Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, Yuejie Chi

informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 72, No. 1, January–February 2024, pp. 222–236
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

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

Gen Li,[a] Changxiao Cai,[b] Yuxin Chen,[a] Yuting Wei,[a,*] Yuejie Chi[c]

[a] Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Department of Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

*Corresponding author

Contact: ligen@wharton.upenn.edu (GL); changxiao.cai@pennmedicine.upenn.edu (CC); yuxinc@wharton.upenn.edu, https://orcid.org/0000-0001-9256-5815 (YuxC); ytwei@wharton.upenn.edu, https://orcid.org/0000-0003-1488-4647 (YW); yuejiechi@cmu.edu, https://orcid.org/0000-0002-6766-5459 (YueC)

Received: November 27, 2021
Revised: October 25, 2022
Accepted: March 16, 2023
Published Online in Articles in Advance: April 21, 2023

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2023.2450

Copyright: © 2023 The Author(s)

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 toward understanding the sample efficiency of Q-learning. Consider a γ-discounted infinite-horizon MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$: to yield an entry-wise ε-approximation of the optimal Q-function, state-of-the-art theory for Q-learning requires a sample size exceeding the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$, which fails to match existing minimax lower bounds. This gives rise to natural questions: What is the sharp sample complexity of Q-learning? Is Q-learning provably suboptimal? This paper addresses these questions for the synchronous setting: (1) when the action space contains a single action (so that Q-learning reduces to TD learning), we prove that the sample complexity of TD learning is minimax optimal and scales as $\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}$ (up to log factor); (2) when the action space contains at least two actions, we settle the sample complexity of Q-learning to be on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ (up to log factor). Our theory unveils the strict suboptimality of Q-learning when the action space contains at least two actions and rigorizes the negative impact of overestimation in Q-learning. Finally, we extend our analysis to accommodate asynchronous Q-learning (i.e., the case with Markovian samples), sharpening the horizon dependency of its sample complexity to be $\frac{1}{(1-\gamma)^4}$.

Funding: Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the Air Force Office of Scientific Research [Grant FA9550-22-1-0198], the Office of Naval Research (ONR) [Grant N00014-22-1-2354], and the National Science Foundation (NSF) [Grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713, and IIS-2218773]. Y. Wei is supported in part by the Google Research Scholar Award and the NSF [Grants CCF-2106778, DMS-2147546/2015447, and CAREER award DMS-2143215]. Y. Chi is supported in part by the ONR [Grants N00014-18-1-2142 and N00014-19-1-2404] and the NSF [Grants CCF-1806154, CCF-2007911, CCF-2106778, ECCS-2126634, and DMS-2134080].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.2450.

Keywords: Q-learning • temporal difference learning • effective horizon • sample complexity • minimax optimality • lower bound • overestimation

1. Introduction

Q-learning is arguably one of the most widely adopted model-free algorithms (Watkins 1989, Watkins and Dayan 1992). Characterizing its sample efficiency lies at the core of the statistical foundation of reinforcement learning (RL) (Sutton and Barto 2018). Whereas classic convergence analyses for Q-learning (Jaakkola et al. 1994, Tsitsiklis 1994, Szepesvári 1998, Borkar and Meyn 2000) focus primarily on the asymptotic regime—in which the number of iterations tends to infinity with other problem parameters held fixed—recent years have witnessed a paradigm shift from asymptotic analyses toward a finite-sample/-time framework (Kearns and Singh 1999; Even-Dar and Mansour 2003; Beck and Srikant 2012; Lee and He 2018; Wainwright 2019b; Chen et al. 2020, 2021; Qu and Wierman 2020; Weng et al. 2020a; Xiong et al. 2020; Li et al. 2022b). Drawing insights from high-dimensional statistics (Wainwright 2019a), a modern nonasymptotic framework unveils more clear and informative impacts of salient problem parameters upon the sample complexity, particularly for those applications with an enormous state/action space and long horizon. Motivated by its practical value, a suite of nonasymptotic theory has been recently developed for Q-learning to accommodate multiple sampling mechanisms (Even-Dar and Mansour 2003, Beck and Srikant 2012, Jin et al. 2018, Wainwright 2019b, Qu and Wierman 2020, Li et al. 2022b).

In this paper, we revisit the sample complexity of Q-learning for tabular Markov decision processes (MDPs). For concreteness, let us consider the synchronous setting, which assumes access to a generative model or a simulator that produces independent samples for all state–action pairs in each iteration (Kearns et al. 2002, Kakade 2003); this setting is termed "synchronous" as the estimates with respect to (w.r.t.) all state–action pairs are updated at once. We investigate the ℓ_∞-based sample complexity, namely, the number of samples needed for synchronous Q-learning to yield an entry-wise ε-accurate estimate of the optimal Q-function. Despite a number of prior works tackling this setting, the dependence of the sample complexity on the effective horizon $\frac{1}{1-\gamma}$ remains unsettled. Take γ-discounted infinite-horizon MDPs for instance: the state-of-the-art sample complexity bounds (Wainwright 2019b, Chen et al. 2020) scale on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ (up to some log factor), where $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively. However, it is unclear whether this scaling is sharp for Q-learning and whether it can be further improved via a more refined theory. On the one hand, the minimax lower limit for this setting is shown to be on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ (up to some log factor) (Azar et al. 2013); this limit is achievable by model-based approaches (Agarwal et al. 2020, Li et al. 2023b) and apparently smaller than prior sample complexity bounds for Q-learning. On the other hand, Wainwright (2019c) argues through numerical experiments that "the usual Q-learning suffers from at least worst-case fourth-order scaling in the discount complexity $\frac{1}{1-\gamma}$, as opposed to the third-order scaling …" although no rigorous justification is provided therein. Given the gap between the achievability bounds and lower bounds in the status quo, it is natural to seek answers to the following questions: What is the tight sample complexity characterization of Q-learning? How does it compare to the minimax sample complexity limit?

1.1. Main Contributions

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

- When $|\mathcal{A}| = 1$, Q-learning coincides with temporal difference (TD) learning in a Markov reward process. For any $0 < \varepsilon < 1$, we prove that a total sample size of

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

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

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

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

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

$$\tilde{\Omega}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right) \tag{3}$$

iterations to achieve ε-accuracy in an ℓ_∞ sense; see Theorem 3. These two theorems taken collectively lead to the first sharp characterization of the sample complexity of Q-learning, strengthening prior theory (Wainwright 2019b, Chen et al. 2020) by a factor of $\frac{1}{1-\gamma}$. In addition, the discrepancy between our sharp characterization and the minimax lower bound makes clear that Q-learning is *not* minimax optimal when $|\mathcal{A}| \geq 2$ and is outperformed by, say, the model-based approaches (Agarwal et al. 2020, Li et al. 2023b) in terms of the sample efficiency.

