On the Sample Complexity of Reinforcement Learning with a Generative Model

Mohammad Gheshlaghi Azar M.AZAR@SCIENCE.RU.NL
Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands

Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille, SequeL Project, 40 avenue, Halley 59650, Villeneuve dAscq, France

Hilbert J. Kappen B.KAPPEN@SCIENCE.RU.NL
Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands

Abstract

We consider the problem of learning the optimal action-value function in the discounted-reward Markov decision processes (MDPs). We prove a new PAC bound on the sample-complexity of model-based value iteration algorithm in the presence of a generative model of the MDP, which indicates that for an MDP with N state-action pairs and the discount factor $\gamma \in [0,1)$ only $O\big(N\log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$ samples are required to find an ε-optimal estimation of the action-value function with the probability $1-\delta$. We also prove a matching lower bound of $\Theta\big(N\log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$ on the sample complexity of estimating the optimal action-value function by every RL algorithm. To the best of our knowledge, this is the first minimax result on the sample complexity of estimating the optimal (action-)value function in which the upper bound matches the lower bound of RL in terms of N, ε, δ and $1-\gamma$. Also, both our lower bound and upper bound improve on the state-of-the-art in terms of $1/(1-\gamma)$.

1. Introduction

Model-based value iteration (VI) (Kearns & Singh, 1999; Buşoniu et al., 2010) is a well-known reinforcement learning (RL) (Szepesvári, 2010; Sutton & Barto, 1998) algorithm which relies on an empirical estimate of the state-transition distributions to estimate the optimal (action-)value function through the Bellman recursion. In the finite state-action problems, it has been shown that an action-value based variant of VI, model-based Q-value iteration (QVI), finds an ε-optimal estimate of the action-value function with high probability using only $T = \tilde{O}(N/\big((1-\gamma)^4\varepsilon^2\big))$ samples (Kearns & Singh, 1999; Kakade, 2004, chap. 9.1), where N and γ denote the size of state-action space and the discount factor, respectively.[1] Although this bound matches the best existing upper bound on the sample complexity of estimating the action-value function (Azar et al., 2011), it has not been clear, so far, whether this bound is a tight bound on the performance of QVI or it can be improved by a more careful analysis of QVI algorithm. This is mainly due to the fact that there is a gap of order $1/(1-\gamma)^4$ between the upper bound of QVI and the state-of-the-art result for lower bound, which is of $\tilde{\Omega}\big(N/\varepsilon^2\big)$ (Strehl et al., 2009).[2] [3]

In this paper, we focus on the problems which are formulated as finite state-action discounted infinite-horizon Markov decision processes (MDPs), and prove a new tight bound of $O\big(N\log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$ on the sample complexity of the QVI algorithm. The new upper bound improves on the existing bound of QVI by

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[1] The notation $g = \tilde{O}(f)$ implies that there are constants c_1 and c_2 such that $g \le c_1 f \log^{c_2}(f)$.

[2] The result of (Strehl et al., 2009) is different from our result since they consider the sample complexity of exploration as opposed to the sample complexity of estimation. However, based on their model, one can prove the same lower-bound of $\Omega\big(N/\varepsilon^2\big)$ on the sample complexity of estimating the optimal action-value function by using the approach that we propose in this paper.

[3] The notation $g = \tilde{\Omega}(f)$ implies that there are constants c_1 and c_2 such that $g \ge c_1 f \log^{c_2}(f)$.

an order of $1/(1-\gamma)$. We also present a new minimax lower bound of $\Theta\big(N\log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$, which also improves on the best existing lower bound of RL by an order of $1/(1-\gamma)^3$. The new results, which close the above-mentioned gap between the lower bound and the upper bound, guarantee that no learning method, given the generative model of the MDP, can be significantly more efficient than QVI in terms of the sample complexity of estimating the action-value function.

The main idea to improve the upper bound of QVI is to express the performance loss of QVI in terms of the variance of the sum of discounted rewards as opposed to the maximum $V_{\max} = R_{\max}/(1-\gamma)$ in the previous results. For this we make use of Bernstein's concentration inequality (Cesa-Bianchi & Lugosi, 2006, appendix, pg. 361), which bounds the estimation error in terms of the variance of the value function as opposed to $V_{\max}$ in previous works. We also rely on the fact that the variance of the sum of discounted rewards, like the expected value of the sum (value function), satisfies a Bellman-like equation, in which the variance of the value function plays the role of the instant reward in the standard Bellman equation (Munos & Moore, 1999). In the case of lower bound, we improve on the result of Strehl et al. (2009) by adding some structure to the class of MDPs for which we prove the lower bound: In the new model, there is a high probability for transition from every intermediate state to itself. This adds to the difficulty of estimating the value function, since even a small estimation error may propagate throughout the recursive structure of the MDP and inflict a big performance loss especially for γ's close to 1.

The rest of the paper is organized as follows. After introducing the notations used in the paper in Section 2, we describe the *model-based Q-value iteration* (QVI) algorithm in Subsection 2.1. We then state our main theoretical results, which are in the form of PAC sample complexity bounds in Section 3. Section 4 contains the detailed proofs of the results of Sections 3, i.e., sample complexity bound of QVI and a general new lower bound for RL. Finally, we conclude the paper and propose some directions for the future work in Section 5.

2. Background

In this section, we review some standard concepts and definitions from the theory of Markov decision processes (MDPs). We then present the model-based Q-value iteration algorithm of Kearns & Singh (1999).

We consider the standard reinforcement learning (RL) framework (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) in which a learning agent interacts with a stochastic environment and this interaction is modeled as a discrete-time discounted MDP. A discounted MDP is a quintuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{X}$ and $\mathcal{A}$ are the set of states and actions, P is the state transition distribution, $\mathcal{R}$ is the reward function, and $\gamma \in (0,1)$ is a discount factor. We denote by $P(\cdot|x,a)$ and $r(x,a)$ the probability distribution over the next state and the immediate reward of taking action a at state x, respectively.

Remark 1. *To keep the representation succinct, in the sequel, we use the notation $\mathcal{Z}$ for the joint state-action space $\mathcal{X}\times\mathcal{A}$. We also make use of the shorthand notations z and β for the state-action pair (x,a) and $1/(1-\gamma)$, respectively.*

Assumption 1 (MDP Regularity)**.** *We assume $\mathcal{Z}$ and, subsequently, $\mathcal{X}$ and $\mathcal{A}$ are finite sets with cardinalities N, $|\mathcal{X}|$ and $|\mathcal{A}|$, respectively. We also assume that the immediate reward $r(x,a)$ is taken from the interval $[0,1]$.*

