Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning

Gen Li [1] **Changxiao Cai** [2] **Yuxin Chen** [2] **Yuantao Gu** [1] **Yuting Wei** [3] **Yuejie Chi** [4]

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. Focusing on the synchronous setting (such that independent samples for all state-action pairs are queried via a generative model in each iteration), substantial progress has been made recently towards understanding the sample efficiency of Q-learning. To yield an entrywise ε-accurate estimate of the optimal Q-function, state-of-the-art theory requires at least an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ samples in the infinite-horizon γ-discounted setting. In this work, we sharpen the sample complexity of synchronous Q-learning to the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ (up to some logarithmic factor) for any $0 < \varepsilon < 1$, leading to an order-wise improvement in $\frac{1}{1-\gamma}$. Analogous results are derived for finite-horizon MDPs as well. Our sample complexity analysis unveils the effectiveness of vanilla Q-learning, which matches that of speedy Q-learning without requiring extra computation and storage. Our result is obtained by identifying novel error decompositions and recursions, which might shed light on how to study other variants of Q-learning.

1. Introduction

Characterizing the sample efficiency of Q-learning (Watkins & 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). While classical convergence analyses for Q-learning (Borkar & 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 & Srikant, 2012; Chen et al., 2020; 2021; Even-Dar & Mansour, 2003; Kearns & Singh, 1999; Lee & He, 2018; Li et al., 2020b; Qu & Wierman, 2020; Wainwright, 2019b; Weng et al., 2020a; Xiong et al., 2020). Drawing on the insights from high-dimensional statistics (Wainwright, 2019a), such a modern non-asymptotic framework unveils more clear and informative dependence of the sample complexity on salient problem parameters, and has been developed for Q-learning under multiple data collection mechanisms (Beck & Srikant, 2012; Even-Dar & Mansour, 2003; Jin et al., 2018; Li et al., 2020b; Qu & Wierman, 2020; Wainwright, 2019b; Wang et al., 2021).

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 (which is often referred to as the synchronous setting) (Kakade, 2003; Kearns et al., 2002). Our focal point is the ℓ_∞-based sample complexity, namely, the number of samples needed for Q-learning to yield an entrywise ε-accurate estimate of the optimal Q-function. Despite a number of prior work tackling this setting, however, the dependency 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 theory Chen et al. (2020); Wainwright (2019b) requires at least an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$ samples (up to some log factor), where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and the action space, respectively. However, it remains unclear whether this scaling is essential for Q-learning algorithms or it is improvable via a more refined theory. In fact, Wainwright (2019b) exhibited a numerical example that hints at the non-sharpness of this scaling, that is, the numerical experiments conducted therein suggested a scaling of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ for certain problem instances. To bridge the gap between the theoretical prediction and the empirical observation, it is natural to investigate

[1]Department of Electronic Engineering, Tsinghua University [2]Department of Electrical and Computer Engineering, Princeton University [3]Department of Statistics and Data Science, Carnegie Mellon University [4]Department of Electrical and Computer Engineering, Carnegie Mellon University. Correspondence to: Yuxin Chen <yuxin.chen@princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

the following question:

Is it possible to tighten the dependence on effective horizon in the sample complexity of Q-learning?

The above-mentioned issue comes up in finite-horizon MDPs as well, in which the scaling of the sample complexity with the horizon H of the MDP remains undetermined.

1.1. Main contributions

In this paper, we develop a refined theoretical framework that allows one to tighten the ℓ_∞-based sample complexity of Q-learning for two different types of MDPs with state space $\mathcal{S}$ and action space $\mathcal{A}$. Here and throughout, the notation $\widetilde{O}(\cdot)$ hides any logarithmic dependencies.

- For γ-discounted infinite horizon MDPs and any $0 < \varepsilon < 1$, we show that a total number of

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

  samples are sufficient to guarantee ε-accuracy. This finding improves prior theory (Chen et al., 2020; Wainwright, 2019b) by a factor of $\frac{1}{1-\gamma}$.

