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

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

February 2021; Revised: October 2022

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 $\gamma$-discounted infinite-horizon MDP with state space $S$ and action space $A$: to yield an entrywise $\epsilon$-approximation of the optimal Q-function, state-of-the-art theory for Q-learning requires a sample size exceeding the order of $|S||A|(1-\gamma)^{5\epsilon^2}$, 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 $|A| = 1$ (so that Q-learning reduces to TD learning), we prove that the sample complexity of TD learning is minimax optimal and scales as $|S|(1-\gamma)^{3\epsilon^2}$ (up to log factor); (2) when $|A| \geq 2$, we settle the sample complexity of Q-learning to be on the order of $|S||A|(1-\gamma)^{4\epsilon^2}$ (up to log factor). Our theory unveils the strict sub-optimality of Q-learning when $|A| \geq 2$, 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 $\frac{1}{(1-\gamma)^{4\epsilon^2}}$.

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

Contents

1 Introduction ........................... 2
  1.1 Main contributions ............................................. 3
  1.2 Related works ............................................. 4

2 Background and algorithms ........................... 5

3 Main results: sample complexity of synchronous Q-learning ........................... 7
  3.1 Minimax optimality of TD learning ............................................. 7
  3.2 Tight sample complexity and sub-optimality of Q-learning ............................................. 9

4 Key analysis ideas (the synchronous case) ........................... 11
  4.1 Vector and matrix notation ............................................. 11
  4.2 Proof outline for Theorem 2 ............................................. 12
  4.3 Proof outline for Theorem 3 ............................................. 14

*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., 2022c; 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., 2022c; 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,
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 $\frac{|S||A|}{(1-\gamma)^2 t}$ (up to some log factor), where $S$ and $A$ 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 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 $\gamma$-discounted infinite-horizon MDPs with state space $S$ and action space $A$, this paper settles the $\ell_\infty$-based sample complexity of synchronous Q-learning. Here and throughout, the standard notation $f(\cdot) = O(g(\cdot))$ (resp. $f(\cdot) = \Omega(g(\cdot))$) means that $f(\cdot)$ is orderwise no larger than (resp. no smaller than) $g(\cdot)$.

<table>
<thead>
<tr>
<th>paper</th>
<th>learning rates</th>
<th>sample complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-Dar and Mansour (2003)</td>
<td>linear: $\frac{1}{t}$</td>
<td>$2 \cdot \frac{</td>
</tr>
<tr>
<td>Even-Dar and Mansour (2003)</td>
<td>polynomial: $\frac{1}{t^x}$, $\omega \in (1/2, 1)$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Beck and Srikant (2012)</td>
<td>constant: $(1-\gamma)^{x^2}$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Wainwright (2019b)</td>
<td>rescaled linear: $\frac{1}{t^{1+(1-\gamma)t}}$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Wainwright (2019b)</td>
<td>polynomial: $\frac{1}{t^x}$, $\omega \in (0, 1)$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Chen et al. (2020)</td>
<td>rescaled linear: $\frac{1}{t^{1+(1-\gamma)t}}$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Chen et al. (2020)</td>
<td>constant: $(1-\gamma)^4 t^2$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>this work (Q-learning, $</td>
<td>A</td>
<td>\geq 2$)</td>
</tr>
<tr>
<td>this work (Q-learning, $</td>
<td>A</td>
<td>\geq 2$)</td>
</tr>
<tr>
<td>this work (TD learning, $</td>
<td>A</td>
<td>= 1$)</td>
</tr>
<tr>
<td>this work (TD learning, $</td>
<td>A</td>
<td>= 1$)</td>
</tr>
</tbody>
</table>

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

- When $|A| = 1$, 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
  \[
  \tilde{O}\left(\frac{|S|}{(1-\gamma)^3\varepsilon^2}\right)
  \]
  is sufficient for TD learning to guarantee $\varepsilon$-accuracy in an $\ell_\infty$ sense; see Theorem 1. This is sharp and minimax optimal (up to some log factor).

- Moving on to the case with $|A| \geq 2$, we demonstrate that a sample size of
  \[
  \tilde{O}\left(\frac{|S||A|}{(1-\gamma)^4\varepsilon^2}\right)
  \]
  suffices for Q-learning to yield $\varepsilon$-accuracy in an $\ell_\infty$ 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
  \[
  \tilde{O}\left(\frac{1}{(1-\gamma)^{4\varepsilon^2}}\right)
  \]
  iterations to achieve $\varepsilon$-accuracy in an $\ell_\infty$ 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 $|A| \geq 2$, 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 $\frac{1}{1-\gamma}$ scaling w.r.t. the effective horizon, which is nearly sharp and improves upon the prior state-of-the-art Li et al. (2022c).

### 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 $\ell_\infty$-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 $\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^5\varepsilon^2}\right)$ 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 $\varepsilon$. Beck and Srikant (2012) investigated the case of constant learning rates; however, their result suffered from an additional factor of $|S||A|$, 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 $\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^5\varepsilon^2}\right)$. 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 $|A| = 1$, the recent works Khamaru et al. (2021a); 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 $\ell_\infty$-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. (2022c); 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 $\tilde{O}\left(\frac{1}{(1-\gamma)^{5}}\right)$ also showed up in the state-of-the-art sample complexity bounds for asynchronous Q-learning (Li et al., 2022c), and our theory is the first to sharpen it to $\tilde{O}\left(\frac{1}{(1-\gamma)^{4}}\right)$. When it comes to the special case with $|A| = 1$, 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 $\ell_\infty$-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 $\tilde{O}\left(\frac{|S||A|}{|S||A|}\right)$ 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 $\tilde{O}\left(\frac{|S||A|}{|S||A|}\right)$, which was subsequently generalized to the asynchronous setting by Li et al. (2022c). The $\ell_\infty$ statistical bounds for variance-reduced TD learning have been investigated in Khamaru et al. (2021a) for the synchronous setting, and in Li et al. (2022c) 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 $\ell_\infty$-based non-asymptotic guarantees. For instance, Bhandari et al. (2021); 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) $\ell_2$ 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., 2021; 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. In addition, Q-learning in conjunction with the pessimism principle has proven effective in dealing with offline data (Shi et al., 2022; Yan et al., 2022). 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), whereas the extension to multi-agent scenarios has been looked at in Hu and Wellman (2003); Li et al. (2022a). 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 $S = \{1, \cdots, |S|\}$ and $A = \{1, \cdots, |A|\}$ the state space and the action space of the MDP, respectively, and let $\Delta(S)$ represent the probability simplex over the set $S$. 

5
Basics of discounted infinite-horizon MDPs. Consider an infinite-horizon MDP as represented by a quintuple $\mathcal{M} = (S, A, P, r, \gamma)$, where $\gamma \in (0, 1)$ indicates the discount factor, $P : S \times A \to \Delta(S)$ represents the probability transition kernel (i.e., $P(s' | s, a)$ is the probability of transitioning to state $s'$ from a state-action pair $(s, a) \in S \times A$), and $r : S \times A \to [0, 1]$ stands for the reward function (i.e., $r(s, a)$ is the immediate reward collected in state $s \in S$ when action $a \in A$ is taken). Note that the immediate rewards are assumed to lie within $[0, 1]$ throughout this paper. Moreover, we let $\pi : S \to \Delta(A)$ represent a policy, so that $\pi(\cdot | s) \in \Delta(A)$ specifies the (possibly randomized) action selection rule in state $s$. If $\pi$ is a deterministic policy, then we denote by $\pi(s)$ the action selected by $\pi$ 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 $\pi$, the associated value function and Q-function of $\pi$ are defined respectively by

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

for all $s \in S$, and

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

for all $(s, a) \in S \times A$. Here, $\{(s_k, a_k)\}_{k \geq 0}$ is a trajectory of the MDP induced by the policy $\pi$ (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 $\pi$ and any state-action pair $(s, a)$. The optimal value function $V^*$ and optimal Q-function $Q^*$ are defined respectively as

$$V^*(s) := \max_\pi V^\pi(s), \quad Q^*(s, a) := \max_\pi Q^\pi(s, a)$$

for any state-action pair $(s, a) \in S \times A$. It is well known that there exists a deterministic optimal policy, denoted by $\pi^*$, that attains $V^*(s)$ and $Q^*(s, a)$ simultaneously for all $(s, a) \in S \times A$ (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 $s_t(s, a, a') \sim P(\cdot | s, a)$ for every state-action pair $(s, a) \in S \times A$.

With this sampling model in place, the Q-learning algorithm (Watkins and Dayan, 1992) maintains a Q-function estimate $Q_t : S \times A \to \mathbb{R}$ 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

$$Q_t = (1 - \eta_t)Q_{t-1} + \eta_t T_t(Q_{t-1}).$$

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

$$T_t(Q)(s, a) := r(s, a) + \gamma \max_{a' \in A} Q(s_t(s, a'), a'), \quad s_t \equiv s_t(s, a) \sim P(\cdot | s, a)$$

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

$$\forall (s, a) \in S \times A : \quad T(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{a' \in A} Q(s', a') \right].$$

Note that the optimal Q-function $Q^*$ is the unique fixed point of the Bellman operator (Bellman, 1952), that is, $T(Q^*) = Q^*$. 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 $V_t : S \to \mathbb{R}$ in the $t$-th iteration is defined as

$$\forall s \in S : \quad V_t(s) := \max_{a \in A} Q_t(s, a).$$
Algorithm 1: Synchronous Q-learning for infinite-horizon discounted MDPs.

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

Algorithm 2: Synchronous TD learning for infinite-horizon discounted MRP.

1: inputs: learning rates \( \{ \eta_t \} \), number of iterations \( T \), discount factor \( \gamma \), initial estimate \( V_0 \).
2: for \( t = 1, 2, \ldots, T \) do
3: \hspace{1em} Draw \( s_t(s) \sim P(\cdot|s) \) for each \( s \in S \).
4: \hspace{1em} Compute \( V_t \) according to (7).
5: end for

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 \( A \) is a singleton (i.e., \(|A| = 1\)). In such a case, the MDP reduces to a Markov reward process (MRP) (Bertsekas, 2017), and we shall abuse the notation to use \( P : S \to \Delta(S) \) to describe the probability transition kernel, and employ \( r : S \to [0,1] \) to represent the reward function (with \( r(s) \) indicating the immediate reward gained in state \( s \)). The TD learning algorithm maintains an estimate \( V_t : S \to \mathbb{R} \) of the value function in each iteration \( t \), and carries out the following iterative update rule

\[
V_t(s) = (1 - \eta_t)V_{t-1}(s) + \eta_t(r(s) + \gamma V_{t-1}(s_t)), \quad s_t \equiv s_t(s) \sim P(\cdot|s)
\]

for each state \( s \in S \). 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 \( [0,\frac{1}{\gamma}] \), and in each iteration, the samples \( \{s_t(s) \mid s \in S\} \) 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 the extension to 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 \(|A| = 1\) and characterize the \( l_{\infty}\)-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 \( 0 \leq t \leq T \), the learning rates satisfy

\[
\frac{1}{1 + c_1(1-\gamma)T} \leq \eta_t \leq \frac{1}{1 + c_2(1-\gamma)T} \tag{8a}
\]

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

\[
T \geq \frac{c_3 \log^3 T \log |S|^2}{(1-\gamma)^2 \varepsilon^2} \tag{8b}
\]

There is no need to maintain additional Q-estimates, as the Q-function and the value function coincide when \(|A| = 1\).
for some sufficiently large universal constant $c_3 > 0$. If the initialization obeys $0 \leq V_0(s) \leq \frac{1}{1-\gamma}$ for all $s \in S$, then with probability at least $1 - \delta$, Algorithm 2 achieves

$$\max_{s \in S} |V_T(s) - V^*(s)| \leq \varepsilon.$$  \hspace{1cm} (9)

**Remark 1 (Mean estimation error).** This high-probability bound immediately translates to a mean estimation error guarantee. Recognizing the crude upper bound $|V_T(s) - V^*(s)| \leq \frac{1}{1-\gamma}$ (see (104) in Section C.1) and taking $\delta \leq \varepsilon(1 - \gamma)$, we reach

$$\mathbb{E}\left[ \max_{s} |V_T(s) - V^*(s)| \right] \leq \varepsilon(1 - \delta) + \delta \frac{1}{1-\gamma} \leq 2\varepsilon,$$

provided that $T \geq \frac{c_3(\log^3 T)((\log \frac{|S|}{T})^{\frac{1}{2}} \varepsilon + 1)}{(1-\gamma)^3 \varepsilon^2}$. \hspace{1cm} (10)

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

$$\tilde{O}\left( \frac{|S|}{(1-\gamma)^3 \varepsilon^2} \right)$$

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., 2021a; Lakshminarayanan and Szepesvari, 2018; Wainwright, 2019b), Theorem 1 strengthens prior results by uncovering an improved scaling (i.e., $(1 - \gamma)^3$) 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 $\varepsilon$ approximation of the value function—in a minimax sense—unless the total sample size exceeds

$$\tilde{\Omega}\left( \frac{|S|}{(1-\gamma)^3 \varepsilon^2} \right).$$

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 (12) using model-based methods or variance-reduced model-free algorithms (e.g., Azar et al. (2013); Khamaru et al. (2021a); 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.

**Remark 2 (Runtime-oblivious learning rates).** Careful readers might remark that the choice (8a) of the learning rates might still rely on prior knowledge on $T$ (or log $T$). Fortunately, Theorem 1 immediately leads to convergence guarantees for another choice of $\eta_t$ selected completely independent of $T$. More specifically, suppose that the learning rates obey

$$\frac{1}{1 + \frac{c_1(1-\gamma)t}{\log^2(t+1)}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2(t+1)}}, \quad \forall t \geq 1$$

for some universal constants $\tilde{c}_1, \tilde{c}_2 > 0$. Then the claim (9) remains valid under this choice (13), provided that

$$T \geq \frac{2\tilde{c}_3(\log^3 T)((\log \frac{|S|}{T})^{\frac{1}{2}} \varepsilon + 1)}{(1-\gamma)^3 \varepsilon^2}.$$  \hspace{1cm} (14)

See Appendix C.3 for the proof.
Remark 3 (Polyak-Ruppert averaging). The results claimed in Remark 2 further allow us to control the estimation error of TD learning under Polyak-Ruppert averaging (Polyak and Juditsky, 1992). More precisely, under the choice (13) of learning rates, the averaged iterates satisfy
\[
\max_{s \in S} \left| \frac{1}{T} \sum_{t=1}^{T} V_T(s) - V^*(s) \right| \leq 4 \sqrt{\frac{c_3 (\log^3 T) \left( \log \frac{|S|/T}{\delta} \right)}{(1 - \gamma)^3 T}}
\]  
(15)
with probability exceeding \(1 - \delta\). See Appendix C.3 for the proof.

Remark 4. It is also noteworthy that: while the last iterate of plain TD learning is shown to be minimax optimal (which concerns worst-case optimality), it might not necessarily enjoy local optimality. As recently demonstrated by Khamaru et al. (2021b), additional algorithmic tricks like variance reduction might be needed in order to ensure local optimality.

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

Next, we move on to the more general case with \(|A| \geq 2\) and study the performance of Q-learning. As it turns out, Q-learning with \(|A| \geq 2\) is considerably more challenging to analyze than the TD learning case, due to the presence of the nonsmooth max operator. Our \(\ell_\infty\)-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 \(0 \leq t \leq T\), the learning rates satisfy
\[
\frac{1}{1 + \frac{c_1 (1 - \gamma) T}{\log T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2 (1 - \gamma) T}{\log T}}
\]  
(16a)
for some small enough universal constants \(c_1 \geq c_2 > 0\). Assume that the total number of iterations \(T\) obeys
\[
T \geq \frac{c_3 (\log^4 T) \left( \log \frac{|S||A|}{\varepsilon (1 - \gamma)^2 T} \right)}{(1 - \gamma)^4 \varepsilon^2}
\]  
(16b)
for some sufficiently large universal constant \(c_3 > 0\). If the initialization obeys \(0 \leq Q_0(s, a) \leq \frac{1}{1 - \gamma}\) for any \((s, a) \in S \times A\), then Algorithm 1 achieves
\[
\max_{(s, a) \in S \times A} \left| Q_T(s, a) - Q^*(s, a) \right| \leq \varepsilon
\]  
(17)
with probability at least \(1 - \delta\).

**Remark 5 (Mean estimation error).** Repeating exactly the same argument as in Remark 1, one can readily translate this high-probability bound into the following mean estimation error guarantee:
\[
\mathbb{E} \left[ \max_{s, a} \left| Q_T(s, a) - Q^*(s, a) \right| \right] \leq \varepsilon (1 - \delta) + \delta \frac{1}{1 - \gamma} \leq 2 \varepsilon,
\]  
(18)
holds as long as \(T \geq \frac{c_3 (\log^4 T) \left( \log \frac{|S||A|}{\varepsilon (1 - \gamma)^2} \right)}{(1 - \gamma)^4 \varepsilon^2} \).

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 \(|S||A|\) independent samples are drawn in each iteration, we can see from Theorem 2 the following sample complexity bound
\[
\tilde{O}\left( \frac{|S||A|}{(1 - \gamma)^4 \varepsilon^2} \right)
\]  
(19)
in order for Q-learning to attain \(\varepsilon\)-accuracy \((0 < \varepsilon < 1)\) in an entrywise sense. To the best of our knowledge, this is the first result that breaks the \(\frac{|S||A|}{(1 - \gamma)^4 \varepsilon^2}\) 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., 2022c; Qu and Wierman, 2020; Wainwright, 2019b).
Learning rates. Akin to the TD learning case, our result accommodates two commonly adopted learning rate schemes (cf. (16a)): (i) linearly rescaled learning rates \( \frac{1}{1 + \frac{c_1}{\log^2 t}} \), and (ii) iteration-invariant learning rates \( \frac{1}{1 + \frac{c_1}{\log^2 t}} \) (which depend on the total number of iterations \( T \) but not the iteration number \( t \)). In particular, when \( T = \frac{c_3(\log^4 T)(\log |S| |A|)}{(1 - \gamma)^3} \), the constant learning rates can be taken to be on the order of

\[
\eta_t \equiv O((1 - \gamma)^3 \varepsilon^2), \quad 0 \leq t \leq T,
\]

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

**Remark 6** (Runtime-oblivious learning rates and Polyak-Ruppert averaging). Akin to Remark 2, Theorem 2 can be easily extended to accommodate a family of learning rates chosen without prior knowledge of \( T \). More concretely, suppose that the learning rates obey

\[
\frac{1}{1 + \frac{c_1(1 - \gamma)t}{\log^2 (t+1)}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1 - \gamma)t}{\log^2 (t+1)}}, \quad \forall t \geq 1
\]

for some suitable constants \( c_1, c_2 > 0 \). Then the claim (17) continues to hold under this choice (20), provided that \( T/2 \geq \frac{c_3(\log^4 T)(\log |S| |A|)}{(1 - \gamma)^3} \). Additionally, similar to Remark 3, we can demonstrate that the averaged Q-learning iterates under the choice (20) of learning rates obey

\[
\max_{(s,a) \in S \times A} \left| \frac{1}{T} \sum_{t=1}^{T} Q_T(s,a) - Q^*(s,a) \right| \leq 4 \sqrt{c_3(\log^4 T)(\log |S| |A| T)} \frac{1}{(1 - \gamma)^4 T}
\]

with probability exceeding \( 1 - \delta \). The proofs of these results are identical to those of Remarks 2-3 (see Appendix C.3), and are hence omitted.

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 \( \frac{|S||A|}{(1 - \gamma)^2} \) 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 lower bound for plain Q-learning, with the aim of confirming the sharpness of Theorem 2 and revealing the sub-optimality of Q-learning.

**Theorem 3.** Assume that \( 3/4 \leq \gamma < 1 \) and that \( T \geq \frac{c_3}{(1 - \gamma)^2} \) 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) \( \eta_t = \frac{1}{1 + c_n(1 - \gamma)t} \) for all \( t \geq 0 \), or (ii) \( \eta_t \equiv \eta \) for all \( t \geq 0 \). There exists a \( \gamma \)-discounted MDP with \( |S| = 4 \) and \( |A| = 2 \) such that Algorithm 1—with any \( c_n > 0 \) and any \( \eta \in (0, 1) \)—obeys

\[
\max_{s \in S} \mathbb{E} \left[ \left| V_T(s) - V^*(s) \right|^2 \right] \geq \frac{c_b}{(1 - \gamma)^4 T \log^2 T},
\]

where \( c_b > 0 \) is some universal constant.

**Remark 7.** This theorem constructs a hard MDP instance with no more than 4 states and 2 actions, with the emphasis of unveiling the sub-optimality of horizon dependency. It can be generalized to accommodate larger state/action space, as we shall elucidate in Section 4.3.

**Remark 8.** Theorem 3 concentrates on two families of learning rates—rescaled linear, and constant learning rates—that are most widely used in practice. Note, however, that our current analysis does not readily generalize to arbitrary learning rates, which we leave for future investigation.
Theorem 3 provides an algorithm-dependent lower bound for vanilla Q-learning. As asserted by this theorem, it is impossible for Q-learning to attain $\varepsilon$-accuracy (in the sense that $\max_s \mathbb{E}[(V_T(s) - V^*(s))^2] \leq \varepsilon^2$) unless the number of iterations exceeds the order of

$$\frac{1}{(1 - \gamma)^3\varepsilon^2}$$

up to some logarithmic factor. Consequently, the performance guarantees for Q-learning derived in Theorem 2 are sharp in terms of the dependency on the effective horizon $\frac{1}{1-\gamma}$. On the other hand, it has been shown in prior literature that the minimax sample complexity limit a generative model is on the order of (Azar et al., 2013; Li et al., 2022c)

$$\frac{|S||A|}{(1 - \gamma)^3\varepsilon^2} \quad \text{(up to log factor);}$$

this in turn reveals the sub-optimality of plain Q-learning, whose horizon scaling is larger than the minimax limit by a factor of $\frac{1}{1-\gamma}$. Hence, more sophisticated algorithmic tricks are necessary in order to further reduce the sample complexity. For instance, a variance-reduced variant of Q-learning—namely, leveraging the idea of variance reduction originating from stochastic optimization (Johnson and Zhang, 2013) to accelerate convergence of Q-learning—has been shown to attain minimax optimality (23) for any $\varepsilon \in (0, 1]$; see Wainwright (2019c) for more details.

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 $M$, the notation $\|M\|_1 := \max_j \sum_i |M_{i,j}|$ is defined as the largest row-wise $\ell_1$ norm of $M$. For any vector $a = [a_i]_{i=1}^n \in \mathbb{R}^n$, we define $\sqrt{a}$ and $|\cdot|$ in a coordinate-wise manner, i.e. $\sqrt{a} := [\sqrt{a_i}]_{i=1}^n \in \mathbb{R}^n$ and $|a| := [|a_i|]_{i=1}^n \in \mathbb{R}^n$. For a set of vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ with $a_k = [a_{k,j}]_{j=1}^n$ (1 $\leq$ $k$ $\leq$ $m$), we define the max operator in an entrywise fashion such that $\max_{1 \leq k \leq m} a_k := [\max_{1 \leq j \leq n} a_{k,j}]_{j=1}^n$. For any vectors $a = [a_i]_{i=1}^n \in \mathbb{R}^n$ and $b = [b_i]_{i=1}^n \in \mathbb{R}^n$, the notation $a \leq b$ (resp. $a \geq b$) means $a_i \leq b_i$ (resp. $a_i \geq b_i$) for all $1 \leq i \leq n$. We also let $a \circ b = [a_i b_i]_{i=1}^n$ denote the Hadamard product. In addition, we denote by $1$ (resp. $e_i$) the all-one vector (resp. the $i$-th standard basis vector), and let $I$ be the identity matrix.

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

$$P^\pi := P \Pi^\pi \quad \text{and} \quad P_\pi := \Pi^\pi P,$$

(24)

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

$$\Pi^\pi := \begin{pmatrix} e_{\pi(1)}^\top \\ e_{\pi(2)}^\top \\ \vdots \\ e_{\pi(|S|)}^\top \end{pmatrix}$$

(25)

with $e_i$ the $i$-th standard basis vector. Moreover, for any vector $V \in \mathbb{R}^{|S|}$, we define $\text{Var}_P(V) \in \mathbb{R}^{(|S|\times|A|)}$ as follows:

$$\text{Var}_P(V) = P(V \circ V) - (PV) \circ (PV).$$

(26)

11
In other words, the \((s, a)\)-th entry of \(\text{Var}_P(V)\) corresponds to the variance \(\text{Var}_{s' \sim P(\cdot | s, a)}(V(s'))\) w.r.t. the distribution \(P(\cdot | s, a)\).

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

\[
\pi_t(s) = \min \left\{ a' \mid Q_t(s, a') = \max_{a''} Q_t(s, a'') \right\}.
\]

In other words, for any \(s \in S\), the policy \(\pi_t\) 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

\[
Q_t(s, \pi_t(s)) = V_t(s) \quad \text{and} \quad PV_t = P^\pi Q_t \geq P^\pi Q_t
\]

for any \(\pi\), where \(P^\pi\) is defined in \((24)\). Further, we introduce a matrix \(P_t \in \{0, 1\}^{||S|| \times |A| \times ||S||}\) such that

\[
P_t((s, a), s') := \begin{cases} 1, & \text{if } s' = s_t(s, a) \\ 0, & \text{otherwise} \end{cases}
\]

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

Finally, let \(X := ([||S||], |A|, \frac{1}{1-\gamma}, \frac{1}{\epsilon})\). The notation \(f(X) = O(g(X))\) or \(f(X) \lesssim g(X)\) (resp. \(f(X) \gtrsim g(X)\)) means that there exists a universal constant \(C_0 > 0\) such that \(|f(X)| \leq C_0|g(X)|\) (resp. \(|f(X)| \geq C_0|g(X)|\)). The notation \(f(X) \asymp g(X)\) means \(f(X) \lesssim g(X)\) and \(f(X) \gtrsim g(X)\) hold simultaneously. We define \(O(\cdot)\) 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 \(P_t\) (cf. \((29)\)) to perform the following update

\[
Q_t = (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma PV_{t-1})
\]

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

\[
\Delta_t := Q_t - Q^*
\]

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 \(\Delta_t\). In view of the update rule \((30)\), we arrive at the following elementary decomposition:

\[
\Delta_t = Q_t - Q^* = (1 - \eta_t)Q_{t-1} + \eta_t(r + \gamma PV_{t-1}) - Q^*
\]

\[
= (1 - \eta_t)(Q_{t-1} - Q^*) + \eta_t(r + \gamma PV_{t-1} - Q^*)
\]

\[
= (1 - \eta_t)\Delta_{t-1} + \eta_t( PV_{t-1} - PV^*)
\]

\[
= (1 - \eta_t)\Delta_{t-1} + \eta_t \gamma \{P(V_{t-1} - V^*) + (P_t - P)V_{t-1}\},
\]

where the third line exploits the Bellman equation \(Q^* = r + \gamma PV^*\). Further, the term \(P(V_{t-1} - V^*) \) can be linked with \(\Delta_{t-1}\) using the definition \((27)\) of \(\pi_t\) as follows

\[
P(V_{t-1} - V^*) = P^\pi_{t-1}Q_{t-1} - P^\pi V^* \leq P^\pi_{t-1}Q_{t-1} - P^\pi_{t-1}Q^* = P^\pi_{t-1}\Delta_{t-1},
\]

\[
P(V_{t-1} - V^*) = P^\pi_{t-1}Q_{t-1} - P^\pi V^* \geq P^\pi_{t-1}Q_{t-1} - P^\pi_{t-1}Q^* = P^\pi_{t-1}\Delta_{t-1},
\]

\[
\]
where we have made use of the relation (28). Substitute (33) into (32) to reach
\[
\Delta_t \leq (1 - \eta_t) \Delta_{t-1} + \eta_t \gamma \{ P^{\pi_{i-1}} \Delta_{t-1} + (P_i - P) V_{t-1} \};
\]
\[
\Delta_t \geq (1 - \eta_t) \Delta_{t-1} + \eta_t \gamma \{ P^{\pi_t} \Delta_{t-1} + (P_i - P) V_{t-1} \}.
\]
(34)

Applying these relations recursively, we obtain
\[
\Delta_t \leq \eta_0(t) \Delta_0 + \sum_{i=1}^{t} \eta_i(t) \gamma \{ P^{\pi_{i-1}} \Delta_{i-1} + (P_i - P) V_{i-1} \},
\]
\[
\Delta_t \geq \eta_0(t) \Delta_0 + \sum_{i=1}^{t} \eta_i(t) \gamma \{ P^{\pi_t} \Delta_{i-1} + (P_i - P) V_{i-1} \},
\]
(35)

where we define
\[
\eta_i(t) := \begin{cases} 
\prod_{j=1}^{t} (1 - \eta_j), & \text{if } i = 0, \\
\eta_i \prod_{j=i+1}^{t} (1 - \eta_j), & \text{if } 0 < i < t, \\
\eta_t, & \text{if } i = t.
\end{cases}
\]
(36)

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. (2022c); Wainwright (2019b)) tackled the second term on the right-hand side of the relation (34) via the following crude bounds:
\[
P^{\pi_{i-1}} \Delta_{i-1} \leq \| P^{\pi_{i-1}} \|_1 \| \Delta_{i-1} \|_{\infty} 1 = \| \Delta_{i-1} \|_{\infty} 1,
\]
\[
P^{\pi_t} \Delta_{i-1} \geq -\| P^{\pi_t} \|_1 \| \Delta_{i-1} \|_{\infty} 1 = -\| \Delta_{i-1} \|_{\infty} 1,
\]
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., \( \sum_{i=1}^{t} \eta_i(t) P^{\pi_{i-1}} \Delta_{i-1} \)) play a major role in sharpening the dependence of sample complexity on the effective horizon.

