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

Gen Li* Tsinghua EE
Changxiao Cai† Princeton ECE
Yuxin Chen† Princeton ECE
Yuantao Gu* Tsinghua EE
Yuting Wei‡ CMU Statistics
Yuejie Chi§ CMU ECE

February 2021; Revised: March 2021

Abstract

Q-learning, which seeks to learn the optimal Q-function of a Markov decision process (MDP) in a model-free fashion, lies at the heart of reinforcement learning. When it comes to the synchronous setting (such that independent samples for all state-action pairs are drawn from a generative model in each iteration), substantial progress has been made recently towards understanding the sample efficiency of Q-learning. Take a γ-discounted infinite-horizon MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$: to yield an entrywise ε-accurate estimate of the optimal Q-function, state-of-the-art theory for Q-learning proves that a sample size on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ is sufficient, which, however, fails to match with the existing minimax lower bound. This gives rise to natural questions: what is the sharp sample complexity of Q-learning? Is Q-learning provably sub-optimal? In this work, we settle these questions by (1) demonstrating that the sample complexity of Q-learning is at most on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ (up to some log factor) for any $0 < \varepsilon < 1$, and (2) developing a matching lower bound to confirm the sharpness of our result. Our findings unveil both the effectiveness and limitation of Q-learning: its sample complexity matches that of speedy Q-learning without requiring extra computation and storage, albeit still being considerably higher than the minimax lower bound for problems with long effective horizon.

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

Contents

1 Introduction **1**
1.1 Main contributions . . . 2
1.2 Related work . . . 3

2 Background and algorithm **4**

3 Main results **5**
3.1 Performance guarantees: achievability . . . 5
3.2 A lower bound for Q-learning . . . 6

4 Key analysis ideas **7**
4.1 Vector and matrix notation . . . 7
4.2 Proof outline for Theorem 1 . . . 8
4.3 Proof outline for Theorem 2 . . . 10

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

5 Concluding remarks **11**

A Upper bounds of Q-learning (Theorem 1) **11**
A.1 Preliminaries . . . 12
A.2 Proof of Lemma 1 . . . 13
A.3 Proof of Lemma 2 . . . 18
A.4 Solving the recurrence relation regarding $\mathbf{\Delta}_t$ . . . 18
A.5 Proof of Lemma 5 . . . 20

B Lower bound: sub-optimality of vanilla Q-learning **22**
B.1 Key quantities related to learning rates . . . 22
B.2 Preliminary calculations . . . 23
B.3 Lower bounds for three cases . . . 25
B.4 Proof of Lemma 3 . . . 38

C Freedman's inequality **39**

1 Introduction

Characterizing the sample efficiency of Q-learning [Watkins and Dayan, 1992, Watkins, 1989] — which is arguably one of the most widely adopted model-free algorithms — 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] have been primarily focused 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., 2020b, Qu and Wierman, 2020, Wainwright, 2019b, Weng et al., 2020a, Xiong et al., 2020]. Drawing on insights from high-dimensional statistics [Wainwright, 2019a], such a modern non-asymptotic framework unveils more clear and informative effect of salient problem parameters upon the sample complexity, particularly for those applications with enormous state/action space and long horizon. In fact, non-asymptotic theory has been developed for Q-learning under multiple sampling mechanisms [Beck and Srikant, 2012, Even-Dar and Mansour, 2003, Jin et al., 2018, Li et al., 2020b, Qu and Wierman, 2020, Wainwright, 2019b].

In this paper, we revisit the sample complexity of Q-learning for tabular Markov decision processes (MDPs), assuming access to a generative model or a simulator that produces independent samples for all state-action pairs in each iteration [Kakade, 2003, Kearns et al., 2002]. This is often referred to as the synchronous setting. Our focal point is the ℓ_∞-based sample complexity, namely, the number of samples needed for synchronous Q-learning to yield an entrywise ε-accurate estimate of the optimal Q-function. Despite a number of prior works tackling this setting, the dependence of the sample complexity on the effective horizon of the MDP remains unsettled. Take γ-discounted infinite-horizon MDPs for instance: the state-of-the-art sample complexity bound [Chen et al., 2020, Wainwright, 2019b] scales on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ (up to some log factor), where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and the action space, respectively. However, it is unclear whether this scaling is sharp for Q-learning or improvable via a more refined theory. On the one hand, the minimax lower limit for this setting has been shown to be on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ (up to some log factor) [Azar et al., 2013]. This limit is provably achievable by model-based approaches [Agarwal et al., 2020, Li et al., 2020a], and apparently better than prior Q-learning theory when the effective horizon $\frac{1}{1-\gamma}$ is large. On the other hand, Wainwright [2019b] conjectured the non-sharpness of existing Q-learning theory via numerical examples, although no rigorous analysis has been provided. Given the gap between the achievability bounds and lower bounds in the status quo, it is natural to investigate the following questions:

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

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

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

1.1 Main contributions

Focusing on γ-discounted infinite-horizon MDPs with state space $\mathcal{S}$ and action space $\mathcal{A}$, the current paper develops a refined theoretical framework that establishes the ℓ_∞-based sample complexity of Q-learning in a sharp manner. Here and throughout, the notation $\widetilde{O}(\cdot)$ hides logarithmic factors when describing the orderwise scaling.

- For any $0<\varepsilon<1$, we demonstrate that a total number of
$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) \tag{1}$$
samples are sufficient for Q-learning to guarantee ε-accuracy in an ℓ_∞ sense; see Theorem 1. This finding improves upon prior theory [Chen et al., 2020, Wainwright, 2019b] by a factor of $\frac{1}{1-\gamma}$.

- Conversely, we construct an MDP instance with 4 states and 2 actions, for which Q-learning provably requires at least
$$\widetilde{\Omega}\Big(\frac{1}{(1-\gamma)^4\varepsilon^2}\Big) \tag{2}$$
samples to achieve ε-accuracy in an ℓ_∞ sense; see Theorem 2.

Our results accommodate both rescaled linear and constant learning rates; see Table 1 for more detailed comparisons with previous literature. In sum, our theoretical guarantees provide a decisive answer to the sample complexity of Q-learning, which is on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ (up to some logarithmic factor). The gap between this scaling and the minimax lower bound makes clear that Q-learning is *not* minimax optimal, and is outperformed by the model-based approaches [Agarwal et al., 2020, Li et al., 2020a] in terms of the sample efficiency. As a worthy note, our work also puts the conjecture in Wainwright [2019c] on a rigorous footing, which, through numerical experiments, suggests "*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* ... ". Indeed, our construction of the hard MDP instance is inspired by Azar et al. [2013], Wainwright [2019c], with some

simplifications and modifications to facilitate analysis. In addition, it is also worth emphasizing that our sample complexity bound matches the theory for speedy Q-learning [Azar et al., 2011a] without requiring extra computation and storage. Our analysis framework uncovers a sort of crucial error decompositions and recursions that are previously unexplored, which might shed light on how to pin down the finite-sample efficacy of other variants of Q-learning like asynchronous Q-learning and double Q-learning.

1.2 Related work

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

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

Finite-sample ℓ_∞ guarantees for asynchronous Q-learning. Moving beyond the synchronous model considered herein, Beck and Srikant [2012], Chen et al. [2021], Even-Dar and Mansour [2003], Li et al. [2020b], Qu and Wierman [2020] developed non-asymptotic convergence guarantees for the asynchronous setting, where the data samples take the form of a single Markovian trajectory (following some behavior policy) and only a single state-action pair is updated in each iteration. A similar scaling of $\widetilde{O}\big(\frac{1}{(1-\gamma)^5\varepsilon^2}\big)$ — in addition to dependency on other salient parameters of the sample trajectory — also showed up in the state-of-the-art sample complexity bound for asynchronous Q-learning [Li et al., 2020b]. The analysis framework developed in this paper might have implications on how to sharpen the dependency of sample complexity on $\frac{1}{1-\gamma}$ for asynchronous Q-learning.

Finite-sample ℓ_∞ guarantees of other Q-learning variants. With the aim of alleviating the suboptimal 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. [2011b] proposed speedy Q-learning, which achieves a sample complexity of $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\big)$ at the expense of doubling the computation and storage complexity. Our result on vanilla Q-learning matches that of speedy Q-learning in an order-wise sense. In addition, Wainwright [2019c] proposed a variance-reduced Q-learning algorithm that is shown to be minimax optimal in the range $\epsilon \in (0, 1)$ with a sample complexity $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\big)$, which was subsequently generalized to the asynchronous setting by Li et al. [2020b]. The ℓ_∞ statistical bounds for variance-reduced TD learning have been investigated in Khamaru et al. [2020], Mou et al. [2020]. 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. In addition, another strand of recent work [Jin et al., 2018, Wang et al., 2020] considered the sample efficiency of Q-learning type algorithms paired with proper exploration strategies (e.g., UCB) under the framework of regret analysis, which is beyond the reach of the current paper.

Others. There are also several other streams of related papers that tackle model-free algorithms but do not pursue ℓ_∞-based non-asymptotic guarantees. For instance, Bhandari et al. [2018], Chen et al. [2019], Doan et al. [2019], Gupta et al. [2019], Srikant and Ying [2019], Wu et al. [2020], Xu et al. [2019a,b] developed finite-sample (weighted) ℓ_2 convergence guarantees for several model-free algorithms, accommodating linear

function approximation as well as off-policy evaluation. Another line of work investigated the asymptotic behavior of some variants of Q-learning, e.g., double Q-learning [Weng et al., 2020b] and relative Q-learning [Devraj and Meyn, 2020]. The effect of more general function approximation schemes (e.g., certain families of neural network approximations) has been studied in Cai et al. [2019], Fan et al. [2019], Murphy [2005], Wai et al. [2019], Xu and Gu [2020] as well. These are beyond the scope of the present paper.

2 Background and algorithm

This paper concentrates on tabular MDPs in the discounted infinite-horizon setting [Bertsekas, 2017]. Denote by $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$ and $\mathcal{A} = \{1, \cdots, |\mathcal{A}|\}$ the state space and the action space of the MDP, respectively. Here and throughout, $\Delta(\mathcal{S})$ stands for the probability simplex over the state space $\mathcal{S}$.

Discounted infinite-horizon MDPs. Consider an infinite-horizon MDP as represented by a quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\gamma \in (0,1)$ indicates the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the probability transition kernel (i.e., $P(s' \,|\, s, a)$ is the probability of transiting to state s' from a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$), and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ stands for the reward function (i.e., $r(s,a)$ is the immediate reward collected in state $s \in \mathcal{S}$ when action $a \in \mathcal{A}$ is taken). Note that the immediate rewards are assumed to lie within $[0,1]$ throughout this paper.

Value function and Q-function. A common objective in RL is to maximize a sort of long-term rewards called value functions or Q-functions. Specifically, given a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ (so that $\pi(s) \in \mathcal{A}$ specifies which action to select in state s), the associated value function and Q-function of π are defined respectively by

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

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

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

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

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

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Q-learning. In this work, we assume access to a generative model [Kearns and Singh, 1999, Sidford et al., 2018]: in each iteration t, we collect an independent sample $s_t(s,a) \sim P(\cdot \,|\, s,a)$ for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. The synchronous Q-learning algorithm maintains a Q-function estimate $Q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all $t \geq 0$; in each iteration t, the algorithm updates *all* entries of the Q-function estimate at once via the following update rule

$$Q_t = (1 - \eta_t) Q_{t-1} + \eta_t \mathcal{T}_t(Q_{t-1}). \tag{3}$$

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

$$\mathcal{T}_t(Q)(s,a) \coloneqq r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'), \qquad s_t \equiv s_t(s,a) \sim P(\cdot \,|\, s,a) \tag{4}$$

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

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

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

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

Noteworthily, the optimal Q-function $Q^\star$ is the unique fixed point of the Bellman operator [Bellman, 1952], that is, $\mathcal{T}(Q^\star) = Q^\star$. Viewed in this light, synchronous Q-learning can be interpreted as a stochastic approximation algorithm [Robbins and Monro, 1951] aimed at solving this fixed-point equation.

Throughout this work, we initialize the algorithm so that $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for every state-action pair (s,a). In addition, the corresponding value function estimate $V_t : \mathcal{S} \to \mathbb{R}$ in the t-th iteration is defined as

$$\forall s \in \mathcal{S}: \qquad V_t(s) := \max_{a \in \mathcal{A}} Q_t(s,a). \tag{5}$$

The complete algorithm is summarized in Algorithm 1.