- Analogously, for finite-horizon MDPs with horizon H, we demonstrate that

$$\widetilde{O}\Big(\frac{H^4}{\varepsilon^2}\Big) \quad \text{samples } \textit{per state-action pair} \tag{2}$$

  are sufficient for Q-learning to attain ε-accuracy.

We consider both rescaled linear and constant learning rates; see Table 1 for more detailed comparisons with existing literature. For both of the above scenarios, our theoretical guarantees are the tightest known to date for vanilla Q-learning.

Encouragingly, the sample complexity (1) is sharp in a minimax sense (up to some logarithmic factor). In fact, our companion paper (see Li et al. (2021a, Theorem 2)) constructs a hard MDP instance to show that the sample complexity of plain Q-learning at least exceeds $\frac{1}{(1-\gamma)^4}$, thereby demonstrating the sharpness of our sample complexity upper bound (1). In addition, it is also worth emphasizing that our sample complexity bound matches the theory for speedy Q-learning (Ghavamzadeh et al., 2011) without requiring extra computation and storage. Our analysis framework uncovers a sort of crucial error decompositions and recursions that are previously unavailable, which might shed light on how to pin down the sample efficiency 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 non-asymptotic behavior of value-based RL algorithms in various scenarios. In the discussion below, we subsample the literature and focus on the papers that are the closest to ours.

Finite-sample ℓ_∞ guarantees for synchronous Q-learning. The sample complexities derived in the literature often depend crucially on the choices of learning rates. Even-Dar & 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 ε. Beck & 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 $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2})$. However, this result remains suboptimal in terms of its scaling in $\frac{1}{1-\gamma}$. See Table 1 for details.

Finite-sample ℓ_∞ guarantees for asynchronous Q-learning. Moving beyond the synchronous model considered herein, (Beck & Srikant, 2012; Even-Dar & Mansour, 2003; Li et al., 2020b; Qu & 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. The state-of-the-art sample complexity bound for asynchronous Q-learning (Li et al., 2020b) scales as $\widetilde{O}\big(\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}\big)$, where $\mu_{\min}$ stands for the minimum state-action occupancy probability of the sample trajectory and t_{mix} represents the mixing time. Clearly, this sample complexity bound also exhibits a scaling of $\widetilde{O}\big(\frac{1}{(1-\gamma)^5}\big)$ w.r.t. the effective horizon. The analysis framework developed in this paper might be applicable to help 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. (2011) proposed speedy Q-learning, which achieves a sample complexity of $\widetilde{O}\big(\frac{1}{(1-\gamma)^4\varepsilon^2}\big)$ at the expense of doubling the computation and storage complexity. Our result on vanilla Q-

paper	learning rates	sample complexity
(Even-Dar & 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 & 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 & 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 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 $\max_{s,a}|Q_T(s,a)-Q^\star(s,a)|\le\varepsilon$ with high probability or $\mathbb{E}\big[\max_{s,a}|Q_T(s,a)-Q^\star(s,a)|\big]\le\varepsilon$, where T is the total number of iterations. All logarithmic factors are omitted in the table to simplify the expressions.

learning order-wise matches that of speedy Q-learning. 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 $\widetilde{O}\big(\frac{1}{(1-\gamma)^3\varepsilon^2}\big)$, which was subsequently generalized to the asynchronous setting by Li et al. (2020b). The ℓ_∞ 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 & Mansour, 2003); however, it remains 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; Zhang et al., 2020a;b) considered the sample efficiency of Q-learning type algorithms paired with proper exploration strategies (e.g., UCB exploration) under the framework of regret analysis, which is beyond the reach of the current paper.

Others. There are also several other strands of related papers that tackle model-free algorithms but do not pursue ℓ_∞-based non-asymptotic guarantees. For instance, Bhandari et al. (2018); Chen et al. (2019); Doan et al. (2019); Gupta et al. (2019); Srikant & 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;c) and relative Q-learning (Devraj & Meyn, 2020). More general function approximation schemes (e.g., certain families of neural network approximations) have been studied in Cai et al. (2019); Fan et al. (2019); Li et al. (2021b); Wai et al. (2019); Xu & Gu (2020) as well. These are beyond the scope of the present paper.