4.2.2 Key intertwined relations underlying \( \{ \| \Delta_t \|_{\infty} \} \)

By exploiting the crucial relations (35) derived above, we proceed to upper and lower bound \( \Delta_t \) separately. To be more specific, defining
\[
\beta := \frac{c_4(1 - \gamma)}{\log T}
\]
(37)
for some constant \( c_4 > 0 \), one can further decompose the upper bound in (35) into several terms:
\[
\Delta_t \leq \eta_0(t) \Delta_0 + \sum_{i=1}^{(1-\beta)t} \eta_i(t) \gamma \{ P^{\pi_{i-1}} \Delta_{i-1} + (P_i - P) V_{i-1} \}
\]
\[
= \zeta_t
\]
(38)
\[
+ \sum_{i=(1-\beta)t+1}^{t} \eta_i(t) \gamma (P_i - P) V_{i-1} + \sum_{i=(1-\beta)t+1}^{t} \eta_i(t) P^{\pi_{i-1}} \Delta_{i-1}.
\]
\( =: \xi_t \)
(39)

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

- Each component in the first term \( \zeta_t \) is fairly small, given that \( \eta_i(t) \) is sufficiently small for any \( i \leq (1 - \beta)t \) (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 \( \zeta_t \) becomes somewhat negligible.
- The second term \( \xi_t \), 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 \( \{ \Delta_i \} \).
In summary, the right-hand side of the above inequality can be further decomposed into some weighted superposition of $\{\Delta_t\}$ in addition to some negligible effect. This is formalized in the following two lemmas, which make apparent the key intertwined relations underlying $\{\Delta_t\}$.

**Lemma 1.** Suppose that $c_1c_2 \leq c_4/8$. With probability at least $1 - \delta$,

$$\Delta_t \leq 30 \sqrt{\left(\frac{\log^4 T}{\gamma^2 (1 - \gamma)^4 T}\right)} \left(1 + \max_{\frac{1}{2} \leq i < t} \|\Delta_i\|_{\infty}\right) 1$$

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

**Lemma 2.** Suppose that $c_1c_2 \leq c_4/8$. With probability at least $1 - \delta$,

$$\Delta_t \geq -30 \sqrt{\left(\frac{\log^4 T}{\gamma^2 (1 - \gamma)^4 T}\right)} \left(1 + \max_{\frac{1}{2} \leq i < t} \|\Delta_i\|_{\infty}\right) 1$$

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

**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

$$\|\Delta_t\|_{\infty} \leq 30 \sqrt{\left(\frac{\log^4 T}{\gamma^2 (1 - \gamma)^4 T}\right)} \left(1 + \max_{\frac{1}{2} \leq i < t} \|\Delta_i\|_{\infty}\right)$$

for all $t \geq \frac{T}{c_2 \log T}$ 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

$$\|\Delta_T\|_{\infty} \leq O\left(\frac{\left(\frac{\log^4 T}{\gamma^2 (1 - \gamma)^4 T}\right)}{1 - \gamma} + \frac{\log^4 T}{\gamma^2 (1 - \gamma)^4 T}\right)$$

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 with 4 states and 2 actions.** Let us construct an MDP $M_{\text{hard}}$ with state space $S = \{0, 1, 2, 3\}$ (see a pictorial illustration in Figure 4.3). We shall denote by $A_s$ the action space associated with state $s$. The probability transition kernel and reward function of $M_{\text{hard}}$ are specified as follows

- $A_0 = \{1\}$, $P(0 \mid 0, 1) = 1$, $r(0, 1) = 0$ (42a)
- $A_1 = \{1, 2\}$, $P(1 \mid 1, 1) = p$, $P(0 \mid 1, 1) = 1 - p$, $r(1, 1) = 1$ (42b)
  $P(1 \mid 1, 2) = p$, $P(0 \mid 1, 2) = 1 - p$, $r(1, 2) = 1$ (42c)
- $A_2 = \{1\}$, $P(2 \mid 2, 1) = p$, $P(0 \mid 2, 1) = 1 - p$, $r(2, 1) = 1$ (42d)
- $A_3 = \{1\}$, $P(3 \mid 3, 1) = 1$, $r(3, 1) = 1$ (42e)

where the parameter $p$ is taken to be

$$p = \frac{4\gamma - 1}{3\gamma}.$$  (43)

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.
As can be easily verified (which we omit here for the sake of brevity), our analysis developed for the smaller 

The latter case—corresponding to the learning rates adopted in establishing the upper bounds—is the most 

when the learning rates are either too small or too large, and analyze the dynamics w.r.t. state 1 to cope 

established a general sample complexity lower bound 

MDP (42) is directly applicable to studying the more general 

depends on the effective horizon). As it turns out, one can straightforwardly extend it to cover larger 

constructed above contains no more than 4 states and 2 actions (as the focus has been to unveil sub-optimal 

Extension: lower bounds for larger |S| and |A|. For pedagogical reasons, the hard instance (42) 

constructed above contains no more than 4 states and 2 actions (as the focus has been to unveil sub-optimal 

dependency on the effective horizon). As it turns out, one can straightforwardly extend it to cover larger 

state and action spaces, with a more general hard instance constructed as follows.

- We begin by generating the following sub-MDP, denoted by $\mathcal{M}_{\text{sub}}$, which comprises 4 states \{1, 2, 3, 4\} 

and no more than $|A| \geq 2$ actions:

\[
\begin{align*}
\mathcal{A}_0 &= \{1\}, & P(0 | 0, 1) &= 1, & r(0, 1) &= 0, \\
\mathcal{A}_1 &= \{1, \ldots, |A|\}, & P(1 | 1, a) &= p, & P(0 | 1, a) &= 1 - p, & r(1, a) &= 1, & \forall a \in \mathcal{A}_1 \\
\mathcal{A}_2 &= \{1\}, & P(2 | 2, 1) &= p, & P(0 | 2, 1) &= 1 - p, & r(2, 1) &= 1, \\
\mathcal{A}_3 &= \{1\}, & P(3 | 3, 1) &= 1, & r(3, 1) &= 1,
\end{align*}
\]

where $p$ is still set according to (43).

- The full MDP $\mathcal{M}_{\text{full}}$ is then constructed by generating $|S|/4$ independent copies of $\mathcal{M}_{\text{sub}}$.

As can be easily verified (which we omit here for the sake of brevity), our analysis developed for the smaller 

MDP (42) is directly applicable to studying the more general $\mathcal{M}_{\text{full}}$, revealing that the lower bound (55) 

w.r.t. the iteration number $T$ remains valid. Recognizing that the total sample size scales as $|S||A|T$, we have 

established a general sample complexity lower bound $\frac{|S||A|}{(1-\gamma)^2 \epsilon^2}$ for synchronous Q-learning to yield $\epsilon$-accuracy.
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 \( \pi_b \) is employed; in other words,

\[
    a_t \sim \pi_b(\cdot | s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot | s_t, a_t), \quad t \geq 0.
\]

When \( \pi_b \) 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 \( \mu_{\pi_b} \) the stationary distribution of this Markov chain. Note that the behavior policy \( \pi_b \) 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

\[
    Q_t(s_{t-1}, a_{t-1}) = (1 - \eta_t)Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t \{ r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in A} Q_{t-1}(s_t, a') \},
\]

\[
    Q_t(s, a) = Q_{t-1}(s, a) \quad \text{for all } (s, a) \neq (s_{t-1}, a_{t-1})
\]

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 \( t \):

\[
    V_t(s) := \max_{a \in A} Q_t(s, a) \quad \text{for all } s \in S.
\]

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 \( \pi_b \) 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

\[
    \mu_{\min} := \min_{(s, a) \in S \times A} \mu_{\pi_b}(s, a).
\]

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

\[
    t_{\text{mix}} := \min \left\{ t \mid \max_{(s, a) \in S \times A} d_{TV}(P^t(\cdot | s, a), \mu_{\pi_b}) \leq \frac{1}{4} \right\}.
\]

Here, \( d_{TV}(\mu, \nu) := \frac{1}{2} \sum_{x \in X} |\mu(x) - \nu(x)| \) indicates the total variation distance between two measures \( \mu \) and \( \nu \) over \( X \) (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. (2022c, Section 2) for a more detailed account of these quantities and assumptions.

---

\(^2\)See Paulin (2015, Section 1.2) for the definition of uniform ergodicity.
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 $0 \leq t \leq T$, the learning rates satisfy
\[
\eta_t \equiv \eta = \frac{c_1 \log^3 T}{(1-\gamma)T \mu_{\min}}
\]
for some universal constants $0 < c_1 \leq 1$. Assume that the total number of iterations $T$ obeys
\[
T \geq \frac{c_2 \log^2 |S||A||T|}{\mu_{\min}} \max\left\{ \frac{\log^3 T}{(1-\gamma)^4 \varepsilon^2}, \frac{t_{\text{mix}}}{1-\gamma} \right\}
\]
for some sufficiently large universal constant $c_2 > 0$. If the initialization obeys $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$ for all $(s,a) \in S \times A$, then asynchronous Q-learning (cf. (48)) satisfies
\[
\max_{(s,a) \in S \times A} |Q_T(s,a) - Q^*(s,a)| \leq \varepsilon
\]
with probability at least $1 - \delta$.

**Remark 9.** Similar to Remark 1 and Remark 5, one can immediately translate the above high-probability result into the following mean estimation error bound:
\[
\mathbb{E}\left[ \max_{s,a} |Q_T(s,a) - Q^*(s,a)| \right] \leq \varepsilon(1-\delta) + \delta \frac{1}{1-\gamma} \leq 2\varepsilon,
\]
which holds as long as $T \geq \frac{c_2 \log^2 |S||A||T|}{\mu_{\min}} \max\left\{ \frac{\log^3 T}{(1-\gamma)^4 \varepsilon^2}, \frac{t_{\text{mix}}}{1-\gamma} \right\}$ for some large enough constant $c_2 > 0$.

This theorem demonstrates that with high probability, the total sample size needed for asynchronous Q-learning to yield entrywise $\varepsilon$ accuracy is
\[
\tilde{O}\left(\frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}\right),
\]
provided that the learning rates are taken to be some proper constant (see (52a)). The first term in (54) resembles our sample complexity characterization of synchronous Q-learning (cf. (19)), except that we replace the number $|S||A|$ of state-action pairs in (19) with $1/\mu_{\min}$ in order to account for non-uniformity across state-action pairs. The second term in (54) 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., 2022c; Qu and Wierman, 2020) scale at least as $\frac{1}{(1-\gamma)^{\alpha}}$ in terms of the dependency on the effective horizon, with Theorem 4 being the first result to sharpen this dependency to $\frac{1}{(1-\gamma)^{\beta}}$. In particular, our sample complexity bound strengthens the state-of-the-art result Li et al. (2022c) by a factor up to $\frac{1}{1-\gamma}$, while improving upon Qu and Wierman (2020) by a factor of at least $\frac{|S||A|}{1-\gamma} \min\left\{ t_{\text{mix}}, \frac{1}{(1-\gamma)^{\alpha}} \right\}$.\(^3\)

Before concluding this section, we note that for a large enough sample size, the first term $\frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2}$ in (54) is essentially unimprovable (up to logarithmic factor). To make precise this statement, we develop a matching algorithm-dependent lower bound as follows, which parallels Theorem 3 previously developed for the synchronous case.

\(^3\)The sample complexity of Li et al. (2022c) scales as $\tilde{O}\left(\frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}\right)$, while the sample complexity of Qu and Wierman (2020) scales as $\tilde{O}\left(\frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)^{\alpha}}\right)$. It is worth noting that $1/\mu_{\min} \geq |S||A|$ and is therefore a large factor.
Theorem 5. Consider any $0.95 \leq \gamma < 1$. Suppose that $\mu_{\min} \leq \frac{1}{c_3 \log^2 T}$ and $T \geq \frac{c_3 \log^2 T}{\mu_{\min}(1-\gamma)}$ for some sufficiently large constant $c_3 > 0$. Assume that the initialization is $Q_0 \equiv 0$, and that the learning rates are taken to be $\eta_t \equiv \eta$ for all $t \geq 0$. Then there exist a $\gamma$-discounted MDP with $|S| = 4$ and $|A| = 3$ and a behavior policy such that (i) the minimum state-action occupancy probability of the sample trajectory is given by $\mu_{\min}$, and (ii) the asynchronous Q-learning update rule (48)—for any $\eta \in (0, 1)$—obeys
\[
\max_{s,a} E \left[ \left| Q_T(s,a) - Q^*(s,a) \right|^2 \right] \geq \frac{c_b}{\mu_{\min}(1-\gamma)^4 T \log^3 T},
\]
where $c_b > 0$ is some universal constant.

In words, Theorem 5 asserts that, for large enough sample size $T$, in general one cannot hope to achieve $\ell_\infty$-based $\varepsilon$-accuracy using fewer than $\tilde{O} \left( \frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2} \right)$ samples, thus confirming the sharpness of our upper bound. The proof of this theorem can be found in Appendix F.

6 Concluding remarks

In this paper, we have settled the sample complexity of synchronous Q-learning in $\gamma$-discounted infinite-horizon MDPs, which is shown to be on the order of $\tilde{O} \left( \frac{|S|}{\mu_{\min}(1-\gamma)^4 \varepsilon^2} \right)$ when $|A| = 1$ and $\tilde{O} \left( \frac{|S||A|}{\mu_{\min}(1-\gamma)^4 \varepsilon^2} \right)$ when $|A| \geq 2$. A matching lower bound has been developed when $|A| \geq 2$ 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 grant 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, CCF-2007911, CCF-2106778, ECCS-2126634, and DMS-2134080. 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.

Theorem 6. Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying
\[
|X_k| \leq R \quad \text{and} \quad \mathbb{E} \left[ X_k \mid \{X_j\}_{j<k} \right] = 0 \quad \text{for all } k \geq 1.
\]

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

18
where we write $\mathbb{E}_{k-1}$ for the expectation conditional on $\{X_j\}_{j<k}$. Then for any given $\sigma^2 \geq 0$, one has

$$
\mathbb{P}\left\{ |Y_n| \geq \tau \text{ and } W_n \leq \sigma^2 \right\} \leq 2 \exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{56}
$$

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

$$
|Y_n| \leq \sqrt{8 \max\left\{ W_n, \frac{\sigma^2}{2K}\right\}} \log \frac{2K}{\delta} + \frac{4}{3} R \log \frac{2K}{\delta}. \tag{57}
$$

**Proof.** See Freedman (1975); Tropp (2011) for the proof of (56). As an immediate consequence of (56), one has

$$
\mathbb{P}\left\{ |Y_n| \geq \sqrt{4\sigma^2 \log \frac{2}{\delta} + \frac{4}{3} R \log \frac{2}{\delta}} \text{ and } W_n \leq \sigma^2 \right\} \leq \delta. \tag{58}
$$

Next, we turn attention to (57). Consider any positive integer $K$. As can be easily seen, the event

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

is contained within the union of the following $K$ events

$$
\mathcal{H}_K \subseteq \bigcup_{0 \leq k < K} B_k,
$$

where we define

$$
B_k := \left\{ |Y_n| \geq \sqrt{4\sigma^2 \log \frac{2K}{\delta} + \frac{4}{3} R \log \frac{2K}{\delta}} \text{ and } \sigma^2 \leq W_n \leq \frac{\sigma^2}{2^{k-1}} \right\}, \quad 1 \leq k \leq K-1,
$$

$$
B_0 := \left\{ |Y_n| \geq \sqrt{4\sigma^2 \log \frac{2K}{\delta} + \frac{4}{3} R \log \frac{2K}{\delta}} \text{ and } W_n \leq \frac{\sigma^2}{2^{K-1}} \right\}.
$$

Invoking inequality (58) with $\sigma^2$ set to be $\frac{\sigma^2}{2^{k-1}}$ and $\delta$ set to be $\frac{\delta}{K}$, we arrive at $\mathbb{P}\{B_k\} \leq \delta/K$. Taken this fact together with the union bound gives

$$
\mathbb{P}\{\mathcal{H}_K\} \leq \sum_{k=0}^{K-1} \mathbb{P}\{B_k\} \leq \delta.
$$

This concludes the proof. \hfill \Box

### 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 $\varepsilon$-range $\varepsilon \in (0, \frac{1}{1-\gamma}]$, as stated below.

**Theorem 7.** Consider any $\gamma \in (0,1)$ and any $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$. Theorem 2 continues to hold if

$$
T \geq \frac{c_3 \left(\log^4 T\right) \left(\log \frac{|S| |A| T}{\delta}\right)}{\gamma^2 (1-\gamma) \min\{\varepsilon^2, \varepsilon\}} \tag{59}
$$

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

**Remark** 10. Clearly, Theorem 7 subsumes Theorem 2 as a special case.

As one can anticipate, the proof of Theorem 7 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 \( Q_t \) and \( V_t \). 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 \( 0 \leq Q_0 \leq \frac{1}{1-\gamma} \). Then for any \( t \geq 0 \),

\[
0 \leq Q_t \leq \frac{1}{1-\gamma} 1 \quad \text{and} \quad 0 \leq V_t \leq \frac{1}{1-\gamma} 1. \tag{60}
\]

**Proof.** We shall prove this by induction. First, our initialization trivially obeys (60) for \( t = 0 \). Next, suppose that (60) is true for the \((t-1)\)-th iteration, namely,

\[
0 \leq Q_{t-1} \leq \frac{1}{1-\gamma} 1 \quad \text{and} \quad 0 \leq V_{t-1} \leq \frac{1}{1-\gamma} 1, \tag{61}
\]

and we intend to justify the claim for the \(t\)-th iteration. Recognizing that \( 0 \leq r \leq 1 \), \( P_t \geq 0 \) and \( \| P_t \|_1 = 1 \), one can straightforwardly see from the update rule (30) and the induction hypothesis (61) that

\[
Q_t = (1 - \eta_t)Q_{t-1} + \eta_t (r + \gamma P_t V_{t-1}) \geq 0
\]

and

\[
Q_t = (1 - \eta_t)Q_{t-1} + \eta_t (r + \gamma P_t V_{t-1}) \\
\leq (1 - \eta_t) \| Q_{t-1} \|_\infty 1 + \eta_t (\| r \|_\infty + \gamma \| P_t \|_1 \| V_{t-1} \|_\infty) 1 \\
\leq (1 - \eta_t) \frac{1}{1-\gamma} 1 + \eta_t \left( 1 + \frac{\gamma}{1-\gamma} \right) 1 = \frac{1}{1-\gamma} 1.
\]

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

\[
0 \leq V_t \leq \frac{1}{1-\gamma} 1,
\]

thus establishing (60) for the \(t\)-th iteration. Applying the induction argument then concludes the proof. \( \square \)

As a result of Lemma 4 and the fact \( 0 \leq Q^* \leq \frac{1}{1-\gamma} 1 \), we have

\[
\| Q_t - Q^* \|_\infty \leq \frac{1}{1-\gamma} \quad \text{for all} \ t \geq 0, \tag{62}
\]

which also confirms that \( 0 \leq \varepsilon \leq \frac{1}{1-\gamma} \) is the full \( \varepsilon \)-range we need to consider. Further, we make note of a direct consequence of the claimed iteration number (59) when \( \varepsilon \leq \frac{1}{1-\gamma} \):

\[
T = \frac{c_3 (\log^4 T)(\log \frac{|S|\cdot|A|}{\delta})}{(1-\gamma)^2 \min\{\varepsilon, \varepsilon^2\}} \geq \frac{c_3 (\log^4 T)(\log \frac{|S|\cdot|A|}{\delta})}{(1-\gamma)^3}, \tag{63}
\]

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 (36):

\[
\eta_t^{(t)} := \begin{cases} 
\prod_{j=1}^t (1 - \eta_j), & \text{if } i = 0, \\
\eta_i \prod_{j=i+1}^t (1 - \eta_j), & \text{if } 0 < i < t, \\
\eta_t, & \text{if } i = t. 
\end{cases} \tag{64}
\]
We now take a moment to bound $\eta_i^{(t)}$. From our assumption (16a) and the condition (63), we know that the learning rate obeys
\[
\frac{1}{2c_1(1 - \gamma)T/\log^3 T} \leq \frac{1}{1 + c_1(1 - \gamma)T/\log^3 T} \leq \eta_i \leq \frac{1}{1 + c_2(1 - \gamma)t/\log^3 T} \leq \frac{1}{c_2(1 - \gamma)t/\log^3 T}
\] (65)
for some constants $c_1, c_2 > 0$. Recalling that
\[
\beta := \frac{c_4(1 - \gamma)}{\log T}
\] (66)
for some universal constant $c_4 > 0$ and considering any $t$ obeying
\[
t \geq \frac{T}{c_2 \log T},
\] (67)
we shall bound $\eta_i^{(t)}$ by looking at two cases separately.

- For any $0 \leq i \leq (1 - \beta)t$, we can use (65) to show that
  \[
  \eta_i^{(t)} \leq \left(1 - \frac{1}{2c_1(1 - \gamma)T/\log^3 T}\right)^{\beta t} \leq \left(1 - \frac{1}{2c_1(1 - \gamma)T/\log^3 T}\right)^{\frac{c_4(1 - \gamma)T}{c_2 \log^3 T}} \times \left(1 - \frac{1}{2c_1(1 - \gamma)T/\log^3 T}\right)^{\frac{c_4 \log T}{c_2 \log^3 T}} < \frac{1}{2T^2},
  \] (68a)
  where the last inequality holds as long as $c_1c_2 \leq c_4/8$.

- When it comes to the case with $i > (1 - \beta)t \geq t/2$, one can upper bound
  \[
  \eta_i^{(t)} \leq \eta_i \leq \frac{1}{c_2(1 - \gamma)i/\log^3 T} < \frac{2}{c_2(1 - \gamma)t/\log^3 T} \leq \frac{2 \log^4 T}{(1 - \gamma)T},
  \] (68b)
  where we have used the constraint (67).

Moreover, the sum of $\eta_i^{(t)}$ over $i$ obeys
\[
\sum_{i=0}^{t} \eta_i^{(t)} = \prod_{j=1}^{t} (1 - \eta_j) + \eta_1 \prod_{j=2}^{t} (1 - \eta_j) + \eta_2 \prod_{j=3}^{t} (1 - \eta_j) + \cdots + \eta_{t-1}(1 - \eta_t) + \eta_t
= \prod_{j=2}^{t} (1 - \eta_j) + \eta_2 \prod_{j=3}^{t} (1 - \eta_j) + \cdots + \eta_{t-1}(1 - \eta_t) + \eta_t = \cdots
= (1 - \eta_t) + \eta_t = 1.
\] (69)

Repeating the same argument further allows us to derive
\[
\sum_{i=\tau}^{t} \eta_i^{(t)} = 1 - \prod_{j=\tau}^{t} (1 - \eta_j)
\] (70)
for any $\tau \leq t$.

**B.2 Proof of Lemma 1**

We shall exploit the relation (39) to prove this lemma. One of the key ingredients of our analysis lies in controlling the terms $\zeta_t$ and $\xi_t$ introduced in (39), which in turn enables us to apply (39) recursively to control $\Delta_t$.  

Step 1: bounding $\zeta_t$. We start by developing an upper bound on $\zeta_t$ (cf. (39)) for any $t$ obeying $\frac{T}{c_2 \log T} \leq t \leq T$. Invoking the preceding upper bounds (68) on $\eta^{(t)}_i$ implies that
\[
\|\zeta_t\|_\infty \leq \eta^{(t)}_0 \|\Delta_0\|_\infty + t \max_{i \leq (1-\beta)t} \eta^{(t)}_i \max_{1 \leq i \leq (1-\beta)t} \left( \|P^{\pi_{i-1}} \Delta_{i-1}\|_\infty + \|P_i V_{i-1}\|_\infty + \|P V_{i-1}\|_\infty \right)
\leq \eta^{(t)}_0 \|\Delta_0\|_\infty + t \max_{i \leq (1-\beta)t} \eta^{(t)}_i \max_{1 \leq i \leq (1-\beta)t} \left\{ \|P^{\pi_{i-1}} 1_1 \|_\infty + (\|P_i\|_1 + \|P\|_1) \|V_{i-1}\|_\infty \right\}
\leq \frac{1}{2 T^2} \frac{1}{1 - \gamma} + \frac{1}{2 T^2} \cdot t \cdot \frac{3}{1 - \gamma}
\leq \frac{2}{(1 - \gamma)^T}.
\]
Here, (i) holds since $\|P^{\pi_{i-1}} 1_1\|_1 = \|P_i\|_1 = \|P\|_1 = 1$ (as they are all probability transition matrices), whereas (ii) arises from the previous bound (68a).

Step 2: bounding $\xi_t$. Moving on to the term $\xi_t$, let us express it as
\[
\xi_t = \sum_{i=(1-\beta)t+1}^t z_i \quad \text{with} \quad z_i : = \eta^{(t)}_i (P_i - P)V_{i-1},
\]
where the $z_i$'s satisfy
\[
\mathbb{E}[z_i | V_{i-1}, \ldots, V_0] = 0.
\]
This motivates us to invoke Freedman’s inequality (see Theorem 6) to control $\xi_t$ for any $t$ obeying $\frac{T}{c_2 \log T} \leq t \leq T$. Towards this, we need to calculate several quantities.