3 Main results

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

3.1 Performance guarantees: achievability

We start by presenting our strengthened ℓ_∞-based sample complexity of Q-learning.

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

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}} \tag{6a}$$

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

$$T \ge \frac{c_3 \big(\log^4 T\big) \big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 \varepsilon^2} \tag{6b}$$

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

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

with probability at least $1 - \delta$.

Remark 1. This high-probability bound immediately translates to a mean estimation error guarantee. Recognizing the crude upper bound $\big|Q_T(s,a) - Q^\star(s,a)\big| \le \frac{1}{1-\gamma}$ (see (33) in Section A.1) and taking $\delta \le \varepsilon(1-\gamma)$, we reach

$$\mathbb{E}\Big[\max_{s,a} \big|Q_T(s,a) - Q^\star(s,a)\big| \Big] \le \varepsilon(1-\delta) + \delta \frac{1}{1-\gamma} \le 2\varepsilon, \tag{7}$$

provided that $T \ge \frac{c_3 (\log^4 T) \big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\varepsilon(1-\gamma)}\big)}{(1-\gamma)^4 \varepsilon^2}$.

Theorem 1 develops a non-asymptotic bound on the iteration complexity of Q-learning in the presence of a generative model. A few remarks and implications are in order.

Sample complexity and sharpened dependency on $\frac{1}{1-\gamma}$. Given that we draw $|\mathcal{S}||\mathcal{A}|$ independent samples in each iteration, the iteration complexity derived in Theorem 1 uncovers to the following sample complexity bound:

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) \tag{8}$$

in order for Q-learning to attain ε-accuracy ($0 < \epsilon < 1$) in an entrywise sense. To the best of our knowledge, this is the first result that breaks the $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ barrier that is present in all prior analyses for vanilla Q-learning [Beck and Srikant, 2012, Chen et al., 2020, Li et al., 2020b, Qu and Wierman, 2020, Wainwright, 2019b]. In addition, the dependence of our result on the effective horizon (i.e., $\frac{1}{(1-\gamma)^4}$) matches the scaling of a numerical example exhibited in Wainwright [2019b, Fig. 2], thus potentially corroborating its sharpness.

Learning rates. In view of the assumption (6a), our result accommodates two commonly adopted learning rate schemes: (i) linearly rescaled learning rates $\frac{1}{1+\frac{c_2(1-\gamma)}{\log^2 T}t}$, and (ii) iteration-invariant learning rates $\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}}$ (which depend on the total number of iterations T but not the iteration number t). In particular, when the sample size is $T = \frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4\varepsilon^2}$, the constant learning rates can be taken to be on the order of

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

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

3.2 A lower bound for Q-learning

The careful reader might remark that there remains a gap between our sample complexity bound for Q-learning and the minimax lower bound [Azar et al., 2013]. More specifically, the minimax lower bound scales on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ and is achievable — up to some logarithmic factor — by the model-based approach Agarwal et al. [2020], Azar et al. [2013], Li et al. [2020a]. We demonstrate the sharpness of our analysis by establishing the following lower bound of vanilla Q-learning. This result confirms the sub-optimality of vanilla Q-learning.

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

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

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

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

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

up to some logarithmic factor. This in turn confirms the sharpness of Theorem 1 in terms of its dependency on the effective horizon $\frac{1}{1-\gamma}$, which is larger than the minimax limit [Azar et al., 2013] by a factor of $\frac{1}{1-\gamma}$. Fortunately, while vanilla Q-learning falls short of achieving minimax optimality, the dependency of its sample complexity on the effective horizon $\frac{1}{1-\gamma}$ can be improved or optimized with the assistance of variance reduction; see Li et al. [2020b], Wainwright [2019c].

4 Key analysis ideas

This section outlines the key ideas for the establishment of our main results. Before delving into the proof details, we first introduce convenient vector and matrix notation that shall be used frequently.

4.1 Vector and matrix notation

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

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

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \qquad \text{and} \qquad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}, \tag{9}$$

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

$$\boldsymbol{\Pi}^\pi = \begin{pmatrix} \boldsymbol{e}_{\pi(1)}^\top & & & \\ & \boldsymbol{e}_{\pi(2)}^\top & & \\ & & \ddots & \\ & & & \boldsymbol{e}_{\pi(|\mathcal{S}|)}^\top \end{pmatrix} \tag{10}$$

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

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V})\circ(\boldsymbol{P}\boldsymbol{V}). \tag{11}$$

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

Moreover, we use the vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent the reward function r, so that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, the (s,a)-th entry of $\boldsymbol{r}$ is given by $r(s,a)$. Analogously, we shall employ the vectors $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent V^π, $V^\star$, V_t, Q^π, $Q^\star$ and Q_t, respectively. Additionally, we denote by π_t the policy such that for any state-action pair (s,a), $\pi_t(s) = \min\big\{a' \,|\, Q_t(s,a') = \max_{a''} Q_t(s,a'')\big\}$. In other words, for any $s \in \mathcal{S}$, the policy π_t picks out the smallest indexed action that attains the largest Q-value in the estimate $Q_t(s,\cdot)$. As an immediate consequence, one can easily verify

$$Q_t\big(s,\pi_t(s)\big) = V_t(s) \qquad \text{and} \qquad \boldsymbol{P}\boldsymbol{V}_t = \boldsymbol{P}^{\pi_t}\boldsymbol{Q}_t \ge \boldsymbol{P}^\pi \boldsymbol{Q}_t \tag{12}$$

for any π, where $\boldsymbol{P}^\pi$ is defined in (9). Further, we introduce a matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ such that

$$\boldsymbol{P}_t\big((s,a),s'\big) := \begin{cases} 1, & \text{if } s' = s_t(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

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

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

4.2 Proof outline for Theorem 1

We are now positioned to describe how to establish Theorem 1, towards which we first express the Q-learning update rule (3) and (4) using the above matrix notation. As can be easily verified, Q-learning employs the samples in $\boldsymbol{P}_t$ (cf. (13)) to perform the following update

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \tag{14}$$

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

$$\boldsymbol{\Delta}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star \tag{15}$$

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

4.2.1 Key decomposition

We start by decomposing the estimation error term $\boldsymbol{\Delta}_t$. In view of the update rule (14), we arrive at the following elementary decomposition:

$$\begin{aligned}
\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\
&= (1-\eta_t)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big(\boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\},
\end{aligned} \tag{16}$$

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

$$\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \le \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1}, \tag{17a}$$

$$\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \ge \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1}, \tag{17b}$$

where we have made use of the relation (12). Substitute (17) into (16) to reach

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}; \\
\boldsymbol{\Delta}_t &\ge (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}.
\end{aligned} \tag{18}$$

Applying these relations recursively, we obtain

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t} \eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\}, \\
\boldsymbol{\Delta}_t &\ge \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t} \eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\},
\end{aligned} \tag{19}$$

where we define

$$\eta_i^{(t)} \coloneqq \begin{cases} \prod_{j=1}^{t}(1-\eta_j), & \text{if } i = 0, \\ \eta_i \prod_{j=i+1}^{t}(1-\eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t. \end{cases} \tag{20}$$

Comparisons to prior approaches. We take a moment to discuss how prior analyses handle the above elementary decomposition. Several prior work (e.g., Li et al. [2020b], Wainwright [2019b]) tackled the second term on the right-hand side of the relation (18) via the following crude bounds:

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

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

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

By exploiting the crucial relations (19) derived above, we proceed to upper and lower bound $\boldsymbol{\Delta}_t$ separately. To be more specific, defining

$$\beta \coloneqq \frac{c_4(1-\gamma)}{\log T} \tag{21}$$

for some constant $c_4 > 0$, one can further decompose the upper bound in (19) into several terms:

$$\boldsymbol{\Delta}_t \le \underbrace{\eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{(1-\beta)t} \eta_i^{(t)}\gamma\big(\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big)}_{=:\,\boldsymbol{\zeta}_t} \tag{22}$$

$$+ \underbrace{\sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)}\gamma(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}}_{=:\,\boldsymbol{\xi}_t} + \sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)}\gamma\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}. \tag{23}$$

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

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

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

Lemma 1. *Suppose that $c_1c_2 \le c_4/4$. With probability at least $1-\delta$,*

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

holds simultaneously for all $t \ge \frac{T}{c_2\log\frac{1}{1-\gamma}}$.

Lemma 2. *Suppose that $c_1c_2 \le c_4/4$. With probability at least $1-\delta$,*

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

holds simultaneously for all $t \ge \frac{T}{c_2\log\frac{1}{1-\gamma}}$.

Proof. The proofs of Lemma 1 and Lemma 2 are deferred to Appendix A.2 and A.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 individual terms leveraging the structure of the discounted MDP. □

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

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\Big(1 + \max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)} \tag{24}$$

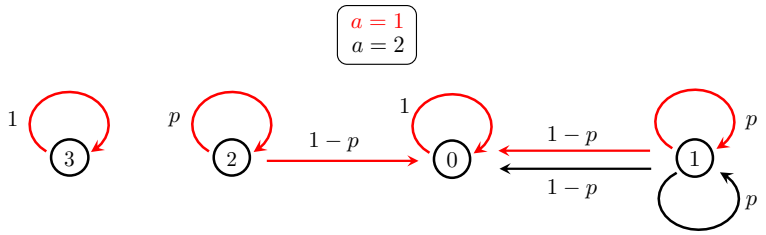

Figure 1: The constructed hard MDP instance used in the analysis of Theorem 2, where $p = \frac{4\gamma-1}{3\gamma}$ and the specifications are described in (26).

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

$$\|\boldsymbol{\Delta}_T\|_\infty \leq O\left(\sqrt{\frac{\left(\log^4 T\right)\left(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4 T}} + \frac{\left(\log^4 T\right)\left(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4 T}\right) \tag{25}$$

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 A.4.

4.3 Proof outline for Theorem 2

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

$$\mathcal{A}_0 = \{1\}, \qquad P(0\,|\,0,1) = 1, \qquad r(0,1) = 0, \tag{26a}$$

$$\mathcal{A}_1 = \{1,2\}, \qquad P(1\,|\,1,1) = p, \qquad P(0\,|\,1,1) = 1-p, \qquad r(1,1) = 1, \tag{26b}$$

$$P(1\,|\,1,2) = p, \qquad P(0\,|\,1,2) = 1-p, \qquad r(1,2) = 1, \tag{26c}$$

$$\mathcal{A}_2 = \{1\}, \qquad P(2\,|\,2,1) = p, \qquad P(0\,|\,2,1) = 1-p, \qquad r(2,1) = 1, \tag{26d}$$

$$\mathcal{A}_3 = \{1\}, \qquad P(3\,|\,3,1) = 1, \qquad r(3,1) = 1, \tag{26e}$$

where the parameter p is taken to be

$$p = \frac{4\gamma - 1}{3\gamma}. \tag{27}$$

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 B.4.

Lemma 3. *Consider the MDP $\mathcal{M}_{\mathsf{hard}}$ constructed in (26). One has*

$$V^\star(0) = Q^\star(0,1) = 0; \tag{28a}$$

$$V^\star(1) = Q^\star(1,1) = Q^\star(1,2) = V^\star(2) = Q^\star(2,1) = \frac{1}{1-\gamma p} = \frac{3}{4(1-\gamma)}; \tag{28b}$$

$$V^\star(3) = Q^\star(3,1) = \frac{1}{1-\gamma}. \tag{28c}$$

Recognizing the elementary decomposition

$$\mathbb{E}\left[\left(V^\star(s) - V_T(s)\right)^2\right] = \left(\mathbb{E}\left[V^\star(s) - V_T(s)\right]\right)^2 + \mathsf{Var}\left(V_T(s)\right) \tag{29}$$

for any state s, our proof consists of lower bounding either the squared bias term $\left(\mathbb{E}[V^\star(s) - V_T(s)]\right)^2$ or the variance term $\mathsf{Var}\left(V_T(s)\right)$. In short, we shall primarily analyze the dynamics w.r.t. state 2 to handle the case

when the learning rates are either too small or too large, and analyze the dynamics w.r.t. state 1 to cope with the case with medium learning rates (with state 3 serving as a helper state to simplify the analysis). The latter case — corresponding to the learning rates adopted in establishing the upper bounds — is the most challenging: critically, from state 1 the agent can take one of two identical actions, whose value tends to be estimated with a high positive bias due to maximizing over the empirical state-action values, highlighting the well-recognized "over-estimation" issue of Q-learning in practice [Hasselt, 2010]. The complete proof is deferred to Appendix B.

