le k<K} \mathcal{B}_k,$$

where we define

$$\mathcal{B}_k := \left\{|Y_n| \ge \sqrt{\frac{4\sigma^2}{2^{k-1}}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta} \text{ and } \frac{\sigma^2}{2^k} \le W_n \le \frac{\sigma^2}{2^{k-1}}\right\}, \qquad 1 \le k \le K-1,$$

$$\mathcal{B}_0 := \left\{|Y_n| \ge \sqrt{\frac{4\sigma^2}{2^{K-1}}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta} \text{ and } W_n \le \frac{\sigma^2}{2^{K-1}}\right\}.$$

Invoking inequality (47) with σ^2 set to be $\frac{\sigma^2}{2^{k-1}}$ and δ set to be $\frac{\delta}{K}$, we arrive at $\mathbb{P}\{\mathcal{B}_k\} \le \delta/K$. Taken this fact together with the union bound gives

$$\mathbb{P}\{\mathcal{H}_K\} \le \sum_{k=0}^{K-1}\mathbb{P}\{\mathcal{B}_k\} \le \delta.$$

This concludes the proof. □

B Upper bounds for Q-learning (Theorem 2)

In this section, we fill in the details for the proof idea outlined in Section 4.2 for synchronous Q-learning. In fact, our proof strategy leads to a more general version that accounts for the full ε-range $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$, as stated below.

Theorem 6. *Consider any $\gamma \in (0,1)$ and any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Theorem 2 continues to hold if*

$$T \geq \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 \min\{\varepsilon^2, \varepsilon\}} \tag{48}$$

for some large enough universal constant $c_3 > 0$.

Remark 4. Clearly, Theorem 6 subsumes Theorem 2 as a special case.

As one can anticipate, the proof of Theorem 6 for Q-learning includes many key ingredients for establishing Theorem 1 for TD learning. We will elaborate on how to modify the proof argument to establish Theorem 1 in Section C.

B.1 Preliminaries

To begin with, we gather a few elementary facts that shall be used multiple times in the proof.

Ranges of $\boldsymbol{Q}_t$ and $\boldsymbol{V}_t$. When properly initialized, the Q-function estimates and the value function estimates always fall within a suitable range, as asserted by the following lemma.

Lemma 4. *Suppose that $0 \leq \eta_t \leq 1$ for all $t \geq 0$. Assume that $\mathbf{0} \leq \boldsymbol{Q}_0 \leq \frac{1}{1-\gamma}\mathbf{1}$. Then for any $t \geq 0$,*

$$\mathbf{0} \leq \boldsymbol{Q}_t \leq \frac{1}{1-\gamma}\mathbf{1} \qquad \text{and} \qquad \mathbf{0} \leq \boldsymbol{V}_t \leq \frac{1}{1-\gamma}\mathbf{1}. \tag{49}$$

Proof. We shall prove this by induction. First, our initialization trivially obeys (49) for $t = 0$. Next, suppose that (49) is true for the $(t-1)$-th iteration, namely,

$$\mathbf{0} \leq \boldsymbol{Q}_{t-1} \leq \frac{1}{1-\gamma}\mathbf{1} \qquad \text{and} \qquad \mathbf{0} \leq \boldsymbol{V}_{t-1} \leq \frac{1}{1-\gamma}\mathbf{1}, \tag{50}$$

and we intend to justify the claim for the t-th iteration. Recognizing that $\mathbf{0} \leq \boldsymbol{r} \leq \mathbf{1}$, $\boldsymbol{P}_t \geq \mathbf{0}$ and $\|\boldsymbol{P}_t\|_1 = 1$, one can straightforwardly see from the update rule (21) and the induction hypothesis (50) that

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \geq \mathbf{0}$$

and

$$\begin{aligned}
\boldsymbol{Q}_t &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \\
&\leq (1-\eta_t)\,\|\boldsymbol{Q}_{t-1}\|_\infty \mathbf{1} + \eta_t\big(\|\boldsymbol{r}\|_\infty + \gamma \|\boldsymbol{P}_t\|_1 \|\boldsymbol{V}_{t-1}\|_\infty\big)\mathbf{1} \\
&\leq (1-\eta_t)\frac{1}{1-\gamma}\mathbf{1} + \eta_t\Big(1 + \frac{\gamma}{1-\gamma}\Big)\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}.
\end{aligned}$$

In addition, from the definition $V_t(s) := \max_a Q_t(s,a)$ for all $t \geq 0$ and all $s \in \mathcal{S}$, it is easily seen that

$$\mathbf{0} \leq \boldsymbol{V}_t \leq \frac{1}{1-\gamma}\mathbf{1},$$

thus establishing (49) for the t-th iteration. Applying the induction argument then concludes the proof. □

As a result of Lemma 4 and the fact $\mathbf{0} \leq \boldsymbol{Q}^\star \leq \frac{1}{1-\gamma}\mathbf{1}$, we have

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \leq \frac{1}{1-\gamma} \qquad \text{for all } t \geq 0, \tag{51}$$

which also confirms that $0 \leq \varepsilon \leq \frac{1}{1-\gamma}$ is the full ε-range we need to consider. Further, we make note of a direct consequence of the claimed iteration number (48) when $\varepsilon \leq \frac{1}{1-\gamma}$:

$$T = \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 \min\{\varepsilon, \varepsilon^2\}} \geq \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^3}, \tag{52}$$

which will be useful for subsequent analysis.

Several facts regarding the learning rates. Next, we gather a couple of useful bounds regarding the learning rates $\{\eta_t\}$. To begin with, we find it helpful to introduce the following related quantities introduced previously in (27):

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^t (1-\eta_j), & \text{if } i = 0, \\ \eta_i \prod_{j=i+1}^t (1-\eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t. \end{cases} \tag{53}$$

We now take a moment to bound $\eta_i^{(t)}$. From our assumption (12a) and the condition (52), we know that the learning rate obeys

$$\frac{1}{2c_1(1-\gamma)T/\log^3 T} \le \frac{1}{1+c_1(1-\gamma)T/\log^3 T} \le \eta_t \le \frac{1}{1+c_2(1-\gamma)t/\log^3 T} \le \frac{1}{c_2(1-\gamma)t/\log^3 T} \tag{54}$$

for some constants $c_1, c_2 > 0$. Recalling that

$$\beta := \frac{c_4(1-\gamma)}{\log T} \tag{55}$$

for some universal constant $c_4 > 0$ and considering any t obeying

$$t \ge \frac{T}{c_2 \log T}, \tag{56}$$

we shall bound $\eta_i^{(t)}$ by looking at two cases separately.

- For any $0 \le i \le (1-\beta)t$, we can use (54) to show that

$$\begin{aligned} \eta_i^{(t)} &\le \Big(1 - \frac{1}{2c_1(1-\gamma)T/\log^3 T}\Big)^{\beta t} \le \Big(1 - \frac{1}{2c_1(1-\gamma)T/\log^3 T}\Big)^{\frac{c_4(1-\gamma)T}{c_2 \log^2 T}} \\ &= \left(\Big(1 - \frac{1}{2c_1(1-\gamma)T/\log^3 T}\Big)^{\frac{2c_1(1-\gamma)T}{\log^3 T}}\right)^{\frac{c_4 \log T}{2c_1 c_2}} < \frac{1}{2T^2}, \end{aligned} \tag{57a}$$

  where the last inequality holds as long as $c_1 c_2 \le c_4/8$.

- When it comes to the case with $i > (1-\beta)t \ge t/2$, one can upper bound

$$\eta_i^{(t)} \le \eta_i \le \frac{1}{c_2(1-\gamma)i/\log^3 T} < \frac{2}{c_2(1-\gamma)t/\log^3 T} \le \frac{2\log^4 T}{(1-\gamma)T}, \tag{57b}$$

  where we have used the constraint (56).

Moreover, the sum of $\eta_i^{(t)}$ over i obeys

$$\begin{aligned} \sum_{i=0}^t \eta_i^{(t)} &= \prod_{j=1}^t (1-\eta_j) + \eta_1 \prod_{j=2}^t (1-\eta_j) + \eta_2 \prod_{j=3}^t (1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t \\ &= \prod_{j=2}^t (1-\eta_j) + \eta_2 \prod_{j=3}^t (1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t = \cdots \\ &= (1-\eta_t) + \eta_t = 1. \end{aligned} \tag{58}$$

Repeating the same argument further allows us to derive

$$\sum_{i=\tau}^t \eta_i^{(t)} = 1 - \prod_{j=\tau}^t (1-\eta_j) \tag{59}$$

for any $\tau \le t$.

B.2 Proof of Lemma 1

We shall exploit the relation (30) to prove this lemma. One of the key ingredients of our analysis lies in controlling the terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ introduced in (30), which in turn enables us to apply (30) recursively to control $\boldsymbol{\Delta}_t$.

Step 1: bounding $\boldsymbol{\zeta}_t$. We start by developing an upper bound on $\boldsymbol{\zeta}_t$ (cf. (30)) for any t obeying $\frac{T}{c_2 \log T} \le t \le T$. Invoking the preceding upper bounds (57) on $\eta_i^{(t)}$ implies that

$$\begin{aligned}
\|\boldsymbol{\zeta}_t\|_\infty &\le \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{i \le (1-\beta)t} \eta_i^{(t)} \max_{1 \le i \le (1-\beta)t} \big(\|\boldsymbol{P}^{\pi_{i-1}} \boldsymbol{\Delta}_{i-1}\|_\infty + \|\boldsymbol{P}_i \boldsymbol{V}_{i-1}\|_\infty + \|\boldsymbol{P} \boldsymbol{V}_{i-1}\|_\infty \big) \\
&\le \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{i \le (1-\beta)t} \eta_i^{(t)} \max_{1 \le i \le (1-\beta)t} \Big\{ \|\boldsymbol{P}^{\pi_{i-1}}\|_1 \|\boldsymbol{\Delta}_{i-1}\|_\infty + \big(\|\boldsymbol{P}_i\|_1 + \|\boldsymbol{P}\|_1 \big) \|\boldsymbol{V}_{i-1}\|_\infty \Big\} \\
&\overset{\text{(i)}}{=} \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{i \le (1-\beta)t} \eta_i^{(t)} \max_{1 \le i \le (1-\beta)t} \big(\|\boldsymbol{\Delta}_{i-1}\|_\infty + 2 \|\boldsymbol{V}_{i-1}\|_\infty \big) \\
&\overset{\text{(ii)}}{\le} \frac{1}{2T^2} \cdot \frac{1}{1-\gamma} + \frac{1}{2T^2} \cdot t \cdot \frac{3}{1-\gamma} \\
&\le \frac{2}{(1-\gamma)T}.
\end{aligned}$$

Here, (i) holds since $\|\boldsymbol{P}^{\pi_{i-1}}\|_1 = \|\boldsymbol{P}_i\|_1 = \|\boldsymbol{P}\|_1 = 1$ (as they are all probability transition matrices), whereas (ii) arises from the previous bound (57a).

Step 2: bounding $\boldsymbol{\xi}_t$. Moving on to the term $\boldsymbol{\xi}_t$, let us express it as

$$\boldsymbol{\xi}_t = \sum_{i=(1-\beta)t+1}^{t} \boldsymbol{z}_i \qquad \text{with } \boldsymbol{z}_i \coloneqq \eta_i^{(t)} \gamma (\boldsymbol{P}_i - \boldsymbol{P}) \boldsymbol{V}_{i-1},$$

where the $\boldsymbol{z}_i$'s satisfy

$$\mathbb{E}\left[\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0 \right] = \boldsymbol{0}.$$

This motivates us to invoke Freedman's inequality (see Theorem 5) to control $\boldsymbol{\xi}_t$ for any t obeying $\frac{T}{c_2 \log T} \le t \le T$. Towards this, we need to calculate several quantities.

- First, it is seen that

$$\begin{aligned}
B &\coloneqq \max_{(1-\beta)t < i \le t} \|\boldsymbol{z}_i\|_\infty \le \max_{(1-\beta)t < i \le t} \|\eta_i^{(t)} (\boldsymbol{P}_i - \boldsymbol{P}) \boldsymbol{V}_{i-1}\|_\infty \\
&\le \max_{(1-\beta)t < i \le t} \eta_i^{(t)} \big(\|\boldsymbol{P}_i\|_1 + \|\boldsymbol{P}\|_1 \big) \|\boldsymbol{V}_{i-1}\|_\infty \le \frac{4 \log^4 T}{(1-\gamma)^2 T},
\end{aligned}$$

  where the last inequality is due to (57b), Lemma 4, and the fact $\|\boldsymbol{P}_i\|_1 = \|\boldsymbol{P}\|_1 = 1$.

- Next, we turn to certain variance terms. For any vector $\boldsymbol{a} = [a_j]$, let us use $\mathsf{Var}\big(\boldsymbol{a} \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big)$ to denote a vector whose j-th entry is given by $\mathsf{Var}\big(a_j \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big)$. With this notation in place, and recalling the notation $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{z})$ in (17), we obtain

$$\begin{aligned}
\boldsymbol{W}_t &\coloneqq \sum_{i=(1-\beta)t+1}^{t} \mathsf{Var}\big(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big) = \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \big(\eta_i^{(t)}\big)^2 \mathsf{Var}\Big((\boldsymbol{P}_i - \boldsymbol{P}) \boldsymbol{V}_{i-1} \,|\, \boldsymbol{V}_{i-1} \Big) \\
&= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \big(\eta_i^{(t)}\big)^2 \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_{i-1}\big) \\
&\le \Big(\max_{(1-\beta)t \le i \le t} \eta_i^{(t)} \Big) \Big(\sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)} \Big) \max_{(1-\beta)t \le i < t} \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_i\big)
\end{aligned}$$

$$\leq \frac{2\log^4 T}{(1-\gamma)T} \max_{(1-\beta)t\leq i<t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i), \tag{60}$$

where the last inequality relies on the previous bounds (57b) and (58).

- In the meantime, Theorem 4 leads us to the following trivial upper bound:

$$\big|\boldsymbol{W}_t\big| \leq \frac{2\log^4 T}{(1-\gamma)T} \cdot \frac{1}{(1-\gamma)^2}\mathbf{1} = \frac{2\log^4 T}{(1-\gamma)^3 T}\mathbf{1} =: \sigma^2\mathbf{1}.$$

By setting $K = \Big\lceil 2\log_2 \frac{1}{1-\gamma} \Big\rceil$, one has

$$\frac{\sigma^2}{2^K} \leq \frac{2\log^4 T}{(1-\gamma)T}. \tag{61}$$

With the above bounds in place, applying the Freedman inequality in Theorem 5 and invoking the union bound over all the $|\mathcal{S}||\mathcal{A}|$ entries of $\boldsymbol{\xi}_t$ demonstrate that

$$\begin{aligned}
|\boldsymbol{\xi}_t| &\leq \sqrt{8\Big(\boldsymbol{W}_t + \frac{\sigma^2}{2^K}\mathbf{1}\Big)\log\frac{8|\mathcal{S}||\mathcal{A}|T\log\frac{1}{1-\gamma}}{\delta}} + \Big(\frac{4}{3}B\log\frac{8|\mathcal{S}||\mathcal{A}|T\log\frac{1}{1-\gamma}}{\delta}\Big)\cdot\mathbf{1} \\
&\leq \sqrt{16\Big(\boldsymbol{W}_t + \frac{2\log^4 T}{(1-\gamma)T}\mathbf{1}\Big)\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \Big(3B\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)\cdot\mathbf{1} \\
&\leq \sqrt{\frac{32\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T}\Big(\max_{(1-\beta)t\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big)} + \frac{12\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2 T}\mathbf{1}
\end{aligned}$$

with probability at least $1-\delta/T$. Here, the second line holds due to (61) and the fact $\log\frac{8|\mathcal{S}||\mathcal{A}|T\log\frac{1}{1-\gamma}}{\delta} \leq 2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ (cf. (52)), whereas the last inequality makes use of the relation (60).

Step 3: using the bounds on $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ to control $\boldsymbol{\Delta}_t$. Let us define

$$\boldsymbol{\varphi}_t \coloneqq 64\frac{\log^4 T\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{(1-\gamma)T}\Big(\max_{\frac{t}{2}\leq i\leq t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big) \tag{62}$$

In view of the upper bounds derived in Steps 1 and 2, and β defined in (55), we have — with probability exceeding $1-\delta$ — that

$$|\boldsymbol{\zeta}_k| + |\boldsymbol{\xi}_k| \leq \sqrt{\boldsymbol{\varphi}_t} \qquad \text{for all } \ 2t/3 \leq k \leq t, \tag{63}$$

provided that $T \geq \frac{c_9(\log^4 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^3}$ for some sufficiently large constant $c_9 > 0$. Substituting (63) into (30), we can upper bound $\boldsymbol{\Delta}_t$ as follows

$$\boldsymbol{\Delta}_k \leq \sqrt{\boldsymbol{\varphi}_t} + \sum_{i=(1-\beta)k+1}^{k}\eta_i^{(k)}\gamma\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} = \sqrt{\boldsymbol{\varphi}_t} + \sum_{i=(1-\beta)k}^{k-1}\eta_{i+1}^{(k)}\gamma\boldsymbol{P}^{\pi_i}\boldsymbol{\Delta}_i \qquad \text{for all } \ 2t/3 \leq k \leq t. \tag{64}$$

Further, we find it convenient to define $\big\{\alpha_i^{(t)}\big\}$ as follows

$$\alpha_i^{(t)} := \frac{\eta_{i+1}^{(t)}}{\sum_{j=(1-\beta)t}^{t-1}\eta_{j+1}^{(t)}}. \tag{65}$$

Clearly, this sequence satisfies

$$\alpha_i^{(t)} \geq \eta_{i+1}^{(t)} \qquad \text{and} \qquad \sum_{i=(1-\beta)t}^{t-1}\alpha_i^{(t)} = 1 \tag{66}$$

for any t, where the first inequality results from (58). With these in place, we can write (64) as

$$\boldsymbol{\Delta}_k \le \sqrt{\boldsymbol{\varphi}_t} + \sum_{i_1=(1-\beta)k}^{k-1} \eta_{i_1+1}^{(k)} \gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{\Delta}_{i_1} = \sum_{i_1=(1-\beta)k}^{k-1} \left(\alpha_{i_1}^{(k)} \sqrt{\boldsymbol{\varphi}_t} + \eta_{i_1+1}^{(k)} \gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{\Delta}_{i_1} \right) \qquad \text{for all } 2t/3 \le k \le t. \tag{67}$$

Given that $(1-\beta)t \ge 2t/3$ (see (55)), we can invoke this relation recursively to yield

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le \sum_{i_1=(1-\beta)t}^{t-1} \left(\alpha_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \eta_{i_1+1}^{(t)} \gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{\Delta}_{i_1} \right) \\
&\le \sum_{i_1=(1-\beta)t}^{t-1} \left[\alpha_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \eta_{i_1+1}^{(t)} \gamma \boldsymbol{P}^{\pi_{i_1}} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \left(\alpha_{i_2}^{(i_1)} \sqrt{\boldsymbol{\varphi}_t} + \eta_{i_2+1}^{(i_1)} \gamma \boldsymbol{P}^{\pi_{i_2}} \boldsymbol{\Delta}_{i_2} \right) \right] \\
&\le \sum_{i_1=(1-\beta)t}^{t-1} \alpha_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} \left(\gamma \boldsymbol{P}^{\pi_{i_1}} \right) \sqrt{\boldsymbol{\varphi}_t} \\
&\qquad + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \prod_{k=1}^{2} \left(\gamma \boldsymbol{P}^{\pi_{i_k}} \right) \boldsymbol{\Delta}_{i_2} \\
&= \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} \left\{ \boldsymbol{I} + \gamma \boldsymbol{P}^{\pi_{i_1}} \right\} \sqrt{\boldsymbol{\varphi}_t} + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \prod_{k=1}^{2} \left(\gamma \boldsymbol{P}^{\pi_{i_k}} \right) \boldsymbol{\Delta}_{i_2},
\end{aligned} \tag{68}$$

where the second inequality relies on (67), the third line uses the inequality $\eta_{i_1+1}^{(t)} \le \alpha_{i_1}^{(t)}$ in (66), and the fourth line is valid since $\sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_2}^{(i_1)} = 1$ (see (66)).