A mapping $\pi : \mathcal{X} \to \mathcal{A}$ is called a stationary and deterministic Markovian policy, or just a policy in short. Following a policy π in an MDP means that at each time step t the control action $A_t \in \mathcal{A}$ is given by $A_t = \pi(X_t)$, where $X_t \in \mathcal{X}$. The *value* and the *action-value functions* of a policy π, denoted respectively by $V^\pi : \mathcal{X} \to \mathbb{R}$ and $Q^\pi : \mathcal{Z} \to \mathbb{R}$, are defined as the expected sum of discounted rewards that are encountered when the policy π is executed. Given an MDP, the goal is to find a policy that attains the best possible values, $V^*(x) \triangleq \sup_\pi V^\pi(x)$, $\forall x \in \mathcal{X}$. Function V^* is called the *optimal value function*. Similarly the *optimal action-value function* is defined as $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$. We say that a policy π^* is optimal if it attains the optimal $V^*(x)$ for all $x \in \mathcal{X}$. The policy π defines the state transition kernel P_π as: $P_\pi(y|x) \triangleq P(y|x,\pi(x))$ for all $x \in \mathcal{X}$. The right-linear operators $P^\pi\cdot$, $P\cdot$ and $P_\pi\cdot$ are then defined as $(P^\pi Q)(z) \triangleq \sum_{y\in\mathcal{X}} P(y|z)Q(y,\pi(y))$, $(PV)(z) \triangleq \sum_{y\in\mathcal{X}} P(y|z)V(y)$ for all $z \in \mathcal{Z}$ and $(P_\pi V)(x) \triangleq \sum_{y\in\mathcal{X}} P_\pi(y|x)V(y)$ for all $x \in \mathcal{X}$, respectively. The optimal action-value function Q^* is the unique fixed-point of the *Bellman optimality operator* defined as $(\mathcal{T}Q)(z) \triangleq r(z) + \gamma(P^{\pi^*}Q)(z)$ for all $z \in \mathcal{Z}$. Also, the action-value function Q^π is the unique fixed-point of the *Bellman operator* $\mathcal{T}^\pi$ which is defined as $(\mathcal{T}^\pi Q)(z) \triangleq r(z) + \gamma(P^\pi Q)(z)$ for all $z \in \mathcal{Z}$. One can also define the Bellman optimality operator and the Bellman operator on the value function as $(\mathcal{T}V)(x) \triangleq r(x,\pi^*(x)) + \gamma(P_{\pi^*}V)(x)$ and $(\mathcal{T}^\pi V)(x) \triangleq r(x,\pi(x)) + \gamma(P_\pi V)(x)$ for all $x \in \mathcal{X}$, re-

spectively.[4] [5]

2.1. Model-based Q-value Iteration (QVI)

The algorithm makes n transition samples from each state-action pair $z \in \mathcal{Z}$ for which it makes n calls to the generative model.[6] It then builds an empirical model of the transition probabilities as: $\widehat{P}(y|z) \triangleq m(y,z)/n$, where $m(y,z)$ denotes the number of times that the state $y \in \mathcal{X}$ has been reached from $z \in \mathcal{Z}$. The algorithm then makes an empirical estimate of the optimal action-value function Q^* by iterating some action-value function Q_k, with the initial value of Q_0, through the empirical Bellman optimality operator $\widehat{\mathcal{T}}$.[7]

3. Main Results

Our main results are in the form of PAC (probably approximately correct) bounds on the ℓ_∞-norm of the difference of the optimal action-value function Q^* and its sample estimate:

Theorem 1 (PAC-bound for model-based Q-value iteration)**.** *Let Assumption 1 hold and T be a positive integer. Then, there exist some constants c and c_0 such that for all $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ a total sampling budget of*

$$T = \lceil \frac{c\beta^3 N}{\varepsilon^2} \log \frac{c_0 N}{\delta} \rceil,$$

suffices for the uniform approximation error $\|Q^* - Q_k\| \le \varepsilon$, w.p. (with the probability) at least $1-\delta$, after only $k = \lceil \log(6\beta/\varepsilon)/\log(1/\gamma) \rceil$ iteration of QVI algorithm.[8] *In particular, one may choose $c = 68$ and $c_0 = 12$.*

The following general result provides a tight lower bound on the number of transitions T for every RL algorithm to achieve an ε-optimal estimate of the action-value function w.p. $1-\delta$, under the assumption that the algorithm is (ε,δ,T)-correct:

Definition 1 ((ε,δ,T)*-correct algorithm*)**.** *Let $Q_T^{\mathfrak{A}}$ be the estimate of Q^* by an RL algorithm $\mathfrak{A}$ after observing $T \ge 0$ transition samples. We say that $\mathfrak{A}$ is (ε,δ,T)-correct on the class of MDPs $\mathbb{M}$ if $\|Q^* - Q_T^{\mathfrak{A}}\| \le \varepsilon$ with probability at least $1-\delta$ for all $M \in \mathbb{M}$.*[9]

Theorem 2 (Lower bound on the sample complexity of estimating the optimal action-value function)**.** *There exists some constants ε_0, δ_0, c_1, c_2, and a class of MDPs $\mathbb{M}$, such that for all $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0/N)$, and every (ε,δ,T)-correct RL algorithm $\mathfrak{A}$ on the class of MDPs $\mathbb{M}$ the number of transitions needs to be at least*

$$T = \lceil \frac{\beta^3 N}{c_1 \varepsilon^2} \log \frac{N}{c_2 \delta} \rceil.$$

4. Analysis

In this section, we first provide the full proof of the finite-time PAC bound of QVI, reported in Theorem 1, in Subsection 4.1. We then prove Theorem 2, the RL lower bound, in Subsection 4.2.

4.1. Poof of Theorem 1

We begin by introducing some new notation. Consider the stationary policy π. We define $\mathbb{V}^\pi(z) \triangleq \mathbb{E}[|\sum_{t\ge 0}\gamma^t r(Z_t) - Q^\pi(z)|^2]$ as the variance of the sum of discounted rewards starting from $z \in \mathcal{Z}$ under the policy π. Also, define $\sigma^\pi(z) \triangleq \gamma^2 \sum_{y\in\mathcal{Z}} P^\pi(y|z)|Q^\pi(y) - P^\pi Q^\pi(z)|^2$ as the immediate variance at $z \in \mathcal{Z}$, i.e., $\gamma^2 \mathbb{V}_{Y\sim P^\pi(\cdot|z)}[Q^\pi(Y)]$. Also, we shall denote v^π and v^* as the immediate variance of the value function V^π and V^* defined as $v^\pi(z) \triangleq \gamma^2 \mathbb{V}_{Y\sim P(\cdot|z)}[V^\pi(Y)]$ and $v^*(z) \triangleq \gamma^2 \mathbb{V}_{Y\sim P(\cdot|z)}[V^*(Y)]$, for all $z \in \mathcal{Z}$, respectively. Further, we denote the immediate variance of the action-value function $\widehat{Q}^\pi$, $\widehat{V}^\pi$ and $\widehat{V}^*$ by $\widehat{\sigma}^\pi$, $\widehat{v}^\pi$ and $\widehat{v}^*$, respectively.

We now prove our first result which indicates that Q_k is very close to $\widehat{Q}^*$ up to an order of $O(\gamma^k)$. Therefore, to prove bound on $\|Q^* - Q_k\|$, one only needs to bound

[4] It is important to note that $\mathcal{T}$ and $\mathcal{T}^\pi$ are contraction with factor γ, i.e., for any pair of value functions V and V' and any policy π, we have $\|\mathcal{T}V - \mathcal{T}V'\| \le \gamma \|V - V'\|$ and $\|\mathcal{T}^\pi V - \mathcal{T}^\pi V'\| \le \gamma \|V - V'\|$ (Bertsekas, 2007, Chap. 1), where $\|\cdot\|$ shall denote the supremum (ℓ_∞) norm which is defined as $\|g\| \triangleq \max_{y\in\mathcal{Y}} |g(y)|$, where $\mathcal{Y}$ is a finite set and $g : \mathcal{Y} \to \mathbb{R}$ is a real-valued function.

[5] For simplicity of the notations, in the sequel, we remove the dependence on z and x , e.g., writing Q for $Q(z)$ and V for $V(x)$, when there is no possible confusion.

[6] The total number of calls to the generative model is given by $T = nN$.