- First, it is seen that
  \[
  B := \max_{(1-\beta)t \leq i \leq t} \|z_i\|_\infty \leq \max_{(1-\beta)t \leq i \leq t} \eta^{(t)}_i (P_i - P)V_{i-1}\|_\infty
  \leq \max_{(1-\beta)t \leq i \leq t} \eta^{(t)}_i (\|P_i\|_1 + \|P\|_1) \|V_{i-1}\|_\infty \leq \frac{4 \log^4 T}{(1 - \gamma)^2 T},
  \]
  where the last inequality is due to (68b), Lemma 4, and the fact $\|P_i\|_1 = \|P\|_1 = 1$.

- Next, we turn to certain variance terms. For any vector $a = [a_j]$, let us use $\text{Var}(a | V_{i-1}, \ldots, V_0)$ to denote a vector whose $j$-th entry is given by $\text{Var}(a_j | V_{i-1}, \ldots, V_0)$. With this notation in place, and recalling the notation $\text{Var}_P(z)$ in (26), we obtain
  \[
  W_t := \sum_{i=(1-\beta)t+1}^t \text{Var}(z_i | V_{i-1}, \ldots, V_0) = \gamma^2 \sum_{i=(1-\beta)t+1}^t (\eta^{(t)}_i)^2 \text{Var}_P((P_i - P)V_{i-1} | V_{i-1})
  = \gamma^2 \sum_{i=(1-\beta)t+1}^t (\eta^{(t)}_i)^2 \text{Var}_P(V_{i-1})
  \leq \left( \max_{(1-\beta)t \leq i \leq t} \eta^{(t)}_i \right) \left( \sum_{i=(1-\beta)t+1}^t \eta^{(t)}_i \right) \max_{(1-\beta)t \leq i \leq t} \text{Var}_P(V_i)
  \leq \frac{2 \log^4 T}{(1 - \gamma)^T} \max_{(1-\beta)t \leq i \leq t} \text{Var}_P(V_i),
  \]
  where the last inequality relies on the previous bounds (68b) and (69).
• In the meantime, Theorem 4 leads us to the following trivial upper bound:

\[
|W_t| \leq \frac{2 \log^4 T}{(1-\gamma)^T} \cdot \frac{1}{(1-\gamma)^2} \leq \frac{2 \log^4 T}{(1-\gamma)^3 T} = \sigma^2 T.
\]

By setting \( K = \left[ 2 \log_2 \left( \frac{1}{1-\gamma} \right) \right] \), one has

\[
\frac{\sigma^2}{2K} \leq \frac{2 \log^4 T}{(1-\gamma)^T}.
\]

With the above bounds in place, applying the Freedman inequality in Theorem 6 and invoking the union bound over all the \(|S||A|\) entries of \( \xi_t \), demonstrate that

\[
|\xi_t| \leq \sqrt{8 \left( W_t + \frac{\sigma^2}{2K} \right) \log \frac{8|S||A|T \log \frac{1}{1-\gamma}}{\delta} + \left( \frac{4}{3} B \log \frac{8|S||A|T \log \frac{1}{1-\gamma}}{\delta} \right) \cdot 1}
\]

\[
\leq \sqrt{16 \left( W_t + \frac{2 \log^4 T}{(1-\gamma)^T} \right) \log \frac{|S||A|T}{\delta} + \left( 3B \log \frac{|S||A|T}{\delta} \right) \cdot 1}
\]

\[
\leq \sqrt{32 \left( \log^4 T \right) \left( \frac{\max_{(1-\beta)\leq i \leq t} \text{Var}_P(V_i) + 1}{(1-\gamma)^T} \right) \cdot 1}
\]

with probability at least \( 1 - \delta/T \). Here, the second line holds due to (72) and the fact \( \log \frac{8|S||A|T \log \frac{1}{1-\gamma}}{\delta} \leq 2 \log \frac{|S||A|T}{\delta} \) (cf. (63)), whereas the last inequality makes use of the relation (71).

**Step 3: using the bounds on \( \xi_t \) and \( \xi_t \) to control \( \Delta_i \).** Let us define

\[
\phi_t := 64 \log^4 T \log \frac{|S||A|T}{(1-\gamma)^T} \left( \max_{\frac{t}{4} \leq i \leq t} \text{Var}_P(V_i) + 1 \right)
\]

In view of the upper bounds derived in Steps 1 and 2, and \( \beta \) defined in (66), we have—with probability exceeding \( 1 - \delta \)—that

\[
|\xi_k| + |\xi_k| \leq \sqrt{\phi_t} \quad \text{for all} \quad 2t/3 \leq k \leq t,
\]

provided that \( T \geq \frac{c_3 (\log^4 T) \left( \log \frac{|S||A|T}{(1-\gamma)^T} \right)}{\delta} \) for some sufficiently large constant \( c_3 > 0 \). Substituting (74) into (39), we can upper bound \( \Delta_i \) as follows

\[
\Delta_k \leq \sqrt{\phi_t} + \sum_{k=1}^{k} \eta_{i_k}^{(k)} \gamma \mathcal{P}_{\pi_i} \Delta_{i-1} = \sqrt{\phi_t} + \sum_{k=1}^{k-1} \eta_{i_k}^{(k)} \gamma \mathcal{P}_{\pi_i} \Delta_i \quad \text{for all} \quad 2t/3 \leq k \leq t.
\]

Further, we find it convenient to define \( \{\alpha_t^{(i)}\} \) as follows

\[
\alpha_t^{(i)} := \frac{\eta_{i+1}^{(i)}}{\sum_{j=(1-\beta)k}^{i} \eta_{i+1}^{(j)}}.
\]

Clearly, this sequence satisfies

\[
\alpha_t^{(i)} \geq \eta_{i+1}^{(i)} \quad \text{and} \quad \sum_{i=(1-\beta)t}^{t-1} \alpha_t^{(i)} = 1
\]

for any \( t \), where the first inequality results from (69). With these in place, we can write (75) as

\[
\Delta_k \leq \sqrt{\phi_t} + \sum_{i=(1-\beta)k}^{k-1} \eta_{i+1}^{(k)} \gamma \mathcal{P}_{\pi_i} \Delta_{i-1} = \sum_{i=(1-\beta)k}^{k-1} \left( \alpha_t^{(i)} \sqrt{\phi_t} + \eta_{i+1}^{(k)} \gamma \mathcal{P}_{\pi_i} \Delta_{i-1} \right) \quad \text{for all} \quad 2t/3 \leq k \leq t.
\]
Given that \((1 - \beta)t \geq 2t/3\) (see (66)), we can invoke this relation recursively to yield

\[
\Delta_t \leq \sum_{i_1 = (1 - \beta)t}^{(t-1)} \left( \alpha_{i_1}^{(t)} \sqrt{\varphi_t} + \eta_{i_1+1}^{(t)} \gamma \mathbf{P}^{\pi_{i_1}} \Delta_{i_1} \right)
\]

\[
\leq \sum_{i_1 = (1 - \beta)t}^{(t-1)} \alpha_{i_1}^{(t)} \sqrt{\varphi_t} + \sum_{i_1 = (1 - \beta)t \equiv (1 - \beta) \beta} \sum_{i_2 = (1 - \beta) i_1}^{i_1-1} \left( \alpha_{i_2}^{(i_1)} \gamma \mathbf{P}^{\pi_{i_2}} \sqrt{\varphi_t} \right)
\]

\[
\geq \sum_{i_1 = (1 - \beta)t \equiv (1 - \beta) \beta} \sum_{i_2 = (1 - \beta) i_1}^{i_1-1} \eta_{i_1+1}^{(i_1)} \eta_{i_2+1}^{(i_1)} \prod_{k=1}^{2} \left( \gamma \mathbf{P}^{\pi_{i_k}} \right) \Delta_{i_2}, \tag{79}
\]

where the second inequality relies on (78), the third line uses the inequality \(\eta_{i_1+1}^{(t)} \leq \alpha_{i_1}^{(t)}\) in (77), and the fourth line is valid since \(\sum_{i_1 \equiv (1 - \beta) \beta} \alpha_{i_1}^{(i_1)} = 1\) (see (77)).

We intend to continue invoking (78) recursively—similar to how we derive (79)—in order to control \(\Delta_t\). To do so, we are in need of some preparation. First, let us define

\[
H := \frac{\log T}{1 - \gamma} \quad \text{and} \quad \alpha_{(i)}^{(H)} := \alpha_{i_1}^{(i_1)} \alpha_{i_2}^{(i_2)} \cdots \alpha_{i_H}^{(i_H)} \geq 0 \tag{80}
\]

for any \(t > i_1 > i_2 > \cdots > i_H\), which clearly satisfies (see (77))

\[
\alpha_{(i)}^{(H)} = \eta_{i_1+1}^{(i_1)} \eta_{i_2+1}^{(i_1)} \cdots \eta_{i_H+1}^{(i_{H-1})}. \tag{81}
\]

In addition, defining the index set

\[
\mathcal{I}_t := \left\{ (i_1, \cdots, i_H) \mid (1 - \beta)t \leq i_1 \leq t - 1, \forall 1 \leq j < H : (1 - \beta)i_j \leq i_{j+1} \leq i_j - 1 \right\}, \tag{82}
\]

we have

\[
\sum_{(i_1, \cdots, i_H) \in \mathcal{I}_t} \alpha_{(i)}^{(H)} = 1. \tag{83}
\]

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

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

for \(c_4\) small enough, thus implying that

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

This is an important property that allows one to invoke the relation (78). With these in place, applying the preceding relation (78) recursively—in a way similar to (79)—further leads to

\[
\Delta_t \leq \sum_{(i_1, \cdots, i_H) \in \mathcal{I}_t} \alpha_{(i)}^{(H)} e^{\left( I + \sum_{h=1}^{H-1} \gamma P^{\pi_{i_h}} \right) \sqrt{\varphi_t} + \sum_{h=1}^{H} \frac{H}{k-1} \prod_{k=1}^{2} \left( \gamma \mathbf{P}^{\pi_{i_k}} \right) \Delta_{i_H}}
\]

\[
\leq \max_{(i_1, \cdots, i_H) \in \mathcal{I}_t} \left\{ e^{\left( I + \sum_{h=1}^{H-1} \gamma P^{\pi_{i_h}} \right) \sqrt{\varphi_t} + \sum_{h=1}^{H} \frac{H}{k-1} \prod_{k=1}^{2} \left( \gamma \mathbf{P}^{\pi_{i_k}} \right) \Delta_{i_H}} \right\} \tag{84}
\]

for all \(t \geq \frac{T}{e^{\log T}}\), where we recall the definition of the entrywise max operator in Section 4.1. Here, the last inequality relies on the fact that \(\sum_{(i_1, \cdots, i_H) \in \mathcal{I}_t} \alpha_{(i)}^{(H)} = 1\) (see (83)). It remains to control \(\beta_1\) and \(\beta_2\), which we shall accomplish separately in the next two steps.
Step 4: bounding $\beta_2$. The term $\beta_2$ defined in (84) is relatively easier to control. Observing that $\prod_{k=1}^{H} P_{\pi_{ik}}$ is still a probability transition matrix, we can derive

$$|\beta_2| = \gamma^H \prod_{1 \leq k \leq h} P_{\pi_{ik}} |\Delta_{iH}| \leq \gamma^H \prod_{1 \leq k \leq h} P_{\pi_{ik}} \frac{\|\Delta_{iH}\|_{\infty}}{\|\Delta_{iH}\|_{1}} = \gamma^H \frac{\|\Delta_{iH}\|_{\infty}}{\|\Delta_{iH}\|_{1}}$$

$$(i) \leq \frac{1}{1-\gamma} \gamma^H \leq \frac{1}{(1-\gamma)T},$$

where (i) results from the crude bound (62). To justify the inequality (ii), we recall the definition (80) of $H$ to see that

$$\gamma^H \leq (1 - (1 - \gamma)) \frac{\sqrt{T}}{\log T} \leq e^{-\log T} = \frac{1}{T},$$

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

Step 5: bounding $\beta_1$. When it comes to the term $\beta_1$ defined in (84), we can upper bound the entrywise square of $\beta_1$—denoted by $|\beta_1|^2$—as follows

$$|\beta_1|^2 = \left| \sum_{h=0}^{H-1} \gamma^h \prod_{1 \leq k \leq h} P_{\pi_{ik}} \varphi_t \right|^2 \leq \sum_{h=0}^{H-1} \gamma^{h/2} \cdot \gamma^{h/2} \sqrt{\prod_{1 \leq k \leq h} P_{\pi_{ik}} \varphi_t}^2$$

$$(ii) \leq \sum_{h=0}^{H-1} \gamma^h \cdot \prod_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} P_{\pi_{ik}} \varphi_t$$

$$(iii) \leq \frac{1}{1 - \gamma} \sum_{h=0}^{H-1} \gamma^h \prod_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} P_{\pi_{ik}} \frac{1}{(1 - \gamma)^2 T} \left( \max_{1 \leq i < t} \text{Var}_P(V_i) + 1 \right)$$

$$(iv) \leq \frac{64 \left( \log^4 T \right) (\log \frac{|S||A|}{\delta})}{(1 - \gamma)^2 T} \max_{1 \leq i < t} \text{Var}_P(V_i) + \frac{64 \left( \log^4 T \right) (\log \frac{|S||A|}{\delta})}{(1 - \gamma)^2 T} \cdot 1.$$

Here, (i) follows from Jensen’s inequality and the fact that $\prod_{k=1}^{h} P_{\pi_{ik}}$ is a probability transition matrix; (ii) holds due to the Cauchy-Schwarz inequality; (iii) utilizes the definition of $\varphi_t$ in (73); (iv) follows since $\prod_{1 \leq k \leq h} P_{\pi_{ik}} 1 = 1$ and $\sum_{0 \leq \gamma \leq H} \gamma^h \leq \frac{1}{1-\gamma}$. 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 $(i_1, \cdots, i_H) \in I_t$, the following holds:

$$\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} P_{\pi_{ik}} \max_{1 \leq i < t} \text{Var}_P(V_i) \leq 4 \gamma^{2(1-\gamma)^2} \left( 1 + 2 \max_{1 \leq i < t} \|\Delta_i\|_{\infty} \right) 1. \quad (85)$$

**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. \qed

Therefore, the above result directly implies that

$$|\beta_1|^2 \leq \frac{320 \left( \log^4 T \right) (\log \frac{|S||A|}{\delta})}{\gamma^2 (1 - \gamma)^4 T} \left( 1 + 2 \max_{1 \leq i < t} \|\Delta_i\|_{\infty} \right) 1. \quad (86)$$

Step 6: putting all this together. Substituting the preceding bounds for $\beta_1$ and $\beta_2$ into (84), we can demonstrate that: with probability at least $1 - \delta$,

$$\Delta_i \leq \frac{1}{1 - \gamma} \cdot 1 + \sqrt{\frac{320 \left( \log^4 T \right) (\log \frac{|S||A|}{\delta})}{\gamma^2 (1 - \gamma)^4 T} \left( 1 + 2 \max_{1 \leq i < t} \|\Delta_i\|_{\infty} \right) 1}.$$
\[
\leq 30 \sqrt{\frac{\log^4 T \log \frac{|S| |A|}{T}}{\gamma^2 (1 - \gamma)^4 T}} \left( 1 + \max_{\frac{1}{2} \leq i < t} \| \Delta_i \|_\infty \right) \cdot 1
\]  

(87)

holds simultaneously for all \( t \geq \frac{T}{c_2 \log T} \), where the second line is valid since \( \frac{1}{(1 - \gamma)^2} \leq \sqrt{\frac{\log^4 T \log \frac{|S| |A|}{T}}{\gamma^2 (1 - \gamma)^4 T}} \) under our sample size condition (63).

### B.3 Proof of Lemma 2

Next, we move forward to develop a lower bound on \( \Delta_t \), which can be accomplished in an analogous manner as for the above upper bound. Applying a similar argument for (84) (except that we need to replace \( \pi_i \) with \( \pi^\star \)), one can deduce that

\[
\Delta_t \geq - \max_{(i_1, \ldots, i_H) \in I_t} \left\{ \sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} P_{i_H}^{\pi^\star} \sqrt{\phi_t} + \gamma^H \prod_{k=1}^{H} P_{i_H}^{\pi^\star} |\Delta_{i_H}| \right\}
\]  

(88)

for any \( t \geq \frac{T}{c_2 \log T} \). It is straightforward to bound the second term on the right-hand side of (88) as

\[
\gamma^H \prod_{1 \leq k \leq H} P^{\pi^\star} \| \Delta_{i_H} \|_\infty 1 \leq \frac{1}{(1 - \gamma)^T} 1,
\]

where the second inequality makes use of (62) as well as the fact that \( \prod_{k} P_{\pi^\star} \) is a probability transition matrix (so that \( \| \prod_{k} P_{\pi^\star} \|_1 = 1 \)). As for the first term on the right-hand side of (88), we can invoke a similar argument for (86) to obtain

\[
\left\| \sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} P_{i_H}^{\pi^\star} \sqrt{\phi_t} \right\|^2 \leq 320 \frac{(\log^4 T \log \frac{|S| |A|}{T}}{\gamma^2 (1 - \gamma)^4 T} \left( 1 + 2 \max_{\frac{1}{2} \leq i < t} \| \Delta_i \|_\infty \right) \cdot 1.
\]

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

\[
\Delta_t \geq -30 \sqrt{\frac{\log^4 T \log \frac{|S| |A|}{T}}{\gamma^2 (1 - \gamma)^4 T}} \left( 1 + \max_{\frac{1}{2} \leq i < t} \| \Delta_i \|_\infty \right) \cdot 1
\]  

(89)

holds simultaneously for all \( t \geq \frac{T}{c_2 \log T} \).

### B.4 Solving the recurrence relation regarding \( \Delta_t \)

Recall from (40) that with probability exceeding \( 1 - 2\delta \), the following recurrence relation

\[
\| \Delta_t \|_\infty \leq 30 \sqrt{\frac{\log^4 T \log \frac{|S| |A|}{T}}{\gamma^2 (1 - \gamma)^4 T}} \left( 1 + \max_{\frac{1}{2} \leq i < t} \| \Delta_i \|_\infty \right) \quad \text{for all} \quad t \geq \frac{T}{c_2 \log T}
\]

(90)

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

\[
u_k := \max \left\{ \| \Delta_t \|_\infty \mid \gamma^k \frac{T}{c_2 \log T} \leq t \leq T \right\}.
\]

(91)

To bound this sequence, we first obtain a crude bound as a result of (62):

\[
u_0 \leq \frac{1}{1 - \gamma}
\]

(92)
Next, it is directly seen from (90) and the definition of $u_k$ that

$$u_k \leq c_6 \sqrt{\frac{(\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} (1 + u_{k-1}), \quad k \geq 1$$

(93)

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

- If $u_k \leq 1$ for some $k \geq 1$, then (93) tells us that

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

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

$$u_j \leq c_6 \sqrt{\frac{(\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} (1 + u_{j-1}) \leq c_6 \sqrt{\frac{2(\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} u_j \quad \text{for all } j > k.$$

- Instead, suppose that $u_j > 1$ for all $0 \leq j \leq k$. Then it is seen from (93) that

$$u_{j+1} \leq c_6 \sqrt{\frac{(\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} (1 + u_j) \leq c_6 \sqrt{\frac{2(\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} u_j \quad \text{for all } j \leq k.$$

This is equivalent to saying that

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

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

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

and hence

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

This is equivalent to saying that

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

Putting the above two cases together and using (92), we conclude that

$$u_k \leq \sqrt{\frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} + \left( \frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T} \right)^{1-1/2^k} u_0^{1/2^k}$$

$$\leq \sqrt{\frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}} + \left( \frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T} \right)^{1-1/2^k} \left( \frac{1}{1 - \gamma} \right)^{1/2^k}, \quad k \geq 1.$$

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

$$\left( \frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{(1 - \gamma)^4 T} \right)^{1-1/2^k} \leq \max \left\{ \sqrt{\frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T}}, \frac{2c_7^2 (\log^4 T) \left( \log \frac{|S| |A|}{\delta} \right)}{\gamma^2 (1 - \gamma)^4 T} \right\}.$$
We first claim that
\[ c \leq c_8 \left( \sqrt{\frac{(\log^4 T)(\log |S||A| \gamma^T)}{\gamma^2(1-\gamma)^4 T}} + \frac{(\log^4 T)(\log |S||A| \gamma^T)}{\gamma^2(1-\gamma)^4 T} \right), \quad k \geq c_7 \log \log \frac{1}{1-\gamma} \]
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 \( 2^k - \frac{T}{c_7 \log \log T} < T \)), we immediately see from the definition (91) of \( u_k \) that
\[ \|\Delta_T\|_{\infty} \leq c_8 \left( \sqrt{\frac{(\log^4 T)(\log |S||A| \gamma^T)}{\gamma^2(1-\gamma)^4 T}} + \frac{(\log^4 T)(\log |S||A| \gamma^T)}{\gamma^2(1-\gamma)^4 T} \right). \quad (94) \]

with probability at least \( 1 - 2\delta \). To finish up, we note that the sample size assumption (59) is equivalent to
\[ \frac{(\log^4 T)(\log |S||A| \gamma^T)}{\gamma^2(1-\gamma)^4 T} \leq \frac{\min\{\varepsilon^2, \varepsilon\}}{c_3} . \]

When \( c_3 > 0 \) is sufficiently large, substituting this relation into (94) gives
\[ \|\Delta_T\|_{\infty} \leq \frac{1}{2} \sqrt{\min\{\varepsilon^2, \varepsilon\}} + \frac{1}{2} \min\{\varepsilon^2, \varepsilon\} = \begin{cases} \frac{1}{2} \varepsilon + \frac{1}{2} \varepsilon^2 & \text{if } \varepsilon \leq 1 \\ \frac{1}{2} \sqrt{\varepsilon} + \frac{1}{2} \varepsilon & \text{if } \varepsilon > 1 \end{cases} \]
\[ \leq \varepsilon \]

as claimed in Theorem 7.

### B.5 Proof of Lemma 5

We first claim that
\[ \max_{\frac{t}{2} \leq i < t} \text{Var}_P(V_t) - \text{Var}_P(V^*) \leq \frac{4}{1 - \gamma} \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_{\infty} 1. \quad (95) \]

If this claim were valid (which we shall justify towards the end of this subsection), then it would lead to
\[ \sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} \max_{\frac{t}{2} \leq i < t} \text{Var}_P(V_t) \leq \sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} \text{Var}_P(V^*) + \frac{4}{(1-\gamma)^2} \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_{\infty} 1. \quad (96) \]

It then boils down to bounding the first term on the right-hand side of (96). Let us first upper bound the variance term involving \( V^* \). For any \( 0 \leq h < H \), one can express (see (26))
\[ \text{Var}_P(V^*) = P(V^* \circ V^*) - (PV^*) \circ (PV^*) \]
\[ \overset{(i)}{=} \mathcal{P}^{|h+1|} (Q^* \circ Q^*) + P(V^* \circ V^*) - \mathcal{P}^{|h+1|} (Q^* \circ Q^*) - \frac{1}{\gamma^2} (Q^* - r) \circ (Q^* - r) \]
\[ = \mathcal{P}^{|h+1|} (Q^* \circ Q^*) + P^x(Q^* \circ Q^*) - \mathcal{P}^{|h+1|} (Q^* \circ Q^*) - \frac{1}{\gamma^2} (Q^* - r) \circ (Q^* - r) \]
\[ \leq \mathcal{P}^{|h+1|} (Q^* \circ Q^*) + \left\| P^x(Q^* \circ Q^*) - \mathcal{P}^{|h+1|} (Q^* \circ Q^*) \right\|_{\infty} 1 - \frac{1}{\gamma^2} (Q^* - r) \circ (Q^* - r) \]
\[ \overset{(ii)}{=} \mathcal{P}^{|h+1|} (Q^* \circ Q^*) + \frac{4}{1 - \gamma} \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_{\infty} 1 - \frac{1}{\gamma^2} (Q^* - r) \circ (Q^* - r) \]
\[ = \frac{1}{\gamma^2} \left( \mathcal{P}^{|h+1|} (Q^* \circ Q^*) - Q^* \circ Q^* \right) + \frac{4}{1 - \gamma} \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_{\infty} 1 - \frac{1}{\gamma^2} r \circ r + \frac{2}{\gamma^2} Q^* \circ r \]
\[ \leq \frac{1}{\gamma} \left( \mathcal{P}^{|h+1|} (Q^* \circ Q^*) - Q^* \circ Q^* \right) + \frac{2}{\gamma^2} Q^* \circ r + \frac{4}{1 - \gamma} \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_{\infty} 1, \quad (97) \]
where (i) relies on the identity $Q^* = r + \gamma PV^*$, and (iii) holds since $0 < \gamma < 1$. To justify (ii), we make the following observation:

$\left\| P^{\pi_{i+h+1}} (Q^* \circ Q^*) - P^{\pi_i} (Q^* \circ Q^*) \right\|_{\infty} = \left\| P \Pi^{\pi_{i+h+1}} (Q^* \circ Q^*) - P \Pi^{\pi_i} (Q^* \circ Q^*) \right\|_{\infty}$

\[(iv) \leq \left\| \Pi^{\pi_{i+h+1}} (Q^* \circ Q^*) - \Pi^{\pi_i} (Q^* \circ Q^*) \right\|_{\infty}
= \left\| (\Pi^{\pi_{i+h+1}} Q^* - \Pi^{\pi_i} Q^*) \circ (\Pi^{\pi_{i+h+1}} Q^* + \Pi^{\pi_i} Q^*) \right\|_{\infty}
\]

\[(v) \leq \frac{2}{1 - \gamma} \left\| \Pi^{\pi_{i+h+1}} Q^* - \Pi^{\pi_i} Q^* \right\|_{\infty}
\]

\[(vi) \leq \frac{2}{1 - \gamma} \left( \left\| \Pi^{\pi_{i+h+1}} Q^* - \Pi^{\pi_{i+h+1}} Q_{i+h+1} \right\|_{\infty} + \left\| \Pi^{\pi_{i+h+1}} Q_{i+h+1} - Q^* \right\|_{\infty} \right)
\]

\[(vii) \leq \frac{4}{1 - \gamma} \max_{i \leq t} \left\| \Delta_i \right\|_{\infty},
\]

where (iv) arises from the fact $\left\| P z \right\|_{\infty} \leq \left\| P 1 \right\|_{\infty} = \left\| z \right\|_{\infty}$, (v) is valid because $\left\| Q^* \right\|_{\infty} \leq 1/(1 - \gamma)$, (vi) follows from the fact that $V_{i+h+1} = \Pi^{\pi_{i+h+1}} Q_{i+h+1}$, and (vi) holds since $\left\| V_{i+h+1} - Q^* \right\|_{\infty} \leq \left\| Q_{i+h+1} - Q^* \right\|_{\infty}$.