We intend to continue invoking (67) recursively — similar to how we derive (68) — in order to control $\boldsymbol{\Delta}_t$. To do so, we are in need of some preparation. First, let us define

$$H \coloneqq \frac{\log T}{1-\gamma} \qquad \text{and} \qquad \alpha_{\{i_k\}_{k=1}^H} \coloneqq \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} \dots \alpha_{i_H}^{(i_{H-1})} \ge 0 \tag{69}$$

for any $t > i_1 > i_2 > \cdots > i_H$, which clearly satisfies (see (66))

$$\alpha_{\{i_k\}_{k=1}^H} \ge \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \dots \eta_{i_H+1}^{(i_{H-1})}. \tag{70}$$

In addition, defining the index set

$$\mathcal{I}_t \coloneqq \Big\{ (i_1, \cdots, i_H) \,|\, (1-\beta)t \le i_1 \le t-1,\ \forall 1 \le j < H : (1-\beta) i_j \le i_{j+1} \le i_j - 1 \Big\}, \tag{71}$$

we have

$$\sum_{(i_1,\cdots,i_H) \in \mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1. \tag{72}$$

Additionally, recalling that $\beta = c_4(1-\gamma)/\log T$, we see that this choice of H satisfies

$$(1-\beta)^H = \left(1 - \frac{c_4(1-\gamma)}{\log T} \right)^{\frac{\log T}{1-\gamma}} \ge \frac{2}{3}$$

for c_4 small enough, thus implying that

$$i_1 > i_2 > \cdots > i_H \ge (1-\beta)^H t \ge 2t/3 \qquad \text{for all } (i_1, \cdots, i_H) \in \mathcal{I}_t.$$

This is an important property that allows one to invoke the relation (67). With these in place, applying the preceding relation (67) recursively — in a way similar to (68) — further leads to

$$\boldsymbol{\Delta}_t \le \sum_{(i_1,\cdots,i_H) \in \mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} \left\{ \left(\boldsymbol{I} + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \right) \sqrt{\boldsymbol{\varphi}_t} + \gamma^H \prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_k}} \left| \boldsymbol{\Delta}_{i_H} \right| \right\}$$

$$\leq \max_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \Bigg\{ \underbrace{\Big(\boldsymbol{I} + \sum_{h=1}^{H-1}\gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}}\Big)\sqrt{\boldsymbol{\varphi}_t}}_{=:\boldsymbol{\beta}_1} + \underbrace{\gamma^H \prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_k}} \big|\boldsymbol{\Delta}_{i_H}\big|}_{=:\boldsymbol{\beta}_2} \Bigg\} \tag{73}$$

for all $t \geq \frac{T}{c_2 \log T}$, where we recall the definition of the entrywise max operator in Section 4.1. Here, the last inequality relies on the fact that $\sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1$ (see (72)). It remains to control $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, which we shall accomplish separately in the next two steps.

Step 4: bounding $\boldsymbol{\beta}_2$. The term $\boldsymbol{\beta}_2$ defined in (73) is relatively easier to control. Observing that $\prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_k}}$ is still a probability transition matrix, we can derive

$$\begin{aligned}
|\boldsymbol{\beta}_2| &= \gamma^H \prod_{1\leq k\leq h} \boldsymbol{P}^{\pi_{i_k}} \big|\boldsymbol{\Delta}_{i_H}\big| \leq \gamma^H \Big\| \prod_{1\leq k\leq h} \boldsymbol{P}^{\pi_{i_k}} \Big\|_1 \|\boldsymbol{\Delta}_{i_H}\|_\infty = \gamma^H \|\boldsymbol{\Delta}_{i_H}\|_\infty \\
&\overset{\text{(i)}}{\leq} \frac{1}{1-\gamma}\gamma^H \overset{\text{(ii)}}{\leq} \frac{1}{(1-\gamma)T},
\end{aligned}$$

where (i) results from the crude bound (51). To justify the inequality (ii), we recall the definition (69) of H to see that

$$\gamma^H \overset{\text{(ii)}}{=} \big(1-(1-\gamma)\big)^{\frac{1}{1-\gamma}\log T} \leq e^{-\log T} = \frac{1}{T},$$

where the inequality comes from the elementary fact that $\gamma^{\frac{1}{1-\gamma}} \leq e^{-1}$ for any $0<\gamma<1$.

Step 5: bounding $\boldsymbol{\beta}_1$. When it comes to the term $\boldsymbol{\beta}_1$ defined in (73), we can upper bound the entrywise square of $\boldsymbol{\beta}_1$ — denoted by $|\boldsymbol{\beta}_1|^2$ — as follows

$$\begin{aligned}
|\boldsymbol{\beta}_1|^2 &= \Big|\Big(\sum_{h=0}^{H-1}\gamma^h \prod_{1\leq k\leq h} \boldsymbol{P}^{\pi_{i_k}}\Big)\sqrt{\boldsymbol{\varphi}_t}\Big|^2 \overset{\text{(i)}}{\leq} \Big|\sum_{h=0}^{H-1}\gamma^{h/2}\cdot\gamma^{h/2}\sqrt{\prod_{1\leq k\leq h}\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{\varphi}_t}\Big|^2 \\
&\overset{\text{(ii)}}{\leq} \sum_{h=0}^{H-1}\gamma^h \cdot \sum_{h=0}^{H-1}\gamma^h \prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{\varphi}_t \\
&\overset{\text{(iii)}}{\leq} \frac{1}{1-\gamma}\sum_{h=0}^{H-1}\gamma^h \prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}} \frac{64\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T}\Big(\max_{\frac{t}{2}\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\boldsymbol{1}\Big) \\
&\overset{\text{(iv)}}{\leq} \frac{64\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2 T}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\frac{t}{2}\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) + \frac{64\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^3 T}\boldsymbol{1}.
\end{aligned}$$

Here, (i) follows from Jensen's inequality and the fact that $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}$ is a probability transition matrix; (ii) holds due to the Cauchy-Schwarz inequality; (iii) utilizes the definition of $\boldsymbol{\varphi}_t$ in (62); (iv) follows since $\prod_{1\leq k\leq h}\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{1}=\boldsymbol{1}$ and $\sum_{0\leq h<H}\gamma^h \leq \frac{1}{1-\gamma}$. To further control the right-hand side of the above inequality, we resort to the following lemma.

Lemma 5. *Suppose that $t \geq \frac{T}{c_2\log T}$. For any $(i_1,\cdots,i_H)\in\mathcal{I}_t$, the following holds:*

$$\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\frac{t}{2}\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \leq \frac{4}{\gamma^2(1-\gamma)^2}\Big(1+2\max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\boldsymbol{1}. \tag{74}$$

Proof. This lemma, which is inspired by but significantly more complicated than Azar et al. (2013, Lemma 8), plays a key role in shaving one $\frac{1}{1-\gamma}$ factor. See Section B.5 for the proof. □

Therefore, the above result directly implies that

$$|\boldsymbol{\beta}_1|^2 \leq \frac{320\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+2\max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\boldsymbol{1}. \tag{75}$$

Step 6: putting all this together. Substituting the preceding bounds for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ into (73), we can demonstrate that: with probability at least $1-\delta$,

$$\begin{aligned}\boldsymbol{\Delta}_t &\le \frac{1}{(1-\gamma)T}\mathbf{1} + \sqrt{\frac{320\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1} \\ &\le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1} \end{aligned} \tag{76}$$

holds simultaneously for all $t \ge \frac{T}{c_2\log T}$, where the second line is valid since $\frac{1}{(1-\gamma)T} \le \sqrt{\frac{(\log^4 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}}$ under our sample size condition (52).

B.3 Proof of Lemma 2

Next, we move forward to develop an lower bound on $\boldsymbol{\Delta}_t$, which can be accomplished in an analogous manner as for the above upper bound. Applying a similar argument for (73) (except that we need to replace π_i with $\pi^\star$), one can deduce that

$$\boldsymbol{\Delta}_t \ge -\max_{(i_1,\cdots,i_H)\in\mathcal{I}_t}\left\{\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi^\star}\sqrt{\boldsymbol{\varphi}_t} + \gamma^H\prod_{k=1}^{H}\boldsymbol{P}^{\pi^\star}\big|\boldsymbol{\Delta}_{i_H}\big|\right\} \tag{77}$$

for any $t \ge \frac{c_2 T}{\log\frac{1}{1-\gamma}}$. It is straightforward to bound the second term on the right-hand side of (77) as

$$\gamma^H\prod_{1\le k\le H}\boldsymbol{P}^{\pi^\star}\big|\boldsymbol{\Delta}_{i_H}\big| \le \gamma^H\Big\|\prod_{1\le k\le H}\boldsymbol{P}^{\pi^\star}\Big\|_1\|\boldsymbol{\Delta}_{i_H}\|_\infty\mathbf{1} \le \frac{1}{(1-\gamma)T}\mathbf{1},$$

where the second inequality makes use of (51) as well as the fact that $\prod_k \boldsymbol{P}^{\pi^\star}$ is a probability transition matrix (so that $\|\prod_k \boldsymbol{P}^{\pi^\star}\|_1 = 1$). As for the first term on the right-hand side of (77), we can invoke a similar argument for (75) to obtain

$$\Big|\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi^\star}\sqrt{\boldsymbol{\varphi}_t}\Big|^2 \le 320\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}.$$

Taking these two bounds together, we see that with probability at least $1-\delta$,

$$\boldsymbol{\Delta}_t \ge -30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1} \tag{78}$$

holds simultaneously for all $t \ge \frac{T}{c_2\log T}$.

B.4 Solving the recurrence relation regarding $\boldsymbol{\Delta}_t$

Recall from (31) that with probability exceeding $1-2\delta$, the following recurrence relation

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)} \qquad \text{for all} \qquad t \ge \frac{T}{c_2\log T} \tag{79}$$

holds, which plays a crucial role in establishing the desired estimation error bound. Specifically, for any $k \ge 0$, let us define

$$u_k := \max\left\{\|\boldsymbol{\Delta}_t\|_\infty \;\Big|\; 2^k\frac{T}{c_2\log T} \le t \le T\right\}. \tag{80}$$

To bound this sequence, we first obtain a crude bound as a result of (51):

$$u_0 \le \frac{1}{1-\gamma}. \tag{81}$$

Next, it is directly seen from (79) and the definition of u_k that

$$u_k \le c_6\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\,(1+u_{k-1})}, \qquad k \ge 1 \tag{82}$$

for some constant $c_6 = 20/\gamma > 0$. In order to analyze the size of u_k, we divide into two cases.

- If $u_k \le 1$ for some $k \ge 1$, then (82) tells us that

$$u_{k+1} \le c_6\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\,(1+u_k)} \le c_6\sqrt{\frac{2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} \le 1,$$

  as long as $T \ge \frac{2c_6^2\log^4 T\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\gamma^2(1-\gamma)^4}$. In other words, once u_{k-1} drops below 1, then all subsequent quantities will remain bounded above by 1, namely, $\max_{j:j\ge k} u_j \le 1$. As a result,

$$u_j \le c_6\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\,(1+u_{j-1})} \le c_6\sqrt{\frac{2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} \qquad \text{for all } j > k.$$

- Instead, suppose that $u_j > 1$ for all $0 \le j \le k$. Then it is seen from (82) that

$$u_{j+1} \le c_6\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\,(1+u_j)} \le c_6\sqrt{\frac{2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\,u_j} \qquad \text{for all } j \le k.$$

  This is equivalent to saying that

$$\log u_{j+1} \le \log\alpha_u + \frac{1}{2}\log u_j \qquad \text{for all } j \le k,$$

  where $\alpha_u = c_6\sqrt{\frac{2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}}$. Invoking a standard analysis strategy for this type of recursive relations yields

$$\log u_{j+1} - 2\log\alpha_u \le \frac{1}{2}\left(\log u_j - 2\log\alpha_u\right) \qquad \text{for all } j \le k,$$

  and hence

$$\log u_{j+1} \le 2\log\alpha_u + \left(\frac{1}{2}\right)^{j+1}\left(\log u_0 - 2\log\alpha_u\right) \qquad \text{for all } j \le k.$$

  This is equivalent to saying that

$$u_j \le \alpha_u^2\left(\frac{u_0}{\alpha_u^2}\right)^{1/2^j} = \left(\alpha_u^2\right)^{1-1/2^j}(u_0)^{1/2^j} \qquad \text{for all } j \le k+1.$$

Putting the above two cases together and using (81), we conclude that

$$\begin{aligned} u_k &\le \sqrt{\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} + \Big(\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big)^{1-1/2^k} u_0^{1/2^k} \\ &\le \sqrt{\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} + \Big(\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big)^{1-1/2^k}\left(\frac{1}{1-\gamma}\right)^{1/2^k}, \qquad k \ge 1. \end{aligned}$$

In particular, as long as $k \geq c_7 \log\log\frac{1}{1-\gamma}$ for some constant $c_7 > 0$, one has $\big(\frac{1}{1-\gamma}\big)^{1/2^k} \leq O(1)$ and

$$\Big(\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\Big)^{1-1/2^k} \leq \max\left\{\sqrt{\frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}}, \frac{2c_6^2\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\right\}.$$

As a result, the above bound simplifies to

$$u_k \leq c_8\bigg(\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\bigg), \qquad k \geq c_7 \log\log\frac{1}{1-\gamma}$$

for some constant $c_8 > 0$.

Consequently, taking $t = T$ and choosing $k = c_7 \log\log\frac{1}{1-\gamma}$ for some appropriate constant $c_7 > 0$ (so as to ensure $2^k \frac{T}{c_2 \log T} < T$), we immediately see from the definition (80) of u_k that

$$\|\boldsymbol{\Delta}_T\|_\infty \leq c_8\bigg(\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\bigg). \tag{83}$$

with probability at least $1-2\delta$. To finish up, we note that the sample size assumption (48) is equivalent to

$$\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T} \leq \frac{\min\{\varepsilon^2, \varepsilon\}}{c_3}.$$

When $c_3 > 0$ is sufficiently large, substituting this relation into (83) gives

$$\begin{aligned}\|\boldsymbol{\Delta}_T\|_\infty &\leq \frac{1}{2}\sqrt{\min\{\varepsilon^2,\varepsilon\}} + \frac{1}{2}\min\{\varepsilon^2,\varepsilon\} = \begin{cases}\frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon^2 & \text{if } \varepsilon \leq 1\\ \frac{1}{2}\sqrt{\varepsilon} + \frac{1}{2}\varepsilon & \text{if } \varepsilon > 1\end{cases}\\ &\leq \varepsilon\end{aligned}$$

as claimed in Theorem 6.

B.5 Proof of Lemma 5

We first claim that

$$\max_{\frac{t}{2}\leq i<t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) - \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) \leq \frac{4}{1-\gamma}\max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}. \tag{84}$$

If this claim were valid (which we shall justify towards the end of this subsection), then it would lead to

$$\sum_{h=0}^{H-1}\gamma^h \prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}} \max_{\frac{t}{2}\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \leq \sum_{h=0}^{H-1}\gamma^h \prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) + \frac{4}{(1-\gamma)^2}\max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}. \tag{85}$$

It then boils down to bounding the first term on the right-hand side of (85). Let us first upper bound the variance term involving $\boldsymbol{V}^\star$. For any $0 \leq h < H$, one can express (see (17))

$$\begin{aligned}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &= \boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - (\boldsymbol{P}\boldsymbol{V}^\star)\circ(\boldsymbol{P}\boldsymbol{V}^\star)\\ &\overset{\text{(i)}}{=} \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r})\circ(\boldsymbol{Q}^\star - \boldsymbol{r})\\ &= \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \boldsymbol{P}^{\pi^\star}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r})\circ(\boldsymbol{Q}^\star - \boldsymbol{r})\\ &\leq \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \big\|\boldsymbol{P}^{\pi^\star}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star)\big\|_\infty \mathbf{1} - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r})\circ(\boldsymbol{Q}^\star - \boldsymbol{r})\\ &\overset{\text{(ii)}}{\leq} \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \frac{4}{1-\gamma}\max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r})\circ(\boldsymbol{Q}^\star - \boldsymbol{r})\end{aligned}$$

$$= \frac{1}{\gamma^2}\big(\gamma^2 \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star\big) + \frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} - \frac{1}{\gamma^2}\boldsymbol{r}\circ\boldsymbol{r} + \frac{2}{\gamma^2}\boldsymbol{Q}^\star \circ \boldsymbol{r}$$

$$\overset{\text{(iii)}}{\le} \frac{1}{\gamma}\big(\gamma \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star\big) + \frac{2}{\gamma^2}\boldsymbol{Q}^\star \circ \boldsymbol{r} + \frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}, \tag{86}$$

where (i) relies on the identity $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$, and (iii) holds since $0<\gamma<1$. To justify (ii), we make the following observation:

$$\begin{aligned}
&\big\|\boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{P}^{\pi^\star}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star)\big\|_\infty = \big\|\boldsymbol{P}\boldsymbol{\Pi}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{P}\boldsymbol{\Pi}^{\pi^\star}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star)\big\|_\infty \\
&\overset{\text{(iv)}}{\le} \big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{\Pi}^{\pi^\star}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star)\big\|_\infty \\
&= \big\|(\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^\star - \boldsymbol{\Pi}^{\pi^\star}\boldsymbol{Q}^\star)\circ(\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^\star + \boldsymbol{\Pi}^{\pi^\star}\boldsymbol{Q}^\star)\big\|_\infty \\
&\overset{\text{(v)}}{\le} \frac{2}{1-\gamma}\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^\star - \boldsymbol{\Pi}^{\pi^\star}\boldsymbol{Q}^\star\big\|_\infty \\
&\le \frac{2}{1-\gamma}\Big(\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^\star - \boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}}\big\|_\infty + \big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}} - \boldsymbol{V}^\star\big\|_\infty\Big) \\
&\overset{\text{(vi)}}{\le} \frac{2}{1-\gamma}\Big(\big\|\boldsymbol{Q}^\star - \boldsymbol{Q}_{i_{h+1}}\big\|_\infty + \big\|\boldsymbol{V}_{i_{h+1}} - \boldsymbol{V}^\star\big\|_\infty\Big) \\
&\overset{\text{(vii)}}{\le} \frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty,
\end{aligned}$$

where (iv) arises from the fact $\|\boldsymbol{P}\boldsymbol{z}\|_\infty \le \|\boldsymbol{P}\|_1\|\boldsymbol{z}\|_\infty = \|\boldsymbol{z}\|_\infty$, (v) is valid because $\|\boldsymbol{Q}^\star\|_\infty \le 1/(1-\gamma)$, (vi) follows from the fact that $\boldsymbol{V}_{i_{h+1}} = \boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}}$, and (vi) holds since $\|\boldsymbol{V}_{i_{h+1}} - \boldsymbol{V}^\star\|_\infty \le \|\boldsymbol{Q}_{i_{h+1}} - \boldsymbol{Q}^\star\|_\infty$.

As it turns out, the first term in (86) allows one to build a telescoping sum. Specifically, invoking (86) allows one to bound

$$\begin{aligned}
\sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &\le \frac{1}{\gamma}\sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\big(\gamma \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star\big) \\
&\qquad + \frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty \sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\mathbf{1} + \frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star \circ \boldsymbol{r}) \\
&\overset{\text{(i)}}{=} \frac{1}{\gamma}\Big(\sum_{h=0}^{H-1}\prod_{k=1}^{h+1}\gamma \boldsymbol{P}^{\pi_{i_k}} - \sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\Big)(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) \\
&\qquad + \frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty \sum_{h=0}^{H-1}\gamma^h\mathbf{1} + \frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star \circ \boldsymbol{r}) \\
&\le \frac{1}{\gamma}\Big(\prod_{k=1}^{H}\gamma \boldsymbol{P}^{\pi_{i_k}} - \boldsymbol{I}\Big)(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \frac{4}{(1-\gamma)^2}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} \\
&\qquad + \frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star \circ \boldsymbol{r}) \\
&\overset{\text{(ii)}}{\le} \left(\frac{2}{\gamma}\|\boldsymbol{Q}^\star\|_\infty^2 + \frac{4}{(1-\gamma)^2}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty + \frac{2}{\gamma^2}\frac{1}{1-\gamma}\|\boldsymbol{Q}^\star\|_\infty\|\boldsymbol{r}\|_\infty\right)\mathbf{1} \\
&\overset{\text{(iii)}}{\le} \frac{1}{(1-\gamma)^2}\left(\frac{2}{\gamma} + 4\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty + \frac{2}{\gamma^2}\right)\mathbf{1} \\
&\le \frac{1}{(1-\gamma)^2}\left(\frac{4}{\gamma^2} + 4\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)\mathbf{1}.
\end{aligned} \tag{87}$$

Here, (i) comes from the identity $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\mathbf{1} = \mathbf{1}$; (ii) holds because each row of $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}$ has unit $\|\cdot\|_1$ norm for any h; (iii) arises from the bound $\|\boldsymbol{Q}^\star\|_\infty \le 1/(1-\gamma)$. This completes the proof, as long as the claim (84) can be justified.