[7] The operator $\widehat{\mathcal{T}}$ is defined on the action-value function Q, for all $z \in \mathcal{Z}$, by $\widehat{\mathcal{T}}Q(z) = r(z) + \gamma \widehat{P}V(z)$, with $V(x) = \max_{a\in\mathcal{A}}(Q(x,a))$ for all $x \in \mathcal{X}$. Also, the empirical operator $\widehat{\mathcal{T}}^\pi$ is defined on the action-value function Q, for every policy π and all $z \in \mathcal{Z}$, by $\widehat{\mathcal{T}}^\pi Q(z) = r(z) + \gamma \widehat{P}^\pi Q(z)$. Likewise, one can also define the empirical Bellman operator $\widehat{\mathcal{T}}$ and $\widehat{\mathcal{T}}^\pi$ for the value function V. The fixed points of the operator $\widehat{\mathcal{T}}$ in $\mathcal{Z}$ and $\mathcal{X}$ domains are denoted by $\widehat{Q}^*$ and $\widehat{V}^*$, respectively. Also, the fixed points of the operator $\widehat{\mathcal{T}}^\pi$ in $\mathcal{Z}$ and $\mathcal{X}$ domains are denoted by $\widehat{Q}^\pi$ and $\widehat{V}^\pi$, respectively. The empirical optimal policy $\widehat{\pi}^*$ is the policy which attains $\widehat{V}^*$ under the model $\widehat{P}$.

[8] For every real number u, $\lceil u \rceil$ is defined as the smallest integer number not less than u.

[9] The algorithm $\mathfrak{A}$, unlike QVI, does not need to generate a same number of transition samples for every state-action pair and can generate samples arbitrarily.

$\|Q^* - \widehat{Q}^*\|$ in high probability.

Lemma 1. *Let Assumption 1 hold and $Q_0(z)$ be in the interval $[0, \beta]$ for all $z \in \mathcal{Z}$. Then we have*

$$\|Q_k - \widehat{Q}^*\| \le \gamma^k \beta.$$

Proof. For all $k \ge 0$, we have

$$\|Q_k - \widehat{Q}^*\| = \|\widehat{\mathcal{T}} Q_{k-1} - \widehat{\mathcal{T}} \widehat{Q}^*\| \le \gamma \|Q_{k-1} - \widehat{Q}^*\|.$$

Thus by an immediate recursion

$$\|Q_k - \widehat{Q}^*\| \le \gamma^k \|Q_0 - \widehat{Q}^*\| \le \gamma^k \beta.$$

□

In the rest of this subsection, we focus on proving a high probability bound on $\|Q^* - \widehat{Q}^*\|$. One can prove a crude bound of $\tilde{O}(\beta^2/\sqrt{n})$ on $\|Q^* - \widehat{Q}^*\|$ by first proving that $\|Q^* - \widehat{Q}^*\| \le \beta \|(P - \widehat{P})V^*\|$ and then using the Hoeffding's tail inequality (Cesa-Bianchi & Lugosi, 2006, appendix, pg. 359) to bound the random variable $\|(P - \widehat{P})V^*\|$ in high probability. Here, we follow a different and more subtle approach to bound $\|Q^* - \widehat{Q}^*\|$, which leads to a tight bound of $\tilde{O}(\beta^{1.5}/\sqrt{n})$: **(i)** We prove in Lemma 2 component-wise upper and lower bounds on the error $Q^* - \widehat{Q}^*$ which are expressed in terms of $(I - \gamma \widehat{P}^{\pi^*})^{-1}\big[P - \widehat{P}\big]V^*$ and $(I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[P - \widehat{P}\big]V^*$, respectively. **(ii)** We make use of the sharp result of Bernstein's inequality to bound $\big[P - \widehat{P}\big]V^*$ in terms of the squared root of the variance of V^* in high probability. **(iii)** We prove the key result of this subsection (Lemma 6) which shows that the variance of the sum of discounted rewards satisfies a Bellman-like recursion, in which the instant reward $r(z)$ is replaced by $\sigma^\pi(z)$. Based on this result we prove an upper-bound of order $O(\beta^{1.5})$ on $(I - \gamma P^\pi)^{-1}\sqrt{\mathbb{V}(Q^\pi)}$ for any policy π, which combined with the previous steps leads to the sharp upper bound of $\tilde{O}(\beta^{1.5}/\sqrt{n})$ on $\|Q^* - \widehat{Q}^*\|$. We now prove the following component-wise bounds on $Q^* - \widehat{Q}^*$ from above and below:

Lemma 2 (Component-wise bounds on $Q^* - \widehat{Q}^*$)**.**

$$Q^* - \widehat{Q}^* \le \gamma (I - \gamma \widehat{P}^{\pi^*})^{-1}\big[P - \widehat{P}\big]V^*, \qquad (1)$$
$$Q^* - \widehat{Q}^* \ge \gamma (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[P - \widehat{P}\big]V^*. \qquad (2)$$

Proof. We have that $\widehat{Q}^* \ge \widehat{Q}^{\pi^*}$. Thus:

$$\begin{aligned} Q^* - \widehat{Q}^* &\le Q^* - \widehat{Q}^{\pi^*} \\ &= (I - \gamma P^{\pi^*})^{-1} r - (I - \gamma \widehat{P}^{\pi^*})^{-1} r \\ &= (I - \gamma \widehat{P}^{\pi^*})^{-1}\big[(I - \gamma \widehat{P}^{\pi^*}) \\ &\quad - (I - \gamma P^{\pi^*})\big](I - \gamma P^{\pi^*})^{-1} r \\ &= \gamma (I - \gamma \widehat{P}^{\pi^*})^{-1}\big[P^{\pi^*} - \widehat{P}^{\pi^*}\big]Q^* \\ &= \gamma (I - \gamma \widehat{P}^{\pi^*})^{-1}\big[P - \widehat{P}\big]V^*. \end{aligned}$$

For Ineq. (2) we have

$$\begin{aligned} Q^* - \widehat{Q}^* &= (I - \gamma P^{\pi^*})^{-1} r - (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1} r \\ &= (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[(I - \gamma \widehat{P}^{\widehat{\pi}^*}) \\ &\quad - (I - \gamma P^{\pi^*})\big](I - \gamma P^{\pi^*})^{-1} r \\ &= \gamma (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[P^{\pi^*} - \widehat{P}^{\widehat{\pi}^*}\big]Q^* \\ &\ge \gamma (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[P^{\pi^*} - \widehat{P}^{\pi^*}\big]Q^* \\ &= \gamma (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}\big[P - \widehat{P}\big]V^*, \end{aligned}$$

in which we make use of the following component-wise inequalities:

$$\widehat{P}^{\pi^*} Q^* \ge \widehat{P}^{\widehat{\pi}^*} Q^*, \qquad \text{and} \qquad (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1} \ge 0.$$

□

We now concentrate on bounding the RHS (right hand sides) of (1) and (2), for that we need the following technical lemmas (Lemma 3 and Lemma 4).

Lemma 3. *Let Assumption 1 hold. Then, for any $0 < \delta < 1$ with probability $1 - \delta$,*

$$\|V^* - \widehat{V}^{\pi^*}\| \le c_v, \qquad \text{and} \qquad \|V^* - \widehat{V}^*\| \le c_v,$$

where $c_v \triangleq \gamma \beta^2 \sqrt{2 \log(2|\mathcal{X}|/\delta)/n}$.