As it turns out, the first term in (97) allows one to build a telescoping sum. Specifically, invoking (97) allows one to bound

$$
\sum_{h=0}^{H-1} P^{\pi_{i+h}} \text{Var}_P (V^*) \leq \frac{1}{\gamma} \sum_{h=0}^{H-1} \gamma P^{\pi_{i+h}} (Q^* \circ Q^*) - Q^* \circ Q^* \right) + \frac{2}{1 - \gamma} \sum_{h=0}^{H-1} \gamma P^{\pi_{i+h}} (Q^* \circ r)
\]

\[(i) \leq \frac{1}{\gamma} \left( \sum_{h=0}^{H-1} \gamma P^{\pi_{i+h}} - \sum_{h=0}^{H-1} \gamma P^{\pi_{i+h}} \right) (Q^* \circ Q^*)
\]

$$
\leq \frac{1}{\gamma} \left( \sum_{h=0}^{H-1} \gamma P^{\pi_{i+h}} - I \right) (Q^* \circ Q^*) + \frac{4}{(1 - \gamma)^2} \max_{i \leq t} \left\| \Delta_i \right\|_{1} \right)
\]

\[(ii) \leq \left( \frac{2}{\gamma} \left\| Q^* \right\|_{\infty} + \frac{4}{(1 - \gamma)^2} \right) \max_{i \leq t} \left\| \Delta_i \right\|_{\infty} \right)
\]

\[(iii) \leq \frac{1}{(1 - \gamma)^2} \left( \frac{2}{\gamma} + \frac{4}{(1 - \gamma)^2} \right) \max_{i \leq t} \left\| \Delta_i \right\|_{\infty}
\]

Here, (i) comes from the identity $\prod_{k=1}^{h} P^{\pi_{i+k}} 1 = 1$; (ii) holds because each row of $\prod_{k=1}^{h} P^{\pi_{i+k}}$ has unit $\| \cdot \|_1$ norm for any $h$; (iii) arises from the bound $\left\| Q^* \right\|_{\infty} \leq 1/(1 - \gamma)$. This completes the proof, as long as the claim (95) can be justified.

**Proof of the inequality (95).** To validate this result, we make the observation that

$$
\text{Var}_P (V_i) - \text{Var}_P (V^*) = [P(V_i \circ V_i) - (PV_i) \circ (PV_i)] - [P(V^* \circ V^*) - (PV^*) \circ (PV^*)]
$$
Theorem 8. Consider any

As it turns out, if Theorem 1 as a special case.

for any set of policies

direct consequence of the claimed iteration complexity (101) when

for some sufficiently large universal constant

general version of Theorem 1 that covers the full

leads to the proof of Theorem 1 for TD learning. Akin to the Q-learning case, we proceed to establish a more

analysis in Section B to yield an improved

different states can be chosen in a separate manner. The proof follows from an identical argument as the

Suppose that

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 \( t \geq \frac{T}{c_2 \log T} \). Then one has

\[
\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^h P_{\tilde{\pi}_k} \max_{\frac{1}{2} \leq \gamma < t} \text{Var}_P(V_i) \leq \frac{4}{\gamma^2(1 - \gamma)^2} \left( 1 + 2 \max_{\frac{1}{2} \leq \gamma < t} \| \Delta_i \|_\infty \right) 1
\]

(99)

for any set of policies \( \{\tilde{\pi}_k\} \) obeying \( \{\tilde{\pi}_k\} \subseteq \Pi \). Here, we define

\[
\Pi := \{ \pi = [\pi(s)]_{s \in S} \mid \pi(s) \in \Pi_s, \forall s \in S \}, \quad \Pi_s := \{ \pi_i(s) \mid i \in [t/2, t) \}.
\]

The key difference between Lemma 6 and Lemma 5 is that: the components of \( \tilde{\pi}_k \) 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 \( \frac{1}{1 - \gamma} \) 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 \( \epsilon \)-range. This is formally stated below, which subsumes Theorem 1 as a special case.

Theorem 8. Consider any \( \gamma \in (0, 1) \) and any \( \epsilon \in (0, \frac{1}{1 - \gamma}] \). Theorem 1 continues to hold if

\[
T \geq \frac{c_3 (\log T) \left( \log \frac{|S| T}{\delta} \right)}{\gamma^2(1 - \gamma)^3 \min\{\epsilon, \epsilon^2\}}
\]

(101)

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 (101) when \( \epsilon \leq \frac{1}{1 - \gamma} \):

\[
T \geq \frac{c_3 (\log T) \left( \log \frac{|S| T}{\delta} \right)}{\gamma^2(1 - \gamma)^3 \min\{\epsilon, \epsilon^2\}} \geq \frac{c_3 (\log T) \left( \log \frac{|S| T}{\delta} \right)}{\gamma^2(1 - \gamma)^2},
\]

(102)
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

\[ V_t = (1 - \eta_t) V_{t-1} + \eta_t (r + \gamma P_t V_{t-1}) \quad \text{for all } t \geq 1, \]

where the matrix \( P_t \in \{0, 1\}^{|S| \times |S|} \) obeys

\[ P_t(s, s') := \begin{cases} 1, & \text{if } s' = s_t(s) \\ 0, & \text{else} \end{cases} \]

for any \( s, s' \in S \). In the sequel, we collect a few other facts concerning the range of \( V_t \) and learning rates.

**Range of \( V_t \).** We claim that: when the initialization \( V_0 \) obeys \( 0 \leq V_0 \leq \frac{1}{1-\gamma} - 1 \), the TD learning iterates obey

\[ 0 \leq V_t \leq \frac{1}{1-\gamma} - 1 \quad \text{and} \quad \| V_t - V^* \|_\infty \leq \frac{1}{1-\gamma} \quad \text{for all } t \geq 0, \]

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 \( |A| = 1 \), 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:

\[ \eta_k^{(t)} := \begin{cases} \prod_{i=1}^{t} (1 - \eta_i (1 - \gamma)), & \text{if } k = 0, \\ \eta_k \prod_{i=k+1}^{t} (1 - \eta_i (1 - \gamma)), & \text{if } 0 < k < t, \\ \eta_t, & \text{if } k = t. \end{cases} \]

Note that this definition (105) differs from the one (64) used for Q-learning, and will only be employed in this section. Consider any iteration number \( t \) satisfying

\[ t \geq \frac{T}{c_2 \log T}. \]

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

\[ (1 - \gamma) \eta_t \geq \frac{1 - \gamma}{1 + \frac{c_1 (1 - \gamma) T}{\log^2 T}} \geq \frac{1 - \gamma}{\frac{2 c_1 (1 - \gamma) T}{\log^2 T}} \frac{\log^2 T}{2 c_1 T} = \frac{\log^2 T}{2 c_1 T}. \]

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 (107) that

\[ \eta_i^{(t)} \leq (1 - \eta_{i/2} (1 - \gamma))^{t/2} \leq \left(1 - \frac{\log^2 T}{2 c_1 T}\right)^{t/2} \leq \left(1 - \frac{\log^2 T}{2 c_1 T}\right)^{\frac{\log^2 T}{2 c_1 T}} \leq \frac{1}{T^2}, \]

where the last inequality holds as long as \( c_1 c_2 \leq 1/8 \) and (102) holds.

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

\[ \eta_i^{(t)} \leq \eta_t \leq \frac{1}{c_2 (1 - \gamma) i / \log^2 T} \leq \frac{2 \log^2 T}{(1 - \gamma) T}, \]

which relies on Assumption (8a).
In addition, given that $P^k 1 = 1$ for any integer $k > 0$, it can be easily verified that

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

and as a result,

$$\left\| \prod_{i=k+1}^t (I - \eta_i (I - \gamma P)) \right\|_1 = \prod_{i=k+1}^t (1 - \eta_i(1 - \gamma)). \tag{109}$$

C.2 Proof of Theorem 8

Step 1: decomposing the error $V_t - V^*$. Taking $\Delta_t := V_t - V^*$, via the basic relation (32), the TD learning update rule can be written as

$$\Delta_t = (1 - \eta_t)\Delta_{t-1} + \eta_t \gamma (P \Delta_{t-1} + (P_t - P) V_{t-1})$$

$$= (I - \eta_t(I - \gamma P))\Delta_{t-1} + \eta_t \gamma(P_t - P)V_{t-1}. \tag{110}$$

Invoking the above relation recursively then leads to

$$\Delta_t = \prod_{i=1}^t (I - \eta_i (I - \gamma P)) \Delta_0 + \sum_{k=1}^t \eta_k \prod_{i=k+1}^t (I - \eta_i (I - \gamma P)) \gamma(P_k - P)V_{k-1} \tag{111}.$$

Step 2: controlling the first term of (111). With regards to the first term of (111), we make the observation that

$$\left\| \prod_{i=1}^t (I - \eta_i (I - \gamma P)) \Delta_0 \right\|_\infty \leq \left\| \prod_{i=1}^t (I - \eta_i (I - \gamma P)) \right\|_1 \left\| \Delta_0 \right\|_\infty$$

$$= \left\{ \prod_{i=1}^t (1 - \eta_i(1 - \gamma)) \right\} \left\| \Delta_0 \right\|_\infty$$

$$\leq \eta_0^{(t)} \frac{1}{1 - \gamma} \leq \frac{1}{(1 - \gamma)T^2}, \tag{112}$$

where the second line arises from (109), and the last inequality holds true due to (108a) as long as $t \geq \frac{T}{\epsilon_T \log T}$.

Step 3: controlling the second term of (111). We then move on to the second term $\xi_t$ in (111), which admits the following expression

$$\xi_t = \sum_{k=1}^t z_k \quad \text{with} \quad z_k := \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P)) \gamma(P_k - P)V_{k-1}. \tag{113}$$

Here, the summands $\{z_k\}$ clearly satisfy

$$E[z_k | V_{k-1}, \ldots, V_0] = 0.$$

We then attempt to invoke the Freedman inequality (see Theorem 6) to control this term. Towards this end, there are several quantities that need to be calculated.

- First of all, we observe that

$$B := \max_{1 \leq k \leq t} \left\| z_k \right\|_\infty \leq \max_{1 \leq k \leq t} \left\| \eta_k \prod_{i=k+1}^t (I - \eta_i(I - \gamma P)) \gamma(P_k - P)V_{k-1} \right\|_\infty.$$
\[
\leq \max_{1 \leq k \leq t} \left\| \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \right\|_1 \|P_k - P\|_V V_{k-1}\|_\infty
\]
\[
= \max_{1 \leq k \leq t} \left\{ \eta_k \prod_{i=k+1}^{t} (1 - \eta_i (1 - \gamma)) \right\} \|P_k - P\|_V V_{k-1}\|_\infty
\]
\[
\leq \max_{1 \leq k \leq t} \eta_k^{(t)} \left( \|P_k\|_1 + \|P\|_1 \right) \|V_{k-1}\|_\infty \leq \frac{4 \log^3 T}{(1 - \gamma)^2 T},
\]
where the third line again makes use of the relation (109) and the last line follows the facts \(\|P_k\|_1 = \|P\|_1 = 1, \|V_{k-1}\|_\infty \leq 1/(1 - \gamma)\), as well as the properties (108).

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

\[
\text{Var}\left( u^T (P_k - P) v \right) = \sum_{i=1}^{|S|} u_i^2 \text{Var}\left( (P_k - P)_{i} \right) \leq \left\{ \max_i |u_i| \right\} [u_1, \ldots, u_{|S|}] \text{Var}_P (v)
\]
\[
\leq \|u\|_1 u^T \text{Var}_P (v),
\]
where we remind the reader of the notation \(\text{Var}_P (v)\) in (26). Additionally, for any vector \(a = [a_j]\), let us employ the notation \(\text{Var}(a | V_{k-1}, \ldots, V_0)\) to represent a vector whose \(j\)-th entry is given by \(\text{Var}(a_j | V_{k-1}, \ldots, V_0)\). Armed with this notation, we obtain

\[
\text{Var}(z_k | V_{k-1}, \ldots, V_0) \leq \gamma^2 \|\eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \|_1 \left\{ \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \right\} \text{Var}_P (V_{k-1})
\]
\[
= \gamma^2 \left\{ \eta_k \prod_{i=k+1}^{t} (1 - \eta_i (1 - \gamma)) \right\} \left\{ \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \right\} \text{Var}_P (V_{k-1})
\]
\[
\leq \eta_k \eta_k^{(t)} \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \text{Var}_P (V_{k-1}),
\]
where the first inequality is a consequence of (115) and the definition of \(z_k\) (cf. (113)), the second line arises from (109), and the last relation results from the definition of \(\eta_k^{(t)}\). This in turn allows us to compute

\[
W_k := \sum_{k=1}^{t} \text{Var}(z_k | V_{k-1}, \ldots, V_0) \leq \sum_{k=1}^{t} \eta_k \eta_k^{(t)} \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \text{Var}_P (V_{k-1})
\]
\[
\leq \frac{t}{2} \left( \eta_k^{(t)} \right)^2 \left( \frac{1}{(1 - \gamma)^2} + \frac{2 \log^3 T}{(1 - \gamma)^2 T} \right) \left( \sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \text{Var}_P (V_{k-1}) \right)
\]
\[
\leq \frac{1}{2(1 - \gamma)^2 T^3} \left( \sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \right) \text{Var}_P (V_k)
\]
\[
\leq \frac{1}{2(1 - \gamma)^2 T^3} \left( \sum_{k=t/2+1}^{t} \eta_k \prod_{i=k+1}^{t} (I - \eta_i (I - \gamma P)) \right) \max_{k: t/2 \leq k < t} \text{Var}_P (V_k),
\]
where the penultimate inequality results from (108); to see why the last inequality holds, observe that

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

\[
= (I - \gamma P)^{-1} \sum_{k = t/2 + 1}^{t} \left[ \prod_{i = k + 1}^{t} (I - \eta_i (I - \gamma P)) - \prod_{i = k}^{t} (I - \eta_i (I - \gamma P)) \right]
\]

\[
= (I - \gamma P)^{-1} - (I - \gamma P)^{-1} \prod_{i = t/2 + 1}^{t} (I - \eta_i (I - \gamma P)) \leq (I - \gamma P)^{-1},
\]

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

- In addition, we also derive the following trivial upper bound based on (117):

\[
|W_t| \leq \frac{1}{2 (1 - \gamma)^2 T^3} \mathbf{1} + \frac{2 \log^3 T}{(1 - \gamma) T} \| (I - \gamma P)^{-1} \|_1 \max_{k: t/2 \leq k < t} \| \text{Var}_P(V_k) \|_\infty \mathbf{1}
\]

\[
\leq \frac{1}{2 (1 - \gamma)^2 T^3} \mathbf{1} + \frac{2 \log^3 T}{(1 - \gamma)^4 T} \leq \frac{3 \log^3 T}{(1 - \gamma)^4 T} \mathbf{1} =: \sigma^2 \mathbf{1},
\]

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

\[
\frac{\sigma^2}{2K} \leq \frac{3 \log^3 T}{(1 - \gamma)^2 T}.
\]

Equipped with the preceding bounds, let us apply the Freedman inequality in Theorem 6 and invoke the union bound over all entries of \(\xi_t\) to show that

\[
|\xi_t| \leq \sqrt{8 \left( W_t + \frac{\sigma^2}{2K} \right) \log \left( 8 |S| T \log \frac{1}{1 - \gamma} \right)} + \left( \frac{1}{3} B \log \left( 8 |S| T \log \frac{1}{1 - \gamma} \right) \right) \mathbf{1}
\]

\[
\leq \sqrt{16 \left( W_t + \frac{3 \log^3 T}{(1 - \gamma)^2 T} \right) \log \left( |S| T \right) \delta} + \left( \frac{1}{3} B \log \left( |S| T \right) \delta \right) \mathbf{1}
\]

\[
\leq \sqrt{32 \left( \log^3 T \right) \left( \log \frac{|S| T \delta}{\delta} \right) \left( (I - \gamma P)^{-1} \max_{k: t/2 \leq k < t} \text{Var}_P(V_k) + \frac{2}{1 - \gamma} \right) + \frac{12 \left( \log^3 T \right) \left( \log \frac{|S| T \delta}{\delta} \right) \mathbf{1}}{(1 - \gamma)^2 T}}
\]

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

\[
\log \frac{8 |S| T \log \frac{1}{1 - \gamma}}{\delta} \leq 2 \log \frac{|S| T}{\delta}
\]

as long as \(\frac{|S| T}{\delta} \geq 8 \log \frac{1}{1 - \gamma}\), whereas the last line holds by using (114), (117) and (119). Further, we make the observation that

\[
(I - \gamma P)^{-1} \text{Var}_P(V^*) = (I - \gamma P)^{-1} \left( P(V^* \circ V^*) - (P V^*) \circ (P V^*) \right)
\]

\[
= (I - \gamma P)^{-1} \left( P(V^* \circ V^*) - \frac{1}{\gamma^2} (V^* - r) \circ (V^* - r) \right)
\]

\[
\leq (I - \gamma P)^{-1} \left( P(V^* \circ V^*) - \frac{1}{\gamma^2} V^* \circ V^* + \frac{2}{\gamma^2} r \circ V^* \right)
\]

\[
\leq (I - \gamma P)^{-1} \left( P(V^* \circ V^*) - \frac{1}{\gamma} V^* \circ V^* + \frac{2}{\gamma^2} r \circ V^* \right)
\]

\[
= \frac{1}{\gamma} (I - \gamma P)^{-1} (\gamma P - I) (V^* \circ V^*) + \frac{2}{\gamma^2} (I - \gamma P)^{-1} (r \circ V^*)
\]

\[
\leq \frac{2}{\gamma^2} (I - \gamma P)^{-1} (r \circ V^*) \leq \frac{2}{\gamma^2 (1 - \gamma)^2} \mathbf{1},
\]
where the second line makes use of the basic relation $V^* = r + \gamma P V^*$. As a consequence, we conclude

$$
(I - \gamma P)^{-1} \max_{k:t/2 \leq k < t} \text{Var}_P (V_k) \leq (I - \gamma P)^{-1} \left( \text{Var}_P (V^*) + \frac{4}{1 - \gamma} \max_{k:t/2 \leq k < t} \|\Delta_k\|_\infty \mathbf{1} \right)
$$

$$
\leq \frac{2}{\gamma(1 - \gamma)^2} \left( 1 + 2 \max_{k:t/2 \leq k < t} \|\Delta_k\|_\infty \right) \mathbf{1}.
$$

(120)

Here, the first inequality arises from (95), while the second inequality holds due to the facts that $\|(I - \gamma P)^{-1}\|_1 = 1/(1 - \gamma)$.

**Step 4: putting everything together.** Consequently, substituting the bounds in Steps 2-3 into (111) yields

$$
\|\Delta_t\|_\infty \leq 30 \sqrt{\frac{(\log^3 T) (\log \frac{|S|T}{\delta})}{\gamma^2(1 - \gamma)^3T}} \left( 1 + \max_{\frac{t}{2} \leq i < t} \|\Delta_i\|_\infty \right) \quad \text{for all } t \geq \frac{T}{c_2 \log T}.
$$

(121)

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

$$
\|\Delta_T\|_\infty \leq c_0 \left( \frac{(\log^3 T) (\log \frac{|S|T}{\delta})}{\gamma^2(1 - \gamma)^3T} + \frac{(\log^3 T) (\log \frac{|S|T}{\delta})}{\gamma^2(1 - \gamma)^3T} \right)
$$

(122)

holds with probability at least $1 - \delta$, where $c_0 > 0$ is some universal constant. As a result, one has

$$
\|\Delta_T\|_\infty \leq \frac{1}{2} \left( \sqrt{\min\{\varepsilon, \varepsilon^2\}} + \min\{\varepsilon, \varepsilon^2\} \right) = \frac{1}{2} (\varepsilon + \varepsilon^2) \mathbf{1}\{\varepsilon \leq 1\} + \frac{1}{2} (\varepsilon + \varepsilon^2) \mathbf{1}\{\varepsilon > 1\} \leq \varepsilon,
$$

as long as the sample size satisfies the following

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

for some constant $c_3 \geq \max\{1, 2c_0\}$. This requirement is equivalent to condition (101) as claimed.

### C.3 Proof for Remarks 2 and 3

**Proof for Remark 2.** Let us divide the dynamics of the algorithm into two parts.

- For any $1 \leq t \leq T/2$, it follows from Lemma 4 that

$$
0 \leq V_{t/2} \leq \frac{1}{1 - \gamma} - 1.
$$

(123)

- Next, let us consider any $T/2 < t \leq T$, and set $\tilde{t} = t - T/2$. It comes from the choice (13) that

$$
\frac{1}{1 + \frac{\tilde{c}_1(1 - \gamma)(t + T/2)}{\log^2(t + T/2) + 1}} \leq \eta_{t+T/2} \leq \frac{1}{1 + \frac{\tilde{c}_2(1 - \gamma)(t + T/2)}{\log^2(t + T/2) + 1}}
$$

$$
\implies \frac{1}{1 + \frac{\tilde{c}_1(1 - \gamma)(t + T/2)}{\log^2(t + T/2)}} \leq \eta_{t+T/2} \leq \frac{1}{1 + \frac{\tilde{c}_2(1 - \gamma)(t + T/2)}{\log^2(t + T/2)}},
$$

provided that $\tilde{c}_1$ and $\tilde{c}_2$ are suitably chosen. By treating $V_{T/2}$ as the initial point and invoking Theorem 1, we immediately establish (9) under the choice (13) and the sample size condition (14).
Proof for Remark 3. Let us define, for each $1 \leq t \leq T$,

$$
\varepsilon_t := \min \left\{ \sqrt{\frac{2c_3(\log^3 T) (\log \frac{|S|}{t})}{(1-\gamma)^3 t}}, \frac{1}{1-\gamma} \right\}.
$$

(124)

Invoking the claim in Remark 2 in conjunction with Lemma 4 reveals that with probability at least $1 - \delta/T$,

$$
\max_{s \in S} |V_t(s) - V^*(s)| \leq \varepsilon_t
$$

for any given $t \leq T$. Taking the union bound over all $1 \leq t \leq T$ implies that with probability exceeding $1 - \delta$,

$$
\max_{s \in S} \left| \frac{1}{T} \sum_{t=1}^{T} V_t(s) - V^*(s) \right| \leq \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t
$$

$$
\leq \sqrt{\frac{2c_3(\log^3 T) (\log \frac{|S|}{T^2})}{(1-\gamma)^3}} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 4 \sqrt{\frac{c_3(\log^3 T)}{(1-\gamma)^3 T}},
$$

where the last inequality results from the elementary inequality $\sum_{t=1}^{T} 1/\sqrt{t} \leq 2\sqrt{T}$. This finishes the proof.

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

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

$$
\log T \leq \frac{1}{1-\gamma}
$$

(125)

throughout the proof; otherwise the lower bound in Theorem 3 is worse than the minimax lower bound in Azar et al. (2013).

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

$$
P_t(s' \mid s,a) := \begin{cases} 
1, & \text{if } s_t(s,a) = s', \\
0, & \text{otherwise},
\end{cases}
$$

(126)

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

$$
P_t(s' \mid 2) := P_t(s' \mid 2,1)
$$

for notational simplicity.

D.1 Key quantities related to learning rates

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

$$
\eta_k^{(t)} := \eta_k \prod_{i=k+1}^{t} (1 - \eta_i (1-\gamma p)) \quad \text{for any } 1 \leq k < t,
$$

(127a)

$$
\eta_0^{(t)} := \prod_{i=1}^{t} (1 - \eta_i (1-\gamma p)),
$$

(127b)

$$
\eta_t^{(t)} := \eta_t.
$$

(127c)
It is helpful to establish several basic properties about these quantities. As can be easily verified,

$$\eta_0^{(t)} + (1 - \gamma p) \sum_{k=1}^{t} \eta_k^{(t)} = \prod_{i=1}^{t} (1 - \hat{\eta}_i) + \hat{\eta}_1 \prod_{i=2}^{t} (1 - \hat{\eta}_i) + \hat{\eta}_2 \prod_{i=3}^{t} (1 - \hat{\eta}_i) + \cdots + \hat{\eta}_{t-1}(1 - \hat{\eta}_t) + \hat{\eta}_t = 1, \quad (128)$$

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

$$\prod_{i=\tau+1}^{t} (1 - \eta_i(1 - \gamma p)) + (1 - \gamma p) \sum_{k=\tau+1}^{t} \eta_k^{(t)} = 1. \quad (129)$$

### 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

$$V_t(0) = Q_t(0, 1) = (1 - \eta_t(1 - \gamma))Q_{t-1}(0, 1) = \prod_{i=1}^{t} (1 - \eta_i(1 - \gamma))Q_0(0, 1) = 0 \quad (130)$$

for all $t \geq 1$. Regarding state 2, the update rule (4) taken together with (130) leads to

$$V_t(2) = Q_t(2, 1) = (1 - \eta_t)Q_{t-1}(2, 1) + \eta_t \{r(2, 1) + \gamma P_t(2 | 2)V_{t-1}(2) + \gamma P_t(0 | 2)V_{t-1}(0)\}$$

$$= (1 - \eta_t)V_{t-1}(2) + \eta_t \{1 + \gamma P_t(2 | 2)V_{t-1}(2)\}, \quad (131)$$

and for state 3,

$$V_t(3) = Q_t(3, 1) = (1 - \eta_t)Q_{t-1}(3, 1) + \eta_t \{r(3, 1) + \gamma V_{t-1}(3)\}$$

$$= (1 - \eta_t(1 - \gamma))V_{t-1}(3) + \eta_t. \quad (132)$$

Similarly, one also has

$$Q_t(1, 1) = (1 - \eta_t)Q_{t-1}(1, 1) + \eta_t \{1 + \gamma P_t(1 | 1, 1)V_{t-1}(1)\}, \quad (133a)$$

$$Q_t(1, 2) = (1 - \eta_t)Q_{t-1}(1, 2) + \eta_t \{1 + \gamma P_t(1 | 1, 2)V_{t-1}(1)\}. \quad (133b)$$

In what follows, we shall first determine a crude range for certain quantities relates to the learning rates $\eta_t$, 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 (131) to obtain

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

$$= \prod_{i=\tau+1}^{t} (1 - \eta_i(1 - \gamma p))V_{\tau+1}(2) + \sum_{k=\tau+1}^{t} \eta_k \prod_{i=k+1}^{t} (1 - \eta_i(1 - \gamma p))\{1 + \gamma (P_k(2 | 2) - p)V_{k-1}(2)\}$$

$$= \prod_{i=\tau+1}^{t} (1 - \eta_i(1 - \gamma p))V_{\tau+1}(2) + \sum_{k=\tau+1}^{t} \eta_k^{(t)} + \sum_{k=\tau+1}^{t} \sum_{i=k+1}^{t} \eta_k^{(t)}(1 - \eta_i(1 - \gamma p))\gamma (P_k(2 | 2) - p)V_{k-1}(2)$$

$$= \prod_{i=\tau+1}^{t} (1 - \eta_i(1 - \gamma p))V_{\tau+1}(2) + \frac{1}{1 - \gamma p} \sum_{k=\tau+1}^{t} \eta_k^{(t)}(1 - \eta_i(1 - \gamma p))\gamma (P_k(2 | 2) - p)V_{k-1}(2)$$

$$= \frac{1}{1 - \gamma p} \prod_{i=\tau+1}^{t} (1 - \eta_i(1 - \gamma p)) \left[ \frac{1}{1 - \gamma p} - V_{\tau+1}(2) \right] + \sum_{k=\tau+1}^{t} \eta_k^{(t)}(1 - \eta_i(1 - \gamma p))\gamma (P_k(2 | 2) - p)V_{k-1}(2), \quad (134)$$

37
where the penultimate line arises from (129). In particular, in the special case where \( \tau = 0 \) (so that \( V_\tau(2) = V_0(2) = 0 \)), this simplifies to

\[
V_1(2) = \frac{1 - \eta_0^{(t)}}{1 - \gamma p} + \sum_{k=1}^t \eta_k^{(t)} \gamma (P_k(2 \mid 2) - p)V_{k-1}(2),
\]

which relies on the definition of \( \eta_0^{(t)} \) in (127). With similar derivation, (132) leads to

\[
V_1(3) = \frac{1}{1 - \gamma} \left[ 1 - \prod_{i=1}^T (1 - \eta_i(1 - \gamma)) \right] = V^*(3) - \frac{1}{1 - \gamma} \prod_{i=1}^T (1 - \eta_i(1 - \gamma)).
\]

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

We start by computing the mean \( V^*(2) - \mathbb{E}[V_T(2)] \). From the construction (42), it is easily seen that \( \mathbb{E}[P_k(2 \mid 2)] = p \), which together with the identity (135) leads to

\[
\mathbb{E}[V_T(2)] = \frac{1 - \eta_0^{(T)}}{1 - \gamma p} \quad \text{and} \quad V^*(2) - \mathbb{E}[V_T(2)] = \frac{\eta_0^{(T)}}{1 - \gamma p}.
\]

Similarly, applying the above argument to (134) and rearranging terms, we immediately arrive at

\[
V^*(2) - \mathbb{E}[V_T(2)] = \prod_{i=\tau+1}^T (1 - \eta_i(1 - \gamma p)) \left[ \frac{1}{1 - \gamma p} - \mathbb{E}[V_T(2)] \right]
\]

for any integer \( 0 \leq \tau < T \).