2. Background and algorithm

This paper concentrates on tabular MDPs and accounts for both the discounted infinite-horizon setting and the finite-horizon counterpart (Bertsekas, 2017). In both settings, we denote by $\mathcal{S}=\{1,\cdots,|\mathcal{S}|\}$ and $\mathcal{A}=\{1,\cdots,|\mathcal{A}|\}$ the state space and the action space of the MDP, respectively. Here and throughout, $\Delta(\mathcal{S})$ stands for the probability simplex over the set $\mathcal{S}$.

2.1. Q-learning for infinite-horizon discounted MDPs

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 each 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 the action selection rule 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 generated by the MDP under 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

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

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 & 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 via the following update rule

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

Here, η_t 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.,

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

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

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

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Noteworthily, the optimal Q-function $Q^{\star}$ is the unique fixed point of the Bellman operator (Bellman, 1952), that is, $\mathcal{T}(Q^{\star}) = Q^{\star}$.

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**

We initialize the algorithm so that $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$ for all (s,a). In addition, the corresponding value function estimate $V_t : \mathcal{S} \to \mathbb{R}$ in the t-th iteration is defined as

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

The complete algorithm is summarized in Algorithm 1.

2.2. Q-learning for finite-horizon MDPs

Finite-horizon MDPs. We now turn attention to a finite-horizon MDP, which can be represented and described by the quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, \{r_h\}_{h=1}^{H}, H)$. Here, H represents the time horizon of the MDP. For any $1 \leq h \leq H$, we let $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denote the probability transition kernel at step h (i.e., $P_h(s' \,|\, s,a)$ is the probability of transiting to s' from (s,a) at step h), and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ indicates the reward function at step h (i.e., $r_h(s,a)$ is the immediate reward at step h in response to (s,a)). As before, we assume normalized rewards such that all the $r_h(s,a)$'s reside within the range $[0,1]$.

Value function and Q-function. In a finite-horizon MDP, the value function and Q-function associated with a deterministic policy $\pi : \mathcal{S} \times \{1, \ldots, H\} \to \mathcal{A}$ are defined by

$$V_h^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{k=h}^{H} r_k(s_k, a_k) \,\middle|\, s_h = s\right]$$

for all $s \in \mathcal{S}$ and all $1 \leq h \leq H$, and

$$Q_h^{\pi}(s,a) \coloneqq \mathbb{E}\left[\sum_{k=h}^{H} r_k(s_k, a_k) \,\middle|\, s_h = s, a_h = a\right]$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and all $1 \leq h \leq H$. As before, the expectations are taken over the randomness of the MDP trajectory $\{(s_k, a_k)\}_{1 \leq k \leq H}$ induced by the transition kernel $\{P_h\}_{h=1}^{H}$, where the policy depends on the step index h as well as the current state s. In view of the assumed bounds for the immediate rewards, it is easily seen that $0 \leq V_h^{\pi}(s) \leq H$ and $0 \leq Q_h^{\pi}(s,a) \leq H$ for any π, any state-action pair (s,a), and any step h. Akin to the infinite-horizon scenario, the optimal value functions $\{V_h^{\star}\}$ and optimal Q-functions $\{Q_h^{\star}\}$ are defined respectively by

$$V_h^{\star}(s) \coloneqq \max_{\pi} V_h^{\pi}(s), \quad Q_h^{\star}(s,a) \coloneqq \max_{\pi} Q_h^{\pi}(s,a)$$

Algorithm 2 Synchronous Q-learning for finite-horizon MDPs

inputs: learning rates $\{\eta_t\}$, number of iterations T, initial estimate $\{Q_{0,h}\}_{h=1}^H$.
for $t = 1, 2, \cdots, T$ **do**
  Set $Q_{t,H+1} = 0$.
  for $h = H, H-1, \cdots, 1$ **do**
    Draw $s_{t,h}(s,a) \sim P_h(\cdot \,|\, s,a)$ for each (s,a).
    Compute Q_t according to (6) and (7).
  end for
end for

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $1 \le h \le H$. A distinguishing feature of finite-horizon MDPs is the non-stationarity of value functions and Q-functions across h.