Proof. We begin by proving bound on $\|V^* - \widehat{V}^{\pi^*}\|$:

$$\begin{aligned} \|V^* - \widehat{V}^{\pi^*}\| &= \|\mathcal{T}^{\pi^*} V^* - \widehat{\mathcal{T}}^{\pi^*} \widehat{V}^{\pi^*}\| \\ &\le \|\mathcal{T}^{\pi^*} V^* - \widehat{\mathcal{T}}^{\pi^*} V^*\| \\ &\quad + \|\widehat{\mathcal{T}}^{\pi^*} V^* - \widehat{\mathcal{T}}^{\pi^*} \widehat{V}^{\pi^*}\| \\ &\le \gamma \|P_{\pi^*} V^* - \widehat{P}_{\pi^*} V^*\| + \gamma \|V^* - \widehat{V}^{\pi^*}\|. \end{aligned}$$

By collecting terms we deduce:

$$\|V^* - \widehat{V}^{\pi^*}\| \le \gamma \beta \|(P_{\pi^*} - \widehat{P}_{\pi^*})V^*\|. \qquad (3)$$

By using a similar argument the same bound can be proven on $\|V^* - \widehat{V}^*\|$:

$$\|V^* - \widehat{V}^*\| \le \gamma \beta \|(P_{\pi^*} - \widehat{P}_{\pi^*})V^*\|. \qquad (4)$$

We then make use of Hoeffding's inequality to bound $|(P_{\pi^*} - \widehat{P}_{\pi^*})V^*(x)|$ for all $x \in \mathcal{X}$ in high probability:

$$\mathbb{P}(|((P_{\pi^*} - \widehat{P}_{\pi^*})V^*)(x)| \geq \varepsilon) \leq 2e^{\frac{-n\varepsilon^2}{2\beta^2}}.$$

By applying the union bound we deduce:

$$\mathbb{P}(\|(P_{\pi^*} - \widehat{P}_{\pi^*})V^*\| \geq \varepsilon) \leq 2|\mathcal{X}|e^{\frac{-n\varepsilon^2}{2\beta^2}}. \tag{5}$$

We then define the probability of failure δ as follows:

$$\delta \triangleq 2|\mathcal{X}|e^{\frac{-n\varepsilon^2}{2\beta^2}}. \tag{6}$$

By plugging (6) into (5) we deduce:

$$\mathbb{P}\left[\|(P_{\pi^*} - \widehat{P}_{\pi^*})V^*\| < \beta\sqrt{2\log\left(2|\mathcal{X}|/\delta\right)/n}\right] \geq 1 - \delta. \tag{7}$$

The results then follow by plugging (7) into (3) and (4). □

Lemma 4 relates v^* to $\widehat{\sigma}^{\pi^*}$ and $\widehat{\sigma}^*$. We make use of this result in 5.

Lemma 4. *Let Assumption 1 hold and $0 < \delta < 1$. Then, w.p. at least $1 - \delta$:*

$$v^* \leq \widehat{\sigma}^{\pi^*} + b_v\mathbf{1}, \tag{8}$$
$$v^* \leq \widehat{\sigma}^* + b_v\mathbf{1}, \tag{9}$$

where b_v is defined as

$$b_v \triangleq \sqrt{\frac{18\gamma^4\beta^4\log\frac{3N}{\delta}}{n}} + \frac{4\gamma^2\beta^4\log\frac{3N}{\delta}}{n},$$

and $\mathbf{1}$ is a function which assigns 1 to all $z \in \mathcal{Z}$.

Proof. Here, we only prove (8). One can prove (9) following similar lines.

$$\begin{aligned}
v^*(z) &= v^*(z) - \gamma^2\mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y)) \\
&\quad + \gamma^2\mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y)) \\
&\leq \gamma^2\big((P - \widehat{P})V^{*2}\big)(z) \\
&\quad - \gamma^2[(PV^*)^2(z) - (\widehat{P}V^*)^2(z)] \\
&\quad + \gamma^2\mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y) - \widehat{V}^{\pi^*}(Y)) \\
&\quad + \gamma^2\mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(\widehat{V}^{\pi*}(Y)),
\end{aligned}$$

It is not difficult to show that $\mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y) - \widehat{V}^{\pi^*}(Y)) \leq \|V^* - \widehat{V}^{\pi^*}\|^2$, which implies that

$$\begin{aligned}
v^*(z) &\leq \gamma^2[P - \widehat{P}]V^{*2}(z) \\
&\quad - \gamma^2[(P - \widehat{P})V^*][(P + \widehat{P})V^*](z) \\
&\quad + \gamma^2\|V^* - \widehat{V}^{\pi^*}\|^2 + \widehat{v}^{\pi^*}(z).
\end{aligned}$$

The following inequality then holds w.p. at least $1-\delta$:

$$v^*(z) \leq \widehat{v}^{\pi^*}(z) + \gamma^2\left[3\beta^2\sqrt{2\frac{\log\frac{3}{\delta}}{n}} + \frac{2\beta^4\log\frac{3N}{\delta}}{n}\right], \tag{10}$$

in which we make use of Hoeffding's inequality as well as Lemma 3 and a union bound to prove the bound on v^* in high probability. It is not then difficult to show that for every policy π and for all $z \in \mathcal{Z}$: $v^\pi(z) \leq \sigma^\pi(z)$. This combined with a union bound on all state-action pairs in Eq.(10) completes the proof. □

The following result proves a sharp bound on $\gamma(P - \widehat{P})V^*$, for which we make use of Bernstein's inequality (Cesa-Bianchi & Lugosi, 2006, appendix, pg. 361) as well as Lemma 4.

Lemma 5. *Let Assumption 1 hold and $0 < \delta < 1$. Define $c_{pv} \triangleq 2\log(2N/\delta)$ and b_{pv} as:*

$$b_{pv} \triangleq \left(\frac{6(\gamma\beta)^{4/3}\log\frac{6N}{\delta}}{n}\right)^{3/4} + \frac{5\gamma\beta^2\log\frac{6N}{\delta}}{n}.$$

Then w.p. $1 - \delta$ we have

$$\gamma(P - \widehat{P})V^* \leq \sqrt{\frac{c_{pv}\widehat{\sigma}^{\pi^*}}{n}} + b_{pv}\mathbf{1}, \tag{11}$$

$$\gamma(P - \widehat{P})V^* \geq -\sqrt{\frac{c_{pv}\widehat{\sigma}^*}{n}} - b_{pv}\mathbf{1}. \tag{12}$$

Proof. For all $z \in \mathcal{Z}$ and all $0 < \delta < 1$, Bernstein's inequality implies that w.p. at least $1 - \delta$:

$$(P - \widehat{P})V^*(z) \leq \sqrt{\frac{2v^*(z)\log\frac{1}{\delta}}{\gamma^2 n}} + \frac{2\beta\log\frac{1}{\delta}}{3n},$$

$$(P - \widehat{P})V^*(z) \geq -\sqrt{\frac{2v^*(z)\log\frac{1}{\delta}}{\gamma^2 n}} - \frac{2\beta\log\frac{1}{\delta}}{3n}.$$

We deduce (using a union bound):

$$\gamma(P - \widehat{P})V^* \leq \sqrt{c'_{pv}\frac{v^*}{n}} + b'_{pv}\mathbf{1}, \tag{13}$$

$$\gamma(P - \widehat{P})V^* \geq -\sqrt{c'_{pv}\frac{v^*}{n}} - b'_{pv}\mathbf{1}, \tag{14}$$

where $c'_{pv} \triangleq 2\log(N/\delta)$ and $b'_{pv} \triangleq 2\gamma\beta\log(N/\delta)/3n$. The result then follows by plugging (8) and (9) into (13) and (14), respectively, and then taking a union bound. □

We now state the key lemma of this section which shows that for any policy π the variance $\mathbb{V}^\pi$ satisfies the following Bellman-like recursion. Later, we use this result, in Lemma 7, to bound $(I - \gamma P^\pi)^{-1}\sigma^\pi$.