Proof of the inequality (84). To validate this result, we make the observation that

$$\begin{aligned}
\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) - \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &= \big[\boldsymbol{P}(\boldsymbol{V}_i \circ \boldsymbol{V}_i) - (\boldsymbol{P}\boldsymbol{V}_i) \circ (\boldsymbol{P}\boldsymbol{V}_i)\big] - \big[\boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - (\boldsymbol{P}\boldsymbol{V}^\star) \circ (\boldsymbol{P}\boldsymbol{V}^\star)\big] \\
&= \boldsymbol{P}(\boldsymbol{V}_i \circ \boldsymbol{V}_i - \boldsymbol{V}^\star \circ \boldsymbol{V}^\star) + (\boldsymbol{P}\boldsymbol{V}^\star) \circ (\boldsymbol{P}\boldsymbol{V}^\star) - (\boldsymbol{P}\boldsymbol{V}_i) \circ (\boldsymbol{P}\boldsymbol{V}_i) \\
&= \boldsymbol{P}\big((\boldsymbol{V}_i - \boldsymbol{V}^\star) \circ (\boldsymbol{V}_i + \boldsymbol{V}^\star)\big) + (\boldsymbol{P}\boldsymbol{V}^\star - \boldsymbol{P}\boldsymbol{V}_i) \circ (\boldsymbol{P}\boldsymbol{V}^\star + \boldsymbol{P}\boldsymbol{V}_i) \\
&\le \Big\{\big\|\boldsymbol{P}\big((\boldsymbol{V}_i - \boldsymbol{V}^\star) \circ (\boldsymbol{V}_i + \boldsymbol{V}^\star)\big)\big\|_\infty + \big\|(\boldsymbol{P}\boldsymbol{V}^\star - \boldsymbol{P}\boldsymbol{V}_i) \circ (\boldsymbol{P}\boldsymbol{V}^\star + \boldsymbol{P}\boldsymbol{V}_i)\big\|_\infty\Big\}\mathbf{1} \\
&\le \frac{4}{1-\gamma}\|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}.
\end{aligned}$$

Here, the last inequality follows since (by applying Lemma 4)

$$\big\|\boldsymbol{P}\big((\boldsymbol{V}_i - \boldsymbol{V}^\star) \circ (\boldsymbol{V}_i + \boldsymbol{V}^\star)\big)\big\|_\infty \le \|\boldsymbol{P}\|_1 \|\boldsymbol{V}_i - \boldsymbol{V}^\star\|_\infty \|\boldsymbol{V}_i + \boldsymbol{V}^\star\|_\infty \le \frac{2}{1-\gamma}\|\boldsymbol{\Delta}_i\|_\infty,$$

$$\text{and} \quad \big\|(\boldsymbol{P}\boldsymbol{V}^\star - \boldsymbol{P}\boldsymbol{V}_i) \circ (\boldsymbol{P}\boldsymbol{V}^\star + \boldsymbol{P}\boldsymbol{V}_i)\big\|_\infty \le \|\boldsymbol{P}\|_1 \|\boldsymbol{V}_i - \boldsymbol{V}^\star\|_\infty \cdot \|\boldsymbol{P}\|_1 \|\boldsymbol{V}_i + \boldsymbol{V}^\star\|_\infty \le \frac{2}{1-\gamma}\|\boldsymbol{\Delta}_i\|_\infty.$$

A useful extension of Lemma 5. Before concluding, we make note of the following extension that proves useful for studying asynchronous Q-learning.

Lemma 6. *Suppose that $t \ge \frac{T}{c_2 \log T}$. Then one has*

$$\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\widehat{\pi}_k} \max_{\frac{t}{2} \le i < t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \le \frac{4}{\gamma^2(1-\gamma)^2}\Big(1 + 2 \max_{\frac{t}{2} \le i < t} \|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1} \tag{88}$$

for any set of policies $\{\widehat{\pi}_k\}$ obeying $\{\widehat{\pi}_k\} \subseteq \Pi$. Here, we define

$$\Pi \coloneqq \big\{\pi = [\pi(s)]_{s\in\mathcal{S}} \,\big|\, \pi(s) \in \Pi_s, \forall s \in \mathcal{S}\big\}, \qquad \Pi_s \coloneqq \big\{\pi_i(s) \mid i \in [t/2, t)\big\}. \tag{89}$$

The key difference between Lemma 6 and Lemma 5 is that: the components of $\widehat{\pi}_k$ corresponding to different states can be chosen in a separate manner. The proof follows from an identical argument as the above proof of Lemma 5, and is hence omitted.

C Analysis for TD learning (Theorem 1)

As it turns out, if $|\mathcal{A}| = 1$ (which reduces to the case of TD learning), we can further modify the previous analysis in Section B to yield an improved $\frac{1}{(1-\gamma)^3}$ scaling. This forms the main content of this section, which leads to the proof of Theorem 1 for TD learning. Akin to the Q-learning case, we proceed to establish a more general version of Theorem 1 that covers the full ε-range. This is formally stated below, which subsumes Theorem 1 as a special case.

Theorem 7. *Consider any $\gamma \in (0,1)$ and any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Theorem 1 continues to hold if*

$$T \ge \frac{c_3\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 \min\{\varepsilon, \varepsilon^2\}} \tag{90}$$

for some sufficiently large universal constant $c_3 > 0$.

C.1 Preliminary facts

Before embarking on the analysis, we begin by presenting several useful preliminary facts. The first one is a direct consequence of the claimed iteration complexity (90) when $\varepsilon \le \frac{1}{1-\gamma}$:

$$T \ge \frac{c_3\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 \min\{\varepsilon, \varepsilon^2\}} \ge \frac{c_3\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^2}, \tag{91}$$

a simple fact that will be used multiple times. In addition, the update rule (7) of TD learning can be expressed using vector/matrix notation as follows

$$\boldsymbol{V}_t = (1-\eta_t)\boldsymbol{V}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \qquad \text{for all } t \geq 1, \tag{92}$$

where the matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}|}$ obeys

$$\boldsymbol{P}_t(s,s') \coloneqq \begin{cases} 1, & \text{if } s' = s_t(s) \\ 0, & \text{else} \end{cases}$$

for any $s, s' \in \mathcal{S}$. In the sequel, we collect a few other facts concerning the range of $\boldsymbol{V}_t$ and learning rates.

Range of $\boldsymbol{V}_t$. We claim that: when the initialization $\boldsymbol{V}_0$ obeys $\boldsymbol{0} \leq \boldsymbol{V}_0 \leq \frac{1}{1-\gamma}\boldsymbol{1}$, the TD learning iterates obey

$$\boldsymbol{0} \leq \boldsymbol{V}_t \leq \frac{1}{1-\gamma}\boldsymbol{1} \qquad \text{and} \qquad \|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_\infty \leq \frac{1}{1-\gamma} \qquad \text{for all } t \geq 0, \tag{93}$$

provided that $0 \leq \eta_t \leq 1$ for all $t \geq 0$. The proof follows immediately by repeating the proof of Lemma 4 (see Section B.1) with $|\mathcal{A}| = 1$, and is hence omitted for brevity.

Learning rates. We shall also collect several useful results concerning the learning rates $\{\eta_t\}$. Let us abuse the notation by defining the following crucial quantities:

$$\eta_k^{(t)} \coloneqq \begin{cases} \prod_{i=1}^{t}\big(1-\eta_i(1-\gamma)\big), & \text{if } k = 0, \\ \eta_k \prod_{i=k+1}^{t}\big(1-\eta_i(1-\gamma)\big), & \text{if } 0 < k < t, \\ \eta_t, & \text{if } k = t. \end{cases} \tag{94}$$

Note that this definition (94) differs from the one (53) used for Q-learning, and will only be employed in this section. Consider any iteration number t satisfying

$$t \geq \frac{T}{c_2 \log T}. \tag{95}$$

Clearly, the learning rate η_t under Assumption (8a) obeys

$$(1-\gamma)\eta_t \geq \frac{1-\gamma}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \geq \frac{1-\gamma}{\frac{2c_1(1-\gamma)T}{\log^2 T}} = \frac{\log^2 T}{2c_1 T}. \tag{96}$$

In what follows, we intend to bound $\eta_k^{(t)}$ for two cases separately.

- For any i obeying $0 \leq i \leq t/2$, it is easily seen from (96) that

$$\begin{aligned} \eta_i^{(t)} &\leq \big(1-\eta_{t/2}(1-\gamma)\big)^{t/2} \leq \Big(1-\frac{\log^2 T}{2c_1 T}\Big)^{t/2} \leq \Big(1-\frac{\log^2 T}{2c_1 T}\Big)^{\frac{T}{2c_2\log T}} \\ &= \left\{\Big(1-\frac{\log^2 T}{2c_1 T}\Big)^{\frac{2c_1 T}{\log^2 T}}\right\}^{\frac{\log T}{4c_1 c_2}} \leq \frac{1}{T^2}, \end{aligned} \tag{97a}$$

  where the last inequality holds as long as $c_1 c_2 \leq 1/8$ and (91) holds.

- When it comes to the case with $i > t/2$, we can develop the following upper bound

$$\eta_i^{(t)} \leq \eta_i \leq \frac{1}{c_2(1-\gamma)i/\log^2 T} < \frac{2\log^3 T}{(1-\gamma)T}, \tag{97b}$$

  which relies on Assumption (8a).

In addition, given that $\boldsymbol{P}^k\mathbf{1} = \mathbf{1}$ for any integer $k > 0$, it can be easily verified that

$$\prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\mathbf{1} = \prod_{i=k+1}^{t} \big(1 - \eta_i(1-\gamma)\big)\mathbf{1},$$

and as a result,

$$\left\| \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big) \right\|_1 = \prod_{i=k+1}^{t} \big(1 - \eta_i(1-\gamma)\big). \tag{98}$$

C.2 Proof of Theorem 7

Step 1: decomposing the error $\boldsymbol{V}_t - \boldsymbol{V}^\star$. Taking $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{V}_t - \boldsymbol{V}^\star$, via the basic relation (23), the TD learning update rule can be written as

$$\begin{aligned}\boldsymbol{\Delta}_t &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big(\boldsymbol{P}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big) \\ &= \big(\boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}. \end{aligned} \tag{99}$$

Invoking the above relation recursively then leads to

$$\boldsymbol{\Delta}_t = \prod_{i=1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\boldsymbol{\Delta}_0 + \underbrace{\sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\gamma(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{V}_{k-1}}_{=:\,\boldsymbol{\xi}_t}. \tag{100}$$

Step 2: controlling the first term of (100). With regards to the first term of (100), we make the observation that

$$\begin{aligned}\left\| \prod_{i=1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\boldsymbol{\Delta}_0 \right\|_\infty &\le \left\| \prod_{i=1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big) \right\|_1 \|\boldsymbol{\Delta}_0\|_\infty \\ &= \left\{ \prod_{i=1}^{t} \big(1 - \eta_i(1-\gamma)\big) \right\} \|\boldsymbol{\Delta}_0\|_\infty \\ &\le \eta_0^{(t)} \cdot \frac{1}{1-\gamma} \le \frac{1}{(1-\gamma)T^2}, \end{aligned} \tag{101}$$

where the second line arises from (98), and the last inequality holds true due to (97a) as long as $t \ge \frac{T}{c_2 \log T}$.

Step 3: controlling the second term of (100). We then move on to the second term $\boldsymbol{\xi}_t$ in (100), which admits the following expression

$$\boldsymbol{\xi}_t = \sum_{k=1}^{t} \boldsymbol{z}_k \qquad \text{with } \boldsymbol{z}_k \coloneqq \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\gamma(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{V}_{k-1}. \tag{102}$$

Here, the summands $\{\boldsymbol{z}_k\}$ clearly satisfy

$$\mathbb{E}\big[\boldsymbol{z}_k \,|\, \boldsymbol{V}_{k-1}, \cdots, \boldsymbol{V}_0\big] = \mathbf{0}.$$

We then attempt to invoke the Freedman inequality (see Theorem 5) to control this term. Towards this end, there are several quantities that need to be calculated.

- First of all, we observe that

$$B \coloneqq \max_{1\le k\le t} \|\boldsymbol{z}_k\|_\infty \le \max_{1\le k\le t} \left\| \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I} - \gamma\boldsymbol{P})\big)\gamma(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{V}_{k-1} \right\|_\infty$$

$$\leq \max_{1\leq k\leq t} \Big\|\eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\Big\|_1 \big\|(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{V}_{k-1}\big\|_\infty$$
$$= \max_{1\leq k\leq t} \left\{\eta_k \prod_{i=k+1}^{t} \big(1 - \eta_i(1-\gamma)\big)\right\} \big\|(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{V}_{k-1}\big\|_\infty$$
$$\leq \max_{1\leq k\leq t} \eta_k^{(t)} \big(\|\boldsymbol{P}_k\|_1 + \|\boldsymbol{P}\|_1\big)\|\boldsymbol{V}_{k-1}\|_\infty \leq \frac{4\log^3 T}{(1-\gamma)^2 T}, \tag{103}$$

where the third line again makes use of the relation (98) and the last line follows the facts $\|\boldsymbol{P}_k\|_1 = \|\boldsymbol{P}\|_1 = 1$, $\|\boldsymbol{V}_{k-1}\|_\infty \leq 1/(1-\gamma)$, as well as the properties (97).

- The next step is to control certain variance terms. Towards this, we first make note of a userful fact. For any given non-negative vector $\boldsymbol{u} = [u_i]_{1\leq i\leq |\mathcal{S}|} \geq \boldsymbol{0}$ and any vector $\boldsymbol{v}$, it is easily seen that

$$\mathsf{Var}\Big(\boldsymbol{u}^\top(\boldsymbol{P}_k - \boldsymbol{P})\boldsymbol{v}\Big) = \sum_{i=1}^{|\mathcal{S}|} u_i^2 \mathsf{Var}\Big((\boldsymbol{P}_k - \boldsymbol{P})_{i,\cdot}\boldsymbol{v}\Big) \leq \Big\{\max_i |u_i|\Big\}\big[u_1, \cdots, u_{|\mathcal{S}|}\big]\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{v})$$
$$\leq \|\boldsymbol{u}\|_1 \boldsymbol{u}^\top \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{v}), \tag{104}$$

where we remind the reader of the notation $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{v})$ in (17). Additionally, for any vector $\boldsymbol{a} = [a_j]$, let us employ the notation $\mathsf{Var}\big(\boldsymbol{a} \,|\, \boldsymbol{V}_{k-1}, \cdots, \boldsymbol{V}_0\big)$ to represent a vector whose j-th entry is given by $\mathsf{Var}\big(a_j \,|\, \boldsymbol{V}_{k-1}, \cdots, \boldsymbol{V}_0\big)$. Armed with this notation, we obtain

$$\mathsf{Var}\big(\boldsymbol{z}_k \,|\, \boldsymbol{V}_{k-1}, \cdots, \boldsymbol{V}_0\big) \leq \gamma^2 \Big\|\eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\Big\|_1 \left\{\eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\right\} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1})$$
$$= \gamma^2 \left\{\eta_k \prod_{i=k+1}^{t} \big(1 - \eta_i(1-\gamma)\big)\right\} \left\{\eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\right\} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1})$$
$$\leq \eta_k \eta_k^{(t)} \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1}), \tag{105}$$

where the first inequality is a consequence of (104) and the definition of $\boldsymbol{z}_k$ (cf. (102)), the second line arises from (98), and the last relation results from the definition of $\eta_k^{(t)}$. This in turn allows us to compute

$$\boldsymbol{W}_t \coloneqq \sum_{k=1}^{t} \mathsf{Var}\big(\boldsymbol{z}_k \,|\, \boldsymbol{V}_{k-1}, \cdots, \boldsymbol{V}_0\big) \leq \sum_{k=1}^{t} \eta_k \eta_k^{(t)} \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1})$$
$$\leq \sum_{k=1}^{t/2} \eta_k^{(t)} \Big\|\eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\Big\|_1 \|\boldsymbol{V}_{k-1}\|_\infty^2 \boldsymbol{1} + \sum_{k=t/2+1}^{t} \eta_k \eta_k^{(t)} \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1})$$
$$\leq \sum_{k=1}^{t/2} \big(\eta_k^{(t)}\big)^2 \frac{1}{(1-\gamma)^2}\boldsymbol{1} + \Big\{\max_{k:\, t/2<k\leq t} \eta_k^{(t)}\Big\} \sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{k-1})$$
$$\leq \frac{1}{2(1-\gamma)^2 T^3}\boldsymbol{1} + \frac{2\log^3 T}{(1-\gamma)T}\left(\sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\right) \max_{k:\, t/2\leq k<t} \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_k\big)$$
$$\leq \frac{1}{2(1-\gamma)^2 T^3}\boldsymbol{1} + \frac{2\log^3 T}{(1-\gamma)T}(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1} \max_{k:\, t/2\leq k<t} \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_k\big), \tag{106}$$

where the penultimate inequality results from (97); to see why the last inequality holds, observe that

$$\sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} \big(\boldsymbol{I} - \eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)$$

$$
\begin{aligned}
&= (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\sum_{k=t/2+1}^{t}\eta_k(\boldsymbol{I}-\gamma\boldsymbol{P})\prod_{i=k+1}^{t}\big(\boldsymbol{I}-\eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\\
&= (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\sum_{k=t/2+1}^{t}\bigg[\prod_{i=k+1}^{t}\big(\boldsymbol{I}-\eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)-\prod_{i=k}^{t}\big(\boldsymbol{I}-\eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\bigg]\\
&= (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}-(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\prod_{i=t/2+1}^{t}\big(\boldsymbol{I}-\eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})\big)\le(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1},
\end{aligned}
$$

where we have used the fact that all entries of $(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}$ and $\boldsymbol{I}-\eta_i(\boldsymbol{I}-\gamma\boldsymbol{P})$ are non-negative.

- In addition, we also derive the following trivial upper bound based on (106):

$$
\begin{aligned}
|\boldsymbol{W}_t| &\le \frac{1}{2(1-\gamma)^2T^3}\mathbf{1}+\frac{2\log^3T}{(1-\gamma)T}\big\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\big\|_1\max_{k:\,t/2\le k<t}\big\|\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_k)\big\|_\infty\mathbf{1}\\
&\le \frac{1}{2(1-\gamma)^2T^3}\mathbf{1}+\frac{2\log^3T}{(1-\gamma)^4T}\mathbf{1}\le\frac{3\log^3T}{(1-\gamma)^4T}\mathbf{1}=:\sigma^2\mathbf{1},
\end{aligned}
\tag{107}
$$

where we have invoked the fact that $\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\|_1=1/(1-\gamma)$. Therefore, by setting $K=\big\lceil 2\log_2\frac{1}{1-\gamma}\big\rceil$, one arrives at

$$
\frac{\sigma^2}{2^K}\le\frac{3\log^3T}{(1-\gamma)^2T}.
\tag{108}
$$

Equipped with the preceding bounds, let us apply the Freedman inequality in Theorem 5 and invoke the union bound over all entries of $\boldsymbol{\xi}_t$ to show that

$$
\begin{aligned}
|\boldsymbol{\xi}_t| &\le \sqrt{8\Big(\boldsymbol{W}_t+\frac{\sigma^2}{2^K}\mathbf{1}\Big)\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}}+\Big(\frac{4}{3}B\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}\Big)\mathbf{1}\\
&\le \sqrt{16\Big(\boldsymbol{W}_t+\frac{3\log^3T}{(1-\gamma)^2T}\mathbf{1}\Big)\log\frac{|\mathcal{S}|T}{\delta}}+\Big(3B\log\frac{|\mathcal{S}|T}{\delta}\Big)\mathbf{1}\\
&\le \sqrt{\frac{32\big(\log^3T\big)\big(\log\frac{|\mathcal{S}|T}{\delta}\big)}{(1-\gamma)T}\Big((\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\max_{k:\,t/2\le k<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_k)+\frac{2}{1-\gamma}\mathbf{1}\Big)}+\frac{12\big(\log^3T\big)\big(\log\frac{|\mathcal{S}|T}{\delta}\big)}{(1-\gamma)^2T}\mathbf{1}
\end{aligned}
$$

with probability at least $1-\delta/T$. Here, the second line follows since

$$
\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}\le 2\log\frac{|\mathcal{S}|T}{\delta}
$$

as long as $\frac{|\mathcal{S}|T}{\delta}\ge 8\log\frac{1}{1-\gamma}$, whereas the last line holds by using (103), (106) and (108). Further, we make the observation that

$$
\begin{aligned}
(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &= (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\Big(\boldsymbol{P}(\boldsymbol{V}^\star\circ\boldsymbol{V}^\star)-(\boldsymbol{P}\boldsymbol{V}^\star)\circ(\boldsymbol{P}\boldsymbol{V}^\star)\Big)\\
&= (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\left(\boldsymbol{P}(\boldsymbol{V}^\star\circ\boldsymbol{V}^\star)-\frac{1}{\gamma^2}(\boldsymbol{V}^\star-\boldsymbol{r})\circ(\boldsymbol{V}^\star-\boldsymbol{r})\right)\\
&\le (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\left(\boldsymbol{P}(\boldsymbol{V}^\star\circ\boldsymbol{V}^\star)-\frac{1}{\gamma^2}\boldsymbol{V}^\star\circ\boldsymbol{V}^\star+\frac{2}{\gamma^2}\boldsymbol{r}\circ\boldsymbol{V}^\star\right)\\
&\le (\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\left(\boldsymbol{P}(\boldsymbol{V}^\star\circ\boldsymbol{V}^\star)-\frac{1}{\gamma}\boldsymbol{V}^\star\circ\boldsymbol{V}^\star+\frac{2}{\gamma^2}\boldsymbol{r}\circ\boldsymbol{V}^\star\right)\\
&= \frac{1}{\gamma}(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}(\gamma\boldsymbol{P}-\boldsymbol{I})(\boldsymbol{V}^\star\circ\boldsymbol{V}^\star)+\frac{2}{\gamma^2}(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}(\boldsymbol{r}\circ\boldsymbol{V}^\star)\\
&\le \frac{2}{\gamma^2}(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}(\boldsymbol{r}\circ\boldsymbol{V}^\star)\le\frac{2}{\gamma^2(1-\gamma)^2}\mathbf{1},
\end{aligned}
$$

where the second line makes use of the basic relation $\boldsymbol{V}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$. As a consequence, we conclude

$$(\boldsymbol{I} - \gamma \boldsymbol{P})^{-1} \max_{k:\, t/2 \le k < t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_k) \le (\boldsymbol{I} - \gamma \boldsymbol{P})^{-1} \Big(\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) + \frac{4}{1-\gamma} \max_{k:\, t/2 \le k < t} \|\boldsymbol{\Delta}_k\|_\infty \mathbf{1} \Big)$$
$$\le \frac{2}{\gamma(1-\gamma)^2} \Big(1 + 2 \max_{k:\, t/2 \le k < t} \|\boldsymbol{\Delta}_k\|_\infty \Big) \mathbf{1}. \tag{109}$$

Here, the first inequality arises from (84), while the second inequality holds due to the facts that $\big\|(\boldsymbol{I} - \gamma \boldsymbol{P})^{-1}\big\|_1 = 1/(1-\gamma)$.