5 Concluding remarks

In this paper, we settle the sample complexity of vanilla Q-learning on an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ for the discounted infinite-horizon setting. Our theoretical bound is shown to be sharp through studying the dynamic of Q-learning on a hard MDP that highlights its over-estimation issue. Our analysis framework — which pinpoints novel error decompositions and recursion relations that are substantially different from prior approaches — might suggest a plausible path towards sharpening the sample complexity of, as well as understanding the algorithmic limits for, other variants of Q-learning (e.g., asynchronous Q-learning and double Q-learning).

Acknowledgements

G. Li and Y. Gu are supported in part by the grant NSFC-61971266. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, DMS-2014279 and IIS-1900140, and the Princeton SEAS Innovation Award. Y. Wei is supported in part by the NSF grants CCF-2007911 and DMS-2015447. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1806154 and CCF-2007911. 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.

A Upper bounds of Q-learning (Theorem 1)

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

Theorem 3. *Consider any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Theorem 1 continues to hold if*

$$T \geq \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 \min\{\varepsilon^2, \varepsilon\}} \tag{30}$$

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

Clearly, Theorem 3 subsumes Theorem 1 as a special case.

A.1 Preliminaries

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

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

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

$$\mathbf{0} \leq \boldsymbol{Q}_t \leq \frac{1}{1-\gamma}\mathbf{1} \qquad \text{and} \qquad \mathbf{0} \leq \boldsymbol{V}_t \leq \frac{1}{1-\gamma}\mathbf{1}. \tag{31}$$

Proof. We shall prove this by induction. First, our initialization trivially obeys (31) for $t = 0$. Next, suppose that (31) is true for the $(t-1)$-th iteration, namely,

$$\mathbf{0} \le \boldsymbol{Q}_{t-1} \le \frac{1}{1-\gamma}\mathbf{1} \qquad \text{and} \qquad \mathbf{0} \le \boldsymbol{V}_{t-1} \le \frac{1}{1-\gamma}\mathbf{1}, \tag{32}$$

and we intend to justify the claim for the t-th iteration. Recognizing that $\mathbf{0} \le \boldsymbol{r} \le \mathbf{1}$, $\boldsymbol{P}_t \ge \mathbf{0}$ and $\|\boldsymbol{P}_t\|_1 = 1$, one can straightforwardly see from the update rule (14) and the induction hypothesis (32) that

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

and

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

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

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

thus establishing (31) for the t-th iteration. Applying the induction argument then concludes the proof. □

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

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_\infty \le \frac{1}{1-\gamma} \qquad \text{for all } t \ge 0, \tag{33}$$

which also confirms that $0 \le \varepsilon \le \frac{1}{1-\gamma}$ is the full ε-range we need to consider. Further, we make note of a direct consequence of the claimed iteration number (30) when $\varepsilon \le \frac{1}{1-\gamma}$:

$$T = \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 \min\{\varepsilon, \varepsilon^2\}} \ge \frac{c_3\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^3}, \tag{34}$$

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 (20):

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^t (1-\eta_j), & \text{if } i = 0, \\ \eta_i \prod_{j=i+1}^t (1-\eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t. \end{cases} \tag{35}$$

We now take a moment to bound $\eta_i^{(t)}$. From our assumption (6a) and the condition (34), we know that the learning rate obeys

$$\frac{1}{2c_1(1-\gamma)T/\log^3 T} \le \frac{1}{1 + c_1(1-\gamma)T/\log^3 T} \le \eta_t \le \frac{1}{1 + c_2(1-\gamma)t/\log^3 T} \le \frac{1}{c_2(1-\gamma)t/\log^3 T} \tag{36}$$

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

$$\beta := \frac{c_4(1-\gamma)}{\log T} \tag{37}$$

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

$$t \ge \frac{T}{c_2 \log \frac{1}{1-\gamma}}, \tag{38}$$

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

- For any $0 \le i \le (1-\beta)t$, we can use (36) and the fact $\log T \ge 2\log\frac{1}{1-\gamma}$ (see Condition (34)) to show that

$$\eta_i^{(t)} \le \Big(1 - \frac{1}{2c_1(1-\gamma)T/\log^3 T}\Big)^{\beta t} \le \Big(1 - \frac{1}{2c_1(1-\gamma)T/\log^3 T}\Big)^{\frac{c_4(1-\gamma)T}{c_2(\log T)(\log\frac{1}{1-\gamma})}} < \frac{1}{2T^2}, \tag{39a}$$

  where the last inequality holds as long as $c_1 c_2 \le c_4/4$.

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

$$\eta_i^{(t)} \le \eta_i \le \frac{1}{c_2(1-\gamma)i/\log^3 T} < \frac{2}{c_2(1-\gamma)t/\log^3 T} \le \frac{2(\log^3 T)(\log\frac{1}{1-\gamma})}{(1-\gamma)T} < \frac{2\log^4 T}{(1-\gamma)T}, \tag{39b}$$

  where we have used the constraint (38).

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

$$\begin{aligned}\sum_{i=0}^{t}\eta_i^{(t)} &= \prod_{j=1}^{t}(1-\eta_j) + \eta_1\prod_{j=2}^{t}(1-\eta_j) + \eta_2\prod_{j=3}^{t}(1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t \\ &= \prod_{j=2}^{t}(1-\eta_j) + \eta_2\prod_{j=3}^{t}(1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t = \cdots \\ &= (1-\eta_t) + \eta_t = 1.\end{aligned} \tag{40}$$

Repeating the same argument further allows us to derive

$$\sum_{i=\tau}^{t}\eta_i^{(t)} = 1 - \prod_{j=\tau}^{t}(1-\eta_j) \tag{41}$$

for any $\tau \le t$.

A.2 Proof of Lemma 1

We shall exploit the relation (23) to prove this lemma. One of the key ingredients of our analysis lies in controlling the terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ introduced in (23), which in turn enables us to apply (23) recursively to control $\boldsymbol{\Delta}_t$.

Step 1: bounding $\boldsymbol{\zeta}_t$. We start by developing an upper bound on $\boldsymbol{\zeta}_t$ (cf. (23)) for any t obeying $\frac{T}{c_2\log\frac{1}{1-\gamma}} \le t \le T$. Invoking the preceding upper bounds (39) on $\eta_i^{(t)}$ implies that

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

Here, (i) holds since $\|\boldsymbol{P}^{\pi_{i-1}}\|_1 = \|\boldsymbol{P}_i\|_1 = \|\boldsymbol{P}\|_1 = 1$ (as they are all probability transition matrices), whereas (ii) arises from the previous bound (39a).

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

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

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

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

This motivates us to invoke Freedman's inequality (see Lemma 4) to control $\boldsymbol{\xi}_t$ for any t obeying $\frac{T}{c_2 \log \frac{1}{1-\gamma}} \le t \le T$. Towards this, we need to calculate several quantities.

- First, it is seen that

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

  where the last inequality is due to (39b), Lemma 4, and the fact $\|\boldsymbol{P}_i\|_1 = \|\boldsymbol{P}\|_1 = 1$.

- Next, we turn to certain variance terms. For any vector $\boldsymbol{a} = [a_j]$, let us use $\mathsf{Var}\big(\boldsymbol{a} \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big)$ to denote a vector whose j-th entry is given by $\mathsf{Var}\big(a_j \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big)$. With this notation in place, and recalling the notation $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{z})$ in (11), we obtain

$$\begin{aligned} \boldsymbol{W}_t &\coloneqq \sum_{i=(1-\beta)t+1}^{t} \mathsf{Var}\big(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\big) = \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \big(\eta_i^{(t)}\big)^2 \mathsf{Var}\Big((\boldsymbol{P}_i - \boldsymbol{P}) \boldsymbol{V}_{i-1} \,|\, \boldsymbol{V}_{i-1}\Big) \\ &= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \big(\eta_i^{(t)}\big)^2 \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_{i-1}\big) \\ &\le \Big(\max_{(1-\beta)t\le i\le t} \eta_i^{(t)}\Big) \Big(\sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)}\Big) \max_{(1-\beta)t\le i<t} \mathsf{Var}_{\boldsymbol{P}}\big(\boldsymbol{V}_i\big) \\ &\le \frac{2\log^4 T}{(1-\gamma)T} \max_{(1-\beta)t\le i<t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i), \end{aligned} \tag{42}$$

  where the last inequality relies on the previous bounds (39b) and (40).

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

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

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

$$\frac{\sigma^2}{2^K} \le \frac{2\log^4 T}{(1-\gamma)T}. \tag{43}$$

With the above bounds in place, applying the Freedman inequality in Lemma 4 and invoking the union bound over all the $|\mathcal{S}||\mathcal{A}|$ entries of $\boldsymbol{\xi}_t$ demonstrate that

$$\begin{aligned} |\boldsymbol{\xi}_t| &\le \sqrt{8\Big(\boldsymbol{W}_t + \frac{\sigma^2}{2^K}\boldsymbol{1}\Big) \log \frac{8|\mathcal{S}||\mathcal{A}|T \log \frac{1}{1-\gamma}}{\delta}} + \Big(\frac{4}{3} B \log \frac{8|\mathcal{S}||\mathcal{A}|T \log \frac{1}{1-\gamma}}{\delta}\Big) \cdot \boldsymbol{1} \\ &\le \sqrt{16\Big(\boldsymbol{W}_t + \frac{2\log^4 T}{(1-\gamma)T}\boldsymbol{1}\Big) \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \Big(3B \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big) \cdot \boldsymbol{1} \end{aligned}$$

$$\leq \sqrt{\frac{32\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)T}\Big(\max_{(1-\beta)t\leq i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big)}+\frac{12\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^2T}\mathbf{1}$$

with probability at least $1-\delta/T$. Here, the second line holds due to (43) and the fact $\log\frac{8|\mathcal{S}||\mathcal{A}|T\log\frac{1}{1-\gamma}}{\delta}\leq 2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ (cf. (34)), whereas the last inequality makes use of the relation (42).

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

$$\boldsymbol{\varphi}_t \coloneqq 64\frac{\log^4 T\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{(1-\gamma)T}\Big(\max_{\frac{t}{2}\leq i\leq t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big) \tag{44}$$

In view of the upper bounds derived in Steps 1 and 2, and β defined in (37), we have — with probability exceeding $1-\delta$ — that

$$|\boldsymbol{\zeta}_k|+|\boldsymbol{\xi}_k|\leq\sqrt{\boldsymbol{\varphi}_t}\qquad\text{for all }\ 2t/3\leq k\leq t, \tag{45}$$

provided that $T\geq\frac{c_9(\log^4 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^3}$ for some sufficiently large constant $c_9>0$. Substituting (45) into (23), we can upper bound $\boldsymbol{\Delta}_t$ as follows

$$\boldsymbol{\Delta}_k\leq\sqrt{\boldsymbol{\varphi}_t}+\sum_{i=(1-\beta)k+1}^{k}\eta_i^{(k)}\gamma\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}=\sqrt{\boldsymbol{\varphi}_t}+\sum_{i=(1-\beta)k}^{k-1}\eta_{i+1}^{(k)}\gamma\boldsymbol{P}^{\pi_i}\boldsymbol{\Delta}_i\qquad\text{for all }\ 2t/3\leq k\leq t. \tag{46}$$

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

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

Clearly, this sequence satisfies

$$\alpha_i^{(t)}\geq\eta_{i+1}^{(t)}\qquad\text{and}\qquad\sum_{i=(1-\beta)t}^{t-1}\alpha_i^{(t)}=1 \tag{47}$$

for any t, where the first inequality results from (40). With these in place, we can write (46) as

$$\boldsymbol{\Delta}_k\leq\sqrt{\boldsymbol{\varphi}_t}+\sum_{i_1=(1-\beta)k}^{k-1}\eta_{i_1+1}^{(k)}\gamma\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\Delta}_{i_1}=\sum_{i_1=(1-\beta)k}^{k-1}\Big(\alpha_{i_1}^{(k)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(k)}\gamma\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\Delta}_{i_1}\Big)\qquad\text{for all } 2t/3\leq k\leq t. \tag{48}$$