Lemma 6. $\mathbb{V}^\pi$ *satisfies the Bellman equation*

$$\mathbb{V}^\pi = \sigma^\pi + \gamma^2 P^\pi \mathbb{V}^\pi.$$

Proof. For all $z \in \mathcal{Z}$ we have

$$\begin{aligned}
\mathbb{V}^\pi(z) &= \mathbb{E}\left[\left|\sum_{t\geq 0}\gamma^t r(Z_t) - Q^\pi(z)\right|^2\right] \\
&= \mathbb{E}_{Z_1\sim P^\pi(.|z)}\mathbb{E}\Bigg[\Bigg|\sum_{t\geq 1}\gamma^t r(Z_t) - \gamma Q^\pi(Z_1) \\
&\quad - (Q^\pi(z) - r(z) - \gamma Q^\pi(Z_1))\Bigg|^2\Bigg] \\
&= \gamma^2 \mathbb{E}_{Z_1\sim P^\pi(.|z)}\mathbb{E}\left[\left|\sum_{t\geq 1}\gamma^{t-1} r(Z_t) - Q^\pi(Z_1)\right|^2\right] \\
&\quad - 2\mathbb{E}_{Z_1\sim P^\pi(.|z)}\Bigg[(Q^\pi(z) - r(z) - \gamma Q^\pi(Z_1)) \\
&\quad \times \mathbb{E}\left(\sum_{t\geq 1}\gamma^t r(Z_t) - \gamma Q^\pi(Z_1)\middle| Z_1\right)\Bigg] \\
&\quad + \mathbb{E}_{Z_1\sim P^\pi(.|z)}(|Q^\pi(z) - r(z) - \gamma Q^\pi(Z_1)|^2) \\
&= \gamma^2 \mathbb{E}_{Z_1\sim P^\pi(.|z)}\mathbb{E}\left[\left|\sum_{t\geq 1}\gamma^{t-1} r(Z_t) - Q^\pi(Z_1)\right|^2\right] \\
&\quad + \gamma^2 \mathbb{V}_{Z_1\sim P^\pi(.|z)}(Q^\pi(Z_1)) \\
&= \gamma^2 \sum_{y\in\mathcal{Z}} P^\pi(y|z)\mathbb{V}^\pi(y) + \sigma^\pi(z),
\end{aligned}$$

in which we rely on $\mathbb{E}(\sum_{t\geq 1}\gamma^t r(Z_t) - \gamma Q^\pi(Z_1)|Z_1) = 0$. □

Based on Lemma 6, one can prove the following result on the immediate variance.

Lemma 7.

$$\|(I - \gamma^2 P^\pi)^{-1}\sigma^\pi\| \leq \beta^2, \tag{15}$$
$$\|(I - \gamma P^\pi)^{-1}\sqrt{\sigma^\pi}\| \leq 2\log(2)\beta^{1.5}. \tag{16}$$

Proof. The first inequality follows from Lemma 6 by solving (6) in terms of $\mathbb{V}^\pi$ and taking the sup-norm over both sides of the resulted equation. In the case of Eq.(16) we have [10]

$$\begin{aligned}
\|(I - \gamma P^\pi)^{-1}\sqrt{\sigma^\pi}\| &= \left\|\sum_{k\geq 0}(\gamma P^\pi)^k\sqrt{\sigma^\pi}\right\| \\
&= \left\|\sum_{l\geq 0}(\gamma P^\pi)^{tl}\sum_{j=0}^{t-1}(\gamma P^\pi)^j\sqrt{\sigma^\pi}\right\| \\
&\leq \sum_{l\geq 0}(\gamma^t)^l\left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j\sqrt{\sigma^\pi}\right\| \\
&= \frac{1}{1-\gamma^t}\left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j\sqrt{\sigma^\pi}\right\|,
\end{aligned} \tag{17}$$

in which we write $k = tl + j$ with t is a positive integer. We now prove a bound on $\left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j\sqrt{\sigma^\pi}\right\|$ by making use of Jensen's inequality as well as Cauchy-Shwarz inequality:

$$\begin{aligned}
\left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j\sqrt{\sigma^\pi}\right\| &\leq \left\|\sum_{j=0}^{t-1}\gamma^j\sqrt{(P^\pi)^j\sigma^\pi}\right\| \\
&\leq \sqrt{t}\left\|\sqrt{\sum_{j=0}^{t-1}(\gamma^2 P^\pi)^j\sigma^\pi}\right\| \\
&\leq \sqrt{t}\left\|\sqrt{(I-\gamma^2 P^\pi)^{-1}\sigma^\pi}\right\| \\
&\leq \beta\sqrt{t},
\end{aligned} \tag{18}$$

where in the last step we rely on (15). The result then follows by plugging (18) into (17) and optimizing the bound in terms of t to achieve the best dependency on β. □

Now, we make use of Lemma 7 and Lemma 5 to bound $\|Q^* - \widehat{Q}^*\|$ in high probability.

Lemma 8. *Let Assumption 1 hold. Then, for any* $0 < \delta < 1$*:*

$$\|Q^* - \widehat{Q}^*\| \leq \varepsilon',$$

w.p. $1-\delta$*, where* ε' *is defined as:*

$$\begin{aligned}
\varepsilon' \triangleq \sqrt{\frac{17\beta^3\log\frac{4N}{\delta}}{n}} + \left(\frac{6(\gamma\beta^2)^{4/3}\log\frac{12N}{\delta}}{n}\right)^{3/4} \\
+ \frac{5\gamma\beta^3\log\frac{12N}{\delta}}{n}.
\end{aligned} \tag{19}$$

Proof. By incorporating the result of Lemma 5 and

[10] For any real-valued function f, $\sqrt{f}$ is defined as a component wise squared-root operator on f.

Lemma 7 into Lemma 2, we deduce that:

$$Q^* - \widehat{Q}^* \leq b\mathbf{1},$$
$$Q^* - \widehat{Q}^* \geq -b\mathbf{1},$$

w.p. $1-\delta$. The scalar b is given by:

$$b \triangleq \sqrt{\frac{17\beta^3 \log\frac{2N}{\delta}}{n}} + \left(\frac{6(\gamma\beta^2)^{4/3}\log\frac{6N}{\delta}}{n}\right)^{3/4} + \frac{5\gamma\beta^3\log\frac{6N}{\delta}}{n}.$$

The result then follows by combining these two bounds and taking the ℓ_∞ norm. □

Proof of Theorem 1. We combine the proof of Lemma 8 and Lemma 1 in order to bound $Q^* - Q_k$ in high probability. We then solve the resulted bound w.r.t. n and k.[11] □

4.2. Proof of the Lower-bound

In this section, we provide the proof of Theorem 2. In our analysis, we rely on the likelihood-ratio method, which has been previously used to prove a lower bound for multi-armed bandits (Mannor & Tsitsiklis, 2004), and extend this approach to RL and MDPs. We begin by defining a class of MDPs for which the proposed lower bound will be obtained (see Figure 1). We define the class of MDPs $\mathbb{M}$ as the set of all MDPs with the state-action space of cardinality $N = 3KL$, where K and L are positive integers. Also, we assume that for all $M \in \mathbb{M}$, the state space $\mathcal{X}$ consists of three smaller sets $\mathcal{S}$, $\mathcal{Y}_1$ and $\mathcal{Y}_2$. The set $\mathcal{S}$ includes K states, each of those states corresponds with the set of actions $\mathcal{A} = \{a_1, a_2, \ldots, a_L\}$, whereas the states in $\mathcal{Y}^1$ and $\mathcal{Y}^2$ are single-action states. By taking the action $a \in \mathcal{A}$ from every state $x \in \mathcal{S}$, we move to the next state $y(z) \in \mathcal{Y}^1$ with the probability 1, where $z = (x, a)$. The transition probability from $\mathcal{Y}^1$ is characterized by the transition probability p_M from every $y(z) \in \mathcal{Y}^1$ to itself and with the probability $1 - p_M$ to the corresponding $y(z) \in \mathcal{Y}^2$.[12] Further, for all $M \in \mathbb{M}$, $\mathcal{Y}^2$ consists of only absorbing states, i.e., for all $y \in \mathcal{Y}^2$, $P(x|x) = 1$. The instant reward r is set to 1 for every state in $\mathcal{Y}^1$ and 0 elsewhere. For this class of MDPs, the optimal action-value function Q^* can be solved in close form from the Bellman equation:

$$Q^*(z) = \gamma V^*(y(z)) = \frac{\gamma}{1-\gamma p_M}, \qquad \forall z \in \mathcal{S}\times\mathcal{A},$$

[11] Note that the total number of samples is then computed by $T = Nn$.