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

Table 1. Comparisons of Existing Sample Complexity Upper Bounds of Synchronous Q-Learning and TD Learning for an Infinite-Horizon γ-Discounted MDP with State Space $\mathcal{S}$ and Action Space $\mathcal{A}$, Where $0 < \varepsilon < 1$ is the Target Accuracy Level

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: $\frac{1}{t^\omega}$, $\omega \in (1/2, 1)$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\left\{\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)^{1/\omega} + \left(\frac{1}{1-\gamma}\right)^{\frac{1}{1-\omega}}\right\}$
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: $\frac{1}{t^\omega}$, $\omega \in (0, 1)$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\left\{\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)^{1/\omega} + \left(\frac{1}{1-\gamma}\right)^{\frac{1}{1-\omega}}\right\}$
Chen et al. (2020)	rescaled linear: $\frac{1}{\frac{1}{(1-\gamma)^2}+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Chen et al. (2020)	constant: $(1-\gamma)^4\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
This work (Q-learning, $\lvert\mathcal{A}\rvert \geq 2$)	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
This work (Q-learning, $\lvert\mathcal{A}\rvert \geq 2$)	constant: $(1-\gamma)^3\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
This work (TD learning, $\lvert\mathcal{A}\rvert = 1$)	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^3\varepsilon^2}$
This work (TD learning, $\lvert\mathcal{A}\rvert = 1$)	constant: $(1-\gamma)^3\varepsilon^2$	$\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^3\varepsilon^2}$

Notes. Here, sample complexity refers to the total number of samples needed to yield either $\max_{s,a}\lvert\hat{Q}(s,a) - Q^\star(s,a)\rvert \leq \varepsilon$ with high probability or $\mathbb{E}[\max_{s,a}\lvert\hat{Q}(s,a) - Q^\star(s,a)\rvert] \leq \varepsilon$, where $\hat{Q}$ is the estimate returned by Q-learning. All logarithmic factors are omitted in the table to simplify the expressions.

negative impact of overestimation on sample efficiency on a rigorous footing.

Finally, we extend our analysis framework to accommodate the asynchronous setting, in which the samples are non–independent and identically distributed (i.i.d.) and take the form of a single Markovian trajectory. To the best of our knowledge, we show for the first time that the sample complexity of asynchronous Q-learning exhibits a $\frac{1}{(1-\gamma)^4}$ scaling w.r.t. the effective horizon, which is nearly sharp and improves upon the prior state of the art Li et al. (2022b).

1.2. Related Works

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

1.2.1. Finite-Sample ℓ_∞-Based Guarantees for Synchronous Q-Learning and TD Learning. The sample complexities derived in prior literature often rely crucially on the choices of learning rates. Even-Dar and Mansour (2003) study the sample complexity of Q-learning with linear learning rates $1/t$ or polynomial learning rates $1/t^\omega$, which scale as $\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^{2.5}}\right)$ when optimized w.r.t. the effective horizon (attained when $\omega = 4/5$). The resulting sample complexity, however, is suboptimal in terms of its dependency on not only $\frac{1}{1-\gamma}$, but also the target accuracy level ε. Beck and Srikant (2012) investigate the case of constant learning rates; however, their result suffers from an additional factor of $\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$, which can be prohibitively large in practice. More recently, Wainwright (2019b) and Chen et al. (2020) further analyze the sample complexity of Q-learning with either constant learning rates or linearly rescaled learning rates, leading to the state-of-the-art bound $\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}\right)$. However, this result remains suboptimal in terms of its scaling with $\frac{1}{1-\gamma}$. See Table 1 for details. In the special case with $\lvert\mathcal{A}\rvert = 1$, the recent works Khamaru et al. (2021b) and Mou et al. (2020) develop instance-dependent results for TD learning with Polyak–Ruppert averaging and study the local (sub)-optimality of TD learning in a different local minimax framework.

1.2.2. Finite-Sample ℓ_∞-Based Guarantees for Asynchronous Q-Learning and TD Learning. Moving beyond the synchronous model, Even-Dar and Mansour (2003), Beck and Srikant (2012), Qu and Wierman (2020), Li et al. (2022b), Shah and Xie (2018), and Chen et al. (2021) develop nonasymptotic convergence guarantees for the asynchronous setting, in which the data samples take the form of a single Markovian trajectory (following some behavior policy) and only a single state–action pair is updated in each iteration. A similar scaling of $\tilde{O}\left(\frac{1}{(1-\gamma)^5}\right)$ also shows up in the state-of-the-art sample complexity bounds for asynchronous Q-learning (Li et al. 2022b), and our theory is the first to sharpen it to $\tilde{O}\left(\frac{1}{(1-\gamma)^4}\right)$. When it comes to the special case with $|\mathcal{A}| = 1$, the nonasymptotic performance guarantees for TD learning with Markovian sample trajectories (assuming that the behavior policy coincides with the target policy) are recently derived by Bhandari et al. (2021), Srikant and Ying (2019), and Mou et al. (2020).

1.2.3. Finite-Sample ℓ_∞-Based 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 are proposed and analyzed. Azar et al. (2011) propose speedy Q-learning, which achieves a sample complexity of $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right)$ at the expense of doubling the computation and storage complexity. Our result on vanilla Q-learning matches that of speedy Q-learning in an order-wise sense. In addition, Wainwright (2019c) proposes a variance-reduced Q-learning algorithm that is shown to be minimax optimal in the range $\epsilon \in (0,1)$ with a sample complexity $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$, which is subsequently generalized to the asynchronous setting by Li et al. (2022b). The ℓ_∞ statistical bounds for variance-reduced TD learning are investigated in Khamaru et al. (2021b) for the synchronous setting and in Li et al. (2022b) for the asynchronous setting. Finally, Xiong et al. (2020) establish 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 sample efficiency.