Given that $(1-\beta)t\geq 2t/3$ (see (37)), we can invoke this relation recursively to yield

$$\begin{aligned}
\boldsymbol{\Delta}_t&\leq\sum_{i_1=(1-\beta)t}^{t-1}\Big(\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(t)}\gamma\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\Delta}_{i_1}\Big)\\
&\leq\sum_{i_1=(1-\beta)t}^{t-1}\left[\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(t)}\gamma\boldsymbol{P}^{\pi_{i_1}}\sum_{i_2=(1-\beta)i_1}^{i_1-1}\Big(\alpha_{i_2}^{(i_1)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_2+1}^{(i_1)}\gamma\boldsymbol{P}^{\pi_{i_2}}\boldsymbol{\Delta}_{i_2}\Big)\right]\\
&\leq\sum_{i_1=(1-\beta)t}^{t-1}\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\sum_{i_1=(1-\beta)t}^{t-1}\sum_{i_2=(1-\beta)i_1}^{i_1-1}\alpha_{i_1}^{(t)}\alpha_{i_2}^{(i_1)}\big(\gamma\boldsymbol{P}^{\pi_{i_1}}\big)\sqrt{\boldsymbol{\varphi}_t}\\
&\qquad+\sum_{i_1=(1-\beta)t}^{t-1}\sum_{i_2=(1-\beta)i_1}^{i_1-1}\eta_{i_1+1}^{(t)}\eta_{i_2+1}^{(i_1)}\prod_{k=1}^{2}\big(\gamma\boldsymbol{P}^{\pi_{i_k}}\big)\boldsymbol{\Delta}_{i_2}\\
&=\sum_{i_1=(1-\beta)t}^{t-1}\sum_{i_2=(1-\beta)i_1}^{i_1-1}\alpha_{i_1}^{(t)}\alpha_{i_2}^{(i_1)}\left\{\boldsymbol{I}+\gamma\boldsymbol{P}^{\pi_{i_1}}\right\}\sqrt{\boldsymbol{\varphi}_t}+\sum_{i_1=(1-\beta)t}^{t-1}\sum_{i_2=(1-\beta)i_1}^{i_1-1}\eta_{i_1+1}^{(t)}\eta_{i_2+1}^{(i_1)}\prod_{k=1}^{2}\big(\gamma\boldsymbol{P}^{\pi_{i_k}}\big)\boldsymbol{\Delta}_{i_2},
\end{aligned} \tag{49}$$

where the second inequality relies on (48), the third line uses the inequality $\eta_{i_1+1}^{(t)} \le \alpha_{i_1}^{(t)}$ in (47), and the fourth line is valid since $\sum_{i_2=(1-\beta)i_1}^{i_1-1} \alpha_{i_2}^{(i_1)} = 1$ (see (47)).

We intend to continue invoking (48) recursively — similar to how we derive (49) — in order to control $\boldsymbol{\Delta}_t$. To do so, we are in need of some preparation. First, let us define

$$H \coloneqq \frac{\log T}{1-\gamma} \qquad \text{and} \qquad \alpha_{\{i_k\}_{k=1}^H} \coloneqq \alpha_{i_1}^{(t)} \alpha_{i_2}^{(i_1)} \dots \alpha_{i_H}^{(i_{H-1})} \ge 0 \tag{50}$$

for any $t > i_1 > i_2 > \cdots > i_H$, which clearly satisfies (see (47))

$$\alpha_{\{i_k\}_{k=1}^H} \ge \eta_{i_1+1}^{(t)} \eta_{i_2+1}^{(i_1)} \dots \eta_{i_H+1}^{(i_{H-1})}. \tag{51}$$

In addition, defining the index set

$$\mathcal{I}_t \coloneqq \Big\{ (i_1, \cdots, i_H) \,|\, (1-\beta)t \le i_1 \le t-1,\ \forall 1 \le j < H : (1-\beta) i_j \le i_{j+1} \le i_j - 1 \Big\}, \tag{52}$$

we have

$$\sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1. \tag{53}$$

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

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

for c_4 small enough, thus implying that

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

This is an important property that allows one to invoke the relation (48). With these in place, applying the preceding relation (48) recursively — in a way similar to (49) — further leads to

$$\begin{aligned} \boldsymbol{\Delta}_t &\le \sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} \left\{ \left(\boldsymbol{I} + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \right) \sqrt{\boldsymbol{\varphi}_t} + \gamma^H \prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_k}} |\boldsymbol{\Delta}_{i_H}| \right\} \\ &\le \max_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \Bigg\{ \underbrace{\left(\boldsymbol{I} + \sum_{h=1}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \right) \sqrt{\boldsymbol{\varphi}_t}}_{=:\boldsymbol{\beta}_1} + \underbrace{\gamma^H \prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_k}} |\boldsymbol{\Delta}_{i_H}|}_{=:\boldsymbol{\beta}_2} \Bigg\} \end{aligned} \tag{54}$$

for all $t \ge \frac{T}{c_2 \log \frac{1}{1-\gamma}}$, where we recall the definition of the entrywise max operator in Section 4.1. Here, the last inequality relies on the fact that $\sum_{(i_1,\cdots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1$ (see (53)).

Step 4: bounding $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. It remains to control $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, which we shall accomplish separately as follows.

- The term $\boldsymbol{\beta}_2$ is relatively easier to control. Observing that $\prod_{k=1}^H \boldsymbol{P}^{\pi_{i_k}}$ is still a probability transition matrix, we can derive

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

  where (i) results from the crude bound (33). To justify the inequality (ii), we recall the definition (50) of H to see that

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

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

- When it comes to $\boldsymbol{\beta}_1$, we can upper bound the entrywise square of $\boldsymbol{\beta}_1$ — denoted by $\boldsymbol{\beta}_1^2$ — as follows

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

Here, (i) follows from Jensen's inequality and the fact that $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}$ is a probability transition matrix; (ii) holds due to the Cauchy-Schwarz inequality; (iii) utilizes the definition of $\boldsymbol{\varphi}_t$ in (44); (iv) follows since $\prod_{1\le k\le h}\boldsymbol{P}^{\pi_{i_k}}\mathbf{1}=\mathbf{1}$ and $\sum_{0\le h<H}\gamma^h\le\frac{1}{1-\gamma}$. To further control the right-hand side of the above inequality, we resort to the following lemma.

Lemma 5. *The following holds:*

$$\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\frac{t}{2}\le i<t}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)\le\frac{4}{\gamma^2(1-\gamma)^2}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}. \tag{55}$$

Proof. See Appendix A.5. □

Therefore, the above result directly implies that

$$\boldsymbol{\beta}_1^2\le\frac{320\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4T}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}. \tag{56}$$

Step 6: putting all this together. Substituting the preceding bounds for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ into (54), we can demonstrate that: with probability at least $1-\delta$,

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le \frac{1}{(1-\gamma)T}\mathbf{1}+\sqrt{\frac{320\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4T}\Big(1+2\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1} \\
&\le 30\sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4T}\Big(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)}\,\mathbf{1}
\end{aligned} \tag{57}$$

holds simultaneously for all $t\ge\frac{T}{c_2\log\frac{1}{1-\gamma}}$, where the second line is valid since $\frac{1}{(1-\gamma)T}\le\sqrt{\frac{(\log^4 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4T}}$ under our sample size condition (34).

A.3 Proof of Lemma 2

Next, we move forward to develop an lower bound on $\boldsymbol{\Delta}_t$, which can be accomplished in an analogous manner as for the above upper bound. Applying a similar argument for (54) (except that we need to replace π_i with $\pi^\star$), one can deduce that

$$\boldsymbol{\Delta}_t\ge-\max_{(i_1,\cdots,i_H)\in\mathcal{I}_t}\left\{\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi^\star}\sqrt{\boldsymbol{\varphi}_t}+\gamma^H\prod_{k=1}^{H}\boldsymbol{P}^{\pi^\star}\big|\boldsymbol{\Delta}_{i_H}\big|\right\} \tag{58}$$

for any $t \geq \frac{c_2 T}{\log \frac{1}{1-\gamma}}$. It is straightforward to bound the second term on the right-hand side of (58) as

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

where the second inequality makes use of (33) as well as the fact that $\prod_k \boldsymbol{P}^{\pi^\star}$ is a probability transition matrix (so that $\|\prod_k \boldsymbol{P}^{\pi^\star}\|_1 = 1$). As for the first term on the right-hand side of (58), we can invoke a similar argument for (56) to obtain

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

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

$$\boldsymbol{\Delta}_t \geq -30\sqrt{\frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\gamma^2(1-\gamma)^4 T} \left(1 + \max_{\frac{t}{2}\leq i<t} \|\boldsymbol{\Delta}_i\|_\infty\right)}\, \mathbf{1} \tag{59}$$

holds simultaneously for all $t \geq \frac{T}{c_2 \log \frac{1}{1-\gamma}}$.

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

Recall from (24) that with probability exceeding $1-2\delta$, the following recurrence relation

$$\|\boldsymbol{\Delta}_t\|_\infty \leq 30\sqrt{\frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T} \left(1 + \max_{\frac{t}{2}\leq i<t} \|\boldsymbol{\Delta}_i\|_\infty\right)} \qquad \text{for all} \qquad t \geq \frac{T}{c_2 \log \frac{1}{1-\gamma}} \tag{60}$$

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

$$u_k := \max\left\{ \|\boldsymbol{\Delta}_t\|_\infty \;\middle|\; 2^k \frac{T}{c_2 \log \frac{1}{1-\gamma}} \leq t \leq T \right\}. \tag{61}$$

To bound this sequence, we first obtain a crude bound as a result of (33):

$$u_0 \leq \frac{1}{1-\gamma}.$$

Next, it is directly seen from (60) and the definition of u_k that

$$u_k \leq c_6\sqrt{\frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\,(1+u_{k-1})}, \qquad k \geq 1 \tag{62}$$

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 (62) tells us that

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

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

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

- Instead, suppose that $u_j > 1$ for all $0 \le j \le k$. Then it is seen from (62) that

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

This is equivalent to saying that

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

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

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

or equivalently,

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

Putting the above two cases together, we conclude that

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

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

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

As a result, the above bound simplifies to

$$u_k \le c_8\left(\sqrt{\frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\right), \qquad k \ge c_7 \log\log \frac{1}{1-\gamma}$$

for some constant $c_8 > 0$.

Consequently, taking $t = T$ and choosing $k = c_7 \log\log \frac{1}{1-\gamma}$ for some appropriate constant $c_7 > 0$ (so as to ensure $2^k \frac{T}{c_2 \log \frac{1}{1-\gamma}} < T$), we immediately see from the definition (61) of u_k that

$$\|\boldsymbol{\Delta}_T\|_\infty \le c_8\left(\sqrt{\frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}\right). \tag{63}$$

with probability at least $1-2\delta$. To finish up, recognize that the sample size assumption (30) is equivalent to saying that

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

When $c_3 > 0$ is sufficiently large, substituting this relation into (63) gives

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

as claimed in Theorem 3.

A.5 Proof of Lemma 5

We first claim that

$$\max_{\frac{t}{2}\le i<t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) - \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) \le \frac{4}{1-\gamma} \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}. \tag{64}$$

If this claim is valid (which we shall justify towards the end of this subsection), then it leads to

$$\sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \max_{\frac{t}{2}\le i<t} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \le \sum_{h=0}^{H-1} \gamma^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) + \frac{4}{(1-\gamma)^2} \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}. \tag{65}$$

It then boils down to bounding the first term on the right-hand side of (65). Towards this, let us first define a subvector of $\boldsymbol{Q}_\pi^\star$ as follows

$$\boldsymbol{Q}_\pi^\star := \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi \boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}, \tag{66}$$

where we denote by $\boldsymbol{r}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}|}$ the reward vector restricted to the actions chosen by the policy π, namely, $r_\pi(s) = r(s, \pi(s))$ for all $s \in \mathcal{S}$. We intend to upper bound the variance term involving $\boldsymbol{V}^\star$. For any $0 \le h < H$, one can express (see (11))

$$\begin{aligned}
\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &= \boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - (\boldsymbol{P}\boldsymbol{V}^\star) \circ (\boldsymbol{P}\boldsymbol{V}^\star) \\
&\overset{\text{(i)}}{=} \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r}) \circ (\boldsymbol{Q}^\star - \boldsymbol{r}) \\
&\le \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \big\|\boldsymbol{P}(\boldsymbol{V}^\star \circ \boldsymbol{V}^\star) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star)\big\|_\infty \mathbf{1} - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r}) \circ (\boldsymbol{Q}^\star - \boldsymbol{r}) \\
&\overset{\text{(ii)}}{\le} \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) + \frac{4}{1-\gamma} \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} - \frac{1}{\gamma^2}(\boldsymbol{Q}^\star - \boldsymbol{r}) \circ (\boldsymbol{Q}^\star - \boldsymbol{r}) \\
&= \frac{1}{\gamma^2}\big(\gamma^2 \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star\big) + \frac{4}{1-\gamma} \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} - \frac{1}{\gamma^2}\boldsymbol{r} \circ \boldsymbol{r} + \frac{2}{\gamma^2}\boldsymbol{Q}^\star \circ \boldsymbol{r} \\
&\overset{\text{(iii)}}{\le} \frac{1}{\gamma}\big(\gamma \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star) - \boldsymbol{Q}^\star \circ \boldsymbol{Q}^\star\big) + \frac{2}{\gamma^2}\boldsymbol{Q}^\star \circ \boldsymbol{r} + \frac{4}{1-\gamma} \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1},
\end{aligned} \tag{67}$$