Q-learning. Assume that we have access to a generative model; in each iteration, we draw an independent sample for each triple (s,a,h) as follows

$$s_{t,h}(s,a) \sim P_h(\cdot \,|\, s,a).$$

In the t-th iteration, the synchronous Q-learning algorithm updates the Q-function estimate $Q_{t,h} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ from $h = H$ to 1 as follows:

$$Q_{t,h} = (1-\eta_t) Q_{t-1,h} + \eta_t \mathcal{T}_{t,h}(Q_{t,h+1}), \tag{6}$$

where η_t denotes the learning rate in the t-th iteration, and $\mathcal{T}_{t,h}$ is the empirical Bellman operator based on samples generated in the t-th iteration for step h, namely,

$$\mathcal{T}_{t,h}(Q)(s,a) \coloneqq r_h(s,a) + \max_{a' \in \mathcal{A}} Q(s_{t,h}, a') \tag{7}$$
$$s_{t,h} \equiv s_{t,h}(s,a) \sim P_h(\cdot \,|\, s,a)$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here, $\mathcal{T}_{t,h}$ can be viewed as an unbiased estimate of the Bellman operator defined for the finite-horizon case.

In addition, the initialization $\{Q_{0,h}\}_{h=1}^H$ of the algorithm is chosen such that $0 \le Q_{0,h}(s,a) \le H - h + 1$ for all (s,a) and all $1 \le h \le H$. Throughout this paper, the value function estimate $V_{t,h} : \mathcal{S} \to \mathbb{R}$ associated with the t-th iterate and step h is defined as

$$\forall s \in \mathcal{S}: \qquad V_{t,h}(s) \coloneqq \max_{a \in \mathcal{A}} Q_{t,h}(s,a). \tag{8}$$

The complete algorithm is summarized in Algorithm 2.

Remark 1. Suppose that the probability transition kernel of the MDP is time-invariant, namely, $P_h \equiv P$ for all $1 \le h \le H$. Then we only need to sample once for each (s,a) — with a total number of $|\mathcal{S}||\mathcal{A}|$ samples — in each iteration t. This should be contrasted with the time-varying case where a total number of $|\mathcal{S}||\mathcal{A}|H$ samples are generated in each iteration. Both cases have been studied for the finite-horizon setting; see, e.g., Jin et al. (2018); Sidford et al. (2018).

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: infinite-horizon MDPs

We start by presenting our strengthened ℓ_∞-based sample complexity of Q-learning for discounted infinite-horizon MDPs — the setting described in Section 2.1.

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*

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}} \tag{9a}$$

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

$$T \ge \frac{c_3 \big(\log^4 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 \varepsilon^2} \tag{9b}$$

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

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

with probability at least $1 - \delta$.

Remark 2. This high-probability bound immediately translates to an expected error bound. Recognizing the crude upper bound $\big|Q_T(s,a) - Q^\star(s,a)\big| \le \frac{1}{1-\gamma}$ and taking $\delta \le \varepsilon(1-\gamma)$, we reach

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

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 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 translates to the following sample complexity bound:

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}\Big) \tag{10}$$

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 & Srikant,

2012; Chen et al., 2020; Li et al., 2020b; Qu & Wierman, 2020; Wainwright, 2019b). In addition, the dependence of our result on the effective horizon (i.e., $\frac{1}{(1-\gamma)^4}$) matches a lower bound developed in our companion paper for plain Q-learning (see Li et al. (2021a, Theorem 2)), thus potentially corroborating its sharpness.