1.2.4. Others.

There are also several other strands of related papers that tackle model-free algorithms but do not pursue ℓ_∞-based nonasymptotic guarantees. For instance, Bhandari et al. (2021), Lakshminarayanan and Szepesvari (2018), Srikant and Ying (2019), Gupta et al. (2019), Doan et al. (2019), Wu et al. (2020), Xu et al. (2019a,b), and Chen et al. (2019) develop finite-sample (weighted) ℓ_2 convergence guarantees for several model-free algorithms, which also allow one to accommodate linear function approximation as well as off-policy evaluation. Another line of recent work (Jin et al. 2018, Bai et al. 2019, Zhang et al. 2020, Li et al. 2023a) considers the sample efficiency of Q-learning-type algorithms paired with proper exploration strategies (e.g., upper confidence bounds) under the framework of regret analysis. The asymptotic behaviors of some variants of Q-learning, for example, double Q-learning (Weng et al. 2020b) and relative Q-learning (Devraj and Meyn 2020) are also studied. In addition, Q-learning in conjunction with the pessimism principle proves effective in dealing with off-line data (Shi et al. 2022, Yan et al. 2022). The effect of more general function approximation schemes (e.g., certain families of neural network approximations) is studied in Fan et al. (2019), Murphy (2005), Cai et al. (2019), Wai et al. (2019), and Xu and Gu (2020), whereas the extension to multiagent scenarios is looked at in Hu and Wellman (2003) and Li et al. (2022a). These are beyond the scope of the present paper.

2. Background and Algorithms

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

2.1. Basics of Discounted Infinite-Horizon MDPs

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

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

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

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

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

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Here, $\{(s_k,a_k)\}_{k\ge 0}$ 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 $0 \le V^\pi(s) \le \frac{1}{1-\gamma}$ and $0 \le Q^\pi(s,a) \le \frac{1}{1-\gamma}$ 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

$$V^\star(s) := \max_\pi V^\pi(s), \qquad Q^\star(s,a) := \max_\pi Q^\pi(s,a)$$

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

2.2. Algorithms: Q-Learning and TD Learning (the Synchronous Setting)

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

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

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

Here, $\eta_t \in (0,1]$ denotes the learning rate or step size in the tth iteration, and $\mathcal{T}_t$ denotes the empirical Bellman operator constructed by samples collected in the tth iteration, that is,

$$\begin{aligned}\mathcal{T}_t(Q)(s,a) &:= r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'),\\ s_t &\equiv s_t(s,a) \sim P(\cdot|s,a)\end{aligned} \tag{5}$$

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

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}:$$

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

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

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

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

Algorithm 1 (Synchronous Q-Learning for Infinite-Horizon Discounted MDPs)

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

As it turns out, TD learning (Sutton 1988, Tsitsiklis and Van Roy 1997, Bhandari et al. 2021) in the synchronous setting can be viewed as a special instance of Q-learning when the action set $\mathcal{A}$ is a singleton (i.e., $|\mathcal{A}| = 1$). In such a case, the MDP reduces to a Markov reward process (MRP) (Bertsekas 2017), and we abuse the notation to use $P : \mathcal{S} \to \Delta(\mathcal{S})$ to describe the probability transition kernel and employ $r : \mathcal{S} \to [0,1]$ to represent the reward function (with $r(s)$ indicating the immediate reward gained in state s). The TD learning algorithm maintains an estimate $V_t : \mathcal{S} \to \mathbb{R}$ of the value function in each iteration t,[1] and carries out the following iterative update rule:

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

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

Algorithm 2 (Synchronous TD Learning for Infinite-Horizon Discounted MRPs)

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

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

3. Main Results: Sample Complexity of Synchronous Q-Learning

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

3.1. Minimax Optimality of TD Learning

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

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

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

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

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

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

$$\max_{s\in\mathcal{S}} |V_T(s) - V^\star(s)| \leq \varepsilon. \tag{9}$$

Remark 1 (Mean Estimation Error). This high-probability bound immediately translates to a mean estimation error guarantee. Recognizing the crude upper bound $|V_T(s) - V^\star(s)| \leq \frac{1}{1-\gamma}$ (see (EC.49) in Online Section E.C.3.1) and taking $\delta \leq \varepsilon(1-\gamma)$, we reach

$$\mathbb{E}\left[\max_s |V_T(s) - V^\star(s)|\right] \leq \varepsilon(1-\delta) + \delta\frac{1}{1-\gamma} \leq 2\varepsilon, \tag{10}$$

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

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

$$\tilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right) \tag{11}$$

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

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

$$\tilde{\Omega}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right). \tag{12}$$

In turn, this, taken together with Theorem 1, unveils the minimax optimality of the sample complexity (modulo some logarithmic factor) of TD learning for the synchronous setting. Whereas prior works demonstrate how to attain the minimax limit (12) using model-based methods or variance-reduced model-free algorithms (e.g., Azar et al. 2013, Pananjady and Wainwright 2020, Khamaru et al. 2021b, Li et al. 2023b), our theory provides the first rigorous evidence that plain TD learning alone is already minimax optimal without the need of Polyak–Ruppert averaging or variance reduction.

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

$$\frac{1}{1+\frac{\tilde{c}_1(1-\gamma)t}{\log^2(t+1)}} \leq \eta_t \leq \frac{1}{1+\frac{\tilde{c}_2(1-\gamma)t}{\log^2(t+1)}}, \qquad \forall t \geq 1 \tag{13}$$

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

$$T \geq \frac{2c_3(\log^3 T)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{(1-\gamma)^3\varepsilon^2}. \tag{14}$$

See Online Appendix EC.3.3 for the proof.

Remark 3 (Polyak-Ruppert Averaging). The results claimed in Remark 2 further allow us to control the estimation error of TD learning under Polyak–Ruppert averaging (Polyak and Juditsky 1992). More precisely, under the choice (13) of learning rates, the averaged iterates satisfy

$$\max_{s\in\mathcal{S}}\left|\frac{1}{T}\sum_{t=1}^{T} V_T(s) - V^\star(s)\right| \leq 4\sqrt{\frac{c_3(\log^3 T)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{(1-\gamma)^3 T}} \tag{15}$$

with probability exceeding $1-\delta$. See Online Appendix EC.3.3 for the proof.

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

3.2. Tight Sample Complexity and Suboptimality of Q-Learning

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

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

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^3 T}} \le \eta_t \le \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^3 T}} \tag{16a}$$

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

$$T \ge \frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4\varepsilon^2} \tag{16b}$$

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

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q_T(s,a) - Q^\star(s,a)| \le \varepsilon \tag{17}$$

with probability at least $1-\delta$.

Remark 5 (Mean Estimation Error). Repeating exactly the same argument as in Remark 1, one can readily translate this high-probability bound into the following mean estimation error guarantee:

$$\mathbb{E}\left[\max_{s,a} |Q_T(s,a) - Q^\star(s,a)|\right] \le \varepsilon(1-\delta) + \delta\frac{1}{1-\gamma} \le 2\varepsilon, \tag{18}$$

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

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

3.2.1. Sample Complexity and Sharpened Dependency on $\frac{1}{1-\gamma}$.

Recognizing that $|\mathcal{S}||\mathcal{A}|$ independent samples are drawn in each iteration, we can see from Theorem 2 the following sample complexity bound:

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \tag{19}$$

in order for Q-learning to attain ε-accuracy ($0 < \varepsilon < 1$) in an entry-wise sense. To the best of our knowledge, this is the first result that breaks the $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ barrier that is present in all state-of-the-art analyses for vanilla Q-learning (Beck and Srikant 2012, Wainwright 2019b, Chen et al. 2020, Qu and Wierman 2020, Li et al. 2022b).