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

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

where the first inequality arises from the fact $\|\boldsymbol{P}\boldsymbol{z}\|_\infty \le \|\boldsymbol{P}\|_1 \|\boldsymbol{z}\|_\infty = \|\boldsymbol{z}\|_\infty$, the second inequality follows from Lemma 4, and the last line holds since

$$\begin{aligned}
\big\|\boldsymbol{Q}^\star_{\pi_{i_{h+1}}} - \boldsymbol{V}_{i_{h+1}}\big\|_\infty &= \big\|\big(\boldsymbol{r}_{\pi_{i_{h+1}}} + \gamma \boldsymbol{P}_{\pi_{i_{h+1}}} \boldsymbol{V}^\star\big) - \boldsymbol{V}_{i_{h+1}}\big\|_\infty \\
&= \big\|\big(\boldsymbol{r}_{\pi_{i_{h+1}}} + \gamma \boldsymbol{P}_{\pi_{i_{h+1}}} \boldsymbol{V}^\star\big) - \big(\boldsymbol{r}_{\pi_{i_{h+1}}} + \gamma \boldsymbol{P}_{\pi_{i_{h+1}}} \boldsymbol{V}_{i_{h+1}}\big)\big\|_\infty \\
&= \gamma\big\|\boldsymbol{P}_{\pi_{i_{h+1}}}\big(\boldsymbol{V}^\star - \boldsymbol{V}_{i_{h+1}}\big)\big\|_\infty \le \big\|\boldsymbol{V}^\star - \boldsymbol{V}_{i_{h+1}}\big\|_\infty \le \big\|\boldsymbol{Q}^\star - \boldsymbol{Q}_{i_{h+1}}\big\|_\infty = \|\boldsymbol{\Delta}_{i_{h+1}}\|_\infty
\end{aligned}$$

and

$$\|\boldsymbol{V}_{i_{h+1}} - \boldsymbol{V}^\star\|_\infty \le \big\|\boldsymbol{Q}_{i_{h+1}} - \boldsymbol{Q}^\star\big\|_\infty = \|\boldsymbol{\Delta}_{i_{h+1}}\|_\infty.$$

As it turns out, the first term in (67) allows one to build a telescoping sum. Specifically, invoking (67) allows one to bound

$$\begin{aligned}
\sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}^\star) &\le \frac{1}{\gamma}\sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\left(\gamma \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^\star\circ\boldsymbol{Q}^\star)-\boldsymbol{Q}^\star\circ\boldsymbol{Q}^\star\right)\\
&\quad+\frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\mathbf{1}+\frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star\circ\boldsymbol{r})\\
&\overset{\text{(i)}}{=}\frac{1}{\gamma}\Big(\sum_{h=0}^{H-1}\prod_{k=1}^{h+1}\gamma \boldsymbol{P}^{\pi_{i_k}}-\sum_{h=0}^{H-1}\prod_{k=1}^{h}\gamma \boldsymbol{P}^{\pi_{i_k}}\Big)(\boldsymbol{Q}^\star\circ\boldsymbol{Q}^\star)\\
&\quad+\frac{4}{1-\gamma}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\sum_{h=0}^{H-1}\gamma^h\mathbf{1}+\frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star\circ\boldsymbol{r})\\
&\le\frac{1}{\gamma}\Big(\prod_{k=1}^{H}\gamma \boldsymbol{P}^{\pi_{i_k}}-\boldsymbol{I}\Big)(\boldsymbol{Q}^\star\circ\boldsymbol{Q}^\star)+\frac{4}{(1-\gamma)^2}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\mathbf{1}\\
&\quad+\frac{2}{\gamma^2}\sum_{h=0}^{H-1}\gamma^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}(\boldsymbol{Q}^\star\circ\boldsymbol{r})\\
&\overset{\text{(ii)}}{\le}\left(\frac{2}{\gamma}\|\boldsymbol{Q}^\star\|_\infty^2+\frac{4}{(1-\gamma)^2}\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty+\frac{2}{\gamma^2}\frac{1}{1-\gamma}\|\boldsymbol{Q}^\star\|_\infty\|\boldsymbol{r}\|_\infty\right)\mathbf{1}\\
&\overset{\text{(iii)}}{\le}\frac{1}{(1-\gamma)^2}\left(\frac{2}{\gamma}+4\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty+\frac{2}{\gamma^2}\right)\mathbf{1}\\
&\le\frac{1}{(1-\gamma)^2}\left(\frac{4}{\gamma^2}+4\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)\mathbf{1}.
\end{aligned}\tag{68}$$

Here, (i) comes from the identity $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\mathbf{1}=\mathbf{1}$; (ii) holds because each row of $\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}$ has unit $\|\cdot\|_1$ norm for any h; (iii) arises from the bound $\|\boldsymbol{Q}^\star\|_\infty\le 1/(1-\gamma)$. This completes the proof, as long as the claim (64) can be justified.

Proof of the inequality (64). To validate this result, we make the observation that

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

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

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

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

B Lower bound: sub-optimality of vanilla Q-learning

In this section, we establish the lower bound claimed in Theorem 2 by analyzing Q-learning for the MDP instance constructed in Section 4.3. Without loss of generality, we assume

$$\log T\le\frac{1}{1-\gamma}\tag{69}$$

throughout the proof; otherwise the lower bound in Theorem 2 is worse than the minimax lower bound $\frac{1}{(1-\gamma)^3 T}$ in Azar et al. [2013].

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

$$P_t(s' \,|\, s, a) \coloneqq \begin{cases} 1, & \text{if } s_t(s,a) = s', \\ 0, & \text{otherwise,} \end{cases} \tag{70}$$

where $s_t(s,a)$ stands for the sample collected in the t-th iteration (see (4)). Recognizing that state 2 is associated with a singleton action space, we shall often write

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

for notational simplicity.

B.1 Key quantities related to learning rates

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

$$\eta_k^{(t)} \coloneqq \eta_k \prod_{i=k+1}^{t} \big(1 - \eta_i(1-\gamma p)\big) \qquad \text{for any } 1 \le k < t, \tag{71a}$$

$$\eta_0^{(t)} \coloneqq \prod_{i=1}^{t} \big(1 - \eta_i(1-\gamma p)\big), \tag{71b}$$

$$\eta_t^{(t)} \coloneqq \eta_t. \tag{71c}$$

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

$$\eta_0^{(t)} + (1-\gamma p)\sum_{k=1}^{t} \eta_k^{(t)} = \prod_{i=1}^{t}(1-\widehat{\eta}_i) + \widehat{\eta}_1 \prod_{i=2}^{t}(1-\widehat{\eta}_i) + \widehat{\eta}_2 \prod_{i=3}^{t}(1-\widehat{\eta}_i) + \cdots + \widehat{\eta}_{t-1}(1-\widehat{\eta}_t) + \widehat{\eta}_t = 1, \tag{72}$$

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

$$\prod_{i=\tau+1}^{t} \big(1 - \eta_i(1-\gamma p)\big) + (1-\gamma p) \sum_{k=\tau+1}^{t} \eta_k^{(t)} = 1. \tag{73}$$

B.2 Preliminary calculations

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

B.2.1 Basic update rules and expansion

Given that $Q_0 = V_0 = 0$ and that state 0 is absorbing, the update rule (3) gives

$$V_t(0) = Q_t(0,1) = \big(1 - \eta_t(1-\gamma)\big) Q_{t-1}(0,1) = \prod_{i=1}^{t} \big(1 - \eta_i(1-\gamma)\big) Q_0(0,1) = 0 \tag{74}$$

for all $t \ge 1$. Regarding state 2, the update rule (3) taken together with (74) leads to

$$\begin{aligned} V_t(2) = Q_t(2,1) &= \big(1-\eta_t\big) Q_{t-1}(2,1) + \eta_t \big\{ r(2,1) + \gamma P_t(2 \,|\, 2) V_{t-1}(2) + \gamma P_t(0 \,|\, 2) V_{t-1}(0) \big\} \\ &= \big(1-\eta_t\big) V_{t-1}(2) + \eta_t \big\{ 1 + \gamma P_t(2 \,|\, 2) V_{t-1}(2) \big\}, \end{aligned} \tag{75}$$

and for state 3,

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

$$= \big(1 - \eta_t(1-\gamma)\big)V_{t-1}(3) + \eta_t. \tag{76}$$

Similarly, one also has

$$Q_t(1,1) = (1-\eta_t)Q_{t-1}(1,1) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,1)V_{t-1}(1)\big\}, \tag{77a}$$
$$Q_t(1,2) = (1-\eta_t)Q_{t-1}(1,2) + \eta_t\big\{1 + \gamma P_t(1\,|\,1,2)V_{t-1}(1)\big\}. \tag{77b}$$

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

Next, we record some elementary decomposition of $V_t(2)$. For any iteration t and $\tau < t$, one can continue the derivation in (75) to obtain

$$\begin{aligned}
V_t(2) &= \big(1 - \eta_t(1-\gamma p)\big)V_{t-1}(2) + \eta_t\big\{1 + \gamma\big(P_t(2\,|\,2) - p\big)V_{t-1}(2)\big\} \\
&= \prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)V_\tau(2) + \sum_{k=\tau+1}^{t}\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i(1-\gamma p)\big)\big\{1+\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2)\big\} \\
&= \prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)V_\tau(2) + \sum_{k=\tau+1}^{t}\eta_k^{(t)} + \sum_{k=\tau+1}^{t}\eta_k^{(t)}\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2) \\
&= \prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)V_\tau(2) + \frac{1-\prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)}{1-\gamma p} + \sum_{k=\tau+1}^{t}\eta_k^{(t)}\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2) \\
&= \frac{1}{1-\gamma p} - \prod_{i=\tau+1}^{t}\big(1-\eta_i(1-\gamma p)\big)\left[\frac{1}{1-\gamma p} - V_\tau(2)\right] + \sum_{k=\tau+1}^{t}\eta_k^{(t)}\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2),
\end{aligned} \tag{78}$$

where the penultimate line arises from (73). In particular, in the special case where $\tau = 0$ (so that $V_\tau(2) = V_0(2) = 0$), this simplifies to

$$V_t(2) = \frac{1-\eta_0^{(t)}}{1-\gamma p} + \sum_{k=1}^{t}\eta_k^{(t)}\gamma\big(P_k(2\,|\,2)-p\big)V_{k-1}(2), \tag{79}$$

which relies on the definition of $\eta_0^{(t)}$ in (71). With similar derivation, (76) leads to

$$V_t(3) = \frac{1}{1-\gamma}\Big[1 - \prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big)\Big] = V^\star(3) - \frac{1}{1-\gamma}\prod_{i=1}^{T}\big(1-\eta_i(1-\gamma)\big). \tag{80}$$

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

We start by computing the mean $V^\star(2) - \mathbb{E}[V_t(2)]$. From the construction (26), it is easily seen that $\mathbb{E}[P_k(2\,|\,2)] = p$, which together with the identity (79) leads to

$$\mathbb{E}\big[V_T(2)\big] = \frac{1-\eta_0^{(T)}}{1-\gamma p} \qquad \text{and} \qquad V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\eta_0^{(T)}}{1-\gamma p}. \tag{81}$$

Similarly, applying the above argument to (78) and rearranging terms, we immediately arrive at

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \prod_{i=\tau+1}^{T}\big(1-\eta_i(1-\gamma p)\big)\left[\frac{1}{1-\gamma p} - \mathbb{E}\big[V_\tau(2)\big]\right] \tag{82}$$

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

Next, we develop a lower bound on the variance $\mathsf{Var}\big(V_T(2)\big)$. Towards this end, consider first a martingale sequence $\{Z_k\}_{0\le k\le T}$ adapted to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$, namely, $\mathbb{E}[Z_{k+1}\,|\,\mathcal{F}_k] = 0$ and $\mathbb{E}[Z_k\,|\,\mathcal{F}_k] =$