Step 4: putting everything together. Consequently, substituting the bounds in Steps 2-3 into (100) yields

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30 \sqrt{\frac{\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 T} \Big(1 + \max_{\frac{t}{2} \le i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big)} \qquad \text{for all} \quad t \ge \frac{T}{c_2 \log T}. \tag{110}$$

Repeating the same argument as in Section B.4, we see that

$$\|\boldsymbol{\Delta}_T\|_\infty \le c_9 \Bigg(\sqrt{\frac{\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 T}} + \frac{\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 T} \Bigg) \tag{111}$$

holds with probability at least $1-\delta$, where $c_9 > 0$ is some universal constant. As a result, one has

$$\|\boldsymbol{\Delta}_T\|_\infty \le \frac{1}{2} \Big(\sqrt{\min\{\varepsilon, \varepsilon^2\}} + \min\{\varepsilon, \varepsilon^2\} \Big) = \frac{1}{2} \left(\varepsilon + \varepsilon^2\right) \mathbb{1}\{\varepsilon \le 1\} + \frac{1}{2} \left(\varepsilon + \varepsilon^2\right) \mathbb{1}\{\varepsilon > 1\} \le \varepsilon,$$

as long as the sample size satisfies the following

$$\frac{\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}|T}{\delta}\big)}{\gamma^2(1-\gamma)^3 T} \le \frac{\min\{\varepsilon, \varepsilon^2\}}{c_3},$$

for some constant $c_3 \ge \max\{1, 2c_9\}$. This requirement is equivalent to condition (90) as claimed.

D Lower bound: sub-optimality of Q-learning (Theorem 3)

In this section, we establish the lower bound claimed in Theorem 3 by analyzing Q-learning for the MDP instance constructed in Section 4.3. Without loss of generality, we assume

$$\log T \le \frac{1}{1-\gamma} \tag{112}$$

throughout the proof; otherwise the lower bound in Theorem 3 is worse than the minimax lower bound $\frac{1}{(1-\gamma)^3 T}$ in Azar et al. (2013).

Throughout, we shall use P_t to represent the sample transitions such that for any triple (s, a, s'),

$$P_t(s' \,|\, s, a) \coloneqq \begin{cases} 1, & \text{if } s_t(s,a) = s', \\ 0, & \text{otherwise}, \end{cases} \tag{113}$$

where $s_t(s,a)$ stands for the sample collected in the t-th iteration (see (5)). Recognizing that state 2 is associated with a singleton action space, we shall often write

$$P_t(s' \,|\, 2) \coloneqq P_t(s' \,|\, 2, 1)$$

for notational simplicity.

D.1 Key quantities related to learning rates

We find it convenient to define the following quantities (by abuse of notation)

$$\eta_k^{(t)} \coloneqq \eta_k \prod_{i=k+1}^{t} \big(1-\eta_i(1-\gamma p)\big) \qquad \text{for any } 1 \le k < t, \tag{114a}$$

$$\eta_0^{(t)} \coloneqq \prod_{i=1}^{t} \big(1-\eta_i(1-\gamma p)\big), \tag{114b}$$

$$\eta_t^{(t)} \coloneqq \eta_t. \tag{114c}$$

It is helpful to establish several basic properties about these quantities. As can be easily verified,

$$\eta_0^{(t)} + (1-\gamma p)\sum_{k=1}^{t} \eta_k^{(t)} = \prod_{i=1}^{t}(1-\widehat{\eta}_i) + \widehat{\eta}_1 \prod_{i=2}^{t}(1-\widehat{\eta}_i) + \widehat{\eta}_2 \prod_{i=3}^{t}(1-\widehat{\eta}_i) + \cdots + \widehat{\eta}_{t-1}(1-\widehat{\eta}_t) + \widehat{\eta}_t = 1, \tag{115}$$

where we denote $\widehat{\eta}_i \coloneqq \eta_i(1-\gamma p)$ to simplify notation. Similarly, for any given integer $0 \le \tau < t$ one has

$$\prod_{i=\tau+1}^{t} \big(1-\eta_i(1-\gamma p)\big) + (1-\gamma p)\sum_{k=\tau+1}^{t} \eta_k^{(t)} = 1. \tag{116}$$

D.2 Preliminary calculations

Before moving forward, we record several basic relations as a result of the Q-learning update rule.

D.2.1 Basic update rules and expansion

Given that $Q_0 = V_0 = 0$ and that state 0 is absorbing, the update rule (4) gives

$$V_t(0) = Q_t(0,1) = \big(1-\eta_t(1-\gamma)\big)Q_{t-1}(0,1) = \prod_{i=1}^{t} \big(1-\eta_i(1-\gamma)\big)Q_0(0,1) = 0 \tag{117}$$

for all $t \ge 1$. Regarding state 2, the update rule (4) taken together with (117) leads to

$$\begin{aligned} V_t(2) = Q_t(2,1) &= \big(1-\eta_t\big)Q_{t-1}(2,1) + \eta_t\big\{r(2,1) + \gamma P_t(2\,|\,2)V_{t-1}(2) + \gamma P_t(0\,|\,2)V_{t-1}(0)\big\} \\ &= \big(1-\eta_t\big)V_{t-1}(2) + \eta_t\big\{1 + \gamma P_t(2\,|\,2)V_{t-1}(2)\big\}, \end{aligned} \tag{118}$$

and for state 3,

$$\begin{aligned} V_t(3) = Q_t(3,1) &= \big(1-\eta_t\big)Q_{t-1}(3,1) + \eta_t\big\{r(3,1) + \gamma V_{t-1}(3)\big\} \\ &= \big(1-\eta_t(1-\gamma)\big)V_{t-1}(3) + \eta_t. \end{aligned} \tag{119}$$

Similarly, one also has

$$Q_t(1,1) = (1-\eta_t)Q_{t-1}(1,1) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,1)V_{t-1}(1)\big\}, \tag{120a}$$

$$Q_t(1,2) = (1-\eta_t)Q_{t-1}(1,2) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,2)V_{t-1}(1)\big\}. \tag{120b}$$

In what follows, we shall first determine a crude range for certain quantities relates to the learning rates η_t, and then combine this with the above relations to establish the desired result.

Next, we record some elementary decomposition of $V_t(2)$. For any iteration t and $\tau < t$, one can continue the derivation in (118) to obtain

$$V_t(2) = \big(1-\eta_t(1-\gamma p)\big)V_{t-1}(2) + \eta_t\big\{1 + \gamma\big(P_t(2\,|\,2) - p\big)V_{t-1}(2)\big\}$$

$$
\begin{aligned}
&= \prod_{i=\tau+1}^{t} \big(1-\eta_i(1-\gamma p)\big) V_\tau(2) + \sum_{k=\tau+1}^{t} \eta_k \prod_{i=k+1}^{t} \big(1-\eta_i(1-\gamma p)\big)\big\{1+\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2)\big\} \\
&= \prod_{i=\tau+1}^{t} \big(1-\eta_i(1-\gamma p)\big) V_\tau(2) + \sum_{k=\tau+1}^{t} \eta_k^{(t)} + \sum_{k=\tau+1}^{t} \eta_k^{(t)} \gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2) \\
&= \prod_{i=\tau+1}^{t} \big(1-\eta_i(1-\gamma p)\big) V_\tau(2) + \frac{1-\prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)}{1-\gamma p} + \sum_{k=\tau+1}^{t} \eta_k^{(t)} \gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2) \\
&= \frac{1}{1-\gamma p} - \prod_{i=\tau+1}^{t} \big(1-\eta_i(1-\gamma p)\big)\left[\frac{1}{1-\gamma p} - V_\tau(2)\right] + \sum_{k=\tau+1}^{t} \eta_k^{(t)} \gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2), \qquad (121)
\end{aligned}
$$

where the penultimate line arises from (116). In particular, in the special case where $\tau = 0$ (so that $V_\tau(2) = V_0(2) = 0$), this simplifies to

$$
V_t(2) = \frac{1-\eta_0^{(t)}}{1-\gamma p} + \sum_{k=1}^{t} \eta_k^{(t)} \gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2), \qquad (122)
$$

which relies on the definition of $\eta_0^{(t)}$ in (114). With similar derivation, (119) leads to

$$
V_t(3) = \frac{1}{1-\gamma}\Big[1 - \prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big)\Big] = V^\star(3) - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big). \qquad (123)
$$

D.2.2 Mean and variance of $V^\star(2) - V_T(2)$

We start by computing the mean $V^\star(2) - \mathbb{E}[V_t(2)]$. From the construction (33), it is easily seen that $\mathbb{E}[P_k(2\,|\,2)] = p$, which together with the identity (122) leads to

$$
\mathbb{E}\big[V_T(2)\big] = \frac{1-\eta_0^{(T)}}{1-\gamma p} \qquad \text{and} \qquad V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\eta_0^{(T)}}{1-\gamma p}. \qquad (124)
$$

Similarly, applying the above argument to (121) and rearranging terms, we immediately arrive at

$$
V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\left[\frac{1}{1-\gamma p} - \mathbb{E}\big[V_\tau(2)\big]\right] \qquad (125)
$$

for any integer $0 \le \tau < T$.

Next, we develop a lower bound on the variance $\mathsf{Var}\big(V_T(2)\big)$. Towards this end, consider first a martingale sequence $\{Z_k\}_{0\le k\le T}$ adapted to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$, namely, $\mathbb{E}[Z_{k+1}\,|\,\mathcal{F}_k] = 0$ and $\mathbb{E}[Z_k\,|\,\mathcal{F}_k] = Z_k$ for all $0 \le k \le T$. In addition, consider any $0 \le \tau < T$, and let W_0 be a random variable such that $\mathbb{E}[W_0\,|\,\mathcal{F}_\tau] = W_0$. Then the law of total variance together with basic martingale properties tells us that

$$
\begin{aligned}
\mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T} Z_k\right) &= \mathbb{E}\left[\mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T} Z_k \mid \mathcal{F}_{T-1}\right)\right] + \mathsf{Var}\left(\mathbb{E}\left[W_0 + \sum_{k=\tau+1}^{T} Z_k \mid \mathcal{F}_{T-1}\right]\right) \\
&= \mathbb{E}\big[\,\mathsf{Var}\left(Z_T \mid \mathcal{F}_{T-1}\right)\big] + \mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T-1} Z_k\right) = \cdots \\
&= \sum_{k=\tau+1}^{T} \mathbb{E}\big[\,\mathsf{Var}\left(Z_k \mid \mathcal{F}_{k-1}\right)\big] + \mathsf{Var}\left(W_0\right) \ge \sum_{k=\tau+1}^{T} \mathbb{E}\big[\,\mathsf{Var}\left(Z_k \mid \mathcal{F}_{k-1}\right)\big]. \qquad (126)
\end{aligned}
$$

Consequently, for any $0 \le \tau < T-1$, it follows from the decomposition (121) (with τ replaced by $\tau+1$) that

$$
\mathsf{Var}\big(V_T(2)\big) \ge \mathbb{E}\left[\sum_{k=\tau+2}^{T} \mathsf{Var}\Big(\eta_k^{(T)}\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2)\,|\,V_{k-1}(2)\Big)\right]
$$

$$= \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\gamma\big)^2 p(1-p)\mathbb{E}\big[\big(V_{k-1}(2)\big)^2\big]$$
$$\geq \frac{(1-\gamma)\,(4\gamma-1)}{9} \cdot \frac{1}{4(1-\gamma)^2} \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2$$
$$= \frac{4\gamma-1}{36\,(1-\gamma)} \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2, \tag{127}$$

where the first identity relies on the fact that $P_k(2\,|\,2)$ is a Bernoulli random variable with mean p, and the inequality comes from the definition of τ (see (133)) and the choice of p (see (34)). As an implication, the sum of squares of $\eta_k^{(T)}$ plays a crucial role in determining the variance of $V_T(2)$.

D.3 Lower bounds for three cases

D.3.1 Case 1: small learning rates ($c_\eta \geq \log T$ or $0 \leq \eta \leq \frac{1}{(1-\gamma)T}$)

In this case, we focus on lower bounding $V^\star(2) - \mathbb{E}\big[V_T(2)\big]$. In view of this identity (124), this boils down to controlling $\eta_0^{(T)}$.

Suppose that $c_\eta > \log T$ (for rescaled linear learning rates) or $0 \leq \eta < \frac{1}{(1-\gamma)T}$ (for constant learning rates). A little algebra then gives

$$\eta_t(1-\gamma p) \leq \begin{cases} \frac{1-\gamma p}{(1-\gamma)t\log T} = \frac{4}{3t\log T} \leq \frac{1}{2}, & \text{if } \eta_t = \frac{1}{1+c_\eta(1-\gamma)t} \\ \frac{1-\gamma p}{(1-\gamma)T} = \frac{4}{3T} \leq \frac{1}{2}, & \text{if } \eta_t = \eta \end{cases} \tag{128}$$

for any $t \geq 1$, provided that $T \geq 15$. Consequently, one can derive

$$\log \eta_0^{(T)} = \sum_{i=1}^{T} \log\big(1-\eta_i(1-\gamma p)\big) \geq -1.5 \sum_{i=1}^{T} \eta_i(1-\gamma p) \geq -2, \tag{129}$$

where the first inequality holds due to the elementary fact $\log(1-x) \geq -1.5x$ for all $0 \leq x \leq 0.5$, and the last inequality follows from the following bound (which makes use of (128))

$$\sum_{i=1}^{T} \eta_i(1-\gamma p) \leq \begin{cases} \frac{3}{4\log T}\sum_{i=1}^{T}\frac{1}{i} \leq 1, & \text{if } \eta_t = \frac{1}{1+c_\eta(1-\gamma)t} \\ \frac{4}{3T}\sum_{i=1}^{T} 1 = \frac{4}{3}, & \text{if } \eta_t = \eta. \end{cases}$$

Combining the above result with the properties (124) and (129) then yields

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\eta_0^{(T)}}{1-\gamma p} \geq \frac{e^{-2}}{1-\gamma p} = \frac{3}{4e^2(1-\gamma)}. \tag{130}$$

This taken together with (36) gives

$$\mathbb{E}\big[\big(V^\star(2) - V_T(2)\big)^2\big] \geq \big(V^\star(2) - \mathbb{E}\big[V_T(2)\big]\big)^2 \geq \frac{9}{16e^4(1-\gamma)^2}. \tag{131}$$

D.3.2 Case 2: large learning rates ($c_\eta \leq 1-\gamma$ or $\eta \geq \frac{1}{(1-\gamma)^2 T}$)

By virtue of (125), the mean gap $V^\star(2) - \mathbb{E}\big[V_T(2)\big]$ depends on two factors: (i) the choice of the learning rates, and (ii) the gap between $\frac{1}{1-\gamma p}$ and $\mathbb{E}\big[V_\tau(2)\big]$, where τ is an integer obeying $0 \leq \tau < T$. To control the factor (ii), we need to choose τ properly. Let us start by considering the simple scenario with $\mathbb{E}\big[\big(V_T(2)\big)^2\big] < \frac{1}{4(1-\gamma)^2}$, for which we have

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] \geq \frac{3}{4(1-\gamma)} - \sqrt{\mathbb{E}\big[\big(V_T(2)\big)^2\big]} \geq \frac{1}{4(1-\gamma)}. \tag{132}$$

Here, we have used (35) and the elementary fact $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$. Consequently, it remains to look at the scenario obeying $\mathbb{E}\big[\big(V_T(2)\big)^2\big] \geq \frac{1}{4(1-\gamma)^2}$, towards which we propose to set τ as follows

$$\tau := \min\left\{0 \leq \tau' \leq T-1 \;\middle|\; \mathbb{E}\big[\big(V_t(2)\big)^2\big] \geq \frac{1}{4(1-\gamma)^2} \text{ for all } \tau'+1 \leq t \leq T\right\}. \tag{133}$$

Clearly, τ is well-defined in this scenario and obeys (in view of both (133) and the initialization $V_0 = 0$)

$$\mathbb{E}\big[\big(V_\tau(2)\big)^2\big] < \frac{1}{4(1-\gamma)^2}. \tag{134}$$

Our analysis for this scenario is divided into three subcases based on the size of the learning rates.

Case 2.1. Consider the case where

$$\prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big) \geq \frac{1}{2}. \tag{135}$$

Invoke (125) to deduce that

$$\begin{aligned} V^\star(2) - \mathbb{E}\big[V_T(2)\big] &= \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\Big[\frac{1}{1-\gamma p} - \mathbb{E}\big[V_\tau(2)\big]\Big] \\ &\geq \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\Big[\frac{3}{4(1-\gamma)} - \sqrt{\mathbb{E}\big[\big(V_\tau(2)\big)^2\big]}\Big] \\ &\geq \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\frac{1}{4(1-\gamma)} \geq \frac{1}{8(1-\gamma)}, \end{aligned}$$

where the second line makes use of the definition (34) and the elementary fact $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$, and the last line relies on the inequalities (134) and (135).

Case 2.2. We now move on to the case where

$$0 \leq \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big) \leq \frac{1}{2}. \tag{136}$$

We intend to demonstrate that the variance of $V_T(2)$ — and hence the typical size of its fluctuation — is too large. In view of the observation (127), it boils down to lower bounding $\sum_{k=\tau+2}^{T}\big(\eta_k^{(T)}\big)^2$, which we accomplish as follows.

- Consider constant learning rates $\eta_k = \eta$, and suppose that η obeys $\frac{1}{(1-\gamma)^2 T} < \eta \leq 1 < \frac{1}{1-\gamma p}$. It is readily seen that $\eta_k^{(T)} = \eta\big(1-\eta(1-\gamma p)\big)^{T-k}$ for any $k \geq 1$. We claim that it suffices to focus on the scenario where

$$\tau \leq T-2. \tag{137}$$

  In fact, if $\tau \geq T-1$, then the definition (133) of τ necessarily requires that

$$\mathbb{E}\big[V_{T-1}(2)\big] \leq \sqrt{\mathbb{E}\big[\big(V_{T-1}(2)\big)^2\big]} < \frac{1}{2(1-\gamma)}.$$

  In view of (124) (with T replaced by $T-1$), a little algebra shows that this is equivalent to $\big(1-\eta(1-\gamma p)\big)^{T-1} \geq 1/3$, and hence $\big(1-\eta(1-\gamma p)\big)^{T} \geq 1/9$. In turn, this combined with (124) leads to

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\big(1-\eta(1-\gamma p)\big)^T}{1-\gamma p} = \frac{3\big(1-\eta(1-\gamma p)\big)^T}{4(1-\gamma)} \geq \frac{1}{12(1-\gamma)}, \tag{138}$$

which already suffices for our purpose.