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

\[
\text{Var} \left( W_0 + \sum_{k=\tau+1}^T Z_k \right) = \mathbb{E} \left[ \text{Var} \left( W_0 + \sum_{k=\tau+1}^T Z_k \mid \mathcal{F}_{\tau-1} \right) \right] + \mathbb{E} \left[ \text{Var} \left( W_0 \mid \mathcal{F}_{\tau-1} \right) \right] = \cdots
\]

\[
= \sum_{k=\tau+1}^T \mathbb{E} \left[ \text{Var} \left( Z_k \mid \mathcal{F}_{k-1} \right) \right] + \text{Var} \left( W_0 \right) \geq \sum_{k=\tau+1}^T \mathbb{E} \left[ \text{Var} \left( Z_k \mid \mathcal{F}_{k-1} \right) \right].
\]

Consequently, for any \( 0 \leq \tau < T - 1 \), it follows from the decomposition (134) (with \( \tau \) replaced by \( \tau + 1 \)) that

\[
\text{Var}(V_T(2)) \geq \mathbb{E} \left[ \sum_{k=\tau+2}^T \text{Var} \left( \eta_k^{(T)} \gamma (P_k(2 \mid 2) - p)V_{k-1}(2) \mid V_{k-1}(2) \right) \right]
\]

\[
= \sum_{k=\tau+2}^T (\eta_k^{(T)})^2 p(1 - p) \mathbb{E} \left[ (V_{k-1}(2))^2 \right]
\]

\[
\geq \left( 1 - \gamma \right) \left( 4 \gamma - 1 \right) \frac{1}{9} \cdot \frac{1}{4(1 - \gamma)^2} \sum_{k=\tau+2}^T (\eta_k^{(T)})^2
\]

\[
= \frac{4 \gamma - 1}{36 (1 - \gamma)} \sum_{k=\tau+2}^T (\eta_k^{(T)})^2,
\]

where the first identity relies on the fact that \( P_k(2 \mid 2) \) is a Bernoulli random variable with mean \( p \), and the inequality comes from the definition of \( \tau \) (see (146)) and the choice of \( p \) (see (43)). 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 \((c_\eta \geq \log T \text{ or } 0 \leq \eta \leq \frac{1}{(1-\gamma)T})\)

In this case, we focus on lower bounding \(V^*(2) - \mathbb{E}[V_T(2)]\). In view of this identity (137), this boils down to controlling \(\eta_0(T)\).

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

\[
\eta_t(1 - \gamma p) \leq \begin{cases} 
\frac{1-\gamma p}{(1-\gamma)\log T} = \frac{1}{2} & \text{if } \eta_t = \frac{1}{1+c_\eta(1-\gamma)t} \\
\frac{1}{(1-\gamma)T} = \frac{1}{2} & \text{if } \eta_t = \eta 
\end{cases}
\]

(141)

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

\[
\log \eta_0(T) = \sum_{i=1}^{T} \log \left(1 - \eta_i(1 - \gamma p)\right) \geq -1.5 \sum_{i=1}^{T} \eta_i(1 - \gamma p) \geq -2, 
\]

(142)

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 (141))

\[
\sum_{i=1}^{T} \eta_i(1 - \gamma p) \leq \begin{cases} 
\frac{3}{4(1-\gamma)T} \sum_{i=1}^{T} \frac{1}{7} \leq 1, & \text{if } \eta_t = \frac{1}{1+c_\eta(1-\gamma)t} \\
\frac{4}{7} \sum_{i=1}^{T} = \frac{4}{7}, & \text{if } \eta_t = \eta.
\end{cases}
\]

Combining the above result with the properties (137) and (142) then yields

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

(143)

This taken together with (45) gives

\[
\mathbb{E}[(V^*(2) - V_T(2))^2] \geq (V^*(2) - \mathbb{E}[V_T(2)])^2 \geq \frac{9}{16e^4(1 - \gamma)^2}. 
\]

(144)

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

By virtue of (138), the mean gap \(V^*(2) - \mathbb{E}[V_T(2)]\) depends on two factors: (i) the choice of the learning rates, and (ii) the gap between \(\frac{1}{(1-\gamma)p}\) and \(\mathbb{E}[V_T(2)]\), where \(\tau\) is an integer obeying \(0 \leq \tau < T\). To control the factor (ii), we need to choose \(\tau\) properly. Let us start by considering the simple scenario with \(\mathbb{E}[(V_T(2))^2] < \frac{1}{4(1-\gamma)^2}\), for which we have

\[
V^*(2) - \mathbb{E}[V_T(2)] \geq \frac{3}{4(1-\gamma)} - \sqrt{\mathbb{E}[(V_T(2))^2]} \geq \frac{1}{4(1 - \gamma)}. 
\]

(145)

Here, we have used (44) and the elementary fact \(\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}\). Consequently, it remains to look at the scenario obeying \(\mathbb{E}[(V_T(2))^2] \geq \frac{1}{4(1-\gamma)^2}\), towards which we propose to set \(\tau\) as follows

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

(146)

Clearly, \(\tau\) is well-defined in this scenario and obeys (in view of both (146) and the initialization \(V_0 = 0\))

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

(147)

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
\[
\prod_{i=\tau+1}^{T} (1 - \eta_i(1 - \gamma p)) \geq \frac{1}{2}. \tag{148}
\]
Invoke (138) to deduce that
\[
V^*(2) - \mathbb{E}[V_T(2)] = \prod_{i=\tau+1}^{T} (1 - \eta_i(1 - \gamma p)) \left[ \frac{1}{1 - \gamma p} - \mathbb{E}[V_\tau(2)] \right]
\geq \prod_{i=\tau+1}^{T} (1 - \eta_i(1 - \gamma p)) \left[ \frac{3}{4(1 - \gamma)} - \sqrt{\mathbb{E}[\mathbb{E}[V_\tau(2)]^2]} \right]
\geq \prod_{i=\tau+1}^{T} (1 - \eta_i(1 - \gamma p)) \frac{1}{4(1 - \gamma)} \geq \frac{1}{8(1 - \gamma)},
\]
where the second line makes use of the definition (43) and the elementary fact \(\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}\), and the last line relies on the inequalities (147) and (148).

Case 2.2. We now move on to the case where
\[
0 \leq \prod_{i=\tau+1}^{T} (1 - \eta_i(1 - \gamma p)) \leq \frac{1}{2}. \tag{149}
\]
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 (140), it boils down to lower bounding \(\sum_{k=\tau+2}^{T} (\eta_k^{(T)})^2\), which we accomplish as follows.

- Consider constant learning rates \(\eta_k = \eta\), and suppose that \(\eta\) obeys \(\frac{1}{(1 - \gamma)|\tau|} < \eta \leq 1 - \frac{1}{1 - \gamma p}\). It is readily seen that \(\eta_k^{(T)} = \eta(1 - \eta(1 - \gamma p))^{T-k}\) for any \(k \geq 1\). We claim that it suffices to focus on the scenario where
\[
\tau \leq T - 2. \tag{150}
\]
In fact, if \(\tau \geq T - 1\), then the definition (146) of \(\tau\) necessarily requires that
\[
\mathbb{E}[V_{T-1}(2)] \leq \sqrt{\mathbb{E}[(V_{T-1}(2))^2]} < \frac{1}{2(1 - \gamma)}.
\]
In view of (137) (with \(T\) replaced by \(T-1\)), a little algebra shows that this is equivalent to \((1 - \eta(1 - \gamma p))^{T-1} \geq 1/3\), and hence \((1 - \eta(1 - \gamma p))^T \geq 1/9\). In turn, this combined with (137) leads to
\[
V^*(2) - \mathbb{E}[V_T(2)] = \frac{(1 - \eta(1 - \gamma p))^T}{1 - \gamma p} = \frac{3(1 - \eta(1 - \gamma p))^T}{4(1 - \gamma)} \geq \frac{1}{12(1 - \gamma)}, \tag{151}
\]
which already suffices for our purpose.

Next, assuming that (150) holds, one can derive
\[
\sum_{k=\tau+2}^{T} (\eta_k^{(T)})^2 = \sum_{k=\tau+2}^{T} \eta^2 (1 - \eta(1 - \gamma p))^{2(T-k)} = \frac{\eta^2 [1 - (1 - \eta(1 - \gamma p))^{2(T-\tau-1)}]}{1 - (1 - \eta(1 - \gamma p))^2}
\geq \frac{\eta^2/2}{1 - (1 - \eta(1 - \gamma p))^2} \geq \frac{3\eta}{16(1 - \gamma)}, \tag{152}
\]
where the first inequality holds since (from the assumptions (149) and \( \tau \leq T - 2 \))

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

and the last inequality follows since

\[
0 \leq 1 - (1 - \eta(1 - \gamma p))^2 = 1 - \left(1 - \frac{4\eta(1 - \gamma)}{3}\right)^2 \leq \frac{8\eta(1 - \gamma)}{3}.
\]

Substituting (152) into (140), we obtain

\[
\text{Var}(V_T(2)) \geq \frac{4\gamma - 1}{36(1 - \gamma)} \sum_{\tau+1}^{T} (\eta_k^{(T)})^2 \geq \frac{2}{36(1 - \gamma)} \cdot \frac{3\eta}{16(1 - \gamma)} = \frac{\eta}{96(1 - \gamma)^2} \geq \frac{1}{96(1 - \gamma)^2\tau},
\]

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

- 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

\[
\frac{\eta_{k+1}^{(T)}}{\eta_k^{(T)}} = \frac{\eta_k - \eta_k(1 - \gamma p)}{\eta_k} = \frac{1 - \frac{4}{3}(1 - \gamma)\eta_k}{1 - (\frac{1}{\eta_k} - \frac{1}{\eta_{k+1}})\eta_k} = \frac{1 - \frac{4}{3}(1 - \gamma)\eta_k}{1 - c_\eta(1 - \gamma)\eta_k} = 1 - \frac{1}{1 - c_\eta(1 - \gamma)\eta_k} \leq 1 - (1 - \gamma)\eta_k \leq 1 - (1 - \gamma)^{\tau'},
\]

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

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

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

where the penultimate inequality comes from the Cauchy-Schwarz inequality. In addition, recognizing that \( \eta_k^{(T)} \leq \left(1 - (1 - \gamma)^{\eta_T}\right)^{k_2} \eta_{k_2}^{(T)} \) for any \( k_2 \geq k_1 \) (see (154)), one has

\[
\sum_{k=\tau'+1}^{T} \eta_k^{(T)} = \sum_{k=\tau'+1}^{T} \eta_k^{(T)},
\]

\[
\sum_{k=\max\{2\tau'-T+1,1\}}^{\tau'} \eta_k^{(T)} \leq \left(1 - (1 - \gamma)^{\eta_T}\right)^{\tau'-\tau} \sum_{k=\tau'+1}^{T} \eta_k^{(T)},
\]

\[
\sum_{k=\max\{3\tau'-2T+1,1\}}^{2\tau'-T} \eta_k^{(T)} \leq \left(1 - (1 - \gamma)^{\eta_T}\right)^{3(T - \tau')} \sum_{k=\tau'+1}^{T} \eta_k^{(T)},
\]

\[
\vdots.
\]

Summing these inequalities up and rearranging terms, we reach

\[
\sum_{k=\tau'+1}^{T} \eta_k^{(T)} \geq \frac{\sum_{k=1}^{T} \eta_k^{(T)}}{1 + (1 - (1 - \gamma)^{\eta_T})^{T - \tau'}} + (1 - (1 - \gamma)^{\eta_T})^{2(T - \tau')} + \ldots \geq \frac{\sum_{k=1}^{T} \eta_k^{(T)}}{1 - (1 - (1 - \gamma)^{\eta_T})^{T - \tau'}}
\]
Throughout this case, we assume that

\[
(1 - (1 - (1 - \gamma)\eta_T)^{T - \tau'})^{T - \tau'} \geq (1 - \eta_T)^{T - \tau'} \geq e^{-1}
\]

(see (153)). These bounds taken collectively with (45) provide that

\[
\sum_{k=\max\{\tau + 2, \tau' + 1\}}^{T} \eta_k^{(T)} = \min \left\{ \sum_{k=\tau + 2}^{T} \eta_k^{(T)}, \sum_{k=\tau' + 1}^{T} \eta_k^{(T)} \right\} \geq (1 - \eta_T) \sum_{k=\tau + 2}^{T} \eta_k^{(T)}
\]

(151)

\[
\geq (1 - e^{-1}) \left[ 1 - \prod_{i=\tau + 2}^{T} (1 - \eta_i(1 - \gamma p)) \right] \frac{1}{1 - \gamma p}
\]

(1 - e^{-1}) \left[ 1 - \prod_{i=\tau + 2}^{T} (1 - \eta_i(1 - \gamma p)) \right] \frac{1}{1 - \gamma p} \geq \frac{1 - e^{-1}}{4(1 - \gamma p)} \geq \frac{3}{32(1 - \gamma)}.
\]

Here, (i) and (ii) follow from (129) and (149), respectively, while (iii) holds since

\[
\eta_{\tau + 1}(1 - \gamma p) \leq 1 - \gamma p = \frac{4(1 - \gamma)}{3} \leq \frac{1}{3}
\]

as long as \(\gamma \geq 3/4\). Substitution into (155) yields

\[
\sum_{k=\tau + 2}^{T} (\eta_k^{(T)})^2 \geq \frac{9\eta_T}{1024(1 - \gamma)}.
\]

(156)

Substituting the above bound into (140), we obtain

\[
\operatorname{Var}(V_T(2)) \geq \frac{4\gamma - 1}{36(1 - \gamma)} \sum_{k=\tau + 1}^{T} (\eta_k^{(T)})^2 \geq \frac{2}{36(1 - \gamma)} \cdot \frac{9\eta_T}{1024(1 - \gamma)}
\]

\[
= \frac{\eta_T}{2048(1 - \gamma)^2} \geq \frac{1}{4096(1 - \gamma)^4 T},
\]

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

**Putting all this together.** With the above bounds in place, it is readily seen that either the bias is too large (see (151)) or the variance is too large (see (153) and (157)). These bounds taken collectively with (45) yield

\[
\mathbb{E}[(V^*(2) - V_T(2))^2] \geq (V^*(2) - \mathbb{E}[V_T(2)])^2 + \operatorname{Var}(V_T(2)) \geq \min \left\{ \frac{1}{144(1 - \gamma)^2}, \frac{1}{96(1 - \gamma)^4 T}, \frac{1}{4096(1 - \gamma)^4 T} \right\} = \frac{1}{4096(1 - \gamma)^4 T},
\]

(158)

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

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

Throughout this case, we assume that

\[
\eta_0^{(T)} \leq \frac{1}{75}.
\]

(159)

In fact, if \(\eta_0^{(T)} > 1/75\), then the scenario becomes much easier to cope with. To see this, applying the previous result (143) and recalling the choice (43) of \(p\) immediately yield

\[
V^*(2) - \mathbb{E}[V_T(2)] \geq \frac{\eta_0^{(T)}}{1 - \gamma p} > \frac{1}{100(1 - \gamma)},
\]

(160)
which together with (45) and the assumption $T \geq \frac{1}{(1-\gamma)^2}$ yields

$$\mathbb{E}[(V^*(2) - V_T(2))^2] \geq (V^*(2) - \mathbb{E}[V_T(2)])^2 \geq \frac{1}{10000(1-\gamma)^2} \geq \frac{1}{10000(1-\gamma)^2T}. \quad (161)$$

We now turn our attention to the dynamics w.r.t. state 1 and its associated value function $V_1(1)$ under the condition (159).

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

$$\tilde{Q}_t(a) = (1 - \eta_t)\tilde{Q}_{t-1}(a) + \eta_t \{1 + \gamma P_t(1 \mid 1, a)\tilde{V}_{t-1}\}, \quad (162)$$

with

$$\tilde{V}_{t-1} := \max_a \tilde{Q}_{t-1}(a) \quad \text{and} \quad \tilde{Q}_0(a) := Q^*(1,a) = \frac{1}{1-\gamma P},$$

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

$$Q_T(1,a) \geq \tilde{Q}_t(a) - \frac{1}{1-\gamma} \prod_{i=1}^t (1 - \eta_i(1-\gamma)), \quad a \in \{1,2\}, \quad (163)$$

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

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

$$\overline{Q}_t = (1 - \eta_t)\overline{Q}_{t-1} + \eta_t \{1 + \gamma P_t(1 \mid 1, 1)\overline{Q}_{t-1}\} \quad \text{and} \quad \overline{Q}_0 = V^*(1) = \frac{1}{1-\gamma P} \quad (164)$$

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 $\tilde{V}_t = \max_a \tilde{Q}_t(a) \geq \tilde{Q}_t(1)$, we can easily verify that

$$\tilde{Q}_t(1) \geq (1 - \eta_t)\tilde{Q}_{t-1}(1) + \eta_t \{1 + \gamma P_t(1 \mid 1, 1)\tilde{Q}_{t-1}(1)\} \geq \overline{Q}_t, \quad (165)$$

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

**A useful lower bound on the auxiliary sequence** (162). In what follows, let us establish a useful lower bound on the sequence $\{\tilde{Q}_t(a)\}$ introduced above. Then we claim that there exists some $\tau \leq T$ (see (179) and (181)) such that

$$\mathbb{P}\left\{\tilde{V}_t \geq \frac{1}{4(1-\gamma)}\right\} \geq \frac{1}{2}, \quad \text{for} \ t \geq \tau. \quad (166)$$

The auxiliary sequence constructed in (164) plays a crucial role in establishing this claim.

**Proof of the claim** (166). We intend to employ the sequence $\overline{Q}_t$ (cf. (164)) to help control $\tilde{V}_t$. It is first observed that the sequence $\overline{Q}_t$ admits the following decomposition (akin to the derivation in (135))

$$\overline{Q}_t = \left(1 - \eta_t(1-\gamma P)\right)\overline{Q}_{t-1} + \eta_t \{1 + \gamma (P_t(1 \mid 1, 1) - p)\overline{Q}_{t-1}\}$$

$$= \prod_{i=1}^t \left(1 - \eta_i(1-\gamma P)\right)\overline{Q}_0 + \sum_{k=1}^t \eta_k \prod_{i=k+1}^t \left(1 - \eta_i(1-\gamma P)\right)\left\{1 + \gamma (P_k(1 \mid 1, 1) - p)\overline{Q}_{k-1}\right\}$$

$$= \eta_0(t) \frac{1}{1-\gamma p} + \sum_{k=1}^t \eta_k(t)\left(1 - \eta_k(1-\gamma p)\right)\overline{Q}_{k-1}$$

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

43
\[
= \frac{1}{1 - \gamma p} + \sum_{k=1}^{t} \eta_k(t) \gamma (P_k(1 | 1, 1) - p) \overline{Q}_{k-1},
\]

(167)

where the last line results from (128). 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

\[
E[Z_k | P_{k-1}(1 | 1, 1), \ldots, P_{t}(1 | 1, 1)] = 0
\]

and

\[
|z_k| \leq \max_{1 \leq k \leq t} \eta_k(t) \cdot \frac{\gamma p}{1 - \gamma},
\]

where the last inequality follows from the basic property \( 0 \leq \overline{Q}_{k-1} \leq \frac{1}{1 - \gamma} \) (akin to Lemma 4) and the fact that \( |P_k(1 | 1, 1) - p| \leq \max\{p, 1 - p\} = p \) since \( p = (4\gamma - 1)/(3\gamma) \) and \( \gamma \geq 3/4 \). We intend to invoke Freedman’s inequality to control (167). Armed with these properties and the fact that \( P_k(1 | 1, 1) \) is a Bernoulli random variable with mean \( p \), we obtain

\[
\sum_{k=1}^{t} \text{Var}(z_k | P_{k-1}(1 | 1, 1), \ldots, P_{t}(1 | 1, 1)) = \sum_{k=1}^{t} (\eta_k(t))^2 p(1 - p)(\gamma \overline{Q}_{k-1})^2
\]

\[
\leq \max_{1 \leq k \leq t} \eta_k(t) \cdot \sum_{k=1}^{t} \eta_k(t) \cdot \frac{1}{3(1 - \gamma)} \leq \frac{\max_{1 \leq k \leq t} \eta_k(t)}{4(1 - \gamma)^2}.
\]

Here, the penultimate inequality relies on the fact \( 0 \leq \overline{Q}_{k-1} \leq \frac{1}{1 - \gamma} \) (akin to Lemma 4) and the choice of \( p \) (see definition (43)), whereas the last inequality results from the following condition (derived through (128))

\[
\sum_{k=1}^{t} \eta_k(t) = (1 - \eta_0(t)) \frac{1}{1 - \gamma} = \frac{1}{1 - \gamma} \leq \frac{3}{4(1 - \gamma)}.
\]

Applying Freedman’s inequality (see (58)) then yields

\[
\mathbb{P}\left\{ \left| \sum_{k=1}^{t} z_k \right| \geq \sqrt{\frac{4 \max_{1 \leq k \leq t} \eta_k(t)}{(1 - \gamma)^2} \log \frac{2}{\delta} + \frac{4 \max_{1 \leq k \leq t} \eta_k(t)}{3(1 - \gamma)} \log \frac{2}{\delta}} \right\} \leq \delta.
\]

(168)

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

\[
\max_{1 \leq k \leq t} \eta_k(t) \leq \frac{1}{50}, \quad \text{for } t \geq \tau,
\]

(169)

whose proof is postponed to Section D.3.4. In light of this claim, setting \( \delta = 1/2 \) in the expression (168) yields

\[
\sum_{k=1}^{t} z_k \geq -\frac{1}{2(1 - \gamma)}
\]

with probably at least 1/2. Combining this with the decomposition (167) and the property (165), we arrive at

\[
\hat{V}_t \geq \hat{Q}_t(1) \geq \overline{Q}_t \geq \frac{1}{1 - \gamma} - \frac{1}{2(1 - \gamma)} = \frac{1}{4(1 - \gamma)}
\]

with probability at least 1/2, where the last identity relies on the choice of \( p \) (see the definition (43)). This establishes the advertised claim (166).
Main proof. With the property (166) in place, we are positioned to prove our main result. Towards this, we find it convenient to define

\[ \Delta_t(a) := \hat{Q}_t(a) - Q^*(1,a), \quad a = 1, 2; \]  
\[ \Delta_{t,\max} := \max_a \Delta_t(a). \]  

The goal is thus to control \( \Delta_{t,\max} \); in fact, we intend to show that \( \Delta_{t,\max} \) is in expectation excessively large, resulting in an “over-estimation” issue that hinders convergence. Towards this, it follows from the iterative update rule (162) that

\[ \Delta_t(a) = (1 - \eta_t)\Delta_{t-1}(a) + \eta_t (1 + \gamma P_t(1|1, a)\hat{V}_{t-1} - Q^*(1,a)) \]
\[ = (1 - \eta_t)\Delta_{t-1}(a) + \eta_t \gamma (P_t(1|1, a)\hat{V}_{t-1} - pV^*(1)) \]
\[ = (1 - \eta_t)\Delta_{t-1}(a) + \eta_t \gamma (p\hat{V}_{t-1} - V^*(1)) + (P_t(1|1, a) - p)\hat{V}_{t-1} \]
\[ = (1 - \eta_t)\Delta_{t-1}(a) + \eta_t \gamma (p\Delta_{t-1,\max} + (P_t(1|1, a) - p)\hat{V}_{t-1}). \]

Here, the second line comes from the Bellman equation \( Q^*(1,a) = 1 + \gamma pV^*(1) \), whereas the last line holds since \( \hat{V}_{t-1} - V^*(1) = \max_a (\hat{Q}_t(a) - V^*(1)) \) (as a consequence of the relation (44)).

Applying the above relation recursively leads to

\[ \Delta_t(a) = \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1 - \eta_i) \gamma (p\Delta_{k-1,\max} + (P_k(1|1, a) - p)\hat{V}_{k-1}), \]  

(171)

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

\[ \xi_t(a) := \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1 - \eta_i) \gamma (P_k(1|1, a) - p)\hat{V}_{k-1}, \]  

(172a)

\[ \xi_{t,\max} := \max_a \xi_t(a), \]  

(172b)

one can express the above relation as follows

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

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

\[ \mathbb{E}[\xi_{t,\max}] \geq \frac{c}{\sqrt{(1-\gamma)^2T\log T}} \quad \text{for all } t \geq \hat{t} \]

(173)

for some universal constant \( c > 0 \), where

\[ \hat{t} := \max \left\{ T' \left| \prod_{i=T'}^{T} (1 - \eta_i (1 - \gamma p)) \leq \frac{6}{7} \right. \right\}, \]

(174)

whose existence is ensured under the condition (159). Given the validity of this claim (which we shall justify in Section D.3.4), we immediately arrive at

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

(175)

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

\[ x_t = (1 - \eta_t)x_{t-1} + \eta_t \left( \gamma px_{t-1} + \frac{c}{\sqrt{(1-\gamma)^2T\log T}} \right) \]

(176)
with \( x_\hat{p} = 0 \), where we recall the definition of \( \hat{p} \) in (174). In comparison to the iterative relation (175) which starts from \( \mathbb{E}[\Delta_{0, \text{max}}] = 0 \) (and hence \( \mathbb{E}[\Delta_{i, \text{max}}] \geq 0 \)), we let the sequence \( x \), start from \( x_\hat{p} = 0 \), where \( \hat{p} \) is defined in (174). It is straightforward to verify that

\[
\mathbb{E}[\Delta_{T, \text{max}}] \geq x_T, \tag{177}
\]

recognizing that

\[
x_t = \sum_{k=\hat{p}}^{t} \eta_k \prod_{i=k+1}^{t} (1 - \eta_i) \gamma^{px_{k-1}} + \sum_{k=\hat{p}}^{T} \eta_k \prod_{i=k+1}^{T} (1 - \eta_i) \frac{c}{\sqrt{(1 - \gamma)^2 T \log T}}.
\]

A little algebra reveals that the sequence (176) obeys

\[
x_T = \frac{c}{\sqrt{(1 - \gamma)^2 T \log T}} \sum_{k=\hat{p}}^{T} \eta_k \prod_{i=k+1}^{T} (1 - \eta_i(1 - \gamma p)) = \frac{c}{\sqrt{(1 - \gamma)^2 T \log T}} \frac{1}{1 - \prod_{i=\hat{p}}^{T} (1 - \eta_i(1 - \gamma p))}
\]

\[
= \frac{3c}{4\sqrt{(1 - \gamma)^4 T \log T}} \left[ 1 - \prod_{i=\hat{p}}^{T} (1 - \eta_i(1 - \gamma p)) \right] \geq \frac{3c}{28\sqrt{(1 - \gamma)^4 T \log T}}.
\]

where the second equality arises from (129), and the last inequality holds as long as \( \prod_{i=\hat{p}}^{T} (1 - \eta_i(1 - \gamma p)) \leq 6/7 \) (see (189)). This taken together with (177) leads to

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

Combining the above bound with (163) leads to

\[
\mathbb{E}\left[ V_T(1) - V^*(1) \right] \geq \mathbb{E}\left[ \Delta_{T, \text{max}} - \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)) \right] \geq \frac{3c}{28\sqrt{(1 - \gamma)^4 T \log T}} - \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)).
\]

Taking this together with (136), we arrive at

\[
\max \left\{ \mathbb{E}\left[ V_T(3) - V^*(3) \right], \mathbb{E}\left[ V_T(1) - V^*(1) \right] \right\} \geq \max \left\{ \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)), \frac{3c}{28\sqrt{(1 - \gamma)^4 T \log T}} - \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)) \right\}
\]

\[
\geq \frac{1}{2} \cdot \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)) + \frac{1}{2} \left[ \frac{3c}{28\sqrt{(1 - \gamma)^4 T \log T}} - \frac{1}{1 - \gamma} \prod_{i=1}^{T} (1 - \eta_i(1 - \gamma)) \right]
\]

\[
= \frac{3c}{56\sqrt{(1 - \gamma)^4 T \log T}}.
\]

This combined with (45) establishes the following desired lower bound:

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

### D.3.4 Proofs of auxiliary results

**Proof of the inequality (163).** We shall establish this claim by induction. To begin with, the inequality (163) 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 (163) for the $(t-1)$-th iteration and the definitions of $V_{t-1}$ and $\hat{V}_{t-1}$, we have

$$V_{t-1}(1) = \max_a Q_{t-1}(1, a) \geq \max_a \hat{Q}_{t-1}(a) - \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma))$$

$$= \hat{V}_{t-1} - \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma)).$$

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

$$Q_t(1, a) - \hat{Q}_t(a) = (1 - \eta_t)(Q_{t-1}(1, a) - \hat{Q}_{t-1}(a)) + \eta_t \gamma P_t(1 \mid 1, a)(V_{t-1}(1) - \hat{V}_{t-1})$$

$$\geq -(1 - \eta_t) \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma)) - \eta_t \gamma P_t(1 \mid 1, a) \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma))$$

$$\geq -(1 - \eta_t) \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma)) - \eta_t \frac{1}{1-\gamma} \prod_{i=1}^{t-1} (1 - \eta_i(1 - \gamma))$$

$$= - \frac{1}{1-\gamma} \prod_{i=1}^{t} (1 - \eta_i(1 - \gamma)),$$

where the first inequality invokes the induction hypothesis for the $(t-1)$-th iteration. This establishes (163) for the $t$-th iteration, and hence the proof is complete via an induction argument.