Z_k for all $0 \le k \le T$. In addition, consider any $0 \le \tau < T$, and let W_0 be a random variable such that $\mathbb{E}[W_0 \,|\, \mathcal{F}_\tau] = W_0$. Then the law of total variance together with basic martingale properties tells us that

$$\begin{aligned}
\mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T} Z_k\right) &= \mathbb{E}\left[\mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T} Z_k \mid \mathcal{F}_{T-1}\right)\right] + \mathsf{Var}\left(\mathbb{E}\left[W_0 + \sum_{k=\tau+1}^{T} Z_k \mid \mathcal{F}_{T-1}\right]\right) \\
&= \mathbb{E}\big[\mathsf{Var}\left(Z_T \mid \mathcal{F}_{T-1}\right)\big] + \mathsf{Var}\left(W_0 + \sum_{k=\tau+1}^{T-1} Z_k\right) = \cdots \\
&= \sum_{k=\tau+1}^{T} \mathbb{E}\big[\mathsf{Var}\left(Z_k \mid \mathcal{F}_{k-1}\right)\big] + \mathsf{Var}\left(W_0\right) \ge \sum_{k=\tau+1}^{T} \mathbb{E}\big[\mathsf{Var}\left(Z_k \mid \mathcal{F}_{k-1}\right)\big]. \qquad (83)
\end{aligned}$$

Consequently, for any $0 \le \tau < T-1$, it follows from the decomposition (78) (with τ replaced by $\tau+1$) that

$$\begin{aligned}
\mathsf{Var}\big(V_T(2)\big) &\ge \mathbb{E}\left[\sum_{k=\tau+2}^{T} \mathsf{Var}\Big(\eta_k^{(T)}\gamma\big(P_k(2\,|\,2) - p\big)V_{k-1}(2) \,|\, V_{k-1}(2)\Big)\right] \\
&= \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\gamma\big)^2 p(1-p)\mathbb{E}\big[\big(V_{k-1}(2)\big)^2\big] \\
&\ge \frac{(1-\gamma)\,(4\gamma-1)}{9} \cdot \frac{1}{4(1-\gamma)^2} \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2 \\
&= \frac{4\gamma-1}{36\,(1-\gamma)} \sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2, \qquad (84)
\end{aligned}$$

where the first identity relies on the fact that $P_k(2\,|\,2)$ is a Bernoulli random variable with mean p, and the inequality comes from the definition of τ (see (90)) and the choice of p (see (27)). As an implication, the sum of squares of $\eta_k^{(T)}$ plays a crucial role in determining the variance of $V_T(2)$.

B.3 Lower bounds for three cases

B.3.1 Case 1: small learning rates ($c_\eta \ge \log T$ or $0 \le \eta \le \frac{1}{(1-\gamma)T}$)

In this case, we focus on lower bounding $V^\star(2) - \mathbb{E}\big[V_T(2)\big]$. In view of this identity (81), this boils down to controlling $\eta_0^{(T)}$.

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

$$\eta_t(1-\gamma p) \le \begin{cases} \frac{1-\gamma p}{(1-\gamma)t\log T} = \frac{4}{3t\log T} \le \frac{1}{2}, & \text{if } \eta_t = \frac{1}{1+c_\eta(1-\gamma)t} \\ \frac{1-\gamma p}{(1-\gamma)T} = \frac{4}{3T} \le \frac{1}{2}, & \text{if } \eta_t = \eta \end{cases} \qquad (85)$$

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

$$\log \eta_0^{(T)} = \sum_{i=1}^{T} \log\big(1 - \eta_i(1-\gamma p)\big) \ge -1.5 \sum_{i=1}^{T} \eta_i(1-\gamma p) \ge -2, \qquad (86)$$

where the first inequality holds due to the elementary fact $\log(1-x) \ge -1.5x$ for all $0 \le x \le 0.5$, and the last inequality follows from the following bound (which makes use of (85))

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

Combining the above result with the properties (81) and (86) then yields

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\eta_0^{(T)}}{1-\gamma p} \geq \frac{e^{-2}}{1-\gamma p} = \frac{3}{4e^2(1-\gamma)}. \tag{87}$$

This taken together with (29) gives

$$\mathbb{E}\big[\big(V^\star(2) - V_T(2)\big)^2\big] \geq \big(V^\star(2) - \mathbb{E}\big[V_T(2)\big]\big)^2 \geq \frac{9}{16e^4(1-\gamma)^2}. \tag{88}$$

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

By virtue of (82), the mean gap $V^\star(2) - \mathbb{E}\big[V_T(2)\big]$ depends on two factors: (i) the choice of the learning rates, and (ii) the gap between $\frac{1}{1-\gamma p}$ and $\mathbb{E}\big[V_\tau(2)\big]$, where τ is an integer obeying $0 \leq \tau < T$. To control the factor (ii), we need to choose τ properly. Let us start by considering the simple scenario with $\mathbb{E}\big[\big(V_T(2)\big)^2\big] < \frac{1}{4(1-\gamma)^2}$, for which we have

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] \geq \frac{3}{4(1-\gamma)} - \sqrt{\mathbb{E}\big[\big(V_T(2)\big)^2\big]} \geq \frac{1}{4(1-\gamma)}. \tag{89}$$

Here, we have used (28) and the elementary fact $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$. Consequently, it remains to look at the scenario obeying $\mathbb{E}\big[\big(V_T(2)\big)^2\big] \geq \frac{1}{4(1-\gamma)^2}$, towards which we propose to set τ as follows

$$\tau := \min\left\{0 \leq \tau' \leq T-1 \;\middle|\; \mathbb{E}\big[\big(V_t(2)\big)^2\big] \geq \frac{1}{4(1-\gamma)^2} \text{ for all } \tau'+1 \leq t \leq T\right\}. \tag{90}$$

Clearly, τ is well-defined in this scenario and obeys (in view of both (90) and the initialization $V_0 = 0$)

$$\mathbb{E}\big[\big(V_\tau(2)\big)^2\big] < \frac{1}{4(1-\gamma)^2}. \tag{91}$$

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

Case 2.1. Consider the case where

$$\prod_{i=\tau+1}^{T} \big(1 - \eta_i(1-\gamma p)\big) \geq \frac{1}{2}. \tag{92}$$

Invoke (82) to deduce that

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

where the second line makes use of the definition (27) and the elementary fact $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$, and the last line relies on the inequalities (91) and (92).

Case 2.2. We now move on to the case where

$$0 \le \prod_{i=\tau+1}^{T} \big(1-\eta_i(1-\gamma p)\big) \le \frac{1}{2}. \tag{93}$$

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 (84), it boils down to lower bounding $\sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2$, which we accomplish as follows.

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

$$\tau \le T-2. \tag{94}$$

In fact, if $\tau \ge T-1$, then the definition (90) of τ necessarily requires that

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

In view of (81) (with T replaced by $T-1$), a little algebra shows that this is equivalent to $\big(1-\eta(1-\gamma p)\big)^{T-1} \ge 1/3$, and hence $\big(1-\eta(1-\gamma p)\big)^{T} \ge 1/9$. In turn, this combined with (81) leads to

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] = \frac{\big(1-\eta(1-\gamma p)\big)^T}{1-\gamma p} = \frac{3\big(1-\eta(1-\gamma p)\big)^T}{4(1-\gamma)} \ge \frac{1}{12(1-\gamma)}, \tag{95}$$

which already suffices for our purpose.

Next, assuming that (94) holds, one can derive

$$\begin{aligned}\sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2 &= \sum_{k=\tau+2}^{T} \eta^2\big(1-\eta(1-\gamma p)\big)^{2(T-k)} = \frac{\eta^2\big[1-\big(1-\eta(1-\gamma p)\big)^{2(T-\tau-1)}\big]}{1-\big(1-\eta(1-\gamma p)\big)^2} \\ &\ge \frac{\eta^2/2}{1-\big(1-\eta(1-\gamma p)\big)^2} \ge \frac{3\eta}{16(1-\gamma)},\end{aligned} \tag{96}$$

where the first inequality holds since (from the assumptions (93) and $\tau \le T-2$)

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

and the last inequality follows since

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

Substituting (96) into (84), we obtain

$$\begin{aligned}\mathsf{Var}\big(V_T(2)\big) &\ge \frac{4\gamma-1}{36(1-\gamma)} \sum_{k=\tau+1}^{T} \big(\eta_k^{(T)}\big)^2 \ge \frac{2}{36(1-\gamma)} \cdot \frac{3\eta}{16(1-\gamma)} \\ &= \frac{\eta}{96(1-\gamma)^2} \ge \frac{1}{96(1-\gamma)^4 T},\end{aligned} \tag{97}$$

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

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

$$\begin{aligned}\frac{\eta_{k-1}^{(T)}}{\eta_k^{(T)}} &= \frac{\eta_{k-1}\big(1-\eta_k(1-\gamma p)\big)}{\eta_k} = \frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-\big(\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}\big)\eta_k} = \frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k} = 1-\frac{\big(\frac{4}{3}-c_\eta\big)(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k} \\ &\le 1-(1-\gamma)\eta_k \le 1-(1-\gamma)\eta_T, \end{aligned} \tag{98}$$

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

$$\begin{aligned}\sum_{k=\tau+2}^{T}\big(\eta_k^{(T)}\big)^2 &\ge \sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\big(\eta_k^{(T)}\big)^2 \ge \frac{1}{T-\max\{\tau+1,\tau'\}}\left[\sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\eta_k^{(T)}\right]^2 \\ &\ge (1-\gamma)\eta_T\left[\sum_{k=\max\{\tau+2,\tau'+1\}}^{T}\eta_k^{(T)}\right]^2, \end{aligned} \tag{99}$$

where the penultimate inequality comes from the Cauchy-Schwarz inequality. In addition, recognizing that $\eta_{k_1}^{(T)} \le \big(1-(1-\gamma)\eta_T\big)^{k_2-k_1}\eta_{k_2}^{(T)}$ for any $k_2 \ge k_1$ (see (98)), one has

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

Summing these inequalities up and rearranging terms, we reach

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

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

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

Here, (i) and (ii) follow from (73) and (93), respectively, while (iii) holds since

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

as long as $\gamma \geq 3/4$. Substitution into (99) yields

$$\sum_{k=\tau+2}^{T} \big(\eta_k^{(T)}\big)^2 \geq \frac{9\eta_T}{1024(1-\gamma)}. \tag{100}$$

Substituting the above bound into (84), we obtain

$$\begin{aligned}\mathsf{Var}\big(V_T(2)\big) &\geq \frac{4\gamma-1}{36(1-\gamma)} \sum_{k=\tau+1}^{T} \big(\eta_k^{(T)}\big)^2 \geq \frac{2}{36(1-\gamma)} \cdot \frac{9\eta_T}{1024(1-\gamma)} \\ &= \frac{\eta_T}{2048(1-\gamma)^2} \geq \frac{1}{4096(1-\gamma)^4 T},\end{aligned} \tag{101}$$

provided that $\gamma \geq 3/4$ (so that $4\gamma - 1 \geq 2$). Here, the last inequality is valid since $\eta_T = \frac{1}{1+c_\eta(1-\gamma)T} \geq \frac{1}{1+(1-\gamma)^2 T} \geq \frac{1}{2(1-\gamma)^2 T}$ 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 (95)) or the variance is too large (see (97) and (101)). These bounds taken collectively with (29) yield

$$\begin{aligned}\mathbb{E}\big[\big(V^\star(2) - V_T(2)\big)^2\big] &\geq \big(V^\star(2) - \mathbb{E}\big[V_T(2)\big]\big)^2 + \mathsf{Var}\big(V_T(2)\big) \\ &\geq \min\left\{\frac{1}{144(1-\gamma)^2}, \frac{1}{96(1-\gamma)^4 T}, \frac{1}{4096(1-\gamma)^4 T}\right\} = \frac{1}{4096(1-\gamma)^4 T},\end{aligned} \tag{102}$$

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

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

Throughout this case, we assume that

$$\eta_0^{(T)} \leq \frac{1}{75}. \tag{103}$$

In fact, if $\eta_0^{(T)} > 1/75$, then the scenario becomes much easier to cope with. To see this, applying the previous result (87) and recalling the choice (27) of p immediately yield

$$V^\star(2) - \mathbb{E}\big[V_T(2)\big] \geq \frac{\eta_0^{(T)}}{1-\gamma p} > \frac{1}{100(1-\gamma)}, \tag{104}$$

which together with (29) and the assumption $T \geq \frac{1}{(1-\gamma)^2}$ yields

$$\mathbb{E}\big[\big(V^\star(2) - V_T(2)\big)^2\big] \geq \big(V^\star(2) - \mathbb{E}\big[V_T(2)\big]\big)^2 \geq \frac{1}{10000(1-\gamma)^2} \geq \frac{1}{10000(1-\gamma)^4 T}. \tag{105}$$

We now turn our attention to the dynamics w.r.t. state 1 and its associated value function $V_t(1)$ under the condition (103).