Next, assuming that (137) holds, one can derive

$$\sum_{k=\tau+2}^{T}\big(\eta_k^{(T)}\big)^2=\sum_{k=\tau+2}^{T}\eta^2\big(1-\eta(1-\gamma p)\big)^{2(T-k)}=\frac{\eta^2\big[1-\big(1-\eta(1-\gamma p)\big)^{2(T-\tau-1)}\big]}{1-\big(1-\eta(1-\gamma p)\big)^2}$$
$$\geq\frac{\eta^2/2}{1-\big(1-\eta(1-\gamma p)\big)^2}\geq\frac{3\eta}{16(1-\gamma)},\tag{139}$$

where the first inequality holds since (from the assumptions (136) and $\tau\leq T-2$)

$$0\leq\big(1-\eta(1-\gamma p)\big)^{2(T-\tau-1)}\leq\big(1-\eta(1-\gamma p)\big)^{T-\tau}=\prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\leq\frac{1}{2},$$

and the last inequality follows since

$$0\leq 1-\big(1-\eta(1-\gamma p)\big)^2=1-\Big(1-\frac{4\eta(1-\gamma)}{3}\Big)^2\leq\frac{8\eta(1-\gamma)}{3}.$$

Substituting (139) into (127), we obtain

$$\mathsf{Var}\big(V_T(2)\big)\geq\frac{4\gamma-1}{36(1-\gamma)}\sum_{k=\tau+1}^{T}\big(\eta_k^{(T)}\big)^2\geq\frac{2}{36(1-\gamma)}\cdot\frac{3\eta}{16(1-\gamma)}$$
$$=\frac{\eta}{96(1-\gamma)^2}\geq\frac{1}{96(1-\gamma)^4T},\tag{140}$$

provided that $\gamma\geq 3/4$ (so that $4\gamma-1\geq 2$). Here, the last inequality is valid since either $\eta\geq\frac{1}{(1-\gamma)^2T}$.

- We then move on to linearly rescaled learning rates with $\eta_t=\frac{1}{1+c_\eta(1-\gamma)t}$ for some $0\leq c_\eta<1-\gamma$. Towards this, we first make the observation that

$$\frac{\eta_{k-1}^{(T)}}{\eta_k^{(T)}}=\frac{\eta_{k-1}\big(1-\eta_k(1-\gamma p)\big)}{\eta_k}=\frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-\big(\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}\big)\eta_k}=\frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k}=1-\frac{\big(\frac{4}{3}-c_\eta\big)(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k}$$
$$\leq 1-(1-\gamma)\eta_k\leq 1-(1-\gamma)\eta_T,\tag{141}$$

with the proviso that $c_\eta<1-\gamma\leq 1/3$ (as long as $\gamma\geq 2/3$). By defining $\tau'\coloneqq T-\frac{1}{(1-\gamma)\eta_T}$, one can deduce that

$$\sum_{k=\tau+2}^{T}\big(\eta_k^{(T)}\big)^2\geq\sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\big(\eta_k^{(T)}\big)^2\geq\frac{1}{T-\max\{\tau+1,\tau'\}}\left[\sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\eta_k^{(T)}\right]^2$$
$$\geq(1-\gamma)\eta_T\left[\sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\eta_k^{(T)}\right]^2,\tag{142}$$

where the penultimate inequality comes from the Cauchy-Schwarz inequality. In addition, recognizing that $\eta_{k_1}^{(T)}\leq\big(1-(1-\gamma)\eta_T\big)^{k_2-k_1}\eta_{k_2}^{(T)}$ for any $k_2\geq k_1$ (see (141)), one has

$$\sum_{k=\tau'+1}^{T}\eta_k^{(T)}=\sum_{k=\tau'+1}^{T}\eta_k^{(T)},$$
$$\sum_{k=\max\{2\tau'-T+1,1\}}^{\tau'}\eta_k^{(T)}\leq\big(1-(1-\gamma)\eta_T\big)^{T-\tau'}\sum_{k=\tau'+1}^{T}\eta_k^{(T)},$$

$$\sum_{k=\max\{3\tau'-2T+1,1\}}^{2\tau'-T} \eta_k^{(T)} \le \big(1-(1-\gamma)\eta_T\big)^{2(T-\tau')} \sum_{k=\tau'+1}^{T} \eta_k^{(T)},$$

$$\cdots$$

Summing these inequalities up and rearranging terms, we reach

$$\sum_{k=\tau'+1}^{T} \eta_k^{(T)} \ge \frac{\sum_{k=1}^{T} \eta_k^{(T)}}{1+\big(1-(1-\gamma)\eta_T\big)^{T-\tau'}+\big(1-(1-\gamma)\eta_T\big)^{2(T-\tau')}+\ldots} \ge \frac{\sum_{k=1}^{T} \eta_k^{(T)}}{1-\big(1-(1-\gamma)\eta_T\big)^{T-\tau'}}$$
$$= \Big(1-\big(1-(1-\gamma)\eta_T\big)^{T-\tau'}\Big) \sum_{k=1}^{T} \eta_k^{(T)} \ge (1-e^{-1}) \sum_{k=1}^{T} \eta_k^{(T)},$$

which relies on the fact $\big(1-(1-\gamma)\eta_T\big)^{T-\tau'} = \big(1-1/(T-\tau')\big)^{T-\tau'} \le e^{-1}$ (using the definition of τ'). Consequently, it is easily seen that

$$\sum_{k=\max\{\tau+2,\tau'+1\}}^{T} \eta_k^{(T)} = \min\left\{\sum_{k=\tau+2}^{T} \eta_k^{(T)}, \sum_{k=\tau'+1}^{T} \eta_k^{(T)}\right\} \ge (1-e^{-1}) \sum_{k=\tau+2}^{T} \eta_k^{(T)}$$
$$\overset{\text{(i)}}{=} (1-e^{-1}) \left[1-\prod_{i=\tau+2}^{t} \big(1-\eta_i(1-\gamma p)\big)\right] \frac{1}{1-\gamma p}$$
$$\overset{\text{(ii)}}{\ge} \left[1-\frac{1}{2\big(1-\eta_{\tau+1}(1-\gamma p)\big)}\right] \frac{1-e^{-1}}{1-\gamma p} \overset{\text{(iii)}}{\ge} \frac{1-e^{-1}}{4(1-\gamma p)} \ge \frac{3}{32(1-\gamma)}.$$

Here, (i) and (ii) follow from (116) and (136), respectively, while (iii) holds since

$$\eta_{\tau+1}(1-\gamma p) \le 1-\gamma p = \frac{4(1-\gamma)}{3} \le \frac{1}{3}$$

as long as $\gamma \ge 3/4$. Substitution into (142) yields

$$\sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2 \ge \frac{9\eta_T}{1024(1-\gamma)}. \tag{143}$$

Substituting the above bound into (127), we obtain

$$\mathsf{Var}\big(V_T(2)\big) \ge \frac{4\gamma-1}{36(1-\gamma)} \sum_{k=\tau+1}^{T} \big(\eta_k^{(T)}\big)^2 \ge \frac{2}{36(1-\gamma)} \cdot \frac{9\eta_T}{1024(1-\gamma)}$$
$$= \frac{\eta_T}{2048(1-\gamma)^2} \ge \frac{1}{4096(1-\gamma)^4 T}, \tag{144}$$

provided that $\gamma \ge 3/4$ (so that $4\gamma-1 \ge 2$). Here, the last inequality is valid since $\eta_T = \frac{1}{1+c_\eta(1-\gamma)T} \ge \frac{1}{1+(1-\gamma)^2 T} \ge \frac{1}{2(1-\gamma)^2 T}$ as long as $T \ge \frac{1}{(1-\gamma)^2}$.

Putting all this together. With the above bounds in place, it is readily seen that either the bias is too large (see (138)) or the variance is too large (see (140) and (144)). These bounds taken collectively with (36) yield

$$\mathbb{E}\big[\big(V^\star(2)-V_T(2)\big)^2\big] \ge \big(V^\star(2)-\mathbb{E}\big[V_T(2)\big]\big)^2 + \mathsf{Var}\big(V_T(2)\big)$$
$$\ge \min\left\{\frac{1}{144(1-\gamma)^2}, \frac{1}{96(1-\gamma)^4 T}, \frac{1}{4096(1-\gamma)^4 T}\right\} = \frac{1}{4096(1-\gamma)^4 T}, \tag{145}$$

provided $T \ge \frac{1}{(1-\gamma)^2}$.

D.3.3 Case 3: medium learning rates ($1-\gamma < c_\eta < \log T$ or $\frac{1}{(1-\gamma)T} \le \eta \le \frac{1}{(1-\gamma)^2 T}$)

Throughout this case, we assume that

$$\eta_0^{(T)} \le \frac{1}{75}. \tag{146}$$

In fact, if $\eta_0^{(T)} > 1/75$, then the scenario becomes much easier to cope with. To see this, applying the previous result (130) and recalling the choice (34) of p immediately yield

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] \ge \frac{\eta_0^{(T)}}{1-\gamma p} > \frac{1}{100(1-\gamma)}, \tag{147}$$

which together with (36) and the assumption $T \ge \frac{1}{(1-\gamma)^2}$ yields

$$\mathbb{E}\big[\big(V^\star(2) - V_T(2)\big)^2\big] \ge \big(V^\star(2) - \mathbb{E}\big[V_T(2)\big]\big)^2 \ge \frac{1}{10000(1-\gamma)^2} \ge \frac{1}{10000(1-\gamma)^4 T}. \tag{148}$$

We now turn our attention to the dynamics w.r.t. state 1 and its associated value function $V_t(1)$ under the condition (146).

Two auxiliary sequences. Towards this, we first eliminate the effect of initialization on $Q_t(1,a)$ by introducing the following auxiliary sequence

$$\widehat{Q}_t(a) = (1-\eta_t)\widehat{Q}_{t-1}(a) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,a)\widehat{V}_{t-1}\big\}, \tag{149}$$

with

$$\widehat{V}_{t-1} \coloneqq \max_a \widehat{Q}_t(a) \qquad \text{and} \qquad \widehat{Q}_0(a) \coloneqq Q^\star(1,a) = \frac{1}{1-\gamma p},$$

where we recall the value of $Q^\star(1,a)$ from Lemma 3. In other words, $\{\widehat{Q}_t(a)\}$ is essentially a Q-learning sequence when initialized at the ground truth. Despite the difference in initialization, we claim that the discrepancy between $\widehat{Q}_t(a)$ and $Q_t(1,a)$ can be well controlled in the following sense:

$$Q_t(1,a) \ge \widehat{Q}_t(a) - \frac{1}{1-\gamma}\prod_{i=1}^{t}\big(1-\eta_i(1-\gamma)\big), \qquad a \in \{1,2\}, \tag{150}$$

which shall be justified in Section D.3.4. As we shall discuss momentarily, the gap $\frac{1}{1-\gamma}\prod_{i=1}^{t}\big(1-\eta_i(1-\gamma)\big)$ is sufficiently small for this case.

Further, in order to control $\widehat{Q}_t(a)$, we find it convenient to introduce another auxiliary sequence as follows

$$\overline{Q}_t = (1-\eta_t)\overline{Q}_{t-1} + \eta_t\big\{1 + \gamma P_t(1\,|\,1,1)\overline{Q}_{t-1}\big\} \qquad \text{and} \qquad \overline{Q}_0 = V^\star(1) = \frac{1}{1-\gamma p}, \tag{151}$$

which can be interpreted as a Q-learning sequence when there is only a single action (so that there is no max operator involved). In view of the basic fact that $\widehat{V}_t = \max_a \widehat{Q}_t(a) \ge \widehat{Q}_t(1)$, we can easily verify that

$$\widehat{Q}_t(1) \ge (1-\eta_t)\widehat{Q}_{t-1}(1) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,1)\widehat{Q}_{t-1}(1)\big\} \ge \overline{Q}_t, \tag{152}$$

allowing one to lower bound $\widehat{V}_t$ by controlling $\overline{Q}_t$.

A useful lower bound on the auxiliary sequence (149). In what follows, let us establish a useful lower bound on the sequence (149) introduced above. Then we claim that there exists some $\tau \le T$ (see (166) and (168)) such that

$$\mathbb{P}\Big\{\widehat{V}_t \ge \frac{1}{4(1-\gamma)}\Big\} \ge \frac{1}{2}, \qquad \text{for } t \ge \tau. \tag{153}$$

The auxiliary sequence constructed in (151) plays a crucial role in establishing this claim.

Proof of the claim (153). We intend to employ the sequence $\overline{Q}_t$ (cf. (151)) to help control $\widehat{V}_t$. It is first observed that the sequence $\overline{Q}_t$ admits the following decomposition (akin to the derivation in (122))

$$\begin{aligned}
\overline{Q}_t &= \big(1-\eta_t(1-\gamma p)\big)\overline{Q}_{t-1} + \eta_t\big\{1+\gamma\big(P_t(1\,|\,1,1)-p\big)\overline{Q}_{t-1}\big\} \\
&= \prod_{i=1}^{t}\big(1-\eta_i(1-\gamma p)\big)\overline{Q}_0 + \sum_{k=1}^{t}\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i(1-\gamma p)\big)\big\{1+\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1}\big\} \\
&= \eta_0^{(t)}\frac{1}{1-\gamma p} + \sum_{k=1}^{t}\eta_k^{(t)} + \sum_{k=1}^{t}\eta_k^{(t)}\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1} \\
&= \frac{1}{1-\gamma p} + \sum_{k=1}^{t}\underbrace{\eta_k^{(t)}\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1}}_{=:\,z_k}, \qquad (154)
\end{aligned}$$

where the last line results from (115). In order to lower bound $\overline{Q}_t$, it boils down to controlling $\sum_k z_k$.

Note that the sequence $\{z_k\}$ defined above is a martingale satisfying

$$\mathbb{E}\big[z_k\,|\,P_{k-1}(1\,|\,1,1),\ldots,P_1(1\,|\,1,1)\big]=0$$
$$\text{and}\qquad |z_k| \le \max_{1\le k\le t}\eta_k^{(t)}\cdot\frac{\gamma p}{1-\gamma},$$

where the last inequality follows from the basic property $0\le\overline{Q}_{k-1}\le\frac{1}{1-\gamma}$ (akin to Lemma 4) and the fact that $\big|P_k(1\,|\,1,1)-p\big|\le\max\{p,1-p\}=p$ since $p=(4\gamma-1)/(3\gamma)$ and $\gamma\ge 3/4$. We intend to invoke Freedman's inequality to control (154). Armed with these properties and the fact that $P_k(1\,|\,1,1)$ is a Bernoulli random variable with mean p, we obtain

$$\begin{aligned}
\sum_{k=1}^{t}\mathsf{Var}\Big(z_k\,|\,P_{k-1}(1\,|\,1,1),\ldots,P_1(1\,|\,1,1)\Big) &= \sum_{k=1}^{t}\big(\eta_k^{(t)}\big)^2 p(1-p)\big(\gamma\overline{Q}_{k-1}\big)^2 \\
&\le \max_{1\le k\le t}\eta_k^{(t)}\cdot\sum_{k=1}^{t}\eta_k^{(t)}\cdot\frac{1}{3(1-\gamma)} \le \frac{\max_{1\le k\le t}\eta_k^{(t)}}{4(1-\gamma)^2}.
\end{aligned}$$

Here, the penultimate inequality relies on the fact $0\le\overline{Q}_{k-1}\le\frac{1}{1-\gamma}$ (akin to Lemma 4) and the choice of p (see definition (34)), whereas the last inequality results from the following condition (derived through (115))

$$\sum_{k=1}^{t}\eta_k^{(t)} = \big(1-\eta_0^{(t)}\big)\frac{1}{1-\gamma p} \le \frac{1}{1-\gamma p} = \frac{3}{4(1-\gamma)}.$$

Applying Freedman's inequality (see (47)) then yields

$$\mathbb{P}\left\{\Big|\sum_{k=1}^{t}z_k\Big| \ge \sqrt{\frac{4\max_{1\le k\le t}\eta_k^{(t)}}{(1-\gamma)^2}\log\frac{2}{\delta}} + \frac{4\max_{1\le k\le t}\eta_k^{(t)}}{3(1-\gamma)}\log\frac{2}{\delta}\right\} \le \delta. \qquad (155)$$

As an implication of the preceding result, a key ingredient towards bounding $\sum_{k=1}^{t}z_k$ lies in controlling the quantity $\max_{1\le k\le t}\eta_k^{(t)}$. To do so, we claim for the moment that there exists some $\tau\le T$ such that

$$\max_{1\le k\le t}\eta_k^{(t)} \le \frac{1}{50}, \qquad \text{for } t\ge\tau, \qquad (156)$$

whose proof is postponed to Section D.3.4. In light of this claim, setting $\delta=1/2$ in the expression (155) yields

$$\sum_{k=1}^{t}z_k \ge -\frac{1}{2(1-\gamma)}$$

with probably at least $1/2$. Combining this with the decomposition (154) and the property (152), we arrive at

$$\widehat{V}_t \geq \widehat{Q}_t(1) \geq \overline{Q}_t \geq \frac{1}{1-\gamma p} - \frac{1}{2(1-\gamma)} = \frac{1}{4(1-\gamma)}$$

with probability at least $1/2$, where the last identity relies on the choice of p (see the definition (34)). This establishes the advertised claim (153). □

Main proof. With the property (153) in place, we are positioned to prove our main result. Towards this, we find it convenient to define

$$\Delta_t(a) \coloneqq \widehat{Q}_t(a) - Q^\star(1,a), \qquad a = 1,2; \tag{157a}$$

$$\Delta_{t,\mathsf{max}} \coloneqq \max_a \Delta_t(a). \tag{157b}$$

The goal is thus to control $\Delta_{T,\mathsf{max}}$; in fact, we intend to show that $\Delta_{T,\mathsf{max}}$ is in expectation excessively large, resulting in an "over-estimation" issue that hinders convergence. Towards this, it follows from the iterative update rule (149) that

$$\begin{aligned}
\Delta_t(a) &= (1-\eta_t)\Delta_{t-1}(a) + \eta_t\big(1 + \gamma P_t(1\,|\,1,a)\widehat{V}_{t-1} - Q^\star(1,a)\big) \\
&= (1-\eta_t)\Delta_{t-1}(a) + \eta_t\gamma\big(P_t(1\,|\,1,a)\widehat{V}_{t-1} - pV^\star(1)\big) \\
&= (1-\eta_t)\Delta_{t-1}(a) + \eta_t\gamma\big(p\big(\widehat{V}_{t-1} - V^\star(1)\big) + \big(P_t(1\,|\,1,a) - p\big)\widehat{V}_{t-1}\big) \\
&= (1-\eta_t)\Delta_{t-1}(a) + \eta_t\gamma\Big(p\Delta_{t-1,\mathsf{max}} + \big(P_t(1\,|\,1,a) - p\big)\widehat{V}_{t-1}\Big).
\end{aligned}$$

Here, the second line comes from the Bellman equation $Q^\star(1,a) = 1 + \gamma p V^\star(1)$, whereas the last line holds since $\widehat{V}_{t-1} - V^\star(1) = \max_a\big(\widehat{Q}_{t-1}(a) - V^\star(1)\big) = \max_a \Delta_{t-1}(a)$ (as a consequence of the relation (35)). Applying the above relation recursively leads to

$$\Delta_t(a) = \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma\Big(p\Delta_{k-1,\mathsf{max}} + \big(P_k(1\,|\,1,a) - p\big)\widehat{V}_{k-1}\Big), \tag{158}$$

where we have used the initialization $\Delta_0(a) = 0$. Letting

$$\xi_t(a) \coloneqq \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma\big(P_k(1\,|\,1,a) - p\big)\widehat{V}_{k-1}, \tag{159a}$$

$$\xi_{t,\mathsf{max}} \coloneqq \max_a \xi_t(a), \tag{159b}$$

one can express the above relation as follows

$$\Delta_{t,\mathsf{max}} = \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma p\Delta_{k-1,\mathsf{max}} + \xi_{t,\mathsf{max}}.$$

Next, we claim that $\mathbb{E}[\xi_{t,\mathsf{max}}]$ satisfies the following property

$$\mathbb{E}[\xi_{t,\mathsf{max}}] \geq \frac{c}{\sqrt{(1-\gamma)^2 T\log T}} \qquad \text{for all } t \geq \widehat{\tau} \tag{160}$$

for some universal constant $c > 0$, where

$$\widehat{\tau} \coloneqq \max\left\{\tau' \;\middle|\; \prod_{i=\tau'}^{T}\big(1 - \eta_i(1-\gamma p)\big) \leq \frac{6}{7}\right\}, \tag{161}$$

whose existence is ensured under the condition (146). Given the validity of this claim (which we shall justify in Section D.3.4), we immediately arrive at

$$\mathbb{E}[\Delta_{t,\mathsf{max}}] \geq \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma p \mathbb{E}[\Delta_{k-1,\mathsf{max}}] + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \qquad \text{for all } t \geq \widehat{\tau}. \tag{162}$$