3.2.2. Learning Rates.

Akin to the TD learning case, our result accommodates two commonly adopted learning rate schemes (cf. (16a)): (i) linearly rescaled learning rates $\frac{1}{1+\frac{c_2(1-\gamma)}{\log^2 T}t}$ 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 $T = \frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4\varepsilon^2}$, the constant learning rates can be taken to be on the order of

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

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.

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

$$\frac{1}{1+\frac{\tilde{c}_1(1-\gamma)t}{\log^3(t+1)}} \le \eta_t \le \frac{1}{1+\frac{\tilde{c}_2(1-\gamma)t}{\log^3(t+1)}}, \qquad \forall t \ge 1 \tag{20}$$

for some suitable constants $\tilde{c}_1, \tilde{c}_2 > 0$. Then, Claim (17) continues to hold under this choice (20) provided that $T/2 \ge \frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4\varepsilon^2}$. Additionally, similar to Remark 3, we can demonstrate that the averaged Q-learning iterates under the choice (20) of learning rates obey

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\frac{1}{T}\sum_{t=1}^{T} Q_T(s,a) - Q^\star(s,a)\right| \le 4\sqrt{\frac{c_3(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4 T}} \tag{21}$$

with probability exceeding $1-\delta$. The proofs of these results are identical to those of Remarks 2 and 3 (see Online Appendix EC.3.3) and are, hence, omitted.

3.2.3. A Matching Lower Bound and Suboptimality.

The careful reader might remark that there remains a gap between our sample complexity bound for Q-learning and the minimax lower bound (Azar et al. 2013). More specifically, the minimax lower bound scales on the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ and is achievable—up to some logarithmic factor—by the model-based approach and variance-reduced methods (Azar et al. 2013, Wainwright 2019c, Agarwal et al. 2020, Li et al. 2023b). This raises natural questions regarding whether our sample complexity bound can be further improved and whether there is any intrinsic bottleneck that prevents vanilla Q-learning from attaining optimal performance. To answer these questions, we develop the following lower bound for plain Q-learning with the aim of confirming the sharpness of Theorem 2 and revealing the suboptimality of Q-learning.

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

$$\max_{s\in\mathcal{S}} \mathbb{E}[|V_T(s) - V^\star(s)|^2] \ge \frac{c_{\mathsf{lb}}}{(1-\gamma)^4 T \log^2 T}, \tag{22}$$

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

Remark 7. This theorem constructs a hard MDP instance with no more than four states and two actions with the emphasis on unveiling the suboptimality of horizon dependency. It can be generalized to accommodate larger state/action space as we elucidate in Section 4.3.

Remark 8. Theorem 3 concentrates on two families of learning rates—rescaled linear and constant learning rates—that are most widely used in practice. Note, however, that our current analysis does not readily generalize to arbitrary learning rates, which we leave for future investigation.

Theorem 3 provides an algorithm-dependent lower bound for vanilla Q-learning. As asserted by this theorem, it is impossible for Q-learning to attain ε-accuracy (in the sense that $\max_s \mathbb{E}[|V_T(s) - V^\star(s)|^2] \le \varepsilon^2$) unless the number of iterations exceeds the order of

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

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

$$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \varepsilon^2} \quad \text{(up to log factor);} \tag{23}$$

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

4. Key Analysis Ideas (the Synchronous Case)

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

4.1. Vector and Matrix Notation

To begin, 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, that is, $\sqrt{\boldsymbol{a}} := [\sqrt{a_i}]_{i=1}^n \in \mathbb{R}^n$ and $|\boldsymbol{a}| := [|a_i|]_{i=1}^n \in \mathbb{R}^n$. For a set of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$ with $\boldsymbol{a}_k = [a_{k,j}]_{j=1}^n$ $(1 \le k \le m)$, we define the max operator in an entry-wise 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 $\boldsymbol{a} \le \boldsymbol{b}$ $(\boldsymbol{a} \ge \boldsymbol{b})$ means $a_i \le b_i$ $(a_i \ge b_i)$ for all $1 \le i \le n$. We also let $\boldsymbol{a} \circ \boldsymbol{b} = [a_i b_i]_{i=1}^n$ denote the Hadamard product. In addition, we denote by $\boldsymbol{1}$ $(\boldsymbol{e}_i)$ the all-one vector (ith standard basis vector), and let $\boldsymbol{I}$ be the identity matrix.

We 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 $\boldsymbol{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ $(\boldsymbol{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|})$ induced by a deterministic policy π over the state–action pairs (states) as follows:

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \quad \text{and} \quad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}, \tag{24}$$

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

$$\boldsymbol{\Pi}^\pi = \begin{pmatrix} \boldsymbol{e}_{\pi(1)}^\top & & & \\ & \boldsymbol{e}_{\pi(2)}^\top & & \\ & & \ddots & \\ & & & \boldsymbol{e}_{\pi(|\mathcal{S}|)}^\top \end{pmatrix} \tag{25}$$

with $\boldsymbol{e}_i$ the ith 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:

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V})\circ(\boldsymbol{P}\boldsymbol{V}). \tag{26}$$

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 employ the vectors $\boldsymbol{V}^{\pi} \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}^{\star} \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^{\pi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{Q}^{\star} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent V^{π}, $V^{\star}$, V_t, Q^{π}, $Q^{\star}$, and Q_t, respectively. Additionally, we define π_t to be the policy associated with Q_t such that, for any state–action pair (s, a),

$$\pi_t(s) = \min\left\{a' | Q_t(s,a') = \max_{a''} Q_t(s,a'')\right\}. \tag{27}$$

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

$$Q_t(s,\pi_t(s)) = V_t(s) \quad \text{and} \quad \boldsymbol{P}\boldsymbol{V}_t = \boldsymbol{P}^{\pi_t}\boldsymbol{Q}_t \geq \boldsymbol{P}^{\pi}\boldsymbol{Q}_t \tag{28}$$

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

$$\boldsymbol{P}_t((s,a),s') := \begin{cases} 1, & \text{if } s' = s_t(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{29}$$

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

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

4.2. Proof Outline for Theorem 2

We are now positioned to describe how to establish Theorem 2, toward which we first express the Q-learning update rules (4) and (5) using the preceding matrix notation. As can be easily verified, Q-learning employs the samples in $\boldsymbol{P}_t$ (cf. (29)) to perform the following update:

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1}) \tag{30}$$

in the tth iteration. In the sequel, we denote by

$$\boldsymbol{\Delta}_t := \boldsymbol{Q}_t - \boldsymbol{Q}^{\star} \tag{31}$$

the error of the Q-function estimate in the tth iteration.