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

$$\widehat{Q}_t(a) = (1-\eta_t)\widehat{Q}_{t-1}(a) + \eta_t\big\{1 + \gamma P_t(1 \,|\, 1, a)\widehat{V}_{t-1}\big\}, \tag{106}$$

with

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

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

$$Q_t(1,a) \geq \widehat{Q}_t(a) - \frac{1}{1-\gamma}\prod_{i=1}^{t}\big(1-\eta_i(1-\gamma)\big), \qquad a\in\{1,2\}, \tag{107}$$

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

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

$$\overline{Q}_t = (1-\eta_t)\overline{Q}_{t-1} + \eta_t\big\{1+\gamma P_t(1\,|\,1,1)\overline{Q}_{t-1}\big\} \qquad \text{and} \qquad \overline{Q}_0 = V^\star(1) = \frac{1}{1-\gamma p}, \tag{108}$$

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

$$\widehat{Q}_t(1) \geq (1-\eta_t)\widehat{Q}_{t-1}(1) + \eta_t\big\{1+\gamma P_t(1\,|\,1,1)\widehat{Q}_{t-1}(1)\big\} \geq \overline{Q}_t, \tag{109}$$

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

A useful lower bound on the auxiliary sequence (106). In what follows, let us establish a useful lower bound on the sequence (106) introduced above. Then we claim that there exists some $\tau \leq T$ (see (123) and (125)) such that

$$\mathbb{P}\Big\{\widehat{V}_t \geq \frac{1}{4(1-\gamma)}\Big\} \geq \frac{1}{2}, \qquad \text{for } t\geq\tau. \tag{110}$$

The auxiliary sequence constructed in (108) plays a crucial role in establishing this claim.

Proof of the claim (110). We intend to employ the sequence $\overline{Q}_t$ (cf. (108)) to help control $\widehat{V}_t$. It is first observed that the sequence $\overline{Q}_t$ admits the following decomposition (akin to the derivation in (79))

$$\begin{aligned}
\overline{Q}_t &= \big(1-\eta_t(1-\gamma p)\big)\overline{Q}_{t-1} + \eta_t\big\{1+\gamma\big(P_t(1\,|\,1,1)-p\big)\overline{Q}_{t-1}\big\} \\
&= \prod_{i=1}^{t}\big(1-\eta_i(1-\gamma p)\big)\overline{Q}_0 + \sum_{k=1}^{t}\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i(1-\gamma p)\big)\big\{1+\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1}\big\} \\
&= \eta_0^{(t)}\frac{1}{1-\gamma p} + \sum_{k=1}^{t}\eta_k^{(t)} + \sum_{k=1}^{t}\eta_k^{(t)}\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1} \\
&= \frac{1}{1-\gamma p} + \sum_{k=1}^{t}\underbrace{\eta_k^{(t)}\gamma\big(P_k(1\,|\,1,1)-p\big)\overline{Q}_{k-1}}_{=:\,z_k},
\end{aligned} \tag{111}$$

where the last line results from (72). 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

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

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 $\big|P_k(1\,|\,1,1)-p\big| \leq \max\{p,1-p\} = p$ since $p=(4\gamma-1)/(3\gamma)$ and $\gamma\geq 3/4$. We intend to invoke Freedman's inequality to control (111). Armed with these properties and the fact that $P_k(1\,|\,1,1)$ is a Bernoulli random variable with mean p, we obtain

$$\sum_{k=1}^{t}\mathsf{Var}\Big(z_k\,|\,P_{k-1}(1\,|\,1,1),\ldots,P_1(1\,|\,1,1)\Big) = \sum_{k=1}^{t}\big(\eta_k^{(t)}\big)^2 p(1-p)\big(\gamma\overline{Q}_{k-1}\big)^2$$

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

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 (27)), whereas the last inequality results from the following condition (derived through (72))

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

Applying Freedman's inequality (see (146)) then yields

$$\mathbb{P}\left\{\left|\sum_{k=1}^{t} z_k\right| \geq \sqrt{\frac{4\max_{1\leq k\leq t}\eta_k^{(t)}}{(1-\gamma)^2}\log\frac{2}{\delta}} + \frac{4\max_{1\leq k\leq t}\eta_k^{(t)}}{3(1-\gamma)}\log\frac{2}{\delta}\right\} \leq \delta. \tag{112}$$

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

$$\max_{1\leq k\leq t} \eta_k^{(t)} \leq \frac{1}{50}, \qquad \text{for } t \geq \tau, \tag{113}$$

whose proof is postponed to Section B.3.4. In light of this claim, setting $\delta = 1/2$ in the expression (112) yields

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

with probably at least $1/2$. Combining this with the decomposition (111) and the property (109), we arrive at

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

with probability at least $1/2$, where the last identity relies on the choice of p (see the definition (27)). This establishes the advertised claim (110). □

Main proof. With the property (110) in place, we are positioned to prove our main result. Towards this, we find it convenient to define

$$\Delta_t(a) \coloneqq \widehat{Q}_t(a) - Q^\star(1,a), \qquad a = 1,2; \tag{114a}$$

$$\Delta_{t,\mathsf{max}} \coloneqq \max_a \Delta_t(a). \tag{114b}$$

The goal is thus to control $\Delta_{T,\mathsf{max}}$; in fact, we intend to show that $\Delta_{T,\mathsf{max}}$ is in expectation excessively large, resulting in an "over-estimation" issue that hinders convergence. Towards this, it follows from the iterative update rule (106) that

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

Here, the second line comes from the Bellman equation $Q^\star(1,a) = 1 + \gamma p V^\star(1)$, whereas the last line holds since $\widehat{V}_{t-1} - V^\star(1) = \max_a\big(\widehat{Q}_{t-1}(a) - V^\star(1)\big) = \max_a \Delta_{t-1}(a)$ (as a consequence of the relation (28)). Applying the above relation recursively leads to

$$\Delta_t(a) = \sum_{k=1}^{t}\eta_k \prod_{i=k+1}^{t}(1-\eta_i)\gamma\Big(p\Delta_{k-1,\mathsf{max}} + \big(P_k(1\,|\,1,a)-p\big)\widehat{V}_{k-1}\Big), \tag{115}$$

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

$$\xi_t(a) \coloneqq \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma\big(P_k(1\,|\,1,a) - p\big)\widehat{V}_{k-1}, \tag{116a}$$

$$\xi_{t,\mathsf{max}} \coloneqq \max_a \xi_t(a), \tag{116b}$$

one can express the above relation as follows

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

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

$$\mathbb{E}[\xi_{t,\mathsf{max}}] \geq \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \qquad \text{for all } t \geq \widehat{\tau} \tag{117}$$

for some universal constant $c > 0$, where

$$\widehat{\tau} \coloneqq \max\left\{\tau' \;\Big|\; \prod_{i=\tau'}^{T} \big(1-\eta_i(1-\gamma p)\big) \leq \frac{6}{7}\right\}, \tag{118}$$

whose existence is ensured under the condition (103). Given the validity of this claim (which we shall justify in Section B.3.4), we immediately arrive at

$$\mathbb{E}[\Delta_{t,\mathsf{max}}] \geq \sum_{k=1}^{t} \eta_k \prod_{i=k+1}^{t} (1-\eta_i)\gamma p \mathbb{E}[\Delta_{k-1,\mathsf{max}}] + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}} \qquad \text{for all } t \geq \widehat{\tau}. \tag{119}$$

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

$$x_t = (1-\eta_t)x_{t-1} + \eta_t\left(\gamma p x_{t-1} + \frac{c}{\sqrt{(1-\gamma)^2 T \log T}}\right) \tag{120}$$

with $x_{\widehat{\tau}} = 0$, where we recall the definition of $\widehat{\tau}$ in (118). In comparison to the iterative relation (119) which starts from $\mathbb{E}[\Delta_{0,\mathsf{max}}] = 0$ (and hence $\mathbb{E}[\Delta_{t,\mathsf{max}}] \geq 0$), we let the sequence x_t start from $x_{\widehat{\tau}} = 0$, where $\widehat{\tau}$ is defined in (118). It is straightforward to verify that

$$\mathbb{E}\big[\Delta_{T,\mathsf{max}}\big] \geq x_T, \tag{121}$$

recognizing that

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

A little algebra reveals that the sequence (120) obeys

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

where the second equality arises from (73), and the last inequality holds as long as $\prod_{i=\widehat{\tau}}^{T} \big(1-\eta_i(1-\gamma p)\big) \leq 6/7$ (see (133)). This taken together with (121) leads to

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

Combining the above bound with (107) leads to

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

Taking this together with (80), we arrive at

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

This combined with (29) establishes the following desired lower bound:

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

B.3.4 Proofs of auxiliary results

Proof of the inequality (107). We shall establish this claim by induction. To begin with, the inequality (107) 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 (107) for the $(t-1)$-th iteration and the definitions of V_{t-1} and $\widehat{V}_{t-1}$, we have

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

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

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

where the first inequality invokes the induction hypothesis for the $(t-1)$-th iteration. This establishes (107) for the t-th iteration, and hence the proof is complete via an induction argument.

Proof of the claim (113). When taking the constant learning rates $\eta_t \equiv \eta \leq \frac{1}{(1-\gamma)^2 T} \leq \frac{1}{50}$ (under the condition $T \geq \frac{50}{(1-\gamma)^2}$), one has

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

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

It then suffices to look at rescaled linear learning rates (i.e., $\eta_t = \frac{1}{1+c_\eta(1-\gamma)t}$). As already calculated in the expression (98), the ratio of two consecutive quantities obeys

$$\frac{\eta_{k-1}^{(t)}}{\eta_k^{(t)}} = \frac{1-\frac{4}{3}(1-\gamma)\eta_k}{1-c_\eta(1-\gamma)\eta_k}. \tag{122}$$

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

$$\tau := \min\left\{\tau' \;\middle|\; \prod_{i=1}^{\tau'} \big(1-\eta_i(1-\gamma p)\big) \leq \frac{1}{50}\right\}, \tag{123}$$

  which clearly satisfies $\tau \leq T$ (in view of (103)). For all $t \geq \tau$, one has

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

  At the same time, we claim that one must have

$$\prod_{i=\tau}^{T} \big(1-\eta_i(1-\gamma p)\big) \leq \frac{2}{3}. \tag{124}$$

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

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

  which contradicts our assumption that $\eta_0^{(T)} > 1/75$ (cf. (103)).

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

$$\tau := \left\lceil \frac{49}{c_\eta(1-\gamma)} \right\rceil < \frac{50}{(1-\gamma)^2} < T, \tag{125}$$

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

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

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

$$\prod_{i=\tau}^{T} \big(1-\eta_i(1-\gamma p)\big) \leq \Big(1-\frac{1-\gamma}{100}\Big)^{T-\tau+1} \leq \Big(1-\frac{1-\gamma}{100}\Big)^{\frac{100}{(1-\gamma)4/3}} \leq \frac{3}{4}. \tag{126}$$

- Putting these two cases together (with τ specified in (123) and (125)), we obtain

$$\max_{1\leq k\leq t} \eta_k^{(t)} \leq \frac{1}{50} \tag{127}$$

  for all $t \geq \tau$, thus establishing the desired inequality (113).