[12] Every state $y \in \mathcal{Y}^2$ is only connected to one state in $\mathcal{Y}^1$ and $\mathcal{S}$, i.e., there is no overlapping path in the MDP.

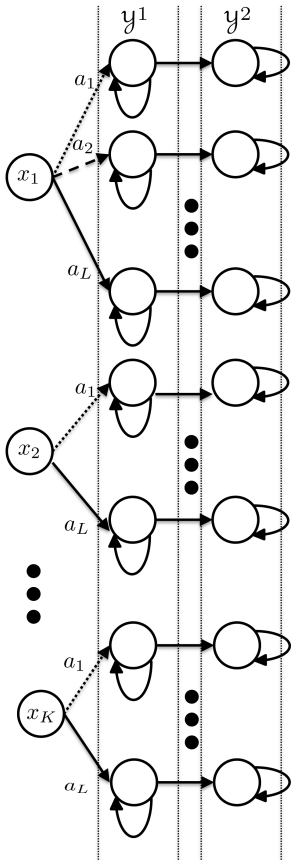

Figure 1. The class of MDPs considered in the proof of Theorem 2. Nodes represent states and arrows show transitions between the states (see the text for details).

In the rest of the proof, we concentrate on proving the lower bound for $\|Q^* - Q_T^{\mathfrak{A}}\|$ for all $z \in \mathcal{S}\times\mathcal{A}$. Now, let us consider a set of two MDPs $\mathbb{M}^* = \{M_0, M_1\}$ in $\mathbb{M}$ with the transition probabilities

$$p_M = \begin{cases} p & M = M_0, \\ p+\alpha & M = M_1, \end{cases}$$

where α and p are some positive numbers such that $0 < p < p + \alpha \leq 1$, which will be quantified later in this section. We assume that the discount factor γ is bounded from below by some positive constant γ_0 for the class of MDP M^*. We denote by $\mathbb{E}_m$ ad $\mathbb{P}_m$ the expectation and the probability under the model M_m in the rest of this section.

We follow the following steps in the proof: **(i)** we prove a lower bound on the sample-complexity of learning the value function for every state $y \in \mathcal{Y}$ on the class of MDP $\mathbb{M}^*$ **(ii)** we then make use of the fact that the

estimates of $Q^*(z)$ for different $z \in \mathcal{S} \times \mathcal{A}$ are independent of each others to combine these bounds and prove the tight result of Theorem 2.

We begin our analysis of the lower bound by the following lemma:

Lemma 9. *Define $\theta \triangleq \exp\big(-c_1'\alpha^2 t/(p(1-p))\big)$ and $Q_t^{\mathfrak{A}}(z)$ as an empirical estimate of the action-value function $Q^*(z)$ by an RL algorithm $\mathfrak{A}$ using $t > 0$ transition samples from the state $y(z) \in \mathcal{Y}^1$ for $z \in \mathcal{X} \times \mathcal{A}$. Then, for every RL algorithm $\mathfrak{A}$, there exists an MDP $M_m \in \mathbb{M}^*$ and constants $c_1' > 0$ and $c_2' > 0$ such that*

$$\mathbb{P}_m(|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) > \varepsilon) > \frac{\theta}{c_2'}, \tag{20}$$

by the choice of $\alpha = 2(1-\gamma p)^2\varepsilon/(\gamma^2)$.

Proof. To prove this result we make use of a contradiction argument, i.e., we assume that there exists an algorithm $\mathfrak{A}$ for which:

$$\begin{aligned} &\mathbb{P}_m((|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) > \varepsilon) \le \frac{\theta}{c_2'}, \quad \text{or} \\ &\mathbb{P}_m((|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) \le \varepsilon) \ge 1 - \frac{\theta}{c_2'}, \end{aligned} \tag{21}$$

for all $M_m \in \mathbb{M}^*$ and show that this assumption leads to a contradiction. To prove this result, we need, first, to introduce some notations: We define the event $\mathcal{E}_1(z) \triangleq \{|Q_0^*(z) - Q_t^{\mathfrak{A}}(z)| \le \varepsilon\}$ for all $z \in \mathcal{S} \times \mathcal{A}$, where $Q_0^* \triangleq \gamma/(1-\gamma p)$ is the optimal action-value function for all $z \in \mathcal{S} \times \mathcal{A}$ under the MDP M_0. We then define $k \triangleq r_1 + r_2 + \cdots + r_t$ as the sum of rewards of making t transitions from $y(z) \in \mathcal{Y}^1$. We also introduce the event $\mathcal{E}_2(z)$, for all $z \in \mathcal{S} \times \mathcal{A}$ as:

$$\mathcal{E}_2(z) \triangleq \left\{ pt - k \le \sqrt{2p(1-p)t \log \frac{c_2'}{2\theta}} \right\}.$$

Further, we define $\mathcal{E}(z) \triangleq \mathcal{E}_1(z) \cap \mathcal{E}_2(z)$. We then state the following technical lemma required for our analysis.

Lemma 10. *For all $p > \frac{1}{2}$:*

$$\mathbb{P}_0(\mathcal{E}_2(z)) > 1 - \frac{2\theta}{c_2'}.$$

Proof. We then make use of the following concentration inequality (Chernoff bound) for Binomial random variables (Hagerup & Rüb, 1990). For $p > \frac{1}{2}$, we have

$$\begin{aligned} \mathbb{P}_0(\mathcal{E}_2(z)) &> 1 - e^{-\frac{2tp(1-p)\log\frac{c_2'}{\theta}}{2tp(1-p)}} \\ &= 1 - e^{-\log\frac{c_2'}{\theta}} = 1 - \frac{2\theta}{c_2'}, \end{aligned} \quad \forall z \in \mathcal{S} \times \mathcal{A}.$$

Now, by the assumption that $\mathbb{P}_m(|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) > \varepsilon) \le \theta/c_2'$ for all $M_m \in \mathbb{M}^*$, we have $\mathbb{P}_0(\mathcal{E}_1(z)) \ge 1 - \theta/c_2' \ge 1 - 1/c_2'$. This combined with Lemma 10 and with the choice of $c_2' = 6$ implies that $\mathbb{P}_0(\mathcal{E}(z)) > 1/2$, for all $z \in \mathcal{S} \times \mathcal{A}$. Based on this result, we prove the following lemma:

Lemma 11. *For all $z \in \mathcal{S} \times \mathcal{A}$: $\mathbb{P}_1(\mathcal{E}_1(z)) > \theta/c_2'$.*

Proof. We define W as the history of all the outcomes of trying z for t times and the likelihood function $L_m(w)$ for all $M_m \in \mathbb{M}^*$ as:

$$L_m(w) \triangleq \mathbb{P}_m(W = w),$$

for every possible history w and $M_m \in \mathbb{M}^*$. This function can be used to define a random variable $L_m(W)$, where W is the sample random path of the process (sequence of observed transitions). The likelihood ratio of the event W between two MDPs M_1 and M_0 can then be written as:

$$\begin{aligned} \frac{L_1(W)}{L_0(W)} &= \frac{(p+\alpha)^k(1-p-\alpha)^{t-k}}{p^k(1-p)^{t-k}} \\ &= \big(1 + \frac{\alpha}{p}\big)^k \big(1 - \frac{\alpha}{1-p}\big)^{t-k} \\ &= \big(1 + \frac{\alpha}{p}\big)^k \big(1 - \frac{\alpha}{1-p}\big)^{k\frac{1-p}{p}} \big(1 - \frac{\alpha}{1-p}\big)^{t-\frac{k}{p}}. \end{aligned}$$