4.2.1. Basic Decomposition.

We start by decomposing the estimation error term $\boldsymbol{\Delta}_t$. In view of the update rule (30), we arrive at the following elementary decomposition:

$$\begin{aligned}\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^{\star} &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1}) - \boldsymbol{Q}^{\star} \\ &= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^{\star}) + \eta_t(\boldsymbol{r} + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{Q}^{\star}) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^{\star}) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\{\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^{\star}) + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\},\end{aligned} \tag{32}$$

where the third line exploits the Bellman equation $\boldsymbol{Q}^{\star} = \boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{V}^{\star}$. Further, the term $\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^{\star})$ can be linked with $\boldsymbol{\Delta}_{t-1}$ using the definition (27) of π_t as follows:

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^{\star}) &= \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^{\star}}\boldsymbol{Q}^{\star} \\ &\leq \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^{\star} = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1},\end{aligned} \tag{33a}$$

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^{\star}) &= \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^{\star}}\boldsymbol{Q}^{\star} \\ &\geq \boldsymbol{P}^{\pi^{\star}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^{\star}}\boldsymbol{Q}^{\star} = \boldsymbol{P}^{\pi^{\star}}\boldsymbol{\Delta}_{t-1},\end{aligned} \tag{33b}$$

where we have made use of Relation (28). Substitute (33) into (32) to reach

$$\begin{aligned}\boldsymbol{\Delta}_t &\leq (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\{\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\}; \\ \boldsymbol{\Delta}_t &\geq (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\{\boldsymbol{P}^{\pi^{\star}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\}.\end{aligned} \tag{34}$$

Applying these relations recursively, we obtain

$$\begin{aligned}\boldsymbol{\Delta}_t &\leq \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\{\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\}, \\ \boldsymbol{\Delta}_t &\geq \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\{\boldsymbol{P}^{\pi^{\star}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\},\end{aligned} \tag{35}$$

where we define

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

4.2.1.1. *Comparisons to Prior Approaches.*

We take a moment to discuss how prior analyses handle the preceding elementary decomposition. Several prior works (e.g., Wainwright 2019b, Li et al. 2022b) tackle the second term on the right-hand side of Relation (34) via the following crude bounds:

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

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

4.2.2. Key Intertwined Relations Underlying $\{\|\boldsymbol{\Delta}_t\|_{\infty}\}$.

By exploiting the crucial relations (35) derived earlier, we proceed to upper and lower bound $\boldsymbol{\Delta}_t$ separately. To be more specific, defining

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

for some constant $c_4 > 0$, one can further decompose the upper bound in (35) into several terms:

$$\boldsymbol{\Delta}_t \le \underbrace{\eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{(1-\beta)t}\eta_i^{(t)}\gamma(\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1})}_{=:\boldsymbol{\zeta}_t} \tag{38}$$

$$+ \underbrace{\sum_{i=(1-\beta)t+1}^{t}\eta_i^{(t)}\gamma(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}}_{=:\boldsymbol{\xi}_t} + \sum_{i=(1-\beta)t+1}^{t}\eta_i^{(t)}\gamma\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}. \tag{39}$$

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 (1975) inequality because of its martingale structure, contributes to the main variance term in the recursion. Note, however, that the resulting variance term also depends on $\{\boldsymbol{\Delta}_i\}$.

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

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

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

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

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

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

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

Proof. The proofs of Lemmas 1 and 2 are deferred to Online Appendices EC.2.2 and EC.2.3, respectively. As a remark, our analysis collects all the error terms accrued through the iterations—instead of bounding them individually—by conducting a high-order nonlinear expansion of the estimation error through recursion, followed by careful control of the main variance term leveraging the structure of the discounted MDP. □

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

$$\|\boldsymbol{\Delta}_t\|_\infty \le 30\sqrt{\frac{(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{\gamma^2(1-\gamma)^4 T}}\left(1 + \max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\right) \tag{40}$$

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

$$\|\boldsymbol{\Delta}_T\|_\infty \le O\left(\sqrt{\frac{(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4 T}} + \frac{(\log^4 T)\left(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{(1-\gamma)^4 T}\right) \tag{41}$$

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 Online Appendix EC.2.4.

4.3. Proof Outline for Theorem 3

4.3.1. Construction of a Hard Instance with Four States and Two Actions.

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

$$\mathcal{A}_0 = \{1\}, \quad P(0\,|\,0,1) = 1, \qquad r(0,1) = 0, \tag{42a}$$

$$\mathcal{A}_1 = \{1,2\}, \; P(1\,|\,1,1) = p, \; P(0\,|\,1,1) = 1-p, \quad r(1,1) = 1, \tag{42b}$$

$$P(1\,|\,1,2) = p, \; P(0\,|\,1,2) = 1-p, \quad r(1,2) = 1, \tag{42c}$$

$$\mathcal{A}_2 = \{1\}, \quad P(2\,|\,2,1) = p, \; P(0\,|\,2,1) = 1-p, \quad r(2,1) = 1, \tag{42d}$$

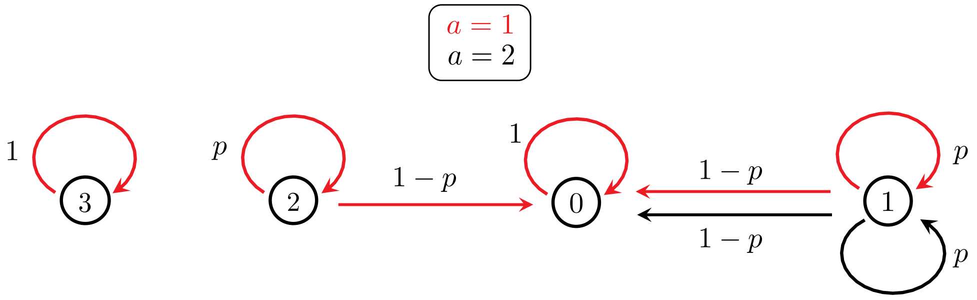
$$\mathcal{A}_3 = \{1\}, \quad P(3\,|\,3,1) = 1, \qquad r(3,1) = 1, \tag{42e}$$

where the parameter p is taken to be

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

Before moving forward to analyze the behavior of Q-learning, we first characterize the optimal value function and Q-function of this MDP; the proof is postponed to Online Section EC.4.4.