Learning rates. In view of the assumption (9a), our result accommodates two commonly seen 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 $T = \frac{c_3(\log^4 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4\varepsilon^2}$, the constant learning rates can be taken to be on the order of

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

depending 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 orderwise sense), making them appealing for practical use.

Comparison with minimax lower bounds. 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 by the model-based approach (up to some logarithmic factor) (Agarwal et al., 2020; Azar et al., 2013; Li et al., 2020a). Fortunately, while vanilla Q-learning might fall short of achieving minimax optimality, the dependency of its sample complexity on the effective horizon can be improved to optimal scaling with the assistance of variance reduction; see Li et al. (2020b); Wainwright (2019c).

3.2. Performance guarantees: finite-horizon MDPs

Next, we move forward to present an analogous ℓ_∞-based sample complexity of Q-learning for finite-horizon MDPs — the setting formulated in Section 2.2.

Theorem 2. *Consider an arbitrary quantity $\delta \in (0,1)$ and $\varepsilon \in (0,1]$. Suppose that the learning rates obey*

$$\frac{1}{1+\frac{c_1 T}{H\log^2 T}} \le \eta_t \le \frac{1}{1+\frac{c_2 t}{H\log^2 T}}, \quad 0 \le t \le T \tag{11a}$$

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

$$T \ge \frac{c_3 H^4\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\varepsilon^2} \tag{11b}$$

for some sufficiently large universal constant $c_3 > 0$. If the initialization obeys $0 \le Q_{0,h}(s,a) \le H+1-h$ for any (s,a) and any $1 \le h \le H$, then Algorithm 2 achieves

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A},\, 1\le h\le H} \big|Q_{T,h}(s,a) - Q_h^\star(s,a)\big| \le \varepsilon$$

with probability exceeding $1-\delta$.

Remark 3. Similar to Remark 2, Theorem 2 implies that

$$\mathbb{E}\Big[\max_{s,a,h}\big|Q_{T,h}(s,a) - Q_h^\star(s,a)\big|\Big] \le 2\varepsilon$$

as soon as $T \ge \frac{c_3 H^4(\log^3 T)\big(\log\frac{|\mathcal{S}||\mathcal{A}|HT}{\varepsilon}\big)}{\varepsilon^2}$.

For the general scenario (namely, $\{P_h\}_{h=1}^H$ vary across h), we collect $|\mathcal{S}||\mathcal{A}|H$ independent samples in each iteration, and hence Theorem 2 uncovers a sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|H^5}{\varepsilon^2}\Big). \tag{12}$$

When it comes to the special case where $P_h \equiv P$ is invariant across all $1 \le h \le H$, our theorem continues to hold if we generate $|\mathcal{S}||\mathcal{A}|$ independent samples in each iteration (with one sample for each state-action pair (s,a) w.r.t. P), leading to a reduced sample complexity of

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|H^4}{\varepsilon^2}\Big). \tag{13}$$

Interestingly, all this is achieved via the same learning rate schedules (i.e., rescaled linear or constant learning rates) as for the discounted infinite-horizon case.

4. Analysis: discounted infinite-horizon MDPs

This section outlines the key ideas for the establishment of our main theorems, focusing on the infinite-horizon setting. The full details of the proof for the infinite-horizon case can be found in our companion paper Li et al. (2021a). The proof for finite-horizon MDPs follows an analogous argument, and is postponed to the supplementary material. Before delving into proof details, we first introduce convenient vector and matrix notations that shall be used frequently.

4.1. Vector and matrix notation

To begin with, for any matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_1 := \max_i \sum_j |M_{i,j}|$ is defined as the largest row-wise ℓ_1 norm of $\boldsymbol{M}$. The matrix $\boldsymbol{I}$ represents the identity matrix. 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 $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m \in \mathbb{R}^n$, we define max in an entrywise fashion such that $\max_{1\le i\le m} \boldsymbol{a}_i := [\max_i a_{i,j}]_{j=1}^n$. 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).