Now, by making use of $\log(1-u) \ge -u - u^2$ for $0 \le u \le 1/2$, and $e^{-u} \ge 1 - u$ for $0 \le u \le 1$, we have

$$\begin{aligned} \big(1 - \frac{\alpha}{1-p}\big)^{(1-p)/p} &\ge e^{\frac{1-p}{p}\left(-\frac{\alpha}{1-p} - (\frac{\alpha}{1-p})^2\right)} \\ &\ge \left(1 - \frac{\alpha}{p}\right)\left(1 - \frac{\alpha^2}{p(1-p)}\right), \end{aligned}$$

for $\alpha \le (1-p)/2$. Thus

$$\begin{aligned} \frac{L_1(W)}{L_0(W)} &\ge \big(1 - \frac{\alpha^2}{p^2}\big)^k \big(1 - \frac{\alpha^2}{p(1-p)}\big)^k \big(1 - \frac{\alpha}{1-p}\big)^{t-\frac{k}{p}} \\ &\ge \big(1 - \frac{\alpha^2}{p^2}\big)^t \big(1 - \frac{\alpha^2}{p(1-p)}\big)^t \big(1 - \frac{\alpha}{1-p}\big)^{t-\frac{k}{p}}, \end{aligned}$$

since $k \le t$.

Using $\log(1-u) \ge -2u$ for $0 \le u \le 1/2$, we have for $\alpha^2 \le p(1-p)$,

$$\big(1 - \frac{\alpha^2}{2p(1-p)}\big)^t \ge \exp\big(-2t\frac{\alpha^2}{p(1-p)}\big) \ge (2\theta/c_2')^{2/c_1'},$$

and for $\alpha^2 \le p^2/2$, we have

$$\big(1 - \frac{\alpha^2}{p^2}\big)^t \ge \exp\big(-t\frac{2\alpha^2}{p^2}\big) \ge (2\theta/c_2')^{2(1-p)/(pc_1')},$$

□

On $\mathcal{E}_2$, we have $t - k/p \leq \sqrt{2\frac{1-p}{p} t \log(c_2/(2\theta))}$, thus for $\alpha \leq (1-p)/2$,

$$\begin{aligned}\big(1-\frac{\alpha}{1-p}\big)^{t-\frac{k}{p}} &\geq \big(1-\frac{\alpha}{1-p}\big)^{\sqrt{2\frac{1-p}{p} t \log(c_2'/2\theta)}} \\ &\geq e^{-\sqrt{2\frac{\alpha^2}{p(1-p)} t \log(c_2'/(2\theta))}} \\ &\geq e^{-\sqrt{2/c_1'}\log(c_2'/\theta)} = (2\theta/c_2')^{\sqrt{2/c_1'}}.\end{aligned}$$

We deduce that

$$\frac{L_1(W)}{L_2(W)} \geq (2\theta/c_2')^{2/c_1' + 2(1-p)/(pc_1') + \sqrt{2/c_1'}} \geq 2\theta/c_2',$$

for the choice of $c_1' = 8$. Thus:

$$\frac{L_1(W)}{L_0(W)} \mathbb{1}_{\mathcal{E}} \geq 2\theta/c_2' \mathbb{1}_{\mathcal{E}},$$

where $\mathbb{1}_{\mathcal{E}}$ is the indicator function of the event $\mathcal{E}(z)$. Then by a change of measure we deduce:

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}_1(z)) \geq \mathbb{P}_1(\mathcal{E}(z)) = \mathbb{E}_1[\mathbb{1}_{\mathcal{E}}] &= \mathbb{E}_0\left(\frac{L_1(W)}{L_0(W)}\mathbb{1}_{\mathcal{E}}\right) \\ &\geq \mathbb{E}_0\left[2\theta/c_2' \mathbb{1}_{\mathcal{E}}\right] = 2\theta/c_2' \mathbb{P}_0(\mathcal{E}(z)) > \theta/c_2',\end{aligned}$$

where we make use of the fact that $\mathbb{P}_0(\mathcal{Q}(z)) > \frac{1}{2}$. □

Now by the choice of $\alpha = 2(1-\gamma p)^2\varepsilon/(\gamma^2)$, we have $\alpha \leq (1-p)/2 \leq p(1-p) \leq p/\sqrt{2}$ whenever $\varepsilon \leq \frac{1-p}{4\gamma^2(1-\gamma p)^2}$. For this choice of α, we have that $Q_1^*(z) - Q_0^*(z) = \frac{\gamma}{1-\gamma(p+\alpha)} - \frac{\gamma}{1-\gamma p} > 2\varepsilon$, thus $Q_0^*(z) + \varepsilon < Q_1^*(z) - \varepsilon$. In words, the random event $\{|Q_0^*(z) - Q(z)| \leq \varepsilon\}$ does not overlap with the event $\{|Q_1^*(z) - Q(z)| \leq \varepsilon\}$.

Now let us return to the assumption of Eq. (21), which states that for all $M_m \in \mathbb{M}^*$, $\mathbb{P}_m(|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) \leq \varepsilon) \geq 1 - \theta/c_2'$ under Algorithm $\mathfrak{A}$. Based on Lemma 11 we have $\mathbb{P}_1(|Q_0^*(z) - Q_t^{\mathfrak{A}}(z)| \leq \varepsilon) > \theta/c_2'$. This combined with the fact that $\{|Q_0^*(y) - Q_t^{\mathfrak{A}}(z)|\}$ and $\{|Q_1^*(z) - Q_t^{\mathfrak{A}}(z)|\}$ do not overlap implies that $\mathbb{P}_1(|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) \leq \varepsilon/\gamma) \leq 1 - \theta/c_2'$, which violates the assumption of Eq. (21). The contradiction between the result of Lemma 11 and the assumption which leads to this result proves the lower bound of Eq. (20). □

Now by the choice of $p = \frac{4\gamma-1}{3\gamma}$ and $c_1 = 8100$, we have that for every $\varepsilon \in (0, 3]$ and for all $0.4 = \gamma_0 \leq \gamma < 1$ there exists an MDP $M_m \in \mathbb{M}^*$ such that:

$$\mathbb{P}_m(|Q^*(z) - Q_t^{\mathfrak{A}}(z)|) > \varepsilon) > \frac{1}{c_2'} e^{\frac{-c_1 T_z \varepsilon^2}{6\beta^3}},$$

This result implies that for any state-action $z \in \mathcal{S} \times \mathcal{A}$, the probability of making an estimation error of ε is at least δ on M_0 or M_1 whenever the number of transition samples T_z from $z \in \mathcal{Z}$ is less that $\xi(\varepsilon, \delta) \triangleq \frac{6\beta^3}{c_1\varepsilon^2} \log \frac{1}{c_2'\delta}$. We now extend this result to the whole state-action space $\mathcal{S} \times \mathcal{A}$.

Lemma 12. *Assume that for every algorithm $\mathfrak{A}$, for every state-action $z \in \mathcal{S} \times \mathcal{A}$ we have*[13]

$$\mathbb{P}_m(|Q^*(z) - Q_{T_z}^{\mathfrak{A}}(z)| > \varepsilon | T_z = t_z) > \delta, \tag{22}$$

Then for any $\delta' \in (0, 1/2)$, for any algorithm $\mathfrak{A}$ using a total number of transition samples less than $T = \frac{N}{6}\xi\big(\varepsilon, \frac{12\delta'}{N}\big)$, there exists an MDP $M_m \in \mathbb{M}^*$ such that

$$\mathbb{P}_m\big(\|Q^* - Q_T^{\mathfrak{A}}\| > \varepsilon\big) > \delta', \tag{23}$$

where $Q_T^{\mathfrak{A}}$ denotes the empirical estimate of the optimal action-value function Q^* by $\mathfrak{A}$ using T transition samples.

