Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model

Mohammad Gheshlaghi Azar, Rémi Munos, Hilbert Kappen

HAL Id: hal-00831875

https://hal.archives-ouvertes.fr/hal-00831875

Submitted on 7 Jun 2013

Minimax PAC Bounds on the Sample Complexity of Reinforcement Learning with a Generative Model

Mohammad Gheshlaghi Azar
· **Rémi Munos** · **Hilbert J. Kappen**

Abstract We consider the problem of learning the optimal action-value function in discounted-reward Markov decision processes (MDPs). We prove new PAC bounds on the sample-complexity of two well-known model-based reinforcement learning (RL) algorithms in the presence of a generative model of the MDP: value iteration and policy iteration. The first result indicates that for an MDP with N state-action pairs and the discount factor $\gamma \in [0,1)$ only $O\big(N \log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$ state-transition samples are required to find an ε-optimal estimation of the action-value function with the probability (w.p.) $1-\delta$. Further, we prove that, for small values of ε, an order of $O\big(N \log(N/\delta)/\big((1-\gamma)^3\varepsilon^2\big)\big)$ samples is required to find an ε-optimal policy w.p. $1-\delta$. 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. To the best of our knowledge, this is the first minimax result on the sample complexity of RL: The upper bound matches the lower bound in terms of N, ε, δ and $1/(1-\gamma)$ up to a constant factor. Also, both our lower bound and upper bound improve on the state-of-the-art in terms of their dependence on $1/(1-\gamma)$.

M. Gheshlaghi Azar
Department of Biophysics
Radboud University Nijmegen
6525 EZ Nijmegen, The Netherlands
E-mail: m.azar@science.ru.nl

R. Munos
INRIA Lille, SequeL Project
40 avenue Halley
59650 Villeneuve d'Ascq, France
E-mail: remi.munos@inria.fr

H. J. Kappen
Department of Biophysics
Radboud University Nijmegen
6525 EZ Nijmegen, The Netherlands
E-mail: b.kappen@science.ru.nl

Keywords sample complexity · Markov decision processes · reinforcement learning · learning theory

1 Introduction

An important problem in the field of reinforcement learning (RL) is to estimate the optimal policy (or the optimal value function) from the observed rewards and the transition samples (Sutton and Barto, 1998; Szepesvári, 2010). To estimate the optimal policy one may use model-free or model-based approaches. In model-based RL, we first learn a model of the MDP using a batch of state-transition samples and then use this model to estimate the optimal policy or the optimal action-value function using the Bellman recursion, whereas model-free methods directly aim at estimating the optimal value function without resorting to learning an explicit model of the dynamical system. The fact that the model-based RL methods decouple the model-estimation problem from the value (policy) iteration problem may be useful in problems with a limited budget of sampling. This is because the model-based RL algorithms, after learning the model, can perform many Bellman recursion steps without any need to make new transition samples, whilst the model-free RL algorithms usually need to generate fresh samples at each step of value (policy) iteration process.

The focus of this article is on model-based RL algorithms for finite state-action problems, where we have access to a generative model (simulator) of the MDP. Especially, we derive tight sample-complexity upper bounds for two well-known model-based RL algorithms, the model-based value iteration and the model-based policy iteration (Wiering and van Otterlo, 2012), It has been shown (Kearns and Singh, 1999; Kakade, 2004, chap. 9.1) that an action-value based variant of model-based value iteration algorithm, Q-value iteration (QVI), finds an ε-optimal estimate of the action-value function with high probability (w.h.p.) using only $\widetilde{O}(N/((1-\gamma)^4\varepsilon^2))$ samples, where N and γ denote the size of state-action space and the discount factor, respectively.[1] One can also prove, using the result of Singh and Yee (1994), that QVI w.h.p. finds an ε-optimal policy using an order of $\widetilde{O}(N/((1-\gamma)^6\varepsilon^2))$ samples. An upper-bound of a same order can be proven for model-based PI. These results match the best upper-bound currently known (Azar et al, 2011b) for the sample complexity of RL. However, there exist gaps with polynomial dependency on $1/(1-\gamma)$ between these upper bounds and the state-of-the-art lower bound, which is of order $\widetilde{\Omega}(N/((1-\gamma)^2\varepsilon^2))$ (Azar et al, 2011a; Even-Dar et al, 2006).[2] It has not been clear, so far, whether the upper bounds or the lower bound can be improved or both.