Proof of the inequality (117). For every t, recalling the definition (116), it is convenient to write

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

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

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

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

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

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

$$\mathbb{P}\left\{|\zeta_t|\geq\sqrt{\frac{8\log\frac{2}{\delta}}{3(1-\gamma)}\sum_{k=1}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2}+\frac{4\eta_t\log\frac{2}{\delta}}{3(1-\gamma)}\right\}\leq\delta. \tag{128}$$

To verify this relation, we first notice that

$$|z_k|\leq\max_{1\leq k\leq t}\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big)\cdot\frac{1}{1-\gamma}\leq\frac{\eta_t}{1-\gamma}, \tag{129}$$

provided that $\max_k \eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big)\leq\eta_t$. To verify the condition $\max_k \eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big)\leq\eta_t$, one can check — similar to (98) — that

$$\frac{\eta_{k-1}\prod_{i=k}^{t}\big(1-\eta_i\big)}{\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big)}=1-\frac{\big(1-c_\eta(1-\gamma)\big)\eta_k}{1-c_\eta(1-\gamma)\eta_k}\leq 1, \tag{130}$$

which indicates that $\eta_k\prod_{i=k+1}^{t}\big(1-\eta_i\big)$ is an increasing sequence as long as $c_\eta\leq\log T\leq\frac{1}{1-\gamma}$ (see (69)). In addition to the boundedness condition (129), we can further calculate

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

where the last inequality comes from the facts that $\widehat{V}_{k-1}\leq\frac{1}{1-\gamma}$ and the choice $p=\frac{4\gamma-1}{3\gamma}$. These bounds taken together with Freedman's inequality (see (146)) validate (128).

By virtue of (128), setting $\delta=\frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big]$ yields that with probability at least $1-\delta$,

$$|\zeta_t|\leq B\coloneqq\sqrt{\frac{8\log\frac{2}{\delta}}{3(1-\gamma)}\sum_{k=1}^{t}\eta_k^2\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]^2}+\frac{4\eta_t\log\frac{2}{\delta}}{3(1-\gamma)}\quad\text{with }\delta=\frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big]. \tag{131}$$

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

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

$$
\begin{aligned}
&= \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \mathbb{E}\big[|\zeta_t|^2 \mathbb{1}\big(|\zeta_t| > B\big)\big]\Big\} \\
&\overset{\text{(i)}}{\geq} \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \frac{1}{(1-\gamma)^2}\mathbb{P}\big\{|\zeta_t| > B\big\}\Big\} \\
&\geq \frac{1}{2B}\Big\{\mathbb{E}[|\zeta_t|^2] - \frac{\delta}{(1-\gamma)^2}\Big\} \overset{\text{(ii)}}{\geq} \frac{1}{4B}\mathbb{E}[|\zeta_t|^2].
\end{aligned} \tag{132}
$$

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 (129) and (40); (ii) holds by the choice of δ. It is thus sufficient to lower bound $\mathbb{E}[|\zeta_t|^2]$. Towards this, let us define

$$
\widehat{\tau} \coloneqq \max\left\{\tau' \;\middle|\; \prod_{i=\tau'}^{T} \big(1-\eta_i(1-\gamma p)\big) \leq \frac{6}{7}\right\}, \tag{133}
$$

which clearly satisfies $\tau \leq \widehat{\tau} \leq T$ (in view of (124) and (126)). Then, for all $t \geq \widehat{\tau}$ one has (which shall be proved towards the end of this subsection)

$$
\sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \geq \frac{1}{8}\sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2. \tag{134}
$$

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

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

where the first line relies on (83), and the last step makes use of the fact (110). To further control the right-hand side of the above inequality, we take $\tau' \coloneqq \max\left\{t - \frac{1}{\eta_{t/2}}, 1\right\}$ and show that

$$
\begin{aligned}
\mathbb{E}\big[|\zeta_t|^2\big] &\overset{\text{(i)}}{\geq} \frac{1}{2}\sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \gamma^2 \cdot 2p(1-p)\frac{1}{16(1-\gamma)^2} \\
&\geq \frac{1}{48(1-\gamma)}\sum_{k=\tau}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \overset{\text{(ii)}}{\geq} \frac{1}{400(1-\gamma)}\sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \\
&\geq \frac{1}{400(1-\gamma)}\sum_{k=\tau'}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \overset{\text{(iii)}}{\geq} \frac{\eta_t}{9600(1-\gamma)}.
\end{aligned} \tag{135}
$$

Here, (i) makes use of the constraint $\widehat{V}_{k-1} \geq \frac{1}{4(1-\gamma)}$, while (ii) makes use of (134), and (iii) are valid if the following property holds (which shall be proved towards the end of this subsection)

$$
\sum_{k=\tau'}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2 \geq \frac{1}{24}\eta_t. \tag{136}
$$

We are now well-equipped to control $\mathbb{E}[\xi_{t,\mathsf{max}}]$ using the property (132). Recall the expression of B in (131), we know that bounding $\mathbb{E}[|\zeta_t|^2]/B$ boils down to controlling

$$
\frac{\mathbb{E}[|\zeta_t|^2]}{\frac{\eta_t}{1-\gamma}\log\frac{2}{\delta}} \qquad \text{and} \qquad \frac{\mathbb{E}[|\zeta_t|^2]}{\sqrt{\frac{\log\frac{2}{\delta}}{1-\gamma}\sum_{k=1}^{t} \eta_k^2 \Big[\prod_{i=k+1}^{t} (1-\eta_i)\Big]^2}}. \tag{137}
$$

- For the first term in (137), recalling that $\delta = \frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big]$, we can demonstrate that

$$\log\frac{1}{\delta} = -\log\frac{(1-\gamma)^2}{2}\mathbb{E}\big[|\zeta_t|^2\big] \le -\log\frac{(1-\gamma)\eta_t}{19200} \le \log\frac{19200(1+(1-\gamma)T\log T)}{1-\gamma} \lesssim \log T, \tag{138}$$

  where the first inequality makes use of the bound (135), and the second inequality arises from the fact $\eta_t \ge \frac{1}{1+(1-\gamma)T\log T}$ (given the range of the learning rates in this case). Combining this with (135), we can guarantee that

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

- Moving to the second term in (137), one can ensure that

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

  Here, (i) follows from (135) and (138) since

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

  (ii) arises from (136); 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 (132) and using the expression of B in (131), we reach at

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

for some constant $c > 0$. Thus, this validates the inequality (117).

Proof of the claim (134). By the definition of $\widehat{\tau}$ in (133), we have $\prod_{i=\widehat{\tau}}^{T}\big(1-\eta_i(1-\gamma p)\big) \le 6/7$. An important observation is that

$$\sum_{k=\tau}^{t}\eta_k\left[\prod_{i=k+1}^{t}(1-\eta_i)\right] \overset{\text{(i)}}{=} 1-\prod_{i=\tau}^{t}(1-\eta_i) \overset{\text{(ii)}}{\ge} \frac{1}{8} \overset{\text{(iii)}}{\ge} \frac{1}{8}\sum_{k=1}^{t}\eta_k\left[\prod_{i=k+1}^{t}(1-\eta_i)\right]. \tag{139}$$

Here, the relations (i) and (iii) arise from (41), and the inequality (ii) follows since

$$\prod_{i=\tau}^{t}(1-\eta_i) \le \prod_{i=\tau}^{\widehat{\tau}}(1-\eta_i) \le \prod_{i=\tau}^{\widehat{\tau}}\big(1-\eta_i(1-\gamma p)\big) = \frac{\prod_{i=\tau}^{T}\big(1-\eta_i(1-\gamma p)\big)}{\prod_{i=\widehat{\tau}+1}^{T}\big(1-\eta_i(1-\gamma p)\big)} \le \frac{3/4}{6/7} \le \frac{7}{8}, \tag{140}$$

where τ is defined in (123) and (125) for linearly rescaled learning rates and $\tau = 1$ for constant learning rates, and we have also made use of (103), (124) and (126) in the penultimate inequality in (140).

With (139) in place, we can continue to prove the claim (134). Recognizing that $\eta_k\big[\prod_{i=k+1}^{t}(1-\eta_i)\big]$ is increasing in k (see (130)), we can obtain

$$\sum_{k=1}^{\tau-1}\eta_k^2\Big[\prod_{i=k+1}^{t}(1-\eta_i)\Big]^2 \le \max_{1\le k<\tau}\left\{\eta_k\Big[\prod_{i=k+1}^{t}(1-\eta_i)\Big]\right\}\sum_{k=1}^{\tau-1}\eta_k\Big[\prod_{i=k+1}^{t}(1-\eta_i)\Big]$$

$$\leq \eta_\tau \Big[\prod_{i=\tau+1}^{t}\big(1-\eta_i\big)\Big]\sum_{k=1}^{\tau-1}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big]$$
$$\leq 7\eta_\tau \Big[\prod_{i=\tau+1}^{t}\big(1-\eta_i\big)\Big]\sum_{k=\tau}^{t}\eta_k\Big[\prod_{i=k+1}^{t}\big(1-\eta_i\big)\Big], \tag{141}$$

where the last inequality comes from (139). With the preceding inequality in place, the claim (134) then follows by observing that

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

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

Proof of the claim (136). Note that for $\tau' \coloneqq \max\Big\{t-\frac{1}{\eta_{t/2}}, 1\Big\}$, one has

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

as long as the following condition holds (recalling the definition of $\widehat{\tau}$ in (133))

$$\eta_{t/2} \leq 2\eta_t \leq 2\eta_{\widehat{\tau}} \leq 1/10. \tag{142}$$

In addition, similar to (73), we can derive

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

where we once again use the condition (142), and the last inequality comes from the derivation in (140). Putting these two bounds together yields

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

To finish up, it remains to justify (142). This condition is obvious for constant learning rates. As for rescaled learning rates, one can see that

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

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

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

provided that $T\geq\frac{c_1}{(1-\gamma)^2}$ for some sufficiently large constant $c_1>0$ and $1-\gamma<c_\eta<\log T$. Taking this together with (133) implies that $\widehat{\tau}\geq\overline{\tau}$ and hence $\eta_{\widehat{\tau}}\leq\eta_{\overline{\tau}}=\frac{1}{1+(1-\gamma)c_\eta\overline{\tau}}=1/20$. □

B.4 Proof of Lemma 3

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

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

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

$$Q^{\star}(1,1) = Q^{\star}(1,2) = V^{\star}(1). \tag{143}$$

From Bellman's equation, we can thus deduce that

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

which in conjunction with (143) and a little algebra leads to

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

Here, the second identity follows since $V^{\star}(0) = 0$, and the third identity makes use of (27). The calculation for $V^{\star}(2)$ and $Q^{\star}(2,1)$ follows from an identical argument and is hence omitted.

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

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

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

C 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 4. *Suppose that $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

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

Define

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

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

$$\mathbb{P}\left\{|Y_n| \ge \tau \text{ and } W_n \le \sigma^2\right\} \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{144}$$

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

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}. \tag{145}$$

Proof. See Freedman [1975], Tropp [2011] for the proof of (144). As an immediate consequence of (144), one has

$$\mathbb{P}\left\{|Y_n| \ge \sqrt{4\sigma^2\log\frac{2}{\delta}} + \frac{4}{3}R\log\frac{2}{\delta} \text{ and } W_n \le \sigma^2\right\} \le \delta. \tag{146}$$

Next, we turn attention to (145). Consider any positive integer K. As can be easily seen, the event

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

is contained within the union of the following K events

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

where we define

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

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

Invoking inequality (146) with σ^2 set to be $\frac{\sigma^2}{2^{k-1}}$ and δ set to be $\frac{\delta}{K}$, we arrive at $\mathbb{P}\{\mathcal{B}_k\} \leq \delta/K$. Taken this fact together with the union bound gives

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

This concludes the proof. □

References

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory*, pages 67–83, 2020.

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

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

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

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

R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

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

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

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

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

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

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

Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

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

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

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

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

D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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

H. Hasselt. Double Q-learning. *Advances in neural information processing systems*, 23:2613–2621, 2010.

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

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

S. Kakade. *On the sample complexity of reinforcement learning.* PhD thesis, University of London, 2003.

M. Kearns, Y. Mansour, and A. Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.

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

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

D. Lee and N. He. Stochastic primal-dual Q-learning. *arXiv preprint arXiv:1810.08298*, 2018.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020a.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

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

S. Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.

G. Qu and A. Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conference on Learning Theory*, pages 3185–3205, 2020.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

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

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

J. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16: 262–270, 2011.

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

H.-T. Wai, M. Hong, Z. Yang, Z. Wang, and K. Tang. Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32:5784–5795, 2019.

M. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019a.

M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019b.

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

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

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

C. J. C. H. Watkins. Learning from delayed rewards. 1989.

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

W. Weng, H. Gupta, N. He, L. Ying, and R. Srikant. The mean-squared error of double Q-learning. *Advances in Neural Information Processing Systems*, 33, 2020b.

Y. Wu, W. Zhang, P. Xu, and Q. Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

H. Xiong, L. Zhao, Y. Liang, and W. Zhang. Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, 33, 2020.

P. Xu and Q. Gu. A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020.

T. Xu, Z. Wang, Y. Zhou, and Y. Liang. Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*, 2019a.

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