Figure 1. (Color online) The Constructed Hard MDP Instance Used in the Analysis of Theorem 3, Where $p = \frac{4\gamma-1}{3\gamma}$ and the Specifications are Described in (42)



Lemma 3. *Consider the MDP $\mathcal{M}_{\mathsf{hard}}$ constructed in* (42). *One has*

$$V^\star(0) = Q^\star(0,1) = 0; \tag{44a}$$

$$V^\star(1) = Q^\star(1,1) = Q^\star(1,2) = V^\star(2) = Q^\star(2,1) = \frac{1}{1-\gamma p} = \frac{3}{4(1-\gamma)}; \tag{44b}$$

$$V^\star(3) = Q^\star(3,1) = \frac{1}{1-\gamma}. \tag{44c}$$

Recognizing the elementary decomposition

$$\mathbb{E}\left[(V^\star(s) - V_T(s))^2\right] = (\mathbb{E}[V^\star(s) - V_T(s)])^2 + \mathsf{Var}(V_T(s)) \tag{45}$$

for any state s, our proof consists of lower bounding either the squared bias term $(\mathbb{E}[V^\star(s) - V_T(s)])^2$ or the variance term $\mathsf{Var}(V_T(s))$. In short, we 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 because of maximizing over the empirical state–action values, highlighting the well-recognized "overestimation" issue of Q-learning in practice (Hasselt 2010). The complete proof is deferred to Online Appendix EC.4.

4.3.2. Extension: Lower Bounds for Larger $|\mathcal{S}|$ and $|\mathcal{A}|$.

For pedagogical reasons, the hard instance (42) constructed contains no more than four states and two actions (as the focus has been to unveil suboptimal dependency on the effective horizon). As it turns out, one can straightforwardly extend it to cover larger state and action spaces with a more general hard instance constructed as follows.

- We begin by generating the following sub-MDP, denoted by $\mathcal{M}_{\mathsf{sub}}$, which comprises four states $\{1,2,3,4\}$ and no more than $|\mathcal{A}| \geq 2$ actions:

$$\mathcal{A}_0 = \{1\}, \quad P(0|0,1) = 1, \quad r(0,1) = 0, \tag{46a}$$

$$\mathcal{A}_1 = \{1,\ldots,|\mathcal{A}|\}, \quad P(1|1,a) = p, \quad P(0|1,a) = 1-p, \quad r(1,a) = 1, \quad \forall a \in \mathcal{A}_1, \tag{46b}$$

$$\mathcal{A}_2 = \{1\}, \quad P(2|2,1) = p, \quad P(0|2,1) = 1-p, \quad r(2,1) = 1, \tag{46c}$$

$$\mathcal{A}_3 = \{1\}, \quad P(3|3,1) = 1, \quad r(3,1) = 1, \tag{46d}$$

where p is still set according to (43).

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

As can be easily verified (which we omit here for the sake of brevity), our analysis developed for the smaller MDP (42) is directly applicable to studying the more general $\mathcal{M}_{\mathsf{full}}$, revealing that the lower bound (55) w.r.t. the iteration number T remains valid. Recognizing that the total sample size scales as $|\mathcal{S}||\mathcal{A}|T$, we establish a general sample complexity lower bound $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ for synchronous Q-learning to yield ε-accuracy.

5. Extension: Sample Complexity of Asynchronous Q-Learning

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

5.1. Markovian Samples and Asynchronous Q-Learning

5.1.1. Markovian Sample Trajectory.

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

$$a_t \sim \pi_{\mathsf{b}}(\cdot|s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot|s_t, a_t), \quad t \geq 0. \tag{47}$$

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

5.1.2. Asynchronous Q-Learning.

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

$$Q_t(s_{t-1}, a_{t-1}) = (1-\eta_t)Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_t\left\{r(s_{t-1}, a_{t-1}) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t, a')\right\}, \tag{48a}$$

$$Q_t(s,a) = Q_{t-1}(s,a) \text{for all } (s,a) \neq (s_{t-1}, a_{t-1}) \tag{48b}$$

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

$$V_t(s) := \max_{a \in \mathcal{A}} Q_t(s,a) \qquad \text{for all } s \in \mathcal{S}. \tag{49}$$

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

5.1.3. Assumptions.

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

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

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

$$\mu_{\mathsf{min}} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mu_{\pi_{\mathsf{b}}}(s,a). \tag{50}$$

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

$$t_{\mathsf{mix}} := \min\left\{t \,\middle|\, \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mathsf{TV}}\left(P^t(\cdot|s,a), \mu_{\pi_{\mathsf{b}}}\right) \leq \frac{1}{4}\right\}. \tag{51}$$

Here, $d_{\mathsf{TV}}(\mu, \nu) := \frac{1}{2}\sum_{x \in \mathcal{X}}|\mu(x) - \nu(x)|$ indicates the total variation distance between two measures μ and ν over $\mathcal{X}$ (Tsybakov and Zaiats 2009), whereas $P^t(\cdot|s,a)$ stands for the distribution of (s_t, a_t) when the sample trajectory is initialized at $(s_0, a_0) = (s,a)$. In words, the mixing time reflects the time required for the Markov chain to become nearly independent of the initial states. See Li et al. (2022b, section 2) for a more detailed account of these quantities and assumptions.

5.2. Sample Complexity of Asynchronous Q-Learning

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

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

$$\eta_t \equiv \eta = \frac{c_1 \log^3 T}{(1-\gamma)T\mu_{\mathsf{min}}} \tag{52a}$$

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

$$T \geq \frac{c_2 \log^2 \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\mu_{\mathsf{min}}} \max\left\{\frac{\log^3 T}{(1-\gamma)^4 \varepsilon^2}, \frac{t_{\mathsf{mix}}}{1-\gamma}\right\} \tag{52b}$$

for some sufficiently large universal constant $c_2 > 0$. If the initialization obeys $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then asynchronous Q-learning (cf. (48)) satisfies

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

with probability at least $1 - \delta$.

Remark 9. Similar to Remarks 1 and 5, one can immediately translate this high-probability result into the following mean estimation error bound:

$$\mathbb{E}\left[\max_{s,a} |Q_T(s,a) - Q^\star(s,a)|\right] \leq \varepsilon(1-\delta) + \delta\frac{1}{1-\gamma} \leq 2\varepsilon, \tag{53}$$

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

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

$$\tilde{O}\left(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\right), \tag{54}$$

provided that the learning rates are taken to be some proper constant (see (52a)). The first term in (54) resembles our sample complexity characterization of synchronous Q-learning (cf. (19)) except that we replace the number $|\mathcal{S}||\mathcal{A}|$ of state–action pairs in (19) with $1/\mu_{\mathsf{min}}$ in order to account for nonuniformity across state–action pairs. The second term in (54) is nearly independent of the target accuracy (except for some logarithmic scaling) and can be viewed as the burn-in time taken for asynchronous Q-learning to mimic synchronous Q-learning despite Markovian data.