In order to study the above recursion, it is helpful to look at the following sequence

$$x_t = (1-\eta_t)x_{t-1} + \eta_t\left(\gamma p x_{t-1} + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}\right) \tag{163}$$

with $x_{\widehat{\tau}} = 0$, where we recall the definition of $\widehat{\tau}$ in (161). In comparison to the iterative relation (162) which starts from $\mathbb{E}[\Delta_{0,\mathsf{max}}] = 0$ (and hence $\mathbb{E}[\Delta_{t,\mathsf{max}}] \geq 0$), we let the sequence x_t start from $x_{\widehat{\tau}} = 0$, where $\widehat{\tau}$ is defined in (161). It is straightforward to verify that

$$\mathbb{E}\big[\Delta_{T,\mathsf{max}}\big] \geq x_T, \tag{164}$$

recognizing that

$$x_t = \sum_{k=\widehat{\tau}}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma p x_{k-1} + \sum_{k=\widehat{\tau}}^{T} \eta_k \prod_{i=k+1}^{T} (1-\eta_i) \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}.$$

A little algebra reveals that the sequence (163) obeys

$$\begin{aligned} x_T &= \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \sum_{k=\widehat{\tau}}^{T} \eta_k \prod_{i=k+1}^{T} \big(1-\eta_i(1-\gamma p)\big) = \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \frac{1}{1-\gamma p}\left[1 - \prod_{i=\widehat{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big)\right] \\ &= \frac{3c}{4\sqrt{(1-\gamma)^4 T \log T}}\left[1 - \prod_{i=\widehat{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big)\right] \geq \frac{3c}{28\sqrt{(1-\gamma)^4 T \log T}}, \end{aligned}$$

where the second equality arises from (116), and the last inequality holds as long as $\prod_{i=\widehat{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big) \leq 6/7$ (see (176)). This taken together with (164) leads to

$$\mathbb{E}\big[\Delta_{T,\mathsf{max}}\big] \geq x_T \geq \frac{3c}{28\sqrt{(1-\gamma)^4 T \log T}}.$$

Combining the above bound with (150) leads to

$$\begin{aligned} \mathbb{E}\Big[V_T(1) - V^\star(1)\Big] &\geq \mathbb{E}\Big[\Delta_{T,\mathsf{max}} - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big)\Big] \\ &\geq \frac{3c}{28\sqrt{(1-\gamma)^4 T \log T}} - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big). \end{aligned}$$

Taking this together with (123), we arrive at

$$\begin{aligned} &\max\Big\{\mathbb{E}\Big[\big|V_T(3) - V^\star(3)\big|\Big],\ \mathbb{E}\Big[\big|V_T(1) - V^\star(1)\big|\Big]\Big\} \\ &\quad\geq \max\left\{\frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big),\ \frac{3c}{28\sqrt{(1-\gamma)^4 T \log T}} - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big)\right\} \\ &\quad\geq \frac{1}{2}\cdot\frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big) + \frac{1}{2}\left[\frac{3c}{28\sqrt{(1-\gamma)^4 T \log T}} - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big)\right] \\ &\quad= \frac{3c}{56\sqrt{(1-\gamma)^4 T \log T}}. \end{aligned}$$

This combined with (36) establishes the following desired lower bound:

$$\max_s \mathbb{E}\Big[\big|V_T(s) - V^\star(s)\big|^2\Big] \geq \left(\frac{3c}{56\sqrt{(1-\gamma)^4 T\log T}}\right)^2 = \frac{9c^2}{56^2(1-\gamma)^4 T\log^2 T}.$$

D.3.4 Proofs of auxiliary results

Proof of the inequality (150). We shall establish this claim by induction. To begin with, the inequality (150) holds trivially for the base case with $t = 0$. Now, let us assume that the claim holds up to the $(t-1)$-th iteration, and we would like to justify it for the t-th iteration. As an immediate consequence of the claim (150) for the $(t-1)$-th iteration and the definitions of V_{t-1} and $\widehat{V}_{t-1}$, we have

$$\begin{aligned} V_{t-1}(1) = \max_a Q_{t-1}(1,a) &\geq \max_a \widehat{Q}_{t-1}(a) - \frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big) \\ &= \widehat{V}_{t-1} - \frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big). \end{aligned}$$

By virtue of the respective update rules of $Q_t(1,a)$ and $\widehat{Q}_t(a)$, we can express their difference as follows:

$$\begin{aligned} Q_t(1,a) - \widehat{Q}_t(a) &= (1-\eta_t)\big(Q_{t-1}(1,a) - \widehat{Q}_{t-1}(a)\big) + \eta_t\gamma P_t(1\,|\,1,a)\big(V_{t-1}(1) - \widehat{V}_{t-1}\big) \\ &\geq -(1-\eta_t)\frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big) - \eta_t\gamma P_t(1\,|\,1,a)\frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big) \\ &\geq -(1-\eta_t)\frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big) - \eta_t\gamma\frac{1}{1-\gamma}\prod_{i=1}^{t-1}\big(1-\eta_i(1-\gamma)\big) \\ &= -\frac{1}{1-\gamma}\prod_{i=1}^{t}\big(1-\eta_i(1-\gamma)\big), \end{aligned}$$

where the first inequality invokes the induction hypothesis for the $(t-1)$-th iteration. This establishes (150) for the t-th iteration, and hence the proof is complete via an induction argument.

Proof of the claim (156). When taking the constant learning rates $\eta_t \equiv \eta \leq \frac{1}{(1-\gamma)^2 T} \leq \frac{1}{50}$ (under the condition $T \geq \frac{50}{(1-\gamma)^2}$), one has

$$\max_{1\leq k\leq t}\eta_k^{(t)} \leq \eta_t = \eta \leq \frac{1}{50},$$

thus allowing us to take $\tau = 1$ for this case.

It then suffices to look at rescaled linear learning rates (i.e., $\eta_t = \frac{1}{1+c_\eta(1-\gamma)t}$). As already calculated in the expression (141), the ratio of two consecutive quantities obeys

$$\frac{\eta_{k-1}^{(t)}}{\eta_k^{(t)}} = \frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k}. \tag{165}$$

In what follows, we divide into two cases, depending on whether this sequence is decreasing or increasing.

- *The case with* $4/3 \leq c_\eta < \log T$. In this scenario, the ratio in (165) is larger than 1, and hence the sequence $\{\eta_k^{(t)}\}$ decreases with k. Let us define

$$\tau := \min\left\{\tau' \;\Big|\; \prod_{i=1}^{\tau'}\big(1-\eta_i(1-\gamma p)\big) \leq \frac{1}{50}\right\}, \tag{166}$$

which clearly satisfies $\tau \le T$ (in view of (146)). For all $t \ge \tau$, one has

$$\max_{1\le k\le t} \eta_k^{(t)} \le \prod_{i=1}^{\tau} \big(1-\eta_i(1-\gamma p)\big) \le \frac{1}{50}.$$

At the same time, we claim that one must have

$$\prod_{i=\tau}^{T} \big(1-\eta_i(1-\gamma p)\big) \le \frac{2}{3}. \tag{167}$$

Otherwise, recalling $\eta_0^{(T)} = \prod_{i=1}^{T} \big(1-\eta_i(1-\gamma p)\big)$, we have

$$\eta_0^{(T)} = \left\{\prod_{i=1}^{\tau-1} \big(1-\eta_i(1-\gamma p)\big)\right\}\left\{\prod_{i=\tau}^{T} \big(1-\eta_i(1-\gamma p)\big)\right\} > \frac{1}{50}\cdot\frac{2}{3} = \frac{1}{75},$$

which contradicts our assumption that $\eta_0^{(T)} > 1/75$ (cf. (146)).

- *The case with* $1-\gamma < c_\eta < 4/3$. In this case, the sequence $\eta_k^{(t)}$ increases with k. If we set

$$\tau \coloneqq \left\lceil \frac{49}{c_\eta(1-\gamma)} \right\rceil < \frac{50}{(1-\gamma)^2} < T, \tag{168}$$

then for all $t \ge \tau$ we have

$$\max_{1\le k\le t} \eta_k^{(t)} = \eta_t^{(t)} = \eta_t \le \eta_\tau \le \frac{1}{1+c_\eta(1-\gamma)\frac{49}{c_\eta(1-\gamma)}} = \frac{1}{50}.$$

Under the condition $T \ge \frac{150}{(1-\gamma)^2} \ge \frac{150}{c_\eta(1-\gamma)}$ (so that $T-\tau+1 \ge \frac{100}{c_\eta(1-\gamma)} \ge \frac{100}{(1-\gamma)4/3}$), one can show that

$$\prod_{i=\tau}^{T} \big(1-\eta_i(1-\gamma p)\big) \le \Big(1-\frac{1-\gamma}{100}\Big)^{T-\tau+1} \le \Big(1-\frac{1-\gamma}{100}\Big)^{\frac{100}{(1-\gamma)4/3}} \le \frac{3}{4}. \tag{169}$$

- Putting these two cases together (with τ specified in (166) and (168)), we obtain

$$\max_{1\le k\le t} \eta_k^{(t)} \le \frac{1}{50} \tag{170}$$

for all $t \ge \tau$, thus establishing the desired inequality (156).

Proof of the inequality (160). For every t, recalling the definition (159), it is convenient to write

$$\begin{aligned}\mathbb{E}[\xi_{t,\mathsf{max}}] &= \mathbb{E}\left[\frac{\xi_t(1)+\xi_t(2)+|\xi_t(1)-\xi_t(2)|}{2}\right] = \mathbb{E}\left[\frac{|\xi_t(1)-\xi_t(2)|}{2}\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left|\sum_{k=1}^{t}\eta_k \prod_{i=k+1}^{t}(1-\eta_i)\gamma\big(P_k(1\,|\,1,1)-P_k(1\,|\,1,2)\big)\widehat{V}_{k-1}\big)\right|\right],\end{aligned}$$

where we have used the fact that $\mathbb{E}[\xi_t(a)] = 0$. To control the right-hand side of the above equation, let us define

$$\zeta_t \coloneqq \sum_{k=1}^{t} z_k, \qquad z_k \coloneqq \eta_k \prod_{i=k+1}^{t}(1-\eta_i)\gamma\big(P_k(1\,|\,1,1)-P_k(1\,|\,1,2)\big)\widehat{V}_{k-1}$$

for any $k \ge 1$, where $\{z_k\}$ also forms a martingale sequence since

$$\mathbb{E}\Big[z_k \,|\, \big\{P_j(1\,|\,1,1), P_j(1\,|\,1,2)\big\}_{1\le j<k}\Big] = 0.$$

As a consequence of Freedman's inequality, we claim that ζ_t satisfies

$$\mathbb{P}\left\{|\zeta_t| \geq \sqrt{\frac{8\log\frac{2}{\delta}}{3(1-\gamma)}\sum_{k=1}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}(1-\eta_i)\Big]^2} + \frac{4\eta_t\log\frac{2}{\delta}}{3(1-\gamma)}\right\} \leq \delta. \tag{171}$$

To verify this relation, we first notice that

$$|z_k| \leq \max_{1\leq k\leq t}\eta_k \prod_{i=k+1}^{t}(1-\eta_i)\cdot\frac{1}{1-\gamma} \leq \frac{\eta_t}{1-\gamma}, \tag{172}$$

provided that $\max_k \eta_k\prod_{i=k+1}^{t}(1-\eta_i) \leq \eta_t$. To verify the condition $\max_k \eta_k\prod_{i=k+1}^{t}(1-\eta_i) \leq \eta_t$, one can check — similar to (141) — that

$$\frac{\eta_{k-1}\prod_{i=k}^{t}(1-\eta_i)}{\eta_k\prod_{i=k+1}^{t}(1-\eta_i)} = 1 - \frac{\big(1-c_\eta(1-\gamma)\big)\eta_k}{1-c_\eta(1-\gamma)\eta_k} \leq 1, \tag{173}$$

which indicates that $\eta_k\prod_{i=k+1}^{t}(1-\eta_i)$ is an increasing sequence as long as $c_\eta \leq \log T \leq \frac{1}{1-\gamma}$ (see (112)). In addition to the boundedness condition (172), we can further calculate

$$\begin{aligned}
&\sum_{k=1}^{t}\mathsf{Var}\big(z_k \,|\, P_{k-1}(1\,|\,1,1), P_{k-1}(1\,|\,1,2), \ldots, P_1(1\,|\,1,1), P_1(1\,|\,1,2)\big) \\
&\quad = \sum_{k=1}^{t}\eta_k^2\left[\prod_{i=k+1}^{t}(1-\eta_i)\right]^2\cdot 2p(1-p)\cdot\big(\gamma\widehat{V}_{k-1}\big)^2 \leq \sum_{k=1}^{t}\eta_k^2\left[\prod_{i=k+1}^{t}(1-\eta_i)\right]^2\cdot\frac{2}{3(1-\gamma)},
\end{aligned}$$

where the last inequality comes from the facts that $\widehat{V}_{k-1} \leq \frac{1}{1-\gamma}$ and the choice $p = \frac{4\gamma-1}{3\gamma}$. These bounds taken together with Freedman's inequality (see (47)) validate (171).

By virtue of (171), setting $\delta = \frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big]$ yields that with probability at least $1-\delta$,

$$|\zeta_t| \leq B \coloneqq \sqrt{\frac{8\log\frac{2}{\delta}}{3(1-\gamma)}\sum_{k=1}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}(1-\eta_i)\Big]^2} + \frac{4\eta_t\log\frac{2}{\delta}}{3(1-\gamma)} \quad \text{with } \delta = \frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big]. \tag{174}$$

When $T \geq \frac{1}{(1-\gamma)^2}$, one can ensure that

$$\begin{aligned}
\mathbb{E}[\xi_{t,\mathsf{max}}] &= \frac{1}{2}\mathbb{E}\big[|\zeta_t|\big] \geq \frac{1}{2}\mathbb{E}\big[|\zeta_t|\mathbb{1}\big(|\zeta_t| \leq B\big)\big] \geq \frac{1}{2B}\mathbb{E}\big[|\zeta_t|^2\mathbb{1}\big(|\zeta_t| \leq B\big)\big] \\
&= \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \mathbb{E}\big[|\zeta_t|^2\mathbb{1}\big(|\zeta_t| > B\big)\big]\Big\} \\
&\overset{\text{(i)}}{\geq} \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \frac{1}{(1-\gamma)^2}\mathbb{P}\big\{|\zeta_t| > B\big\}\Big\} \\
&\geq \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \frac{\delta}{(1-\gamma)^2}\Big\} \overset{\text{(ii)}}{\geq} \frac{1}{4B}\mathbb{E}[|\zeta_t|^2].
\end{aligned} \tag{175}$$

Here, (i) holds since

$$|\zeta_t| \leq \sum_{k=1}^{t}|z_k| \leq \left[\sum_{k=1}^{t}\eta_k\prod_{i=k+1}^{t}(1-\eta_i)\right]\cdot\frac{1}{1-\gamma} \leq \frac{1}{1-\gamma}$$

as a consequence of (172) and (58); (ii) holds by the choice of δ. It is thus sufficient to lower bound $\mathbb{E}[|\zeta_t|^2]$. Towards this, let us define

$$\widehat{\tau} \coloneqq \max\left\{\tau' \;\Big|\; \prod_{i=\tau'}^{T}\big(1-\eta_i(1-\gamma p)\big) \leq \frac{6}{7}\right\}, \tag{176}$$

which clearly satisfies $\tau \le \widehat{\tau} \le T$ (in view of (167) and (169)). Then, for all $t \ge \widehat{\tau}$ one has (which shall be proved towards the end of this subsection)

$$\sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \ge \frac{1}{8} \sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2. \tag{177}$$

We now proceed to lower bound $\mathbb{E}[|\zeta_t|^2]$ for $t \ge \widehat{\tau}$. We first observe that for any $t \ge \widehat{\tau}$,

$$\begin{aligned}
\mathbb{E}\big[|\zeta_t|^2\big] &\ge \sum_{k=1}^{t} \mathbb{E}\left[\mathsf{Var}\Big(\eta_k \prod_{i=k+1}^{t} (1-\eta_i) \gamma \big(P_k(1 \,|\, 1,1) - P_k(1 \,|\, 1,2) \big) \widehat{V}_{k-1} \,|\, \widehat{V}_{k-1} \Big) \right] \\
&\ge \frac{1}{2} \sum_{k=\tau}^{t} \mathbb{E}\left[\mathsf{Var}\Big(\eta_k \prod_{i=k+1}^{t} (1-\eta_i) \gamma \big(P_k(1 \,|\, 1,1) - P_k(1 \,|\, 1,2) \big) \widehat{V}_{k-1} \,|\, \widehat{V}_{k-1} \ge \frac{1}{4(1-\gamma)} \Big) \right],
\end{aligned}$$

where the first line relies on (126), and the last step makes use of the fact (153). To further control the right-hand side of the above inequality, we take $\tau' \coloneqq \max\Big\{ t - \frac{1}{\eta_{t/2}}, 1 \Big\}$ and show that

$$\begin{aligned}
\mathbb{E}\big[|\zeta_t|^2\big] &\overset{\text{(i)}}{\ge} \frac{1}{2} \sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \gamma^2 \cdot 2p(1-p) \frac{1}{16(1-\gamma)^2} \\
&\ge \frac{1}{48(1-\gamma)} \sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \overset{\text{(ii)}}{\ge} \frac{1}{400(1-\gamma)} \sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \\
&\ge \frac{1}{400(1-\gamma)} \sum_{k=\tau'}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \overset{\text{(iii)}}{\ge} \frac{\eta_t}{9600(1-\gamma)}.
\end{aligned} \tag{178}$$

Here, (i) makes use of the constraint $\widehat{V}_{k-1} \ge \frac{1}{4(1-\gamma)}$, while (ii) makes use of (177), and (iii) are valid if the following property holds (which shall be proved towards the end of this subsection)

$$\sum_{k=\tau'}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2 \ge \frac{1}{24} \eta_t. \tag{179}$$

We are now well-equipped to control $\mathbb{E}[\xi_{t,\mathsf{max}}]$ using the property (175). Recall the expression of B in (174), we know that bounding $\mathbb{E}[|\zeta_t|^2]/B$ boils down to controlling

$$\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta_t}{1-\gamma} \log \frac{2}{\delta}} \qquad \text{and} \qquad \frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log \frac{2}{\delta}}{1-\gamma} \sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2}}. \tag{180}$$

- For the first term in (180), recalling that $\delta = \frac{(1-\gamma)^2}{2} \mathbb{E}\big[|\zeta_t|^2\big]$, we can demonstrate that

$$\log \frac{1}{\delta} = -\log \frac{(1-\gamma)^2}{2} \mathbb{E}\big[|\zeta_t|^2\big] \le -\log \frac{(1-\gamma)\eta_t}{19200} \le \log \frac{19200(1+(1-\gamma)T\log T)}{1-\gamma} \lesssim \log T, \tag{181}$$

  where the first inequality makes use of the bound (178), and the second inequality arises from the fact $\eta_t \ge \frac{1}{1+(1-\gamma)T\log T}$ (given the range of the learning rates in this case). Combining this with (178), we can guarantee that

$$\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta_t}{1-\gamma} \log \frac{2}{\delta}} \gtrsim \frac{1}{\log T}.$$

- Moving to the second term in (180), one can ensure that

$$\frac{\mathbb{E}\big[|\zeta_t|^2\big]}{\sqrt{\frac{\log \frac{2}{\delta}}{1-\gamma} \sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2}} \overset{\text{(i)}}{\gtrsim} \sqrt{\frac{1}{(1-\gamma)\log T} \sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i) \Big]^2}$$

$$\overset{\text{(ii)}}{\gtrsim} \sqrt{\frac{\eta_t}{(1-\gamma)\log T}} \overset{\text{(iii)}}{\gtrsim} \frac{1}{\sqrt{(1-\gamma)^2 T \log T}}.$$

Here, (i) follows from (178) and (181) since

$$\mathbb{E}\big[|\zeta_t|^2\big] \gtrsim \frac{1}{1-\gamma}\sum_{k=1}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2 \qquad \text{and} \qquad \log\frac{2}{\delta} \lesssim \log T;$$

(ii) arises from (179); and (iii) relies on the fact $\eta_t \gtrsim \frac{1}{(1-\gamma)T\log T}$ (given the range of the learning rates in this case).

Substituting the above relations into (175) and using the expression of B in (174), we reach at

$$\mathbb{E}[\xi_{t,\mathsf{max}}] \geq \frac{1}{4B}\mathbb{E}\big[|\zeta_t|^2\big] \geq \frac{c}{\sqrt{(1-\gamma)^2 T\log T}},$$

for some constant $c>0$. Thus, this validates the inequality (160).