Proof. First note that if the total number of observed transitions is less than $KL/2\xi(\varepsilon, \delta) = (N/6)\xi(\varepsilon, \delta)$, then there exists at least $KL/2 = N/6$ state-action pairs that are sampled at most $\xi(\varepsilon, \delta)$ times. Indeed, if this was not the case, then the total number of transitions would be strictly larger than $N/6\xi(\varepsilon, \delta)$, which implies a contradiction). Now let us denote those states as $z_{(1)}, \ldots, z_{(N/6)}$.

We consider the specific class of MDPs described in Figure 1. In order to prove that (23) holds for any algorithm, it is sufficient to prove it for the class of algorithms that return an estimate $Q_{T_z}^{\mathfrak{A}}(z)$ for each state-action z based on the transition samples observed from z only (indeed, since the samples from z and z' are independent, the samples collected from z' do not bring more information about $Q^*(z)$ than the information brought by the samples collected from z). Thus, by defining $\mathcal{Q}(z) \triangleq \{|Q^*(z) - Q_{T_z}^{\mathfrak{A}}(z)| > \varepsilon\}$, we have that for such algorithms, the events $\mathcal{Q}(z)$ and $\mathcal{Q}(z')$ are conditionally independent given T_z and $T_{z'}$. Thus, there

[13] Note that we allow T_z to be random.

exists an MDP $M_m \in \mathbb{M}^*$ such that:

$$\begin{aligned}
&\mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \cap \{T_{z(i)} \leq \xi(\varepsilon,\delta)\}_{1\leq i\leq N/6}\Big) \\
= &\sum_{t_1=0}^{\xi(\varepsilon,\delta)} \cdots \sum_{t_{N/6}=0}^{\xi(\varepsilon,\delta)} \mathbb{P}_m\Big(\{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big) \\
&\mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \cap \{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big) \\
= &\sum_{t_1=0}^{\xi(\varepsilon,\delta)} \cdots \sum_{t_{N/6}=0}^{\xi(\varepsilon,\delta)} \mathbb{P}_m\Big(\{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big) \\
&\prod_{1\leq i\leq N/6} \mathbb{P}_m\Big(\mathcal{Q}(z_{(i)})^c \cap T_{z(i)} = t_i\Big) \\
\leq &\sum_{t_1=0}^{\xi(\varepsilon,\delta)} \cdots \sum_{t_{N/6}=0}^{\xi(\varepsilon,\delta)} \mathbb{P}_m\Big(\{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big)(1-\delta)^{N/6},
\end{aligned}$$

from Eq. (22), thus

$$\begin{aligned}
&\mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \Big| \{T_{z(i)} \leq \xi(\varepsilon,\delta)\}_{1\leq i\leq N/6}\Big) \\
\leq &(1-\delta)^{N/6}.
\end{aligned}$$

We finally deduce that if the total number of transition samples is less than $\frac{N}{6}\xi(\varepsilon,\delta)$, then

$$\begin{aligned}
&\mathbb{P}_m(\|Q^* - Q_T^{\mathfrak{A}}\| > \varepsilon\Big) \geq \mathbb{P}_m\Big(\bigcup_{z\in\mathcal{S}\times\mathcal{A}} \mathcal{Q}(z)\Big) \\
\geq &\ 1 - \mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \Big| \{T_{z(i)} \leq \xi(\varepsilon,\delta)\}_{1\leq i\leq N/6}\Big) \\
\geq &\ 1 - (1-\delta)^{N/6} \geq \frac{\delta N}{12},
\end{aligned}$$

whenever $\frac{\delta N}{6} \leq 1$. Setting $\delta' = \frac{\delta N}{12}$, we obtain the desired result. □

Lemma 12 implies that if the total number of samples T is less than $\beta^3 N/(c_1\varepsilon^2)\log(N/(c_2\delta))$, with the choice of $c_1 = 8100$ and $c_2 = 72$, then the probability of $\|Q^* - Q_T^{\mathfrak{A}}\| \leq \varepsilon$ is at maximum $1-\delta$ on either M_0 or M_1. This is equivalent to the statement that for every RL algorithm $\mathfrak{A}$ to be (ε,δ,T)-correct on the set $\mathbb{M}^*$, and subsequently on the class of MDPs $\mathbb{M}$, the total number of transitions T needs to satisfy the inequality $T > \beta^3 N/(c_1\varepsilon^2)\log(N/(c_2\delta))$, which concludes the proof of Theorem 2.

5. Conclusion and Future Works

In this paper, we have presented the first minimax bound on the sample complexity of estimating the optimal action-value function in discounted reward MDPs. We have proven that the model-based Q-value iteration algorithm (QVI) is an optimal learning algorithm since it minimizes the dependencies on $1/\varepsilon$, N, δ and $1/(1-\gamma)$. Also, our results have significantly improved on the state-of-the-art in terms of dependency on $1/(1-\gamma)$. Overall, we conclude that QVI is an efficient RL algorithm which completely closes the gap between the lower and upper bound of the sample complexity of RL in the presence of a generative model of the MDP.

In this work, we are only interested in the estimation of the optimal action-value function and not the problem of exploration. Therefore, we did not compare our results with the-state-of-the-art of PAC-MDP (Strehl et al., 2009; Szita & Szepesvári, 2010) and upper-confidence bound based algorithms (Bartlett & Tewari, 2009; Jaksch et al., 2010), in which the choice of the exploration policy has an influence on the behavior of the learning algorithm. However, we believe that it would be possible to improve on the state-of-the-art in PAC-MDP, based on the results of this paper. This is mainly due to the fact that most PAC-MDP algorithms rely on an extended variant of model-based Q-value iteration to estimate the action-value function, but they use the naive result of Hoeffding's inequality for concentration of measure which leads to non-tight sample complexity results. One can improve on those results, in terms of dependency on $1/(1-\gamma)$, using the improved analysis of this paper which makes use of the sharp result of Bernstein's inequality as opposed to the Hoeffding's inequality in the previous works. Also, we believe that the existing lower bound on the *exploration complexity* of any reinforcement learning algorithm (Strehl et al., 2009) can be significantly improved in terms of dependency on $1/(1-\gamma)$ based on the new class of MDPs presented in this paper.

References

Azar, M. Gheshlaghi, Munos, R., Ghavamzadeh, M., and Kappen, H. J. Speedy q-learning. In *Advances in Neural Information Processing Systems 24*, pp. 2411–2419. 2011.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Bertsekas, D. P. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmount, Massachusetts, third edition, 2007.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic*

Programming. Athena Scientific, Belmont, Massachusetts, 1996.

Buşoniu, L., Babuška, R., De Schutter, B., and Ernst, D. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Boca Raton, Florida, 2010.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

Hagerup, L. and Rüb, C. A guided tour of chernoff bounds. *Information Processing Letters*, 33:305–308, 1990.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Kakade, S. M. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, Gatsby Computational Neuroscience Unit, 2004.

Kearns, M. and Singh, S. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 12*, pp. 996–1002. MIT Press, 1999.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Munos, R. and Moore, A. Influence and variance of a Markov chain : Application to adaptive discretizations in optimal control. In *Proceedings of the 38th IEEE Conference on Decision and Control*, 1999.

Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts, 1998.

Szepesvári, Cs. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

Szita, I. and Szepesvári, Cs. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1031–1038. Omnipress, 2010.