We now pause to compare Theorem 4 with prior nonasymptotic theory for asynchronous Q-learning. As far as we know, all existing sample complexity bounds (Even-Dar and Mansour 2003, Beck and Srikant 2012, Qu and Wierman 2020, Chen et al. 2021, Li et al. 2022b) scale at least as $\frac{1}{(1-\gamma)^5}$ in terms of the dependency on the effective horizon with Theorem 4 being the first result to sharpen this dependency to $\frac{1}{(1-\gamma)^4}$. In particular, our sample complexity bound strengthens the state-of-the-art result Li et al. (2022b) by a factor up to $\frac{1}{1-\gamma}$ and improves upon Qu and Wierman (2020) by a factor of at least $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}\min\left\{t_{\mathsf{mix}}, \frac{1}{(1-\gamma)^3\varepsilon^2}\right\}$.[3]

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

Theorem 5. *Consider any $0.95 \le \gamma < 1$. Suppose that $\mu_{\mathsf{min}} \le \frac{1}{c_3 \log^2 T}$ and $T \ge \frac{c_3 \log^3 T}{\mu_{\mathsf{min}}(1-\gamma)^7}$ for some sufficiently large constant $c_3 > 0$. Assume that the initialization is $Q_0 \equiv 0$ and the learning rates are taken to be $\eta_t \equiv \eta$ for all $t \ge 0$. Then, there exists a γ-discounted MDP with $|\mathcal{S}| = 4$ and $|\mathcal{A}| = 3$ and a behavior policy such that (i) the minimum state–action occupancy probability of the sample trajectory is given by μ_{min} and (ii) the asynchronous Q-learning update rule (48)—for any $\eta \in (0, 1)$—obeys*

$$\max_{s,a} \mathbb{E}[|Q_T(s,a) - Q^\star(s,a)|^2] \ge \frac{c_{\mathsf{lb}}}{\mu_{\mathsf{min}}(1-\gamma)^4 T \log^3 T}, \tag{55}$$

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

In words, Theorem 5 asserts that, for large enough sample size T, in general, one cannot hope to achieve ℓ_∞-based ε-accuracy using fewer than $\tilde{O}\left(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}\right)$ samples, thus confirming the sharpness of our upper bound. The proof of this theorem can be found in Online Appendix EC.6.

6. Concluding Remarks

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

Acknowledgments

The authors are grateful to Laixi Shi for helpful discussions about the lower bound and thank Shaocong Ma for pointing out some errors in an early version of this work. Part of this work was done when G. Li, Y. Chen, and Y. Wei were visiting the Simons Institute for the Theory of Computing.

Endnotes

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

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

[3] The sample complexity of Li et al. (2022b) scales as $\tilde{O}\left(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\right)$, whereas the sample complexity of Qu and Wierman (2020) scales as $\tilde{O}\left(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}\right)$. It is worth noting that $1/\mu_{\mathsf{min}} \ge |\mathcal{S}||\mathcal{A}|$ and is, therefore, a large factor.

References

Agarwal A, Kakade S, Yang LF (2020) Model-based reinforcement learning with a generative model is minimax optimal. *Proc. Conf. Learn. Theory* (PMLR, New York), 67–83.

Azar MG, Munos R, Kappen HJ (2013) Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learn.* 91(3):325–349.

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

Bai Y, Xie T, Jiang N, Wang YX (2019) Provably efficient Q-learning with low switching cost. *Adv. Neural Inform. Processing Systems* 33:8002–8011.

Beck CL, Srikant R (2012) Error bounds for constant step-size Q-learning. *Systems Control Lett.* 61(12):1203–1208.

Bellman R (1952) On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA* 38(8):716–719.

Bertsekas DP (2017) *Dynamic Programming and Optimal Control*, 4th ed. (Athena Scientific, Nashua, NH).

Bhandari J, Russo D, Singal R (2021) A finite time analysis of temporal difference learning with linear function approximation. *Oper. Res.* 69(3):950–973.

Borkar VS, Meyn SP (2000) The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* 38(2):447–469.

Cai Q, Yang Z, Lee JD, Wang Z (2019) Neural temporal-difference and Q-learning converges to global optima. *Adv. Neural Inform. Processing Systems* 33:11312–11322.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2020) Finite-sample analysis of stochastic approximation using smooth convex envelopes. Preprint, submitted February 3, https://arxiv.org/abs/2002.00874.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2021) A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. Preprint, submitted February 2, https://arxiv.org/abs/2102.01567.

Chen Z, Zhang S, Doan TT, Maguluri ST, Clarke JP (2019) Performance of Q-learning with linear function approximation: Stability and finite-time analysis. Preprint, submitted May 27, https://arxiv.org/abs/1905.11425.

Devraj AM, Meyn SP (2020) Q-learning with uniformly bounded variance: Large discounting is not a barrier to fast learning. Preprint, submitted February 24, https://arxiv.org/abs/2002.10301.

Doan T, Maguluri S, Romberg J (2019) Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. *Internat. Conf. Machine Learn.* (PMLR, New York), 1626–1635.

Even-Dar E, Mansour Y (2003) Learning rates for Q-learning. *J. Machine Learn. Res.* 5:1–25.

Fan J, Wang Z, Xie Y, Yang Z (2019) A theoretical analysis of deep Q-learning. Preprint, submitted January 1, https://arxiv.org/abs/1901.00137.

Freedman DA (1975) On tail probabilities for martingales. *Ann. Probab.* 3(1):100–118.

Gupta H, Srikant R, Ying L (2019) Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Adv. Neural Inform. Processing Systems* 33:4706–4715.

Hasselt H (2010) Double Q-learning. *Adv. Neural Inform. Processing Systems* 23:2613–2621.

Hu J, Wellman MP (2003) Nash Q-learning for general-sum stochastic games. *J. Machine Learn. Res.* 4:1039–1069.

Jaakkola T, Jordan MI, Singh SP (1994) Convergence of stochastic iterative dynamic programming algorithms. *Adv. Neural Inform. Processing Systems* 6:703–710.

Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? *Adv. Neural Inform. Processing Systems* 31:4863–4873.

Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. *Adv. Neural Inform. Processing Systems* 26:315–323.

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

Kearns MJ, Singh SP (1999) Finite-sample convergence rates for Q-learning and indirect algorithms. *Adv. Neural Inform. Processing Systems* 11:996–1002.