**Proof of the claim (169).** When taking the constant learning rates $\eta_t \equiv \eta \leq \frac{1}{(1-\gamma)^2T} \leq \frac{1}{50}$ (under the condition $T \geq \frac{50}{(1-\gamma)^2}$), one has

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

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

It then suffices to look at rescaled linear learning rates (i.e., $\eta_t = \frac{1}{(1+cn(1-\gamma)p)}$). As already calculated in the expression (154), the ratio of two consecutive quantities obeys

$$\frac{\eta_{k+1}^{(t)}}{\eta_k^{(t)}} = \frac{1 - \frac{4}{3}(1 - \gamma)\eta_k}{1 - c\eta(1 - \gamma)\eta_k}.$$ (178)

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 (178) is larger than 1, and hence the sequence $\{\eta_k^{(t)}\}$ decreases with $k$. Let us define

$$\tau := \min \left\{ \tau' \mid \prod_{i=1}^{\tau'} (1 - \eta_i(1 - \gamma p)) \leq \frac{1}{50} \right\},$$ (179)

which clearly satisfies $\tau \leq T$ (in view of (159)). For all $t \geq \tau$, one has

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

At the same time, we claim that one must have

$$\prod_{i=\tau}^{T} (1 - \eta_i(1 - \gamma p)) \leq \frac{2}{3},$$ (180)

47
Otherwise, recalling \( \eta_0^{(T)} = \prod_{t=1}^{T} (1 - \eta_t(1 - \gamma p)) \), we have

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

which contradicts our assumption that \( \eta_0^{(T)} > 1/75 \) (cf. (159)).

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

\[
\tau := \left\lfloor \frac{49}{c_{\eta}(1 - \gamma)} \right\rfloor < \frac{50}{(1 - \gamma)^2} < T,
\]

then for all \( t \geq \tau \) we have

\[
\max_{1 \leq k \leq t} \eta_k^{(t)} = \eta_t^{(t)} = \eta_t \leq \eta_{c_t} \leq \frac{1}{1 + c_{\eta}(1 - \gamma)} \frac{49}{c_{\eta}(1 - \gamma)} = \frac{1}{50}.
\]

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

\[
\prod_{i=\tau}^{T} (1 - \eta_i(1 - \gamma p)) \leq \left( \frac{1 - \gamma}{100} \right)^{T - \tau + 1} \leq \left( \frac{1 - \gamma}{100} \right)^{\frac{100}{(1 - \gamma)^4}} \leq \frac{3}{4}.
\]

- Putting these two cases together (with \( \tau \) specified in (179) and (181)), we obtain

\[
\max_{1 \leq k \leq t} \eta_k^{(t)} \leq \frac{1}{50}
\]

for all \( t \geq \tau \), thus establishing the desired inequality (169).

**Proof of the inequality (173).** For every \( t \), recalling the definition (172), it is convenient to write

\[
E[\xi_{t,\text{max}}] = E \left[ \frac{\xi_t(1) + \xi_t(2) + |\xi_t(1) - \xi_t(2)|}{2} \right] = E \left[ \frac{\xi_t(1) - \xi_t(2)}{2} \right]
\]

\[
= \frac{1}{2} E \left[ \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1 - \eta_i) \gamma(P_k(1 \mid 1, 1) - P_k(1 \mid 1, 2)) \hat{V}_{k-1} \right],
\]

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

\[
\zeta_t := \sum_{k=1}^{t} z_k, \quad z_k := \eta_k \prod_{i=k+1}^{t} (1 - \eta_i) \gamma(P_k(1 \mid 1, 1) - P_k(1 \mid 1, 2)) \hat{V}_{k-1}
\]

for any \( k \geq 1 \), where \( \{ z_k \} \) also forms a martingale sequence since

\[
E \left[ z_k \mid \{ P_j(1 \mid 1, 1), P_j(1 \mid 1, 2) \}_{1 \leq j < k} \right] = 0.
\]

As a consequence of Freedman’s inequality, we claim that \( \zeta_t \) satisfies

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

(184)

To verify this relation, we first notice that

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

(185)
provided that \( \max_i \eta_k \prod_{i=k+1}^t (1 - \eta_i) \leq \eta_t \). To verify the condition \( \max_i \eta_k \prod_{i=k+1}^t (1 - \eta_i) \leq \eta_t \), one can check—similar to (154)—that

\[
\frac{\eta_{k-1} \prod_{i=k}^t (1 - \eta_i)}{\eta_k \prod_{i=k+1}^t (1 - \eta_i)} = 1 - \frac{(1 - c_\eta (1 - \gamma)) \eta_k}{1 - c_\eta (1 - \gamma) \eta_k} \leq 1,
\]

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 (125)). In addition to the boundedness condition (185), we can further calculate

\[
\sum_{k=1}^t \text{Var}(z_k | P_{k-1}(1 | 1, 1), P_{k-1}(1 | 1, 2), \ldots, P_1(1 | 1, 1), P_t(1 | 1, 2))
\]

\[
= \sum_{k=1}^t \eta_k^2 \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 \cdot 2p(1 - p) \cdot (\gamma \hat{V}_{k-1})^2 \leq \sum_{k=1}^t \eta_k^2 \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 \cdot \frac{2}{3(1 - \gamma)},
\]

where the last inequality comes from the facts that \( \hat{V}_{k-1} \leq \frac{1}{1 - \gamma} \) and the choice \( p = \frac{\gamma - 1}{3\gamma} \). These bounds taken together with Freedman’s inequality (see (58)) validate (184).

By virtue of (184), setting \( \delta = \frac{(1 - \gamma)^2}{2} \mathbb{E}[|\zeta_t|^2] \) yields that with probability at least \( 1 - \delta \),

\[
|\zeta_t| \leq B := \sqrt{\frac{8 \log \frac{2}{3(1 - \gamma)}}{\varphi} \sum_{k=1}^t \eta_k^2 \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 + \frac{4\eta_k \log \frac{2}{3(1 - \gamma)}}{\varphi}} \quad \text{with } \delta = \frac{(1 - \gamma)^2}{2} \mathbb{E}[|\zeta_t|^2].
\]

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

\[
\mathbb{E}[\xi_{t, \max}] = \frac{1}{2} \mathbb{E}[|\zeta_t|] \geq \frac{1}{2} \mathbb{E}[|\zeta_t| I(|\zeta_t| \leq B)] \geq \frac{1}{2B} \mathbb{E}[|\zeta_t|^2 I(|\zeta_t| \leq B)]
\]

\[
= \frac{1}{2B} \left\{ \mathbb{E}[|\zeta_t|^2] - \mathbb{E}[|\zeta_t|^2 I(|\zeta_t| > B)] \right\}
\]

\[
\geq \frac{1}{2B} \left\{ \mathbb{E}[|\zeta_t|^2] - \frac{1}{(1 - \gamma)^2} \mathbb{P}\{|\zeta_t| > B\} \right\}
\]

\[
\geq \frac{1}{4B} \mathbb{E}[|\zeta_t|^2].
\]

(188)

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 (185) and (69); (ii) holds by the choice of \( \delta \). It is thus sufficient to lower bound \( \mathbb{E}[|\zeta_t|^2] \).

Towards this, let us define

\[
\hat{\tau} := \max \left\{ \tau' \mid \prod_{i=\tau'}^t (1 - \eta_i (1 - \gamma p)) \leq \frac{6}{7} \right\},
\]

(189)

which clearly satisfies \( \tau \leq \hat{\tau} \leq T \) (in view of (180) and (182)). Then, for all \( t \geq \hat{\tau} \) one has (which shall be proved towards the end of this subsection)

\[
\sum_{k=\hat{\tau}}^t \eta_k^2 \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 \geq \frac{1}{8} \sum_{k=1}^t \eta_k^2 \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2.
\]

(190)

We now proceed to lower bound \( \mathbb{E}[|\zeta_t|^2] \) for \( t \geq \hat{\tau} \). We first observe that for any \( t \geq \hat{\tau} \),

\[
\mathbb{E}[|\zeta_t|^2] \geq \sum_{k=1}^t \text{Var} \left( \eta_k \prod_{i=k+1}^t (1 - \eta_i) \gamma (P_k(1 | 1, 1) - P_k(1 | 1, 2)) \hat{V}_{k-1} \right)
\]

49
\[
\geq \frac{1}{2} \sum_{k=\tau}^{t} \mathbb{E} \left[ \text{Var} \left( \eta_k \prod_{i=k+1}^{t} (1-\eta_i) \gamma \left( P_k(1 \mid 1,1) - P_k(1 \mid 1,2) \right) \right) \Bigg| \hat{V}_{k-1} \geq \frac{1}{4(1-\gamma)} \right],
\]

where the first line relies on (139), and the last step makes use of the fact (166). To further control the right-hand side of the above inequality, we take \( \tau' := \max \left\{ t - \frac{1}{\eta T}, 1 \right\} \) and show that

\[
\mathbb{E} \left[ |\zeta_t|^2 \right] \geq \frac{1}{2} \sum_{k=\tau}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2 \geq \frac{1}{48(1-\gamma)} \sum_{k=\tau}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2 \geq \frac{1}{400(1-\gamma)} \sum_{k=1}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2 \geq \frac{\eta_t}{9600(1-\gamma)}. \tag{191}
\]

Here, (i) makes use of the constraint \( \hat{V}_{k-1} \geq \frac{1}{4(1-\gamma)} \), while (ii) makes use of (190), and (iii) are valid if the following property holds (which shall be proved towards the end of this subsection)

\[
\sum_{k=\tau}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2 \geq \frac{1}{24} \eta_t. \tag{192}
\]

We are now well-equipped to control \( \mathbb{E}[|\zeta_{t,\text{max}}|^2] \) using the property (188). Recall the expression of \( B \) in (187), we know that bounding \( \mathbb{E}[|\zeta_t|^2]/B \) boils down to controlling

\[
\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta_t}{1-\gamma} \log \frac{2}{\delta}} \text{ and } \frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log \frac{2}{\delta}}{1-\gamma} \sum_{k=1}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2}}. \tag{193}
\]

• For the first term in (193), recalling that \( \delta = \frac{(1-\gamma)^2}{2} \mathbb{E}[|\zeta_t|^2] \), we can demonstrate that

\[
\log \frac{1}{\delta} = -\log \frac{(1-\gamma)^2}{2} \mathbb{E}[|\zeta_t|^2] \leq -\log \frac{(1-\gamma)\eta_t}{19200} \leq \log \frac{19200(1+(1-\gamma)T \log T)}{1-\gamma} \lesssim \log T, \tag{194}
\]

where the first inequality makes use of the bound (191), and the second inequality arises from the fact \( \eta_t \geq \frac{1}{1+(1-\gamma)T \log T} \) (given the range of the learning rates in this case). Combining this with (191), we can guarantee that

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

• Moving to the second term in (193), one can ensure that

\[
\frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log \frac{2}{\delta}}{1-\gamma} \sum_{k=1}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2}} \lesssim \sqrt{\frac{1}{(1-\gamma) \log T} \sum_{k=1}^{t} \eta_k^2 \left[ \prod_{i=k+1}^{t} (1-\eta_i) \right]^2} \gtrsim \frac{\eta_t}{(1-\gamma) \log T} \gtrsim \frac{1}{\sqrt{(1-\gamma)^2 T \log T}}. \tag{195}
\]

Here, (i) follows from (191) and (194) since

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

(ii) arises from (192); 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 (188) and using the expression of $B$ in (187), we reach
\[ \mathbb{E}[\xi_{t,\text{max}}] \geq \frac{1}{4B} \mathbb{E}[(\xi_{t,L}^2)] \geq \frac{c}{\sqrt{(1-\gamma)^2T \log T}}, \]
for some constant $c > 0$. Thus, this validates the inequality (173).

**Proof of the claim (190).** By the definition of $\tilde{\tau}$ in (189), we have $\prod_{i=\tilde{\tau}}^t (1 - \eta_i (1 - \gamma p)) \leq 6/7$. An important observation is that
\[ \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \overset{(i)}{=} 1 - \prod_{i=\tau}^t (1 - \eta_i) \geq \frac{1}{8} \sum_{k=1}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]. \quad (195) \]
Here, the relations (i) and (iii) arise from (70), and the inequality (ii) follows since $c > 0$.

With (195) in place, we can continue to prove the claim (190). Recognizing that $\eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]$ is increasing in $k$ (see (186)), we can obtain
\[ \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 \leq \max_{1 \leq k < t} \left\{ \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \right\} \sum_{k=1}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \]
\[ \leq \eta_{\tilde{\tau}} \left[ \prod_{i=\tilde{\tau}+1}^t (1 - \eta_i) \right] \sum_{k=1}^\tau \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \]
\[ \leq 7 \eta_{\tilde{\tau}} \left[ \prod_{i=\tilde{\tau}+1}^t (1 - \eta_i) \right] \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]. \quad (197) \]
where the last inequality comes from (195). With the preceding inequality in place, the claim (190) then follows by observing that
\[ \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]^2 \geq \min_{\tau \leq k \leq t} \left\{ \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \right\} \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \]
\[ \geq \eta_{\tilde{\tau}} \left[ \prod_{i=\tilde{\tau}+1}^t (1 - \eta_i) \right] \sum_{k=\tau}^t \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right], \]
where we make use of the monotonicity of $\eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right]$ again.

**Proof of the claim (192).** Note that for $\tau' := \max \left\{ t - \frac{1}{\eta_{t/2}}, 1 \right\}$, one has
\[ \eta_k \left[ \prod_{i=k+1}^t (1 - \eta_i) \right] \geq \eta_k (1 - \eta_{t/2})^{t - \tau'} \geq \eta_k (1 - \eta_{t/2})^{1/\eta_{t/2}} \geq \frac{1}{3} \eta_{t/2}, \quad \text{for all } \tau' \leq k \leq t, \]
as long as the following condition holds (recalling the definition of $\tilde{\tau}$ in (189))
\[ \eta_{t/2} \leq 2 \eta_{t/2} \leq 2 \tilde{\tau} \leq 1/10. \quad (198) \]
In addition, similar to (129), we can derive
\[ \sum_{k=\tau'}^t \eta_k \prod_{i=k+1}^t (1 - \eta_i) = 1 - \prod_{i=\tau'}^t (1 - \eta_i) \geq 1 - \max \left\{ (1 - \eta_t)^{1/\eta_{t/2} + 1}, \prod_{i=1}^t (1 - \eta_i) \right\} \]
\[ \geq 1 - \max \left\{ e^{-1/2}, \prod_{i=1}^{\tilde{\tau}} (1 - \eta_i) \right\} \geq \frac{1}{8}, \]

where we once again use the condition (198), and the last inequality comes from the derivation in (196). Putting these two bounds together yields

\[ t \sum_{k=\tau'}^t \eta_k^2 \left[ \prod_{i=k+1}^{t} (1 - \eta_i) \right]^2 \geq \min_{k: \tau' \leq k \leq t} \left\{ \eta_k \left[ \prod_{i=k+1}^{t} (1 - \eta_i) \right] \right\} \sum_{k=\tau'}^t \eta_k \left[ \prod_{i=k+1}^{t} (1 - \eta_i) \right] \]

\[ \geq \frac{1}{3} \eta_{\tau'} \cdot \frac{1}{8} \geq \frac{1}{24} \eta_{\tau'}. \]

To finish up, it remains to justify (198). This condition is obvious for constant learning rates. As for rescaled learning rates, one can see that

\[ \eta_i = \frac{1}{1 + (1 - \gamma)c_\eta i} \geq \frac{19}{20c_\eta (1 - \gamma)i} \]

for all \( i \geq \tau \), where \( \tau := \left\lceil \frac{19}{c_\eta (1 - \gamma)} \right\rceil \). This allows one to obtain

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

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 (189) implies that \( \tilde{\tau} \geq \tau \) and hence \( \eta_{\tilde{\tau}} \leq \eta_{\tau} = \frac{1}{4(1 - \gamma)} = 1/20. \)

**D.4 Proof of Lemma 3**

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

\[ V^\pi(0) = 0 \quad \text{for all } \pi \implies V^*(0, 1) = 0. \]

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,

\[ Q^*(1, 1) = Q^*(1, 2) = V^*(1). \] (199)

From Bellman’s equation, we can thus deduce that

\[ Q^*(1, 1) = r(1, 1) + \gamma P(0|1, 1)V^*(0) + \gamma P(1|1, 1)V^*(1), \]

which in conjunction with (199) and a little algebra leads to

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

Here, the second identity follows since \( V^*(0) = 0 \), and the third identity makes use of (43). The calculation for \( V^*(2) \) and \( Q^*(2, 1) \) follows from an identical argument and is hence omitted.

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

\[ V^*(3) = Q^*(3, 1) = r(3, 1) + \gamma P(3|3, 1)V^*(3) = 1 + \gamma V^*(3), \]

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

52
E Analysis for asynchronous Q-learning (Theorem 4)

E.1 Notation and preliminary facts

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

$$P_t((s,a),s') = \begin{cases} 1, & \text{if } (s,a,s') = (s_{t-1},a_{t-1},s_t); \\ 0, & \text{else.} \end{cases}$$  \hfill (200)

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

$$\Lambda_t((s,a),(s,a)) = \begin{cases} \eta, & \text{if } (s,a) = (s_{t-1},a_{t-1}); \\ 0, & \text{otherwise.} \end{cases}$$  \hfill (201)

Armed with the above notation, the asynchronous Q-learning update rule (48) can be conveniently expressed as follows:

$$Q_t = (I - \Lambda_t)Q_{t-1} + \Lambda_t(r + \gamma P_t V_{t-1}).$$  \hfill (202)

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

Lemma 7. Suppose that $0 < \eta_t \leq 1$ for all $t \geq 0$. Assume that $0 \leq Q_0 \leq \frac{1}{1-\gamma}1$. Then for any $t \geq 0$, one has

$$0 \leq Q_t \leq \frac{1}{1-\gamma}1 \quad \text{and} \quad 0 \leq V_t \leq \frac{1}{1-\gamma}1.$$  \hfill (203)

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

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 (202), we can derive the following elementary decomposition

$$\Delta_t := Q_t - Q^* = (I - \Lambda_t)Q_{t-1} + \Lambda_t(r + \gamma P_t V_{t-1}) - Q^*$$
$$= (I - \Lambda_t)(Q_{t-1} - Q^*) + \Lambda_t(r + \gamma P_t V_{t-1} - Q^*)$$
$$= (I - \Lambda_t)(Q_{t-1} - Q^*) + \gamma \Lambda_t(P_t V_{t-1} - P V^*)$$
$$= (I - \Lambda_t)\Delta_{t-1} + \gamma \Lambda_t(P_t - P)V_{t-1} + \gamma \Lambda_t P(V_{t-1} - V^*).$$  \hfill (204)

where the penultimate identity follows from the Bellman optimality equation $Q^* = r + PV^*$. Combining (204) with the inequalities (33) and using the definition (27) of $\pi_t$ result in

$$\Delta_t \leq (I - \Lambda_t)\Delta_{t-1} + \gamma \Lambda_t(P_t - P)V_{t-1} + \gamma \Lambda_t P^\pi_{t-1} \Delta_{t-1},$$  \hfill (205a)

$$\Delta_t \geq (I - \Lambda_t)\Delta_{t-1} + \gamma \Lambda_t(P_t - P)V_{t-1} + \gamma \Lambda_t P^\pi_t \Delta_{t-1}.$$  \hfill (205b)

Apply the above two relations recursively to obtain

$$\Delta_t \leq \gamma \sum_{i=1}^t \prod_{j=i+1}^t (I - \Lambda_j)\Lambda_i(P_t - P)V_{t-1} + \gamma \sum_{i=1}^t \prod_{j=i+1}^t (I - \Lambda_j)\Lambda_i P^\pi_{t-1} \Delta_{i-1} + \prod_{j=1}^t (I - \Lambda_j)\Delta_0,$$  \hfill (206a)

$$\Delta_t \geq \gamma \sum_{i=1}^t \prod_{j=i+1}^t (I - \Lambda_j)\Lambda_i(P_t - P)V_{t-1} + \gamma \sum_{i=1}^t \prod_{j=i+1}^t (I - \Lambda_j)\Lambda_i P^\pi_t \Delta_{i-1} + \prod_{j=1}^t (I - \Lambda_j)\Delta_0.$$  \hfill (206b)
By defining the following diagonal matrices

$$\Lambda_i^{(t)} := \begin{cases} 
\prod_{j=1}^{t}(I - \Lambda_j), & \text{if } i = 0, \\
\Lambda_i \prod_{j=i+1}^{t}(I - \Lambda_j), & \text{if } 0 < i < t, \\
\Lambda_t, & \text{if } i = t,
\end{cases} \quad (207)$$

and setting

$$\beta = \frac{c_3(1 - \gamma)}{\log T} \quad (208)$$

for some constant $c_3 > 0$, we can rearrange terms in the upper bound (206a) to reach

$$\Delta_t \leq \Lambda_0^{(t)} \Delta_0 + \gamma \sum_{i=1}^{(1-\beta)t} \Lambda_i^{(t)} \left[ (P_i - P)V_{i-1} + P^{\pi_{i-1}} \Delta_{i-1} \right]$$

$$=: \zeta_t$$

$$+ \gamma \sum_{i=(1-\beta)t+1}^{t} \Lambda_i^{(t)} (P_i - P)V_{i-1} + \gamma \sum_{i=(1-\beta)t+1}^{t} \Lambda_i^{(t)} P^{\pi_{i-1}} \Delta_{i-1}. \quad (209)$$

In the subsequent steps, we shall first develop bounds on the sizes of the terms $\zeta_t$ and $\xi_t$ in (209) separately, and then combine these bounds with (209) recursively in order to derive the advertised upper bound on $\Delta_t$.

**Step 2: bounding the terms $\zeta_t$ and $\xi_t$.** The terms $\zeta_t$ and $\xi_t$ defined in (209) can be bounded with high probability by the following lemmas.

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

$$\|\zeta_t\|_\infty \leq \frac{4}{(1 - \gamma)^T} \quad (210)$$

for all $t$ obeying $\frac{T}{c_4 \log T} \leq t \leq T$. Here, $c_4 > 0$ is some constant obeying $c_4 \leq c_1 c_3 / 4$, where the constants $c_1$ and $c_3$ appear in (52a) and (208), respectively.

**Proof.** See Section E.3.1.

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

$$|\xi_t| \leq \sqrt{\frac{16(\log^3 T)(\log \frac{|S|A|T}}{(1 - \gamma)^T \mu_{\min}} \left( \max_{(1-\beta)t \leq i < t} \text{Var}_P(V_i) + 1 \right) + \frac{6(\log^3 T)(\log \frac{|S|A|T}}{(1 - \gamma)^2 T \mu_{\min}} 1} \quad (211)$$

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

**Proof.** See Section E.3.2.

**Step 3: controlling $\Delta_t$.** Consider any $t$ obeying $\frac{T}{c_4 \log T} \leq t \leq T$ and any $k$ obeying $2t/3 \leq k \leq t$. Under the sample size condition (52b), Lemmas 8-9 together with a little algebra lead to

$$|\zeta_t| + |\xi_t| \leq \sqrt{\frac{32(\log^3 T)(\log \frac{|S|A|T}}{(1 - \gamma)^T \mu_{\min}} \left( \max_{(1-\beta)t \leq i < k} \text{Var}_P(V_i) + 1 \right) \leq \sqrt{\varphi_t}}$$

where we define

$$\varphi_t := \frac{32(\log^3 T)(\log \frac{|S|A|T}}{(1 - \gamma)^T \mu_{\min}} \left( \max_{\frac{t}{2} \leq i \leq t} \text{Var}_P(V_i) + 1 \right). \quad (212)$$
Combining this inequality with (209) allows us to obtain

$$\Delta_k \leq \sqrt{\varphi_i} + \sum_{i=(1-\beta)k+1}^{k} \Lambda^{(k)}_{i} \gamma P^{\pi_{i-1}} \Delta_{i-1} = \sqrt{\varphi_i} + \sum_{i=(1-\beta)k}^{k-1} \Lambda^{(k)}_{i+1} \gamma P^{\pi_{i}} \Delta_{i} \quad \text{for all } 2t/3 \leq k \leq t. \quad (213)$$

Similar to the quantity $\alpha^{(t)}_i$ defined in (76), let us define

$$D^{(t)}_i := \left( \sum_{j=(1-\beta)t}^{t-1} \Lambda^{(t)}_{j+1} \right)^{-1} \Lambda^{(t)}_{i+1}, \quad (214)$$

which, according to (69) and the definition (201), clearly satisfies

$$D^{(t)}_i \geq \Lambda^{(t)}_{i+1} \geq 0 \quad \text{and} \quad \sum_{i=(1-\beta)t}^{t-1} D^{(t)}_i = I. \quad (215)$$

Set $i_0 = t$ for notational convenience. With this set of notation and the property (215) in mind, we can derive the following bound

$$\Delta_i \leq \sum_{i_1=(1-\beta)t}^{t-1} \left( D^{(t)}_{i_1} \sqrt{\varphi_i} + \Lambda^{(t)}_{i_1+1} \gamma P^{\pi_{i_1}} \Delta_{i_1} \right)$$

$$\leq \sum_{i_1=(1-\beta)t}^{t-1} \left[ D^{(t)}_{i_1} \sqrt{\varphi_i} + \frac{\sum_{i_2=(1-\beta)i_1}^{t-1} \Lambda^{(t)}_{i+1} \gamma P^{\pi_{i_1}} \Delta_{i_1}}{\sum_{i_2=(1-\beta)i_2}^{t-1} \delta^{(i_2)}_{i_2} \sqrt{\varphi_i}} \right]$$

$$\leq \sum_{i_1=(1-\beta)t}^{t-1} D^{(t)}_{i_1} \sqrt{\varphi_i} + \sum_{i_2=(1-\beta)i_2}^{t-1} \sum_{i_1=(1-\beta)i_1}^{t-1} \delta^{(i_2)}_{i_2} \sqrt{\varphi_i} \left( \sum_{\gamma P^{\pi_{i_1}} \Delta_{i_1}} \delta^{(i_2)}_{i_2} \right)$$

$$= \sum_{i_1=(1-\beta)t}^{t-1} D^{(t)}_{i_1} \left\{ I + \sum_{i_2=(1-\beta)i_2}^{t-1} \gamma P^{\pi_{i_1}} D^{(i_2)}_{i_2} \right\} \sqrt{\varphi_i} + \sum_{i_1=(1-\beta)i_1}^{t-1} \sum_{i_2=(1-\beta)i_2}^{t-1} \gamma P^{\pi_{i_1}} \Delta_{i_1}\Delta_{i_2}. \quad (216)$$

Here, the first relation makes use of the second property in (215), the second relation further expands $\Delta_{i_1}$ in the same way as in the first line of (216), whereas the third inequality relies on the first property in (215).