Proof of the claim (177). By the definition of $\widehat{\tau}$ in (176), we have $\prod_{i=\widehat{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big) \leq 6/7$. An important observation is that

$$\sum_{k=\tau}^{t}\eta_k\left[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\right] \overset{\text{(i)}}{=} 1-\prod_{i=\tau}^{t}\big(1-\eta_i\big) \overset{\text{(ii)}}{\geq} \frac{1}{8} \overset{\text{(iii)}}{\geq} \frac{1}{8}\sum_{k=1}^{t}\eta_k\left[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\right]. \tag{182}$$

Here, the relations (i) and (iii) arise from (59), and the inequality (ii) follows since

$$\prod_{i=\tau}^{t}\big(1-\eta_i\big) \leq \prod_{i=\tau}^{\widehat{\tau}}\big(1-\eta_i\big) \leq \prod_{i=\tau}^{\widehat{\tau}}\big(1-\eta_i(1-\gamma p)\big) = \frac{\prod_{i=\tau}^{T}\big(1-\eta_i(1-\gamma p)\big)}{\prod_{i=\widehat{\tau}+1}^{T}\big(1-\eta_i(1-\gamma p)\big)} \leq \frac{3/4}{6/7} \leq \frac{7}{8}, \tag{183}$$

where τ is defined in (166) and (168) for linearly rescaled learning rates and $\tau=1$ for constant learning rates, and we have also made use of (146), (167) and (169) in the penultimate inequality in (183).

With (182) in place, we can continue to prove the claim (177). Recognizing that $\eta_k\big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\big]$ is increasing in k (see (173)), we can obtain

$$\begin{aligned}
\sum_{k=1}^{\tau-1}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2 &\leq \max_{1\leq k<\tau}\left\{\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]\right\}\sum_{k=1}^{\tau-1}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big] \\
&\leq \eta_\tau\Big[\prod_{i=\tau+1}^{t}\big(1-\eta_i\big)\Big]\sum_{k=1}^{\tau-1}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big] \\
&\leq 7\eta_\tau\Big[\prod_{i=\tau+1}^{t}\big(1-\eta_i\big)\Big]\sum_{k=\tau}^{t}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big],
\end{aligned} \tag{184}$$

where the last inequality comes from (182). With the preceding inequality in place, the claim (177) then follows by observing that

$$\begin{aligned}
\sum_{k=\tau}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2 &\geq \min_{\tau\leq k\leq t}\left\{\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]\right\}\sum_{k=\tau}^{t}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big] \\
&\geq \eta_\tau\Big[\prod_{i=\tau+1}^{t}\big(1-\eta_i\big)\Big]\sum_{k=\tau}^{t}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big],
\end{aligned}$$

where we make use of the monotonicity of $\eta_k\big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\big]$ again. □

Proof of the claim (179). Note that for $\tau' \coloneqq \max\Big\{t - \frac{1}{\eta_{t/2}}, 1\Big\}$, one has

$$\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big] \geq \eta_t\big(1-\eta_{t/2}\big)^{t-\tau'} \geq \eta_t\big(1-\eta_{t/2}\big)^{1/\eta_{t/2}} \geq \frac{1}{3}\eta_t, \qquad \text{for all } \tau' \leq k \leq t,$$

as long as the following condition holds (recalling the definition of $\widehat{\tau}$ in (176))

$$\eta_{t/2} \leq 2\eta_t \leq 2\eta_{\widehat{\tau}} \leq 1/10. \tag{185}$$

In addition, similar to (116), we can derive

$$\begin{aligned}\sum_{k=\tau'}^{t}\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big) &= 1-\prod_{i=\tau'}^{t}\big(1-\eta_i\big) \geq 1-\max\Big\{\big(1-\eta_t\big)^{1/\eta_{t/2}+1}, \prod_{i=1}^{t}\big(1-\eta_i\big)\Big\} \\ &\geq 1-\max\Big\{e^{-1/2}, \prod_{i=1}^{\widehat{\tau}}\big(1-\eta_i\big)\Big\} \geq \frac{1}{8},\end{aligned}$$

where we once again use the condition (185), and the last inequality comes from the derivation in (183). Putting these two bounds together yields

$$\begin{aligned}\sum_{k=\tau'}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2 &\geq \min_{k:\,\tau'\leq k\leq t}\Big\{\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]\Big\}\sum_{k=\tau'}^{t}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big] \\ &\geq \frac{1}{3}\eta_t\cdot\frac{1}{8} \geq \frac{1}{24}\eta_t.\end{aligned}$$

To finish up, it remains to justify (185). This condition is obvious for constant learning rates. As for rescaled learning rates, one can see that

$$\eta_i = \frac{1}{1+(1-\gamma)c_\eta i} \geq \frac{19}{20c_\eta(1-\gamma)i} \qquad \text{for all } i \geq \overline{\tau},$$

where $\overline{\tau} \coloneqq \lceil\frac{19}{c_\eta(1-\gamma)}\rceil$. This allows one to obtain

$$\log\Big[\prod_{i=\overline{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big)\Big] \leq -\sum_{i=\overline{\tau}}^{T}\eta_i(1-\gamma p) \leq -\sum_{i=\overline{\tau}}^{T}\frac{19}{15c_\eta i} \leq -\frac{19\log\frac{T}{\overline{\tau}}}{15c_\eta} \leq -\frac{19\log\frac{c_\eta(1-\gamma)T}{20}}{15c_\eta} \leq -\frac{1}{5},$$

provided that $T \geq \frac{c_1}{(1-\gamma)^2}$ for some sufficiently large constant $c_1 > 0$ and $1-\gamma < c_\eta < \log T$. Taking this together with (176) implies that $\widehat{\tau} \geq \overline{\tau}$ and hence $\eta_{\widehat{\tau}} \leq \eta_{\overline{\tau}} = \frac{1}{1+(1-\gamma)c_\eta\overline{\tau}} = 1/20$. □

D.4 Proof of Lemma 3

Given that state 0 is an absorbing state with zero immediate reward, it is easily seen that

$$V^\pi(0) = 0 \qquad \text{for all } \pi \qquad \Longrightarrow \qquad V^\star(0) = Q^\star(0,1) = 0.$$

Moreover, by construction, taking action 1 and taking action 2 in state 1 result in the same behavior (in terms of both the reward function and the associated transition probability), and as a consequence,

$$Q^\star(1,1) = Q^\star(1,2) = V^\star(1). \tag{186}$$

From Bellman's equation, we can thus deduce that

$$Q^\star(1,1) = r(1,1) + \gamma P(0\,|\,1,1)V^\star(0) + \gamma P(1\,|\,1,1)V^\star(1),$$

which in conjunction with (186) and a little algebra leads to

$$V^\star(1) = \frac{r(1,1) + \gamma P(0\,|\,1,1)V^\star(0)}{1 - \gamma P(1\,|\,1,1)} = \frac{1}{1-\gamma p} = \frac{3}{4(1-\gamma)}.$$

Here, the second identity follows since $V^\star(0) = 0$, and the third identity makes use of (34). The calculation for $V^\star(2)$ and $Q^\star(2,1)$ follows from an identical argument and is hence omitted.

Turning to state 3, by Bellman's equation, we have

$$V^\star(3) = Q^\star(3,1) = r(3,1) + \gamma P(3\,|\,3,1)V^\star(3) = 1 + \gamma V^\star(3),$$

which leads to $V^\star(3) = \frac{1}{1-\gamma}$.

E Analysis for asynchronous Q-learning (Theorem 4)

E.1 Notation and preliminary facts

Vector and matrix notation. We shall adopt the vector notation $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ in the same way as in Section 4.1. The sample transition matrix $\boldsymbol{P}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ in the asynchronous case is defined such that

$$\boldsymbol{P}_t\big((s,a),s'\big) = \begin{cases} 1, & \text{if } (s,a,s') = (s_{t-1}, a_{t-1}, s_t); \\ 0, & \text{else.} \end{cases} \tag{187}$$

It is also handy to introduce the diagonal matrix $\boldsymbol{\Lambda}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ such that

$$\boldsymbol{\Lambda}_t\big((s,a),(s,a)\big) = \begin{cases} \eta, & \text{if } (s,a) = (s_{t-1}, a_{t-1}); \\ 0, & \text{otherwise.} \end{cases} \tag{188}$$

Armed with the above notation, the asynchronous Q-learning update rule (38) can be conveniently expressed as follows:

$$\boldsymbol{Q}_t = \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big). \tag{189}$$

Range of V_t and Q_t. Similar to the synchronous counterpart, we have the following elementary properties.

Lemma 7. *Suppose that $0 < \eta_t \le 1$ for all $t \ge 0$. Assume that $\boldsymbol{0} \le \boldsymbol{Q}_0 \le \frac{1}{1-\gamma}\boldsymbol{1}$. Then for any $t \ge 0$, one has*

$$\boldsymbol{0} \le \boldsymbol{Q}_t \le \frac{1}{1-\gamma}\boldsymbol{1} \qquad \text{and} \qquad \boldsymbol{0} \le \boldsymbol{V}_t \le \frac{1}{1-\gamma}\boldsymbol{1}. \tag{190}$$

Proof. The proof is the same as that of Lemma 4, and is hence omitted for brevity. □

E.2 Main steps for proving Theorem 4

We are now in a position to outline the main steps for the proof of Theorem 4.

Step 1: deriving basic recursions. According to the update rule (189), we can derive the following elementary decomposition

$$\begin{aligned} \boldsymbol{\Delta}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{Q}_{t-1} + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \boldsymbol{\Lambda}_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\ &= \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}_{t-1} + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}\big(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\big), \end{aligned} \tag{191}$$

where the penultimate identity follows from the Bellman optimality equation $\boldsymbol{Q}^\star = \boldsymbol{r} + \boldsymbol{P}\boldsymbol{V}^\star$. Combining (191) with the inequalities (24) and using the definition (18) of π_t result in

$$\boldsymbol{\Delta}_t \le \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}_{t-1} + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1}, \tag{192a}$$

$$\boldsymbol{\Delta}_t \ge \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}_{t-1} + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1}. \tag{192b}$$

Apply the above two relations recursively to obtain

$$\boldsymbol{\Delta}_t \le \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}_{i-1} + \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + \prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0, \tag{193a}$$

$$\boldsymbol{\Delta}_t \ge \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}_{i-1} + \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{i-1} + \prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0. \tag{193b}$$

By defining the following diagonal matrices

$$\boldsymbol{\Lambda}_i^{(t)} := \begin{cases} \prod_{j=1}^{t}(\boldsymbol{I} - \boldsymbol{\Lambda}_j), & \text{if } i = 0, \\ \boldsymbol{\Lambda}_i\prod_{j=i+1}^{t}(\boldsymbol{I} - \boldsymbol{\Lambda}_j), & \text{if } 0 < i < t, \\ \boldsymbol{\Lambda}_t, & \text{if } i = t, \end{cases} \tag{194}$$

and setting

$$\beta = \frac{c_3(1-\gamma)}{\log T} \tag{195}$$

for some constant $c_3 > 0$, we can rearrange terms in the upper bound (193a) to reach

$$\begin{aligned}\boldsymbol{\Delta}_t \le \underbrace{\boldsymbol{\Lambda}_0^{(t)}\boldsymbol{\Delta}_0 + \gamma\sum_{i=1}^{(1-\beta)t}\boldsymbol{\Lambda}_i^{(t)}\Big[\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}_{i-1} + \boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}\Big]}_{=:\,\boldsymbol{\zeta}_t} \\ + \underbrace{\gamma\sum_{i=(1-\beta)t+1}^{t}\boldsymbol{\Lambda}_i^{(t)}\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}_{i-1}}_{=:\,\boldsymbol{\xi}_t} + \gamma\sum_{i=(1-\beta)t+1}^{t}\boldsymbol{\Lambda}_i^{(t)}\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}.\end{aligned} \tag{196}$$

In the subsequent steps, we shall first develop bounds on the sizes of the terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ in (196) separately, and then combine these bounds with (196) recursively in order to derive the advertised upper bound on $\boldsymbol{\Delta}_t$.

Step 2: bounding the terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$. The terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ defined in (196) can be bounded with high probability by the following lemmas.

Lemma 8. *With probability at least $1-\delta$, we have*

$$\|\boldsymbol{\zeta}_t\|_\infty \le \frac{4}{(1-\gamma)T} \tag{197}$$

for all t obeying $\frac{T}{c_4\log T} \le t \le T$. Here, $c_4 > 0$ is some constant obeying $c_4 \le c_1c_3/4$, where the constants c_1 and c_3 appear in (42a) and (195), respectively.

Proof. See Section E.3.1. □

Lemma 9. *Suppose that $0 < \eta \le \frac{\log^3 T}{(1-\gamma)T\mu_{\mathsf{min}}}$. With probability at least $1-\delta$, one has*

$$|\boldsymbol{\xi}_t| \le \sqrt{\frac{16\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\Big(\max_{(1-\beta)t\le i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) + \mathbf{1}\Big)} + \frac{6\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2T\mu_{\mathsf{min}}}\mathbf{1} \tag{198}$$

for all t obeying $\frac{T}{c_4\log T} \le t \le T$ for some constant $c_4 > 0$.

Proof. See Section E.3.2. □

Step 3: controlling $\boldsymbol{\Delta}_t$. Consider any t obeying $\frac{T}{c_4 \log T} \le t \le T$ and any k obeying $2t/3 \le k \le t$. Under the sample size condition (42b), Lemmas 8-9 together with a little algebra lead to

$$|\boldsymbol{\zeta}_k| + |\boldsymbol{\xi}_k| \le \sqrt{\frac{32\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\bigg(\max_{(1-\beta)k \le i \le k} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) + \boldsymbol{1}\bigg)} \le \sqrt{\boldsymbol{\varphi}_t},$$

where we define

$$\boldsymbol{\varphi}_t \coloneqq \frac{32\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\bigg(\max_{\frac{t}{2} \le i \le t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) + \boldsymbol{1}\bigg). \tag{199}$$

Combining this inequality with (196) allows us to obtain

$$\boldsymbol{\Delta}_k \le \sqrt{\boldsymbol{\varphi}_t} + \sum_{i=(1-\beta)k+1}^{k} \boldsymbol{\Lambda}_i^{(k)} \gamma \boldsymbol{P}^{\pi_{i-1}} \boldsymbol{\Delta}_{i-1} = \sqrt{\boldsymbol{\varphi}_t} + \sum_{i=(1-\beta)k}^{k-1} \boldsymbol{\Lambda}_{i+1}^{(k)} \gamma \boldsymbol{P}^{\pi_i} \boldsymbol{\Delta}_i \qquad \text{for all } \ 2t/3 \le k \le t. \tag{200}$$

Similar to the quantity $\alpha_i^{(t)}$ defined in (65), let us define

$$\boldsymbol{D}_i^{(t)} \coloneqq \bigg(\sum_{j=(1-\beta)t}^{t-1} \boldsymbol{\Lambda}_{j+1}^{(t)}\bigg)^{-1} \boldsymbol{\Lambda}_{i+1}^{(t)}, \tag{201}$$

which, according to (58) and the definition (188), clearly satisfies

$$\boldsymbol{D}_i^{(t)} \ge \boldsymbol{\Lambda}_{i+1}^{(t)} \ge \boldsymbol{0} \qquad \text{and} \qquad \sum_{i=(1-\beta)t}^{t-1} \boldsymbol{D}_i^{(t)} = \boldsymbol{I}. \tag{202}$$

Set $i_0 = t$ for notational convenience. With this set of notation and the property (202) in mind, we can derive the following bound

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le \sum_{i_1=(1-\beta)t}^{t-1} \Big(\boldsymbol{D}_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \boldsymbol{\Lambda}_{i_1+1}^{(t)} \gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{\Delta}_{i_1}\Big) \\
&\le \sum_{i_1=(1-\beta)t}^{t-1} \bigg[\boldsymbol{D}_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \boldsymbol{\Lambda}_{i_1+1}^{(t)} \gamma \boldsymbol{P}^{\pi_{i_1}} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \Big(\boldsymbol{D}_{i_2}^{(i_1)} \sqrt{\boldsymbol{\varphi}_t} + \boldsymbol{\Lambda}_{i_2+1}^{(i_1)} \gamma \boldsymbol{P}^{\pi_{i_2}} \boldsymbol{\Delta}_{i_2}\Big)\bigg] \\
&\le \sum_{i_1=(1-\beta)t}^{t-1} \boldsymbol{D}_{i_1}^{(t)} \sqrt{\boldsymbol{\varphi}_t} + \sum_{i_1=(1-\beta)t}^{t-1} \boldsymbol{D}_{i_1}^{(t)} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \big(\gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{D}_{i_2}^{(i_1)}\big) \sqrt{\boldsymbol{\varphi}_t} \\
&\qquad + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \prod_{k=1}^{2} \big(\boldsymbol{\Lambda}_{i_k+1}^{(i_{k-1})} \gamma \boldsymbol{P}^{\pi_{i_k}}\big) \boldsymbol{\Delta}_{i_2} \\
&= \sum_{i_1=(1-\beta)t}^{t-1} \boldsymbol{D}_{i_1}^{(t)} \bigg\{\boldsymbol{I} + \sum_{i_2=(1-\beta)i_1}^{i_1-1} \gamma \boldsymbol{P}^{\pi_{i_1}} \boldsymbol{D}_{i_2}^{(i_1)}\bigg\} \sqrt{\boldsymbol{\varphi}_t} + \sum_{i_1=(1-\beta)t}^{t-1} \sum_{i_2=(1-\beta)i_1}^{i_1-1} \prod_{k=1}^{2} \big(\boldsymbol{\Lambda}_{i_k+1}^{(i_{k-1})} \gamma \boldsymbol{P}^{\pi_{i_k}}\big) \boldsymbol{\Delta}_{i_2}.
\end{aligned} \tag{203}$$

Here, the first relation makes use of the second property in (202), the second relation further expands $\boldsymbol{\Delta}_{i_1}$ in the same way as in the first line of (203), whereas the third inequality relies on the first property in (202).

Next, we intend to invoke the above relation multiple times to reach a simpler relation. Set

$$H \coloneqq \frac{\log T}{1-\gamma}. \tag{204}$$

Similar to the way we derive (73), we can apply the relation (203) recursively and use the basic relation $|\boldsymbol{\Delta}_k| \le \frac{1}{1-\gamma}\boldsymbol{1}$ for any k to show that

$$\boldsymbol{\Delta}_t \le \sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \bigg\{\boldsymbol{D}_{i_1}^{(t)} \bigg(\boldsymbol{I} + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} \big(\boldsymbol{P}^{\pi_{i_k}} \boldsymbol{D}_{i_{k+1}}^{(i_k)}\big)\bigg) \sqrt{\boldsymbol{\varphi}_t} + \gamma^H \prod_{k=1}^{H} \big(\boldsymbol{\Lambda}_{i_k+1}^{(i_{k-1})} \boldsymbol{P}^{\pi_{i_k}}\big) |\boldsymbol{\Delta}_{i_H}|\bigg\}, \tag{205}$$

where $\mathcal{I}_t$ has been defined in (71). To further simplify (205), we need to control the two terms on the right-hand side of (205) separately.