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

Khamaru K, Xia E, Wainwright MJ, Jordan MI (2021a) Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning. Preprint, submitted June 28, https://arxiv.org/abs/2106.14352.

Khamaru K, Pananjady A, Ruan F, Wainwright MJ, Jordan MI (2021b) Is temporal difference learning optimal? An instance-dependent analysis. *SIAM J. Math. Data Sci.* 3(4):1013–1040.

Lakshminarayanan C, Szepesvari C (2018) Linear stochastic approximation: How far does constant step-size and iterate averaging go? *Internat. Conf. Artificial Intelligence Statist.*, 1347–1355.

Lee D, He N (2018) Stochastic primal-dual Q-learning. Preprint, submitted October 18, https://arxiv.org/abs/1810.08298.

Li G, Chi Y, Wei Y, Chen Y (2022a) Minimax-optimal multi-agent RL in Markov games with a generative model. *Adv. Neural Inform. Processing Systems* Forthcoming.

Li G, Shi L, Chen Y, Chi Y (2023a) Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Inform. Inference* 12(2):969–1043.

Li G, Wei Y, Chi Y, Chen Y (2023b) Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Oper. Res.* Forthcoming.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2022b) Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Trans. Inform. Theory* 68(1):448–473.

Mou W, Li CJ, Wainwright MJ, Bartlett PL, Jordan MI (2020) On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. Preprint, submitted April 9, https://arxiv.org/abs/2004.04719.

Murphy S (2005) A generalization error for Q-learning. *J. Machine Learn. Res.* 6:1073–1097.

Pananjady A, Wainwright MJ (2020) Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Trans. Inform. Theory* 67(1):566–585.

Paulin D (2015) Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic J. Probab.* 20:1–32.

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.

Qu G, Wierman A (2020) Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conf. Learn. Theory* 3185–3205.

Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

Shah D, Xie Q (2018) Q-learning with nearest neighbors. *Adv. Neural Inform. Processing Systems* 32:3111–3121.

Shi L, Li G, Wei Y, Chen Y, Chi Y (2022) Pessimistic Q-learning for offline reinforcement learning: Toward optimal sample complexity. *Internat. Conf. Machine Learn.* (PMLR, New York).

Sidford A, Wang M, Wu X, Yang L, Ye Y (2018) Near-optimal time and sample complexities for solving Markov decision processes with a generative model. *Adv. Neural Inform. Processing Systems* 31:5186–5196.

Srikant R, Ying L (2019) Finite-time error bounds for linear stochastic approximation and TD learning. *Conf. Learn. Theory*, 2803–2830.

Sutton RS (1988) Learning to predict by the methods of temporal differences. *Machine Learn.* 3(1):9–44.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Szepesvári C (1998) The asymptotic convergence-rate of Q-learning. *Adv. Neural Inform. Processing Systems* 10:1064–1070.

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

Tsitsiklis J, Van Roy B (1997) An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automatic Control* 42(5):674–690.

Tsybakov AB, Zaiats V (2009) *Introduction to Nonparametric Estimation*, vol. 11 (Springer, New York).

Wai HT, Hong M, Yang Z, Wang Z, Tang K (2019) Variance reduced policy evaluation with smooth function approximation. *Adv. Neural Inform. Processing Systems* 32:5784–5795.

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

Wainwright MJ (2019b) Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. Preprint, submitted May 15, https://arxiv.org/abs/1905.06265.

Wainwright MJ (2019c) Variance-reduced Q-learning is minimax optimal. Preprint, submitted June 11, https://arxiv.org/abs/1906.04697.

Watkins CJCH (1989) Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge, UK.

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

Weng B, Xiong H, Zhao L, Liang Y, Zhang W (2020a) Momentum Q-learning with finite-sample convergence guarantee. Preprint, submitted April 9, https://arxiv.org/abs/2007.15418.

Weng W, Gupta H, He N, Ying L, Srikant R (2020b) The mean-squared error of double Q-learning. *Adv. Neural Inform. Processing Systems* 33:6815–6826.

Wu Y, Zhang W, Xu P, Gu Q (2020) A finite time analysis of two time-scale actor critic methods. Preprint, submitted May 4, https://arxiv.org/abs/2005.01350.

Xiong H, Zhao L, Liang Y, Zhang W (2020) Finite-time analysis for double Q-learning. *Adv. Neural Inform. Processing Systems* 33.

Xu P, Gu Q (2020) A finite-time analysis of Q-learning with neural network function approximation. *Internat. Conf. Machine Learn.* (PMLR, New York), 10555–10565.

Xu T, Zou S, Liang Y (2019a) Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Adv. Neural Inform. Processing Systems* 33:10633–10643.

Xu T, Wang Z, Zhou Y, Liang Y (2019b) Reanalysis of variance reduced temporal difference learning. *Internat. Conf. Learn. Representations*.

Yan Y, Li G, Chen Y, Fan J (2022) The efficacy of pessimism in asynchronous Q-learning. Preprint, submitted March 14, https://arxiv.org/abs/2203.07368.

Zhang Z, Zhou Y, Ji X (2020) Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Adv. Neural Inform. Processing Systems* 33.

Gen Li is a postdoctoral researcher in the department of statistics and data science at the Wharton School, University of Pennsylvania. His research interests include reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization. He has received the excellent graduate award and the excellent thesis award from Tsinghua University.

Changxiao Cai is a postdoctoral researcher at the University of Pennsylvania. His research interests include statistical machine learning, high-dimensional statistics, convex and nonconvex optimization, reinforcement learning, and information theory. He received the School of Engineering and Applied Science Award for Excellence from Princeton University in 2020.

Yuxin Chen is an associate professor of statistics and data science and of electrical and systems engineering at the University of Pennsylvania. His research interests include statistics, optimization, and machine learning. He has received the Alfred P. Sloan Research Fellowship, the International Consortium of Chinese Mathematicians Best Paper Award, the Princeton Graduate Mentoring Award, and was selected as a finalist for the Best Paper Prize for Young Researchers in Continuous Optimization.

Yuting Wei is an assistant professor of statistics and data science at the Wharton School, University of Pennsylvania. She was the recipient of the 2022 National Science Foundation CAREER Award, an honorable mention for the 2023 Bernoulli Society's New Researcher Award, and the 2018 Erich L. Lehmann Citation from the Berkeley Statistics Department. Her research interests include high-dimensional statistics, nonparametric statistics, statistical machine learning, and reinforcement learning.

Yuejie Chi is a professor in the Department of Electrical and Computer Engineering and a faculty affiliate with the Machine Learning Department and CyLab at Carnegie Mellon University. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning, and inverse problems, with applications in sensing, imaging, decision making, and societal systems, broadly defined.