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 *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:

$$\boldsymbol{P}^\pi \coloneqq \boldsymbol{P}\boldsymbol{\Pi}^\pi \qquad \text{and} \qquad \boldsymbol{P}_\pi \coloneqq \boldsymbol{\Pi}^\pi \boldsymbol{P} \tag{14}$$

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

$$\boldsymbol{\Pi}^\pi = \begin{pmatrix} \boldsymbol{e}_{\pi(1)}^\top & & & \\ & \boldsymbol{e}_{\pi(2)}^\top & & \\ & & \ddots & \\ & & & \boldsymbol{e}_{\pi(|\mathcal{S}|)}^\top \end{pmatrix}. \tag{15}$$

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:

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V}) \circ (\boldsymbol{P}\boldsymbol{V}); \tag{16}$$

in other words, the (s,a)-th entry of $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V})$ is the variance of $\{V(s')\}_{1\le s'\le|\mathcal{S}|}$ 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 $r(s,a)$. 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$s$, $\pi_t(s) = \min\big\{a' \,|\, Q_t(s,a') = \max_{a''} Q_t(s,a'')\big\}$. 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 has

$$Q_t\big(s,\pi_t(s)\big) = V_t(s), \qquad \boldsymbol{P}\boldsymbol{V}_t = \boldsymbol{P}^{\pi_t}\boldsymbol{Q}_t \ge \boldsymbol{P}^\pi \boldsymbol{Q}_t \tag{17}$$

for any π, where $\boldsymbol{P}^\pi$ is defined in (14). Further, we introduce a matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ such that

$$\boldsymbol{P}_t\big((s,a),s'\big) \coloneqq \begin{cases} 1, & \text{if } s' = s_t(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

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

Throughout the paper, the notation $f(n) \lesssim g(n)$ (resp. $f(n) \gtrsim g(n)$) means that there exists a constant $C_0 > 0$ such that $|f(n)| \le C_0|g(n)|$ (resp. $|f(n)| \ge C_0|g(n)|$).

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. (18)) to perform the following update

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \tag{19}$$

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

$$\boldsymbol{\Delta}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star \tag{20}$$

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 (19), we arrive at the following elementary decomposition:

$$\begin{aligned} \boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^\star &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}\big) - \boldsymbol{Q}^\star \\ &= (1-\eta_t)\big(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star\big) + \eta_t\big(\boldsymbol{r} + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\big) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\big) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} \\ &\quad + \eta_t\gamma\big\{\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}. \end{aligned} \tag{21}$$

Further, the term $\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star)$ can be linked with $\boldsymbol{\Delta}_{t-1}$ as follows

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) &= \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \\ &\le \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1}, \end{aligned} \tag{22a}$$

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) &= \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star \\ &\ge \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^\star}\boldsymbol{Q}^\star = \boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1}, \end{aligned} \tag{22b}$$

where we have made use of the relation (17). Substitute (22) into (21) to reach

$$\begin{aligned} \boldsymbol{\Delta}_t &\le (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}; \\ \boldsymbol{\Delta}_t &\ge (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big\}. \end{aligned} \tag{23}$$

Applying these relations recursively, we obtain

$$\begin{aligned} \boldsymbol{\Delta}_t &\le \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^t \eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\}, \\ \boldsymbol{\Delta}_t &\ge \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^t \eta_i^{(t)}\gamma\big\{\boldsymbol{P}^{\pi^\star}\boldsymbol{\Delta}_{i-1} + (\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}\big\}, \end{aligned} \tag{24}$$

where we define

$$\eta_i^{(t)} \coloneqq \begin{cases} \prod_{j=1}^t(1-\eta_j), & \text{if } i = 0, \\ \eta_i\prod_{j=i+1}^t(1-\eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t. \end{cases} \tag{25}$$

Comparisons to prior approaches. We take a moment to discuss how previous analyses handle the above decomposition. Several prior work (e.g., Li et al. (2020b); Wainwright (2019b)) tackled the second term on the right-hand side of the relation (23) via the following crude bounds:

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

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

4.2.2. Upper bound and lower bound on $\boldsymbol{\Delta}_t$

We proceed to upper and lower bound $\boldsymbol{\Delta}_t$ separately by exploiting the crucial relations (24) derived above. To be more specific, defining $\beta \coloneqq \frac{c_4(1-\gamma)}{\log T}$ for some sufficiently small constant $c_4 > 0$, one can further decompose the upper bound in (24) into several terms below:

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

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

- Each component in the term $\boldsymbol{\zeta}_t$ is fairly small, given that $\eta_i^{(t)}$ is sufficiently small for any $i \le (1-\beta)t$ (to see this, note that each component has undergone the contraction $(1-\eta_j)$ for sufficiently many times).
- The 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. As it turns out, the resulting variance term due to $\boldsymbol{\xi}_t$ depends also on $\{\boldsymbol{\Delta}_i\}$ for the last few iterations prior to t.

In other words, the right-hand side of the above inequality can be further decomposed into some negligible effect and a certain weighted superposition of several $\{\boldsymbol{\Delta}_i\}$. Viewed in this light, a crucial step then boils down to carefully exploiting such intertwined relations regarding $\{\boldsymbol{\Delta}_i\}$ to develop the following upper bound.

Lemma 1. *With probability at least $1-\delta$, one has*

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

holds simultaneously for all $t \ge \frac{T}{c_2\log\frac{1}{1-\gamma}}$, where $c_{\rm ub} > 0$ is some universal constant.

Similarly, making use of the lower bound in (24) allows one to develop an analogous lower bound as follows.

Lemma 2. *With probability at least $1-\delta$, one has*

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

holds simultaneously for all $t \ge \frac{T}{c_2\log\frac{1}{1-\gamma}}$, where $c_{\rm lb} > 0$ is some universal constant.

The preceding two bounds, which are the crux of our analysis, lead to the improved sample complexity. In principle, 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 by leveraging the structure of the discounted MDP. The proofs of Lemmas 1 and 2 can be found in Li et al. (2021a, Appendix A)

4.2.3. Combining upper and lower bounds on $\boldsymbol{\Delta}_t$

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

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

for all $t \ge \frac{T}{c_2\log\frac{1}{1-\gamma}}$ with probability exceeding $1-2\delta$. Employing elementary analysis tailored to this crucial recurrence relation, we can demonstrate that (where details can be found in Li et al. (2021a, Appendix A))

$$\|\boldsymbol{\Delta}_T\|_\infty \lesssim \sqrt{\frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T}} + \frac{\big(\log^4 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{(1-\gamma)^4 T} \tag{26}$$

with probability at least $1-2\delta$. To finish up, recognize that the sample size assumption (9b) is equivalent to saying that

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

When $c_3 > 0$ is sufficiently large, substituting this relation into (26) leads to the advertised bound

$$\|\boldsymbol{\Delta}_T\|_\infty \le \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon^2 \le \varepsilon.$$

5. Concluding remarks

In this paper, we tighten the sample complexity of vanilla Q-learning to an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ for the discounted infinite-horizon setting, and $\frac{|\mathcal{S}||\mathcal{A}|H^4}{\varepsilon^2}$ for the finite-horizon setting (modulo some logarithmic terms). 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 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-2106739, CCF-1907661, DMS-2014279 and IIS-1900140, and the Princeton SEAS Innovation Award. Y. Wei is supported in part by the grants NSF CCF-2106778, 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, and NSF CCF-2106778, CCF-1806154 and CCF-2007911. We thank Shaocong Ma for pointing out some error in an early version of this work.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y., and Chi, Y. Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021a.

Li, G., Chen, Y., Chi, Y., Gu, Y., and Wei, Y. Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *arXiv preprint arXiv:2105.08024*, 2021b.

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

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

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

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

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

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

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

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

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

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

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

Wang, B., Yan, Y., and Fan, J. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *arXiv preprint arXiv:2105.14016*, 2021.

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

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

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

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

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

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

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

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

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

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

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

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020b.