- We shall begin with the first term on the right-hand side of (205). Towards this end, let us define a collection of policies $\{\widehat{\pi}_k\}$ recursively and backward as follows:

$$\widehat{\pi}_k \coloneqq \begin{cases} \arg\max_{\pi\in\Pi} \boldsymbol{\Pi}^{\pi}\sqrt{\boldsymbol{\varphi}_t}, & \text{if } k = H-1, \\ \arg\max_{\pi\in\Pi} \boldsymbol{\Pi}^{\pi}\Big(\boldsymbol{I} + \sum_{h=k+1}^{H-1}\gamma^h \prod_{j=1}^{h}\boldsymbol{P}^{\widehat{\pi}_j}\Big)\sqrt{\boldsymbol{\varphi}_t}, & \text{if } k = H-2,\cdots,1, \end{cases} \tag{206}$$

or alternatively (in view of the definition (15) of $\boldsymbol{P}^{\pi}$),

$$\widehat{\pi}_k \coloneqq \begin{cases} \arg\max_{\pi\in\Pi} \boldsymbol{P}^{\pi}\sqrt{\boldsymbol{\varphi}_t}, & \text{if } k = H-1; \\ \arg\max_{\pi\in\Pi} \boldsymbol{P}^{\pi}\Big(\boldsymbol{I} + \sum_{h=k+1}^{H-1}\gamma^h \prod_{j=1}^{h}\boldsymbol{P}^{\widehat{\pi}_j}\Big)\sqrt{\boldsymbol{\varphi}_t}, & \text{if } k = H-2,\cdots,1. \end{cases} \tag{207}$$

Here, Π is a policy set satisfying

$$\Pi \coloneqq \big\{\pi = [\pi(s)]_{s\in\mathcal{S}} \,\big|\, \pi(s)\in\Pi_s, \forall s\in\mathcal{S}\big\}, \qquad \Pi_s \coloneqq \big\{\pi_i(s) \mid i\in[t/2,t)\big\}; \tag{208}$$

in words, for any policy π belonging to Π, each $\pi(s)$ coincides with one of the policy iterates $\pi_i(s)$ during the latest βt iterations, although we do not require all $\{\pi(s)\}$ across different states to be associated with the same time stamp i. With this collection of policies in place, we can deduce that

$$\begin{aligned}
&\sum_{i_1,\cdots,i_H} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-1}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&= \sum_{i_1,\cdots,i_{H-1}}\sum_{i_H} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-2}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-2}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\Big(\gamma\boldsymbol{P}^{\pi_{i_{H-1}}}\boldsymbol{D}_{i_H}^{(i_{H-1})}\Big)\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&= \sum_{i_1,\cdots,i_{H-1}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-2}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-2}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\Big(\gamma\boldsymbol{P}^{\pi_{i_{H-1}}}\sum_{i_H}\boldsymbol{D}_{i_H}^{(i_{H-1})}\Big)\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&\overset{\text{(i)}}{=} \sum_{i_1,\cdots,i_{H-1}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-2}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-2}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\gamma\boldsymbol{P}^{\pi_{i_{H-1}}}\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&\overset{\text{(ii)}}{\le} \sum_{i_1,\cdots,i_{H-1}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-2}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-2}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\gamma\boldsymbol{P}^{\widehat{\pi}_{H-1}}\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&= \sum_{i_1,\cdots,i_{H-2}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-3}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-3}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\gamma\boldsymbol{P}^{\pi_{i_{H-2}}}\Big(\sum_{i_{H-1}}\boldsymbol{D}_{i_{H-1}}^{(i_{H-2})}\Big)\big(\boldsymbol{I}+\gamma\boldsymbol{P}^{\widehat{\pi}_{H-1}}\big)\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&\overset{\text{(iii)}}{=} \sum_{i_1,\cdots,i_{H-2}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-3}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-3}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\gamma\boldsymbol{P}^{\pi_{i_{H-2}}}\big(\boldsymbol{I}+\gamma\boldsymbol{P}^{\widehat{\pi}_{H-1}}\big)\Big)\sqrt{\boldsymbol{\varphi}_t} \\
&\overset{\text{(iv)}}{\le} \sum_{i_1,\cdots,i_{H-2}} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-3}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)}) + \prod_{k=1}^{H-3}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\gamma\boldsymbol{P}^{\widehat{\pi}_{H-2}}\big(\boldsymbol{I}+\gamma\boldsymbol{P}^{\widehat{\pi}_{H-1}}\big)\Big)\sqrt{\boldsymbol{\varphi}_t},
\end{aligned}$$

where we abbreviate $\sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t}$ as $\sum_{i_1,\cdots,i_H}$ as long as it is clear from the context. Here, (i) and (iii) arise from the second property in (202), while (ii) and (iv) are due to the construction (207). Continuing the derivation of the above inequality recursively, we arrive at

$$\sum_{i_1,\cdots,i_H} \boldsymbol{D}_{i_1}^{(t)}\Big(\boldsymbol{I} + \sum_{h=1}^{H-1}\prod_{k=1}^{h}(\gamma\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{D}_{i_{k+1}}^{(i_k)})\Big)\sqrt{\boldsymbol{\varphi}_t} \le \Big(\boldsymbol{I} + \sum_{h=1}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\widehat{\pi}_k}\Big)\sqrt{\boldsymbol{\varphi}_t}.$$

- We now turn attention to the second term on the right-hand side of (205). It is seen that

$$\begin{aligned}
\sum_{i_1,\cdots,i_H}\prod_{k=1}^{H}\gamma^H\big(\boldsymbol{\Lambda}_{i_k+1}^{(i_{k-1})}\boldsymbol{P}^{\pi_{i_k}}\big)\big|\boldsymbol{\Delta}_{i_H}\big| &\le \frac{\gamma^H}{1-\gamma}\sum_{i_1,\cdots,i_H}\prod_{k=1}^{H}\big(\boldsymbol{D}_{i_k}^{(i_{k-1})}\boldsymbol{P}^{\pi_{i_k}}\big)\mathbf{1} \\
&= \frac{\gamma^H}{1-\gamma}\sum_{i_1,\cdots,i_{H-1}}\prod_{k=1}^{H-1}\big(\boldsymbol{D}_{i_k}^{(i_{k-1})}\boldsymbol{P}^{\pi_{i_k}}\big)\Big(\sum_{i_H}\boldsymbol{D}_{i_H}^{(i_{H-1})}\big(\boldsymbol{P}^{\pi_{i_H}}\mathbf{1}\big)\Big) \\
&= \frac{\gamma^H}{1-\gamma}\sum_{i_1,\cdots,i_{H-1}}\prod_{k=1}^{H-1}\big(\boldsymbol{D}_{i_k}^{(i_{k-1})}\boldsymbol{P}^{\pi_{i_k}}\big)\Big(\sum_{i_H}\boldsymbol{D}_{i_H}^{(i_{H-1})}\Big)\mathbf{1} \\
&= \frac{\gamma^H}{1-\gamma}\sum_{i_1,\cdots,i_{H-1}}\prod_{k=1}^{H-1}\big(\boldsymbol{D}_{i_k}^{(i_{k-1})}\boldsymbol{P}^{\pi_{i_k}}\big)\mathbf{1} \\
&= \cdots = \frac{\gamma^H}{1-\gamma}\mathbf{1} = \frac{\gamma^{\frac{\log T}{1-\gamma}}}{1-\gamma}\mathbf{1} \\
&\le \frac{1}{(1-\gamma)T}\mathbf{1},
\end{aligned}$$

  where the first line follows from the first property in (202), the third line is due to the fact $\boldsymbol{P}^{\pi}\mathbf{1}=\mathbf{1}$ for any π, and the fourth line arises from the second property in (202).

Substituting the above two bounds into (205) yields

$$\boldsymbol{\Delta}_t \le \underbrace{\Big(\boldsymbol{I}+\sum_{h=1}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\widehat{\pi}_k}\Big)\sqrt{\boldsymbol{\varphi}_t}}_{=:\boldsymbol{\beta}}+\frac{1}{(1-\gamma)T}\mathbf{1}. \tag{209}$$

Step 4: putting all pieces together. Repeating our analysis for the term $\boldsymbol{\beta}_1$ in Section B.2 (i.e., Step 5 of Section B.2 with Lemma 5 replaced by Lemma 6), we arrive at

$$|\boldsymbol{\beta}|^2 \le \frac{320\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}$$

with probability at least $1-\delta$. Substitution into (209) then yields

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le \sqrt{\frac{320\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T\mu_{\mathsf{min}}}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}+\frac{1}{(1-\gamma)T}\mathbf{1} \\
&\le 30\sqrt{\frac{\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T\mu_{\mathsf{min}}}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}
\end{aligned} \tag{210}$$

holds simultaneously for all $t\ge\frac{T}{c_4\log T}$, provided that the sample size condition (42b) is satisfied. Similarly, we can also establish the following lower bound on $\boldsymbol{\Delta}_t$ (which we omit the details for the sake of brevity)

$$\boldsymbol{\Delta}_t \ge -30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T\mu_{\mathsf{min}}}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}$$

with probability at least $1-\delta$. To summarize, it is seen that with probability exceeding $1-2\delta$,

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T\mu_{\mathsf{min}}}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}. \tag{211}$$

This resembles the relation (31) derived for the synchronous case, except that T in the denominator is replaced with $\mu_{\mathsf{min}}T$. As a result, we can readily repeat the argument in Appendix B.4 to reach

$$\|\boldsymbol{\Delta}_T\|_\infty \le O\bigg(\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T\mu_{\mathsf{min}}}} + \frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T\mu_{\mathsf{min}}}\bigg), \tag{212}$$

which in turn establishes the claimed result in Theorem 4.

E.3 Proofs of technical lemmas

E.3.1 Proof of Lemma 8

In view of the definition of $\boldsymbol{\zeta}_t$ in (196) and the fact that $\boldsymbol{\Lambda}_0^{(t)}$ is a diagonal matrix, we can deduce that

$$\begin{aligned}
\|\boldsymbol{\zeta}_t\|_\infty &\le \|\boldsymbol{\Lambda}_0^{(t)}\|\|\boldsymbol{\Delta}_0\|_\infty + t\max_{1\le i\le(1-\beta)t}\|\boldsymbol{\Lambda}_i^{(t)}\|\max_{1\le i\le(1-\beta)t}\big(\|\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}\|_\infty + \|\boldsymbol{P}_i\boldsymbol{V}_{i-1}\|_\infty + \|\boldsymbol{P}\boldsymbol{V}_{i-1}\|_\infty\big)\\
&\le \|\boldsymbol{\Lambda}_0^{(t)}\|\|\boldsymbol{\Delta}_0\|_\infty + t\max_{1\le i\le(1-\beta)t}\|\boldsymbol{\Lambda}_i^{(t)}\|\max_{1\le i\le(1-\beta)t}\Big\{\|\boldsymbol{P}^{\pi_{i-1}}\|_1\|\boldsymbol{\Delta}_{i-1}\|_\infty + \big(\|\boldsymbol{P}_i\|_1 + \|\boldsymbol{P}\|_1\big)\|\boldsymbol{V}_{i-1}\|_\infty\Big\}\\
&\overset{\text{(i)}}{=} \|\boldsymbol{\Lambda}_0^{(t)}\|\|\boldsymbol{\Delta}_0\|_\infty + t\max_{1\le i\le(1-\beta)t}\|\boldsymbol{\Lambda}_i^{(t)}\|\max_{1\le i\le(1-\beta)t}\big(\|\boldsymbol{\Delta}_{i-1}\|_\infty + 2\|\boldsymbol{V}_{i-1}\|_\infty\big)\\
&\overset{\text{(ii)}}{\le} \frac{1}{T^2}\cdot\frac{1}{1-\gamma} + \frac{1}{T^2}\cdot t\cdot\frac{3}{1-\gamma}\\
&\le \frac{4}{(1-\gamma)T}.
\end{aligned}$$

Here, (i) holds true since $\|\boldsymbol{P}^{\pi_{i-1}}\|_1 = \|\boldsymbol{P}_i\|_1 = \|\boldsymbol{P}\|_1 = 1$. To verify (ii), we first define

$$t_k(s,a) \coloneqq \text{the time stamp when the trajectory visits } (s,a) \text{ for the } k\text{-th time} \tag{213}$$

and

$$K_t(s,a) \coloneqq \Big|\big\{k\ge 1 \mid t_k(s,a) < t\big\}\Big|, \tag{214}$$

namely, the total number of times — before the t-th iteration — that the sample trajectory visits (s,a). Then Li et al. (2021b, Lemma 8) tells us that with probability at least $1-\delta$,

$$K_{t_1}(s,a) - K_{t_2}(s,a) \ge \frac{1}{2}(t_1-t_2)\mu_{\mathsf{min}}, \tag{215}$$

holds uniformly for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $0\le t_2\le t_1\le T$ obeying

$$t_1 - t_2 \ge \frac{886 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}.$$

This in turn implies that: if $\beta t \ge \frac{886 t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ and $i\le(1-\beta)t$, then one has

$$\|\boldsymbol{\Lambda}_0^{(t)}\| = \bigg\|\prod_{j=1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\bigg\| = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}(1-\eta)^{K_t(s,a)} \le (1-\eta)^{\frac{1}{2}t\mu_{\mathsf{min}}} \le \frac{1}{T^2} \tag{216a}$$

$$\|\boldsymbol{\Lambda}_i^{(t)}\| = \bigg\|\boldsymbol{\Lambda}_i\prod_{j=i+1}^{t}\big(\boldsymbol{I}-\boldsymbol{\Lambda}_j\big)\bigg\| \le \max_{(s,a)}(1-\eta)^{K_t(s,a)-K_i(s,a)} \le (1-\eta)^{\frac{1}{2}\beta t\mu_{\mathsf{min}}} \le \frac{1}{T^2} \tag{216b}$$

with probability at least $1-\delta$, provided that $\eta\beta t\mu_{\mathsf{min}} > 4\log T$. In other words, (216) holds with probability at least $1-\delta$, as long as

$$t > \max\bigg\{\frac{4\log T}{\eta\beta\mu_{\mathsf{min}}}, \frac{886 t_{\mathsf{mix}}}{\beta\mu_{\mathsf{min}}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\bigg\} = \max\bigg\{\frac{4T}{c_1c_3\log T}, \frac{886 t_{\mathsf{mix}}}{c_3(1-\gamma)\mu_{\mathsf{min}}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\log T\bigg\}.$$

This taken together with the sample size assumption (42b) concludes the proof of Lemma 8.

E.3.2 Proof of Lemma 9

Fix any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let us look at the (s,a)-th entry of $\boldsymbol{\xi}_t$, i.e., $\xi_t(s,a)$. For notational simplicity, 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. $\boldsymbol{P}$).

Using the definition of $\boldsymbol{\xi}_t$ in (196) and the above notation, we can derive

$$\xi_t(s,a) = \gamma \sum_{i=(1-\beta)t+1}^{t} \prod_{j=i+1}^{t} \big(1-\boldsymbol{\Lambda}_j(s,a)\big)\boldsymbol{\Lambda}_i(s,a)\big(\boldsymbol{P}_i(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}_{i-1}. \tag{217}$$

Equipped with the definitions of $t_k(s,a)$ (cf. (213)) and $K_t(s,a)$ (cf. (214)), we can further rewrite (217) as

$$\xi_t(s,a) = \gamma \sum_{k=K_{(1-\beta)t+1}}^{K_t(s,a)} (1-\eta)^{K_t(s,a)-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}_{t_k}. \tag{218}$$

In what follows, we shall suppress the notation and write $t_k = t_k(s,a)$ and $K_t = K_t(s,a)$ to streamline notation.

The main step thus boils down to controlling (218). Towards this, we claim that: with probability at least $1-\delta$,

$$\begin{aligned}&\left|\sum_{k=K_\beta}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}_{t_k}\right| \\ &\quad\le \sqrt{\frac{16\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\Big(\max_{t_{K_\beta}\le i\le t_K}\mathsf{Var}_{\boldsymbol{P}(s,a)}(\boldsymbol{V}_i)+1\Big)} + \frac{6\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2T\mu_{\mathsf{min}}}\end{aligned} \tag{219}$$

holds simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all $1\le K_\beta\le K\le T$, provided that $0<\eta\le\frac{\log^3 T}{(1-\gamma)T\mu_{\mathsf{min}}}$. If this claim were true, then taking $K_\beta = K_{(1-\beta)t+1}$ and $K = K_t$ and substituting the bound (219) into the expression (218) would lead to

$$|\boldsymbol{\xi}_t| \le \sqrt{\frac{16\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\Big(\max_{(1-\beta)t\le i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big)} + \frac{6\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2T\mu_{\mathsf{min}}}\mathbf{1}, \tag{220}$$

thus concluding the proof of this lemma. To finish up, it is sufficient to justify the claim (219), which forms the content of the remainder of this proof.

Proof of the claim (219). Let us use the notation in (53) to express $\eta_k^{(K)} = (1-\eta)^{K-k}\eta$. 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; see Li et al. (2021b, Section B.1). We can then express the term

$$X_K \coloneqq \sum_{k=K_\beta}^{K}(1-\eta)^{K-k}\eta\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}_{t_k},$$

as follows:

$$X_K = \sum_{k=K_\beta}^{K} z_k \qquad \text{with } z_k \coloneqq \eta_k^{(K)}\big(\boldsymbol{P}_{t_k+1}(s,a)-\boldsymbol{P}(s,a)\big)\boldsymbol{V}_{t_k},$$

where the z_k's satisfy

$$\mathbb{E}\left[z_k \,|\, t_k,\cdots,t_1,\boldsymbol{V}_{t_k},\cdots,\boldsymbol{V}_{t_1}\right] = 0.$$

We intend to invoke the Freedman inequality to control X_K for any K obeying $K \le T$. Similar to the synchronous counterpart, we can see that

$$\begin{aligned}
B &\coloneqq \max_{1<k\le K} \|z_k\|_\infty \le \eta_k^K \big(\|\boldsymbol{P}_{t_k+1}\|_1 + \|\boldsymbol{P}\|_1\big)\|\boldsymbol{V}_{t_k}\|_\infty \le \frac{2\eta_k^K}{1-\gamma} \le \frac{2\eta}{1-\gamma}, \\
W &\coloneqq \sum_{k=K_\beta}^{K} \mathsf{Var}\big(z_k \,|\, t_k, \cdots, t_1, \boldsymbol{V}_{t_k}, \cdots, \boldsymbol{V}_{t_1}\big) = \gamma^2 \sum_{k=K_\beta}^{K} \big(\eta_k^{(K)}\big)^2 \mathsf{Var}\big((\boldsymbol{P}_{t_k+1} - \boldsymbol{P})\boldsymbol{V}_{t_k} \mid \boldsymbol{V}_{t_k}\big) \\
&\le \sum_{k=K_\beta}^{K} \big(\eta_k^{(K)}\big)^2 \mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}_{t_k}\big) \le \Big(\max_{K_\beta\le k\le K} \eta_k^{(K)}\Big)\Big(\sum_{k=K_\beta}^{K} \eta_k^{(K)}\Big)\mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}_{t_k}\big) \\
&\le \eta \max_{K_\beta\le k\le K} \mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}_{t_k}\big) \le \eta \max_{t_{K_\beta}\le i\le t_K} \mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}_i\big),
\end{aligned}$$

where we have made use of (58). In addition, we make note of a trivial upper bound on W as follows

$$\sigma^2 := \frac{\eta}{(1-\gamma)^2} \ge \eta \max_{t_{K_\beta}\le i\le t_K} \mathsf{Var}_{\boldsymbol{P}(s,a)}\big(\boldsymbol{V}_i\big) \ge W_t.$$

With the preceding bounds in place, applying the Freedman inequality in Theorem 5 and taking $L = \log_2 \frac{1}{1-\gamma}$ imply that

$$\begin{aligned}
|X_K| &\le \sqrt{8\max\left\{W, \frac{\sigma^2}{2^L}\right\}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2\log_2\frac{1}{1-\gamma}}{\delta}} + \frac{8\eta}{3(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2\log_2\frac{1}{1-\gamma}}{\delta} \\
&\le \sqrt{8\eta\max\Big\{\max_{t_{K_\beta}\le i\le t_K} \mathsf{Var}_{\boldsymbol{P}(s,a)}(\boldsymbol{V}_i), 1\Big\}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2\log_2\frac{1}{1-\gamma}}{\delta}} + \frac{8\eta}{3(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2\log_2\frac{1}{1-\gamma}}{\delta} \\
&\le \sqrt{\frac{16\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T\mu_{\mathsf{min}}}\Big(\max_{t_{K_\beta}\le i\le t_K} \mathsf{Var}_{\boldsymbol{P}(s,a)}(\boldsymbol{V}_i) + 1\Big)} + \frac{6\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2 T\mu_{\mathsf{min}}}
\end{aligned}$$

with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2}$, provided that $\eta \le \frac{\log^3 T}{(1-\gamma)T\mu_{\mathsf{min}}}$. We can thus conclude the proof by taking the union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $1 \le K_\beta \le K \le T$. □

References

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory*, pages 67–83.

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.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716.

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.

Bhandari, J., Russo, D., and Singal, R. (2021). A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*.

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.

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., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*.

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*.

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. PMLR.

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*.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

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.

Hasselt, H. (2010). Double Q-learning. *Advances in neural information processing systems*, 23:2613–2621.

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.

Kakade, S. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, University of London.

Kearns, M., Mansour, Y., and Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3):193–208.

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*.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2021). Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040.

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355.

Lee, D. and He, N. (2018). Stochastic primal-dual Q-learning. *arXiv preprint arXiv:1810.08298*.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021a). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Neural Information Processing Systems (NeurIPS)*.

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. *Advances in Neural Information Processing Systems*, 33.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021b). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*.

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*.

Murphy, S. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097.

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

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*, pages 3185–3205.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

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. (2018). 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.

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.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270.

Tsitsiklis, J. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690.

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

Tsybakov, A. B. and Zaiats, V. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. (2019). Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32:5784–5795.

Wainwright, M. (2019a). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019b). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

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

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). The mean-squared error of double Q-learning. *Advances in Neural Information Processing Systems*, 33.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*.

Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, 33.

Xu, P. and Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2019a). Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*.

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

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.