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

$$H := \log\frac{T}{1-\gamma}. \quad (217)$$

Similar to the way we derive (84), we can apply the relation (216) recursively and use the basic relation $|\Delta_k| \leq \frac{1}{1-\gamma} \mathbf{1}$ for any $k$ to show that

$$\Delta_i \leq \sum_{(i_1, \ldots, i_H) \in I_t} \left\{ D^{(t)}_{i_1} \left( I + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} (P^{\pi_{i_k}} D^{(i_k)}_{i_{k+1}}) \right) \sqrt{\varphi_i} + \gamma^H \prod_{k=1}^{H} (A^{(i_{k+1})}_{i_{k+1}} P^{\pi_{i_k}}) \Delta_{i_H} \right\}, \quad (218)$$

where $I_t$ has been defined in (82). To further simplify (218), we need to control the two terms on the right-hand side of (218) separately.

- We shall begin with the first term on the right-hand side of (218). Towards this end, let us define a collection of policies $\{\pi_k\}$ recursively and backward as follows:

$$\hat{\pi}_k := \begin{cases} \arg\max_{\pi \in \Pi} \prod_{i=1}^{h} \sqrt{\varphi_i}, & \text{if } k = H - 1, \\ \arg\max_{\pi \in \Pi} \sqrt{\varphi_i}, & \text{if } k = H - 2, \ldots, 1, \end{cases} \quad (219)$$
or alternatively (in view of the definition (24) of $P^\pi$),

$$\hat{\pi}_k := \begin{cases} \arg \max_{\pi \in \Pi} P^\pi \sqrt{\varphi_t}, & \text{if } k = H - 1; \\
\arg \max_{\pi \in \Pi} P^\pi \left( I + \sum_{h=1}^{H-1} \gamma^h \prod_{j=1}^{h} P^\pi_j \right) \sqrt{\varphi_t}, & \text{if } k = H - 2, \cdots, 1. \end{cases}$$

(220)

Here, $\Pi$ is a policy set satisfying

$$\Pi := \{ \pi = [\pi(s)]_{s \in S} \mid \pi(s) \in \Pi_s, \forall s \in S \}; \quad \Pi_s := \{ \pi_i(s) \mid i \in [t/2, t) \};$$

(221)
in words, for any policy $\pi$ belonging to $\Pi$, each $\pi(s)$ coincides with one of the policy iterates $\pi_i(s)$ during the latest $\beta t$ 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

$$\sum_{i_1, \cdots, i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) \sqrt{\varphi_t} \right)$$

$$= \sum_{i_1, \cdots, i_{H-1}} \sum_{i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-2} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) + \prod_{k=1}^{H-2} \left( \gamma P^{\pi_{i_H}} D_{i_{H-1}}^{(i)} \right) \sqrt{\varphi_t} \right)$$

$$= \sum_{i_1, \cdots, i_{H-1}} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-2} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) + \prod_{k=1}^{H-2} \left( \gamma P^{\pi_{i_H}} D_{i_{H-1}}^{(i)} \right) \sqrt{\varphi_t} \right)$$

(i)

$$\leq \sum_{i_1, \cdots, i_{H-1}} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-3} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) + \prod_{k=1}^{H-3} \left( \gamma P^{\pi_{i_H}} D_{i_{H-2}}^{(i)} \right) \left( \sum_{i_{H-1}} D_{i_{H-1}}^{(i)} \left( I + \gamma P^{\pi_{i_{H-1}}} \right) \right) \sqrt{\varphi_t} \right)$$

(ii)

$$= \sum_{i_1, \cdots, i_{H-2}} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-3} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) + \prod_{k=1}^{H-3} \left( \gamma P^{\pi_{i_H}} D_{i_{H-2}}^{(i)} \right) \left( \sum_{i_{H-1}} D_{i_{H-1}}^{(i)} \left( I + \gamma P^{\pi_{i_{H-1}}} \right) \right) \sqrt{\varphi_t} \right)$$

(iii)

$$\leq \sum_{i_1, \cdots, i_{H-2}} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-3} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) + \prod_{k=1}^{H-3} \left( \gamma P^{\pi_{i_H}} D_{i_{H-2}}^{(i)} \right) \left( \sum_{i_{H-1}} D_{i_{H-1}}^{(i)} \left( I + \gamma P^{\pi_{i_{H-1}}} \right) \right) \sqrt{\varphi_t} \right)$$

(iv)

where we abbreviate $\sum_{(i_1, \cdots, i_H) \in I_t}$ 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 (215), while (ii) and (iv) are due to the construction (220). Continuing the derivation of the above inequality recursively, we arrive at

$$\sum_{i_1, \cdots, i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) \sqrt{\varphi_t} \right) \leq \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} \right) \sqrt{\varphi_t} \right).$$

• We now turn attention to the second term on the right-hand side of (218). It is seen that

$$\sum_{i_1, \cdots, i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} D_{i_{k+1}}^{(i)} \right) \left| \Delta_{i_H} \right| \right) \leq \frac{\gamma^H}{1 - \gamma} \sum_{i_1, \cdots, i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} \right) \right)$$

$$= \frac{\gamma^H}{1 - \gamma} \sum_{i_1, \cdots, i_H} D_{i_t}^{(i)} \left( I + \sum_{h=1}^{H-1} \prod_{k=1}^{h} \left( \gamma P^{\pi_{i_k}} \right) \right) \left( \sum_{i_H} D_{i_{H-1}}^{(i_H-1)} \left( \gamma P^{\pi_{i_H}} \right) \right) \left( \sum_{i_{H-1}} D_{i_{H-1}}^{(i_{H-1})} \right) \sqrt{\varphi_t}.$$
\[
\begin{align*}
\gamma^H \prod_{k=1}^{H-1} (D_{i_k}^{(i_{k-1})} P^{\pi_{i_k}}) 1 \\
= \cdots = \gamma^H 1 = \frac{\gamma \log T}{1 - \gamma} 1 \\
\leq \frac{1}{(1 - \gamma)T} 1,
\end{align*}
\]

where the first line follows from the first property in (215), the third line is due to the fact \( P\pi 1 = 1 \) for any \( \pi \), and the fourth line arises from the second property in (215).

Substituting the above two bounds into (218) yields

\[
\Delta_t \leq \left( I + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} P^\pi_{i_k} \right) \sqrt{\varphi_t} + \frac{1}{(1 - \gamma)T} 1.  
\tag{222}
\]

**Step 4: putting all pieces together.** Repeating our analysis for the term \( \beta_1 \) in Section B.2 (i.e., Step 5 of Section B.2 with Lemma 5 replaced by Lemma 6), we arrive at

\[
|\beta|^2 \leq \frac{320}{\gamma^2 (1 - \gamma)^4 T \mu_{\min}} \left( 1 + 2 \max_{\frac{T}{2} \leq t < T} \|\Delta_t\|_\infty \right) 1
\]

with probability at least \( 1 - \delta \). Substitution into (222) then yields

\[
\Delta_t \leq \sqrt{\frac{320}{\gamma^2 (1 - \gamma)^4 T \mu_{\min}} \left( 1 + \max_{\frac{T}{2} \leq t < T} \|\Delta_t\|_\infty \right) 1} + \frac{1}{(1 - \gamma)T} 1
\leq 30 \sqrt{\frac{(\log^3 T) (\log \frac{|S| |A| T}{\delta})}{\gamma^2 (1 - \gamma)^4 T \mu_{\min}} \left( 1 + \max_{\frac{T}{2} \leq t < T} \|\Delta_t\|_\infty \right) 1}
\tag{223}
\]

holds simultaneously for all \( t \geq \frac{T}{4} \log T \), provided that the sample size condition (52b) is satisfied. Similarly, we can also establish the following lower bound on \( \Delta_t \) (which we omit the details for the sake of brevity)

\[
\Delta_t \geq -30 \sqrt{\frac{(\log^3 T) (\log \frac{|S| |A| T}{\delta})}{\gamma^2 (1 - \gamma)^4 T \mu_{\min}} \left( 1 + \max_{\frac{T}{2} \leq t < T} \|\Delta_t\|_\infty \right) 1}
\]

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

\[
\|\Delta_t\|_\infty \leq 30 \sqrt{\frac{(\log^3 T) (\log \frac{|S| |A| T}{\delta})}{\gamma^2 (1 - \gamma)^4 T \mu_{\min}} \left( 1 + \max_{\frac{T}{2} \leq t < T} \|\Delta_t\|_\infty \right) 1}
\tag{224}
\]

This resembles the relation (40) derived for the synchronous case, except that \( T \) in the denominator is replaced with \( \mu_{\min} T \). As a result, we can readily repeat the argument in Appendix B.4 to reach

\[
\|\Delta_T\|_\infty \leq O \left( \frac{(\log^4 T) (\log \frac{|S| |A| T}{\delta})}{(1 - \gamma)^4 T \mu_{\min}} + \frac{C_1}{(1 - \gamma)^4 T \mu_{\min}} \right),
\tag{225}
\]

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 $\zeta_t$ in (209) and the fact that $\Lambda_0^{(t)}$ is a diagonal matrix, we can deduce that

$$
\|\zeta_t\|_{\infty} \leq \|\Lambda_0^{(t)}\|_{\infty} + t \max_{1 \leq i \leq |S|} \|\Lambda_i^{(t)}\|_{\infty} \left(\|P_{i-1}^{\pi}\|_{\infty} + \|P_{i}V_{i-1}\|_{\infty} + \|PV_{i-1}\|_{\infty}\right)
$$

$$
\leq \|\Lambda_0^{(t)}\|_{\infty} + t \max_{1 \leq i \leq |S|} \|\Lambda_i^{(t)}\|_{\infty} \left(\|P_{i-1}^{\pi}\|_{\infty} + \|\Lambda_i^{(t)}\|_{\infty} \max_{1 \leq i \leq |S|} \left\{\|P_{i-1}^{\pi}\|_{\infty} + (\|P_1\|_1 + \|P\|_1)\|V_{i-1}\|_1\right\}\right)
$$

$$
(\text{E.3.2 Proof of Lemma 9})
$$

$$
= \|\Lambda_0^{(t)}\|_{\infty} + t \max_{1 \leq i \leq |S|} \|\Lambda_i^{(t)}\|_{\infty} \left(\|\Lambda_i^{(t)}\|_{\infty} + 2\|V_{i-1}\|_{\infty}\right)
$$

Here, (i) holds true since $\|P_{i-1}^{\pi}\|_1 = \|P_i\|_1 = \|P\|_1 = 1$. To verify (ii), we first define

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

and

$$
K_t(s,a) := \left\{k \geq 1 \mid t_k(s,a) < t\right\},
$$

namely, the total number of times — before the $t$-th iteration — that the sample trajectory visits $(s,a)$. Then Li et al. (2022c, Lemma 8) tells us that with probability at least $1 - \delta$,

$$
K_{t_1}(s,a) - K_{t_2}(s,a) \geq \frac{1}{2} (t_1 - t_2) \mu_{\min},
$$

holds uniformly for all $(s,a) \in S \times A$ and $0 \leq t_2 \leq t_1 \leq T$ obeying

$$
t_1 - t_2 \geq \frac{88t_{\text{mix}}}{\mu_{\min}} \log \frac{|S||A|T}{\delta}.
$$

This in turn implies that: if $\beta t \geq \frac{88t_{\text{mix}}}{\mu_{\min}} \log \frac{|S||A|T}{\delta}$ and $i \leq (1 - \beta)t$, then one has

$$
\|\Lambda_0^{(t)}\| = \left\|\prod_{j=1}^t (I - \Lambda_j)\right\| \leq \max_{(s,a) \in S \times A} (1 - \eta)^{K_t(s,a)} \leq (1 - \eta)^{\frac{i}{2}\mu_{\min}} \leq \frac{1}{T^2}
$$

$$
(\text{E.3.2 Proof of Lemma 9})
$$

$$
\|\Lambda_i^{(t)}\| = \left\|\Lambda_i \prod_{j=1}^t (I - \Lambda_j)\right\| \leq \max_{(s,a) \in S \times A} (1 - \eta)^{K_t(s,a) - K_i(s,a)} \leq (1 - \eta)^{\frac{i}{2}\beta \mu_{\min}} \leq \frac{1}{T^2}
$$

with probability at least $1 - \delta$, provided that $\eta \beta t \mu_{\min} > 4 \log T$. In other words, (229) holds with probability at least $1 - \delta$, as long as

$$
t > \max \left\{\frac{4 \log T}{\eta \beta \mu_{\min}}, \frac{88t_{\text{mix}}}{\beta \mu_{\min}} \log \frac{|S||A|T}{\delta}\right\} = \max \left\{\frac{4T}{c_1 \log T}, \frac{88t_{\text{mix}}}{c_3 (1 - \gamma) \mu_{\min}} \log \frac{|S||A|T}{\delta} \log T\right\}.
$$

This taken together with the sample size assumption (52b) concludes the proof of Lemma 8.

E.3.2 Proof of Lemma 9

Fix any state-action pair $(s,a) \in S \times A$, and let us look at the $(s,a)$-th entry of $\xi_t$, i.e., $\xi_t(s,a)$. For notational simplicity, let $\Lambda_j(s,a)$ denote the $(s,a)$-th diagonal entry of the diagonal matrix $\Lambda_j$, and $P_t(s,a)$ (resp. $P(s,a)$) the $(s,a)$-th row of $P_t$ (resp. $P$).
Using the definition of $\xi_t$ in (209) and the above notation, we can derive

$$\xi_t(s, a) = \gamma \sum_{i=(1-\beta)t+1}^t \prod_{j=i+1}^t (1 - \Lambda_j(s, a)) \Lambda_i(s, a) (P_i(s, a) - P(s, a)) V_{i-1}. \tag{230}$$

Equipped with the definitions of $t_k(s, a)$ (cf. (226)) and $K_t(s, a)$ (cf. (227)), we can further rewrite (230) as

$$\xi_t(s, a) = \gamma \sum_{k=K_t(s, a)}^{K_t(s, a)} (1 - \eta)^{k-1} \eta (P_{t_k+1}(s, a) - P(s, a)) V_{t_k}. \tag{231}$$

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

$$\left| \sum_{k=K_{\beta}}^{K} (1 - \eta)^{K-k} \eta (P_{t_k+1}(s, a) - P(s, a)) V_{t_k} \right| \leq \sqrt{\frac{16(\log^3 T)(\log \frac{|S||A|}{\delta})}{(1-\gamma)T\mu_{\min}}} \left( \max_{t_{K_{\beta}} \leq i \leq t_K} \text{Var}_P(V_i) + 1 \right) + \frac{6(\log^3 T)(\log \frac{|S||A|}{\delta})}{(1-\gamma)^2T\mu_{\min}} \tag{232}$$

holds simultaneously for all $(s, a) \in S \times A$ and all $1 \leq K_{\beta} \leq K \leq \frac{\log^3 T}{(1-\gamma)T\mu_{\min}}$. If this claim were true, then taking $K_{\beta} = K_{(1-\beta)t+1}$ and $K = K_t$ and substituting the bound (232) into the expression (231) would lead to

$$|\xi_t| \leq \sqrt{\frac{16(\log^3 T)(\log \frac{|S||A|}{\delta})}{(1-\gamma)T\mu_{\min}}} \left( \max_{(1-\beta)t \leq i \leq t} \text{Var}_P(V_i) + 1 \right) + \frac{6(\log^3 T)(\log \frac{|S||A|}{\delta})}{(1-\gamma)^2T\mu_{\min}}, \tag{233}$$

thus concluding the proof of this lemma. To finish up, it is sufficient to justify the claim (232), which forms the content of the remainder of this proof.

**Proof of the claim (232).** Let us use the notation in (64) to express $\eta^{(K)}_k = (1-\eta)^{K-k} \eta$. For any fixed integer $K > 0$, the following vectors

$$\{P_{t_k+1}(s, a) \mid 1 \leq k \leq K\}$$

are identically and independently distributed; see Li et al. (2022c, Section B.1). We can then express the term

$$X_K := \sum_{k=K_{\beta}}^{K} (1 - \eta)^{K-k} \eta (P_{t_k+1}(s, a) - P(s, a)) V_{t_k},$$

as follows:

$$X_K = \sum_{k=K_{\beta}}^{K} z_k \quad \text{with} \quad z_k := \eta^{(K)}_k (P_{t_k+1}(s, a) - P(s, a)) V_{t_k},$$

where the $z_k$’s satisfy

$$\mathbb{E}[z_k \mid t_k, \cdots, t_1, V_{t_k}, \cdots, V_{t_1}] = 0.$$

We intend to invoke the Freedman inequality to control $X_K$ for any $K$ obeying $K \leq T$. Similar to the synchronous counterpart, we can see that

$$B := \max_{1 < k \leq K} \|z_k\|_\infty \leq \eta^K \|P_{t_k+1} + \|P_1\|_1\|V_{t_k}\|_\infty \leq \frac{2\eta^K}{1-\gamma} \leq \frac{2\eta}{1-\gamma},$$

59
With the preceding bounds in place, applying the Freedman inequality in Theorem 6 and taking
with probability at least
where we have made use of (69). In addition, we make note of a trivial upper bound on

where we have made use of (69). In addition, we make note of a trivial upper bound on

We shall denote by

Let us construct an MDP

\[ F \]

**Lower bound for asynchronous Q-learning (Theorem 5)**

This section establishes Theorem 5 by identifying a hard MDP instance satisfying the assumed conditions.

**F.1 Construction of a hard instance and its values**

Let us construct an MDP \( \mathcal{M}_{\text{hard}} \) with state space \( S = \{0, 1, 2, 3\} \) as follows, which is partly inspired by the idea from Li et al. (2022b). We shall denote by \( \mathcal{A}_s \) the action space associated with state \( s \). The probability transition kernel and the reward function of \( \mathcal{M}_{\text{hard}} \) are specified as follows:

- \( \mathcal{A}_0 = \{1, 2\}, \quad P(\cdot | 0, 1) = [1, 0, 0, 0], \quad r(0, 1) = \frac{2}{3}, \quad (234a) \)

- \( P(\cdot | 0, 2) = \begin{bmatrix} 1 - 2(1 - \gamma)(1 - \mu_0), 2(1 - \gamma)\mu_1, 2(1 - \gamma)\mu_2, 2(1 - \gamma)\mu_3 \end{bmatrix}, \quad r(0, 2) = 0, \quad (234b) \)

- \( \mathcal{A}_1 = \{1, 2, 3\}, \quad P(\cdot | 1, 1) = \begin{bmatrix} \tfrac{3}{2}(1 - \gamma)\mu_0, 1 - \tfrac{3}{2}(1 - \gamma)\mu_0, 0, 0 \end{bmatrix}, \quad r(1, 1) = 1, \quad (234c) \)

- \( P(\cdot | 1, 2) = \begin{bmatrix} \tfrac{3}{2}(1 - \gamma)\mu_0, 1 - \tfrac{3}{2}(1 - \gamma)\mu_0, 0, 0 \end{bmatrix}, \quad r(1, 2) = 1, \quad (234d) \)

- \( P(\cdot | 1, 3) = \begin{bmatrix} 0, 1 - 3(1 - \gamma)(\mu_2 + \mu_3), 3(1 - \gamma)\mu_2, 3(1 - \gamma)\mu_3 \end{bmatrix}, \quad r(1, 3) = 0, \quad (234e) \)

- \( \mathcal{A}_2 = \{1, 2\}, \quad P(\cdot | 2, 1) = \begin{bmatrix} 2(1 - \gamma)\mu_0, 0, 1 - 2(1 - \gamma)\mu_0, 0 \end{bmatrix}, \quad r(2, 1) = 1, \quad (234f) \)

- \( P(\cdot | 2, 2) = \begin{bmatrix} 0, 0, 0, 0 \end{bmatrix}, \quad r(2, 2) = 0, \quad (234g) \)

- \( \mathcal{A}_3 = \{1, 2\}, \quad P(\cdot | 3, 1) = \begin{bmatrix} 0, 0, 0, 1 \end{bmatrix}, \quad r(3, 1) = \frac{3}{4}, \quad (234h) \)
\[ P(3, 2) = [2(1 - \gamma)\mu_0, 2(1 - \gamma)\mu_1, 2(1 - \gamma)\mu_2, 1 - 2(1 - \gamma)(1 - \mu_3)], \quad r(3, 2) = 0, \quad (234i) \]

where the parameter \( \mu = [\mu_0, \mu_1, \mu_2, \mu_3] \in \Delta(S) \) is set as

\[
\mu = \left[ \frac{2}{5}, \frac{2}{5} - \frac{c_\mu}{\log^2 T}, \frac{1}{5}, \frac{c_\mu}{\log^2 T} \right], \quad (235) \]

for some sufficiently small quantity \( 0 < c_\mu = O(1) \). In particular, actions 1 and 2 from state 1 are identical, which will play a similar critical role as the case of synchronous Q-learning in pinpointing the “over-estimation” issue.

In addition, let the behavior policy \( \pi_b \) be uniform distributions such that

\[
\pi_b(a | s) = \frac{1}{|A_s|}, \quad \text{for all } s \in S \text{ and all } a \in A_s. \]

Then the transition probability from state \( s \) to state \( s' \) under the behavior policy \( \pi_b \) is given by

\[
P^{\pi_b}(s' | s) = \sum_{a \in A_s} \pi_b(a | s) \ P(s' | s, a) = \gamma I(s' = s) + (1 - \gamma)\mu(s'). \]

With this in mind, it can be easily verified that the stationary distribution under \( \pi_b \) is given by

\[
\mu_b(s, a) = \frac{1}{|A_s|} \mu(s), \quad \text{for all } s \in S \text{ and all } a \in A_s. \]

This together with (235) implies that

\[
\mu_{\min} = \frac{c_\mu}{3 \log^2 T}. \quad (236) \]

Moreover, if the sample trajectory is initialized with an initial state distribution \( \mu_0 \), then the marginal state distribution at time \( t \) can be calculated as

\[
\mu_t = (P^{\pi_b})^t \mu_0 = (\gamma I + (1 - \gamma)\mu 1^T)^t \mu_0 = \gamma^t \mu_0 + (1 - \gamma^t)\mu, \]

thus indicating that the total variation distance \( d_{TV} \) between \( \mu_t \) and \( \mu \) obeys

\[
d_{TV}(\mu_t, \mu) = \frac{1}{2} \| \mu_t - \mu \|_1 = \frac{1}{2} \gamma^t \| \mu_0 - \mu \|_1. \]

Consequently, the mixing time of the sample trajectory (Paulin, 2015) obeys

\[
t_{\text{mix}} \in \left[ \frac{c_{\text{mix}, 1}}{1 - \gamma}, \frac{c_{\text{mix}, 2}}{1 - \gamma} \right] \quad (237) \]

for some universal constants \( c_{\text{mix}, 1}, c_{\text{mix}, 2} > 0 \).

Before embarking on the analysis of the behavior of asynchronous Q-learning, let us first look at the optimal value function and Q-function of the constructed MDP.

**Lemma 10.** Consider the MDP \( \mathcal{M}_{\text{hard}} \) as constructed in (234). It holds that

\[
V^*(0) = Q^*(0, 1) = \frac{2}{3(1 - \gamma)}; \quad (238a) \\
Q^*(0, 2) < V^*(0) - \frac{4}{15}; \quad (238b) \\
V^*(1) = Q^*(1, 1) = Q^*(1, 2) = \frac{5 + 2\gamma}{(5 + 3\gamma)(1 - \gamma)}; \quad (238c) \\
Q^*(1, 3) < V^*(1) - \frac{7}{8}; \quad (238d) 
\]
We now proceed to analyze the dynamics of asynchronous Q-learning when applied to the above MDP. In order to justify (238), let us begin by defining two vectors $V$ and $Q$ as follows:

$$V = \left[ \frac{2}{3(1-\gamma)^2}, \frac{5+2\gamma}{(5+3\gamma)(1-\gamma)^3}, \frac{15+8\gamma}{3(5+4\gamma)(1-\gamma)}, \frac{3}{4(1-\gamma)} \right]^\top,$$

$$Q = r + \gamma PV,$$

where $r = [r(s,a)]$ denotes the reward vector. Then the claimed expressions of the value function in (238) are valid as long as we can validate $V(s) = \max_a Q(s,a)$ (namely, they satisfy the Bellman equation). These are elementary calculations, which we omit for brevity. Once the expressions of both the value function and the Q-function are settled, the remaining set of advertised inequalities can be validated straightforwardly, which is omitted as well for conciseness.

### F.2 Analysis for the constructed MDP

We now proceed to analyze the dynamics of asynchronous Q-learning when applied to the above MDP instance, for which we divide into three cases based on the magnitudes of the learning rates. Throughout the proof, we denote by $t_k(s,a)$ the iteration number corresponding to the $k$-th time the state-action pair $(s,a)$ is visited, and let $N_T(s,a)$ represent the total number of visits to $(s,a)$ up to time $T$. We shall also reuse the notation of the sample transition matrix $P_t$ defined in (200), as well as the value estimate vector $V_t = [V_t(s)]_{s \in S}$ as in Appendix E.1.

#### F.2.1 Case 1: small learning rates ($\eta < \frac{1}{\mu_{\min}(1-\gamma)^2}$)

In this case, we focus attention on analyzing state 3. To begin with, we claim that if $\eta \leq c_8 \frac{(1-\gamma)^2}{\log T}$ for some sufficiently small constant $c_8 > 0$, then one has, for all $t \leq T$,

$$Q_t(3,1) \leq V^*(3) \quad \text{and} \quad Q_t(3,2) < V^*(3) - \frac{1}{6}. \quad (239)$$

Given the assumption $\mu_{\min} T \geq c_9 \log T$ and $\eta < \frac{1}{\mu_{\min}(1-\gamma)^2}$, the regime $\eta < \frac{1}{\mu_{\min}(1-\gamma)^2}$ clearly satisfies $\eta \leq c_8 \frac{(1-\gamma)^2}{\log T}$.

**Proof of Claim (239).** We would like to prove the second inequality in (239) by induction. Clearly, it suffices to look at those iterations where the value of $Q_t(3,2)$ changes, namely, $\{t_k(3,2)\}_{k \geq 1}$. Assume for the moment that (239) holds for all $t < t_k(3,2)$. Taking $\vec{V} := \left[ \frac{1}{1-\gamma}, \frac{1}{1-\gamma}, \frac{1}{1-\gamma}, V^*(3) \right]^\top$, we can invoke the asynchronous Q-learning update rule iteratively to deduce that

$$Q_{t_k(3,2)}(3,2) = (1 - \eta)^k Q_0(3,2) + \sum_{i=1}^k \gamma(1 - \eta)^{k-i} P_{t_k(3,2)}((3,2), \cdot) V_{t_{i-1}}$$

$$\leq \sum_{i=1}^k \gamma(1 - \eta)^{k-i} P_{t_k(3,2)}((3,2), \cdot) \vec{V}$$

$$\leq \gamma P(\cdot | 3,2) \vec{V} + \sum_{i=1}^k \gamma(1 - \eta)^{k-i} [P_{t_k(3,2)}((3,2), \cdot) - P(\cdot | 3,2)] \vec{V}$$

$$< \gamma P(\cdot | 3,2) \vec{V} + \frac{1}{12}$$
\[\sum_{i=1}^{k} \eta(1-\eta)^{k-i} \left[ P_{t_i(3,2)}((3,2), \cdot) - P(\cdot | 3,2) \right] V \leq \frac{1}{12}\]

holds with probability at least \(1 - 1/T\); and the validity of the last inequality can be shown by observing that

\[
\gamma P(\cdot | 3,2) V = \gamma P(3 | 3,2) V(3) + \frac{\gamma}{1 - \gamma} (1 - P(3 | 3,2))
\]

\[
= \gamma \left( 1 - 2(1 - \gamma) \right) V(3) + \gamma (2(1 - \gamma) \mu_3) V(3) + \frac{\gamma}{1 - \gamma} (2(1 - \gamma) \mu_3)
\]

\[
= \gamma \left( 1 - 2(1 - \gamma) \right) V(3) + 3.5 \gamma \mu_3 \leq \gamma (1 - 2(1 - \gamma)) V(3) + 2 \gamma
\]

\[
= V(3) - \frac{3}{4} (2 \gamma + 1) + 2 \gamma \leq V(3) - \frac{1}{4},
\]

where we have used the facts that \(V(3) = V^*(3) = \frac{3}{4(1 - \gamma)}\), \(P(3 | 3,2) = 1 - 2(1 - \gamma)(1 - \mu_3)\) and \(\gamma < 1\). Thus, standard induction arguments immediately validate the second inequality in the claim (239) for all \(t \leq T\).

Regarding the first inequality of (239), it is seen that for any \(k \geq 1\),

\[
Q_{t_k(3,1)}(3,1) = (1 - \eta)^k Q_0(3,1) + \sum_{i=1}^{k} \eta(1-\eta)^{k-i} \left( r(3,1) + \gamma P_{t_i(3,1)}((3,1), \cdot) V_{t_i(3,1)-1} \right)
\]

\[
\leq \sum_{i=1}^{k} \eta(1-\eta)^{k-i} (r(3,1) + \gamma V(3)) \leq V^*(3),
\]

where we have used the facts that \(Q_0 \equiv 0\) and \(V(3) = V^*(3) = \frac{3}{4(1 - \gamma)}\) and the elementary inequality \(\sum_{i=1}^{k} \eta(1-\eta)^{k-i} \leq 1\). Given that the above bound holds for all \(k \geq 1\) (and \(\{t_k(3,1)\}_{k \geq 1}\) correspond to all iterations when the value of \(Q_t(3,1)\) changes), we have established the first advertised inequality in (239).

Next, let us define

\[
s := \max \left\{ k : Q_{t_k(3,1)}(3,1) < V^*(3) - \frac{1}{8} \right\}.
\]

From this definition and Claim (239), we know that for any \(k > s\), it holds that

\[
Q_{t_k(3,1)}(3,1) \geq V^*(3) - \frac{1}{8}.
\]

Recognizing that

\[
|Q_{t_k(s,a)}(s,a) - Q_{t_{k-1}(s,a)}(s,a)| = \eta |Q_{t_k(s,a)}(s,a) - r(s,a) - \gamma P_{t_k(s,a)}((s,a), \cdot) V_{t_{k-1}(s,a)}| \leq \frac{\eta}{1 - \gamma} < \frac{1}{24},
\]

we can readily obtain

\[
Q_{t_{k-1}(3,1)}(3,1) \geq Q_{t_k(3,1)}(3,1) - \frac{1}{24} \geq V^*(3) - \frac{1}{6} > Q_{t_{k-1}(3,1)}(3,2),
\]

\[
\Rightarrow V_{t_{k-1}(3,1)}(3) = Q_{t_{k-1}(3,1)}(3,1), \quad \forall k > s.
\]

Combine this with the Q-learning update rule to arrive at

\[
Q_{t_k(3,1)}(3,1) = (1 - \eta)Q_{t_{k-1}(3,1)}(3,1) + \eta (r(3,1) + \gamma Q_{t_{k-1}(3,1)}(3,1))
\]

\[\leq V^*(3) - \frac{1}{8}.
\]

Here, the second line holds since \(V\) serves as an entrywise upper bound on \(V_t\); the third line follows since \(\sum_{i=1}^{k} \eta(1-\eta)^{k-i} \leq 1\); the fourth line in (240) is valid since, according to the Bernstein inequality (see Li et al. (2022c, Lemma 1)),

\[\left| \sum_{i=1}^{k} \eta(1-\eta)^{k-i} [P_{t_i(3,2)}((3,2), \cdot) - P(\cdot | 3,2)] V \right| \leq \frac{1}{12}\]
where the last inequality holds since $V^*(3) = \frac{3}{4(1-\gamma)}$. This taken together with (241) (so that $V^*(3) - Q_{t,e(3,1)}(3,1) \geq 1/8$) leads to

$$V^*(3) - Q_T(3,1) = Q^*(3,1) - Q_T(3,1) \geq \frac{1}{8}(1 - \eta(1 - \gamma)) N_T(3,1) \geq c_9$$

for some constant $c_9 > 0$; here, the last inequality holds since, according to Li et al. (2022c, Lemma 8),

$$N_T(3,1) \leq \frac{3}{2} \mu_b(3,1) T = \frac{9}{4} \mu_{\text{min}} T \leq O\left(\frac{1}{\eta(1 - \gamma)}\right)$$

occurs with probability at least $1 - 1/T$, provided that $T \geq \frac{443 \mu_{\text{max}} \log(10T)}{\mu_{\text{min}}^2}$. Note that according to (236) and (237), \(\frac{\mu_{\text{max}}}{\mu_{\text{min}}}\) is on the order of \(\frac{\log^2 T}{\mu_{\text{min}}}\).

Putting (239) and (242) together then reveals that with probability at least $1 - 2/T$,

$$\max_a |Q^*(3, a) - Q_T(3, a)| \geq V^*(3) - V_T(3) = V^*(3) - \max \{Q_T(3,1), Q_T(3,2)\} \geq \min \left\{\frac{1}{6}, c_9\right\}.$$  

(243)

**F.2.2 Case 2: large learning rates (\(\eta > \frac{\log T}{\mu_{\text{min}}(1 - \gamma)^2 T}\))**

**Case 2.1: \(\eta \geq \frac{cs(1 - \gamma)^2}{\log T}\) for some small enough constant \(c_8 > 0\).** Under the condition that \(\eta \geq \frac{cs(1 - \gamma)^2}{\log T}\), we claim that with probability at least \(1 - \gamma\),

$$\exists (s, a) \quad \text{s.t.} \quad |Q_T(s, a) - Q^*(s, a)| \geq c_{10} \frac{1 - \gamma}{\log T}$$

(244)

for some universal constant \(c_{10} > 0\). We shall first prove this claim.

**Proof of Claim (244).** We shall focus attention on the case where \((s_{T-1}, a_{T-1}) = (2, 1)\). Given that the stationary distribution obeys $\mu_b(2,1) = 1/10$ and that $T$ is sufficiently large (so that the empirical distribution approaches the stationary distribution), we know that

$$P\{(s_{T-1}, a_{T-1}) = (2, 1)\} \geq \mu_b(2,1)/2 = 1/20.$$  

Let us first look at the case where $|V_{T-1}(0) - V^*(0)| > \frac{1}{2\gamma(1-\gamma)}$. It follows from $P(-|0,1] = [1,0,0,0]$ that

$$|Q_T(0,1) - Q^*(0,1)| = |r(0,1) + \gamma V_{T-1}(0) - V^*(0)| = \gamma |V_{T-1}(0) - V^*(0)| > \frac{0.03}{1 - \gamma}$$

as long as $\gamma \geq 0.95$, which clearly satisfies (244). When it comes to the complement case where $|V_{T-1}(0) - V^*(0)| \leq \frac{1}{2\gamma(1-\gamma)}$, either of the following two scenarios will happen:

- If $|V_{T-1}(2) - V_{T-1}(0)| \leq \frac{1}{27(1-\gamma)}$, then the assumption $|V_{T-1}(0) - V^*(0)| \leq \frac{1}{27(1-\gamma)}$ yields

$$Q_T(2,2) \leq V_{T-1}(2) \leq V_{T-1}(0) + \frac{1}{27(1-\gamma)} \leq V^*(0) + \frac{2}{27(1-\gamma)} \leq V^*(2) - \frac{1}{27(1-\gamma)} < Q^*(2,2) - \frac{1}{27(1-\gamma)}.$$  

(245)
where the second line holds since, for $\gamma \geq 0.95$,

$$
\gamma \left[1 - 2(1 - \gamma)(\mu_1 + \mu_3)\right] V^*(2) = \gamma \left[1 - \frac{4}{5} (1 - \gamma)\right] \frac{15 + 8\gamma}{3(5 + 4\gamma)(1 - \gamma)} \geq \frac{7}{9(1 - \gamma)} = V^*(0) + \frac{1}{9(1 - \gamma)}.
$$

and the last inequality holds since $Q^*(2, 2) > r(2, 2) + \gamma P(2 | 2, 2)V^*(2)$ (from the Bellman equation).

- Consider instead the scenario with $|V_{T-1}(2) - V_{T-1}(0)| > \frac{1}{2\tau(1 - \gamma)}$. For notational convenience, define

$$
Q_T^*(2, 1) := (1 - \eta)Q_{T-1}(2, 1) + \eta (r(2, 1) + \gamma V_{T-1}(1)), \quad \forall s \in S.
$$

Recognizing that $\min \{ P(0 | 2, 1), P(2 | 2, 1) \} \geq 2(1 - \gamma)\mu_0$, we can show that with probability at least $(1 - \gamma)\mu_0 = \frac{2(1 - \gamma)}{5}$,

$$
|Q_T(2, 1) - Q^*(2, 1)| = \max \left\{ |Q_T^0(2, 1) - Q^*(2, 1)|, |Q_T^2(2, 1) - Q^*(2, 1)| \right\} \geq \frac{1}{2} |Q_T^0(2, 1) - Q_T^2(2, 1)| = \frac{1}{2} \eta \gamma |V_{T-1}(0) - V_{T-1}(2)| \geq \frac{\eta \gamma}{54(1 - \gamma)} \geq \frac{0.95 \epsilon_0 (1 - \gamma)}{54 \log T},
$$

where the last inequality holds since $\eta \geq \frac{c_1 (1 - \gamma)^2}{\log T}$ and $\gamma \geq 0.95$.

We can thus conclude that for all the above scenarios, the claim (244) holds with probability at least $\frac{1 - \gamma}{50}$. \qed

**Case 2.2:** $\frac{\log T}{\mu_{\min}(1 - \gamma)^2 T} < \eta \leq \frac{c_1 (1 - \gamma)^2}{\log T}$. Recall from Claim (239) that: when $\eta \leq \frac{c_1 (1 - \gamma)^2}{\log T}$, one has

$$
Q_t(3, 1) \leq V^*(3) \quad \text{and} \quad Q_t(3, 2) \leq V^*(3) - \frac{1}{6}, \quad \forall t \leq T. \quad (246)
$$

In addition, for any $k \geq \frac{\log T}{\eta (1 - \gamma)}$, we can use $P(\cdot | 3, 1) = [0, 0, 0, 1]$ and the Bellman equation to derive

$$
Q_{t_k(3,1)}(3, 1) := (1 - \eta)Q_{t_{k-1}(3, 1)}(3, 1) + \eta (r(3, 1) + \gamma V_{t_{k-1}(3, 1)}(3, 1)) \geq (1 - \eta)Q_{t_{k-1}(3, 1)}(3, 1) + \eta \left(\frac{3}{4} + \gamma Q_{t_{k-1}(3, 1)}(3, 1)\right) = (1 - \eta (1 - \gamma))Q_{t_{k-1}(3, 1)} + \frac{3}{4} \eta
$$

$$
= (1 - \eta (1 - \gamma))^k Q_0(3, 1) + \frac{3}{4} \sum_{i=1}^k \eta (1 - \eta (1 - \gamma))^{k-i} = \left[1 - (1 - \eta (1 - \gamma))^k\right] V^*(3) \geq V^*(3) - \frac{1}{T(1 - \gamma)},
$$

where the last line is valid since $V^*(3) = \frac{3}{4(1 - \gamma)}$ and $Q_0(3, 1) = 0$. It is also seen from Li et al. (2022c, Lemma 8) that, with probability at least $1 - 1/T$,

$$
(N_T/3)(3, 1) \geq \frac{1}{2} \mu_3(3, 1) T = \frac{3}{4} \mu_{\min} T \geq \frac{\log T}{\eta (1 - \gamma)}
$$

as long as $\eta \geq \frac{\log T}{\mu_{\min}(1 - \gamma)^2 T}$. The above two results taken collectively yield

$$
0 \leq V^*(3) - V_t(3) \leq V^*(3) - Q_t(3, 1) \leq \frac{1}{T(1 - \gamma)}, \quad \forall t \leq T. \quad (247)
$$
Next, we move on to analyze state $2$. Towards this end, let us define

$$s := \max \left\{ k : Q_{t_k(2,1)}(2,2) > Q_{t_k(2,1)}(2,1) - \frac{1}{2} \text{ or } |Q_{t_k(2,1)}(2,1) - V^*(2)| > \frac{1}{4} \right\}. \quad (248)$$

Note that when $\eta \leq \frac{\gamma (1-\gamma)^2}{\log 2}$, one has

$$|Q_{t_k(s,a)}(s,a) - Q_{t_k(s,a)}(s,a)| = \eta |Q_{t_k(s,a)}(s,a) - r(s,a) - \gamma P_{t_k(s,a)}((s,a),\cdot)V_{t_k(s,a)-1}| \leq \frac{\eta}{1-\gamma} < \frac{1}{4},$$

which combined with (248) implies that $Q_{t_k-1(2,1)}(2,2) < Q_{t_k-1(2,1)}(2,1)$ for any $k > s$ and hence

$$V_{t_{k-1}(2,1)}(2) = Q_{t_{k-1}(2,1)}(2,1) \quad \text{for } k > s. \quad (249)$$

This crucial identity together with the construction of $P(\cdot | 2,1)$ in turn allows one to derive, for any $k > s$,

$$Q_{t_k(2,1)}(2,1) = (1-\eta)Q_{t_{k-1}(2,1)}(2,1) + \eta \left( r(2,1) + \gamma P_{t_k(2,1)}((2,1),\cdot)V_{t_k(2,1)-1} \right)$$

$$= (1-\eta)(1-\gamma)(1+2\gamma\mu_0)Q_{t_{k-1}(2,1)}(2,1)$$

$$+ \eta \left( (1+2\gamma)(1-\gamma)\mu_0 V_{t_k(2,1)-1}(0) + \gamma \left( P_{t_k(2,1)}((2,1),\cdot) - P(\cdot | 2,1) \right)V_{t_k(2,1)-1} \right)$$

$$= (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s}Q_{t_{k-1}(2,1)}(2,1) + \eta \sum_{i=s+1}^{k} (1-\eta)(1-\gamma)(1-2\gamma\mu_0)^{k-i}$$

$$\cdot \left( (1+2\gamma)(1-\gamma)\mu_0 V_{t_i(2,1)-1}(0) + \gamma \left( P_{t_i(2,1)}((2,1),\cdot) - P(\cdot | 2,1) \right)V_{t_i(2,1)-1} \right), \quad (250)$$

thus leading to

$$\mathbb{E}[Q_{t_k(2,1)}(2,1) | Q_{t_s(2,1)}(2,1)] \leq (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s}Q_{t_{k-1}(2,1)}(2,1)$$

$$+ \eta \sum_{i=s+1}^{k} (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-i} (1-2\gamma)(1-\gamma)\mu_0 V^*(0))$$

$$\leq (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s}Q_{t_{k-1}(2,1)}(2,1)$$

$$+ \eta \sum_{i=s+1}^{k} (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-i} V^*(2)$$

$$= V^*(2) - (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s} \left( V^*(2) - Q_{t_{k-s}(2,1)}(2,1) \right)$$

$$\leq V^*(2) - \frac{1}{4} (1-\eta)(1+2\gamma\mu_0)^{k-s},$$

where the second inequality arises from the Bellman equation (so that $V^*(2) \geq 1+2\gamma(1-\gamma)\mu_0 V^*(0)$).

- If $(1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s} \geq \frac{1}{2}$, then we can readily see that

$$\mathbb{E}[|Q_T(2,1) - Q^*(2,1)|^2] \geq \left( \mathbb{E}[Q_{t_k(2,1)}(2,1) - V^*(2) | Q_{t_s(2,1)}(2,1)] \right)^2 \geq \frac{1}{64}. \quad (251)$$

- Otherwise, consider the case where $(1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{k-s} < \frac{1}{2}$. The expression (250) allows one to control the variance as follows:

$$\text{Var}(Q_{t_k(2,1)}(2,1) | Q_{t_s(2,1)}(2,1)) = \eta^2 \gamma^2 \sum_{i=s+1}^{k} (1-\eta)(1-\gamma)(1+2\gamma\mu_0)^{2(k-i)}$$

$$\cdot \mathbb{E} \left[ \left( \left( P_{t_i(2,1)}((2,1),\cdot) - P(\cdot | 2,1) \right)V_{t_i(2,1)-1} \right)^2 | V_{t_i(2,1)-1} \right]$$

66
\[= \eta^2 \gamma^2 \sum_{i=s+1}^{k} (1 - \eta(1 - \gamma)(1 + 2\gamma\mu_0))^{2(k-i)}\]
\[\cdot \mathbb{E}\left[\mathbb{E}\left[2(1 - \gamma)\mu_0 (1 - 2(1 - \gamma)\mu_0) \left\{V_{t_i,(2,1)-1}(2) - V_{t_i,(2,1)-1}(0)\right\}^2 \mid V_{t_i,(2,1)-1}\right]\right]\]
\[\geq \eta^2 \gamma^2 \sum_{i=s+1}^{k} (1 - \eta(1 - \gamma)(1 + 2\gamma\mu_0))^{2(k-i)} \left\{(1 - \gamma)\mu_0 \cdot \frac{1}{36(1 - \gamma)^2}\right\}\]
\[= \frac{\eta^2 \gamma^2 \mu_0}{36(1 - \gamma)} \left(1 - (1 - \eta(1 - \gamma)(1 + 2\gamma\mu_0))^{k-s}\right) \geq \frac{\eta}{200(1 + 2\gamma\mu_0)(1 - \gamma)^2} \geq \frac{\eta}{400(1 - \gamma)^2},\]

where the third inequality holds since \(V_{t_i,(2,1)-1}(2) - V_{t_i,(2,1)-1}(0) \geq V^*(2) - \frac{1}{4} - V^*(0) > \frac{5}{27(1 - \gamma)}\) (using the definition of \(s\) in (248)), and the last line uses \(\gamma \geq 0.95\) and \(\mu_0 = 2/5\). As a result,

\[\mathbb{E}\left[|Q_T(2,1) - Q^*(2,1)|^2\right] \geq \mathbb{E}\left[\text{Var}(Q_{t_k,(2,1)}(2,1) \mid Q_{t_i,(2,1)}(2,1))\right] \geq \frac{\eta}{400(1 - \gamma)^2}.\]

Taking the above two results together reveals that

\[\mathbb{E}\left[|Q_T(2,1) - Q^*(2,1)|^2\right] \geq \min\left\{\frac{\eta}{400(1 - \gamma)^2}, \frac{1}{64}\right\} \geq \min\left\{\frac{\log T}{400\mu_{\text{min}}(1 - \gamma)^4 T}, \frac{1}{64}\right\}.\]

Combining Case 2.1 and Case 2.2. Putting together (244) and (253) together leads to

\[\max_{s,a} \mathbb{E}\left[|Q_T(s,a) - Q^*(s,a)|^2\right] \geq \min\left\{\frac{\log T}{400\mu_{\text{min}}(1 - \gamma)^4 T}, \frac{1}{64}, \frac{c_{10}^2(1 - \gamma)^3}{50 \log^2 T}\right\},\]

provided that \(\eta > \frac{\log T}{\mu_{\text{min}}(1 - \gamma)^4 T}\).

### F.2.3 Case 3: medium learning rates \((\frac{1}{\mu_{\text{max}}(1 - \gamma)^2 T} \leq \eta \leq \frac{\log T}{\mu_{\text{min}}(1 - \gamma)^4 T})\)

We now shift attention to the dynamics underlying state 1, and look at its associated value function \(V_t(1)\).

**Two auxiliary sequences.** Before proceeding, we abuse notation by defining

\[t_i := \min\left\{t > t_{i-1} : \text{both pairs (1, 1) and (1, 2) have been visited during } (t_{i-1}, t)\right\},\]

and letting \(N_{t_{k-1},t_k}(1,a,1)\) (resp. \(N_{t_{k-1},t_k}(1,a)\)) denote the number of times the sample trajectory visits \((s, a, s') = (1, a, 1)\) (resp. \((s, a) = (1, a)\)) within the time interval \((t_{k-1}, t_k)\). From standard Bernstein’s inequality (see Li et al. (2022c, Lemma 8)) and the definition of \(t_i\), one can easily see that with probability at least \(1 - 1/T\),

\[1 \leq N_{t_{i-1},t_i}(1, a) \leq c_{11} \log T\]

holds simultaneously for all \(t_i \leq T\), where \(c_{11} > 0\) is some suitable constant. In turn, this bound (255) also implies that

\[K := \max\left\{k : t_k \leq T\right\} \geq c_{12} \frac{T}{\log T}\]

for some constant \(c_{12} > 0\). These follow from fairly standard concentration arguments and are hence omitted.

Under our assumption on \(T\), the sample trajectory mixes well after \(T/3\) iterations. In order to further remove the effect of \(Q_{T/3}(1,a)\), let us introduce the following auxiliary sequence

\[\hat{Q}_k(a) = (1 - \eta)\hat{Q}_{k-1}(a) + \eta \left\{\frac{1}{3} + \gamma P_k(a)\bar{V}_{k-1}\right\},\]
where \( \hat{V}_{k-1} := \max_a \hat{Q}_{k-1}(a) \),
\[
\hat{Q}_0(a) := Q^*(1, a) - V^*(0) = \frac{5}{3(5 + 3\gamma)(1 - \gamma)}, \quad a = 1, 2
\]
\[
P_k(a) := \frac{N_{[t_{k-1}, t_k]}(1, a, 1)}{N_{[t_{k-1}, t_k]}(1, a)}.
\]

Repeating the proof of (163) (which we omit here for brevity), we arrive at the following relation:
\[
\hat{Q}_{t_k(1, a)}(1, a) - V^*(0) \geq \hat{Q}_k(a) - \frac{1}{1 - \gamma} \left(1 - \eta(1 - \gamma)\right)^k - \frac{1}{T(1 - \gamma)^2}, \quad a \in \{1, 2\}. \tag{258}
\]

Furthermore, in order to control \( \hat{Q}_k(a) \), we construct an additional auxiliary sequence as follows
\[
\bar{Q}_k = (1 - \eta)\bar{Q}_{k-1} + \eta \left\{ \frac{1}{3} + \gamma P_k(1)\bar{Q}_{k-1} \right\} \quad \text{and} \quad \bar{Q}_0 = V^*(1) - V^*(0) = \frac{5}{3(5 + 3\gamma)(1 - \gamma)}. \tag{259}
\]

From the basic fact that \( \hat{V}_k = \max_a \hat{Q}_k(a) \geq \hat{Q}_k(1) \), it can be easily verified that
\[
\hat{Q}_k(1) \geq (1 - \eta)\hat{Q}_{k-1}(1) + \eta \left( \frac{1}{3} + \gamma P_k(1)\hat{Q}_{k-1}(1) \right) \geq \bar{Q}_k, \tag{260}
\]
which in turn motivates us to lower bound \( \hat{V}_k \) by controlling \( \bar{Q}_k \). Using similar analysis for (166), we reach
\[
\mathbb{P}\left\{ \hat{V}_k \geq \frac{1}{5(1 - \gamma)} \right\} \geq \frac{1}{2}, \quad \text{for any } k \text{ with } t_k \geq \frac{2T}{3}. \tag{261}
\]

**Main proof.** With the preceding auxiliary sequences in place, let us define (akin to the synchronous case)
\[
\Delta_k(a) := \hat{Q}_k(a) + V^*(0) - Q^*(1, a), \quad a = 1, 2; \tag{262a}
\]
\[
\Delta_{k, \text{max}} := \max_a \Delta_k(a). \tag{262b}
\]

Based on the iterative update rule (257), we can once again derive
\[
\Delta_k(a) = \sum_{i=1}^{k} \eta (1 - \eta)^{k-i} \gamma \left[ p \Delta_{k-1, \text{max}} + (P_k(a) - p)\hat{V}_{k-1} \right], \quad a = 1, 2, \tag{263}
\]

where \( p = 1 - \frac{\delta}{2}(1 - \gamma)\mu_0 \). Then adopting the same analysis as for the synchronous case, we arrive at
\[
\mathbb{E} \left[ \max_a |Q_T(1, a) - Q^*(1, a)| \right] \geq \mathbb{E} \left[ |V_T(1) - V^*(1)| \right] \geq \frac{c_{13}}{\sqrt{\mu_{\min}(1 - \gamma)^4 T \log^3 T}}, \tag{264}
\]
for some constant \( c_{13} > 0 \). Here, we have made use of the fact that
\[
(1 - \eta(1 - \gamma))^k \leq \left( 1 - \frac{3 \log^2 T}{c_\gamma T} \right)^{\frac{c_{12}^2}{\log T}} \leq \frac{1}{T},
\]
an immediate consequence of (256) and the assumption that \( \eta(1 - \gamma) \geq \frac{1}{\mu_{\min}} = \frac{3 \log^2 T}{c_\gamma T} \).

**F.2.4 Putting all this together**

Combining (243), (254), and (264) leads to
\[
\max_{s,a} \mathbb{E} \left[ |Q_T(s, a) - Q^*(s, a)|^2 \right] \geq \min \left\{ c_2^3, \frac{\log T}{400 \mu_{\min}(1 - \gamma)^4 T}, \frac{1}{64}, \frac{c_{10}^2 (1 - \gamma)^3}{\log^2 T}, \frac{c_{13}^2}{\mu_{\min}(1 - \gamma)^4 T \log^3 T} \right\}, \tag{265}
\]
for any \( 0 < \eta < 1 \). Then the conclusion is handy under the proviso that \( T \geq \frac{c_3^3}{\mu_{\min}(1 - \gamma)^4 \log T} \).
References