In this paper, we prove new bounds on the performance of QVI and PI which indicate that for both algorithms with the probability (w.p) $1-\delta$ an order of $O(N\log(N/\delta)/((1-\gamma)^3\varepsilon^2))$ samples suffice to achieve an ε-optimal estimate of action-value function as well as to find an ε-optimal policy. The new upper bound improves on the previous result of AVI and API by an order of $1/(1-\gamma)$. We also

---

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

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

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 $1/(1-\gamma)$. 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 and PI in terms of the sample complexity of estimating the optimal action-value function or the optimal policy.

The main idea to improve the upper bound of the above-mentioned RL algorithms is to express the performance loss $Q^* - Q_k$, where Q_k is the estimate of the action-value function after k iteration of QVI or PI, in terms of Σ^{π^*}, the variance of the sum of discounted rewards under the optimal policy π^*, as opposed to the maximum $V_{\max} = R_{\max}/(1-\gamma)$ as was used before. For this we make use of the Bernstein's concentration inequality (Cesa-Bianchi and Lugosi, 2006, appendix, pg. 361), which is expressed in terms of the variance of the random variables. 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 and Moore, 1999; Sobel, 1982). These results allow us to prove a high-probability bound of order $\widetilde{O}(\sqrt{\Sigma^{\pi^*}/(n(1-\gamma))})$ on the performance loss of both algorithms, where n is the number of samples per state-action. This leads to a tight PAC upper-bound of $\widetilde{O}(N/(\varepsilon^2(1-\gamma)^3))$ on the sample complexity of these methods.

In the case of lower bound, we introduce a new class of "hard" MDPs, which adds some structure to the bandit-like class of MDP used previously by Azar et al (2011a); Even-Dar et al (2006): In the new model, there exist states with high probability of transition to themselves. This adds to the difficulty of estimating the value function, since even a small modeling error may cause a large error in the estimate of the optimal value function, especially when the discount factor γ is 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 and the *model-based policy iteration* (PI) 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 matching 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 two model-based RL algorithms which make use of generative model for sampling: the model-based Q-value iteration and the model-based policy iteration (Wiering and van Otterlo, 2012; Kearns and Singh, 1999).

We consider the standard reinforcement learning (RL) framework (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998), where an RL 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.[3] 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.

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 $[0, 1]$.[4]*

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, $V^*(x) \triangleq \sup_\pi V^\pi(x)$, $\forall x \in \mathcal{X}$. The 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 also 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

$$(\mathcal{T}Q)(z) \triangleq r(z) + \gamma (P^{\pi^*} Q)(z), \qquad \forall z \in \mathcal{Z}.$$

Also, for the policy π, 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 $x \in \mathcal{X}$, respectively.

It is important to note that $\mathcal{T}$ and $\mathcal{T}^\pi$ are γ-contractions, i.e., for any pair of value functions V and V' and any policy π, we have $\|\mathcal{T}V - \mathcal{T}V'\| \leq \gamma \|V - V'\|$ and $\|\mathcal{T}^\pi V - \mathcal{T}^\pi V'\| \leq \gamma \|V - V'\|$ (Bertsekas, 2007, Chap. 1). $\|\cdot\|$ shall denote the supremum (ℓ_∞) norm, 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. We also define the ℓ_1-norm on the function g as $\|g\|_1 = \sum_{y \in \mathcal{Y}} |g(y)|$.

For ease of exposition, 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.

---

[3] For simplicity, here we assume that the reward $r(x, a)$ is a deterministic function of state-action pairs (x, a). Nevertheless, It is straightforward to extend our results to the case of stochastic rewards under some mild assumption, e.g., boundedness of the absolute value of the rewards.

[4] Our results also hold if the rewards are taken from some interval $[r_{\min}, r_{\max}]$ instead of $[0, 1]$, in which case the bounds scale with the factor $r_{\max} - r_{\min}$.

2.1 Algorithms

We begin by describing the procedure which is used by both PI and QVI to make an empirical estimate of the state-transition distributions.

The model estimator makes n transition samples for each state-action pair $z \in \mathcal{Z}$ for which it makes n calls to the generative model, i.e., the total number of calls to the generative model is $T = nN$. 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 the state-action pair $z \in \mathcal{Z}$ (see Algorithm 3). Based on the empirical model $\widehat{P}$ 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 $\widehat{P}$.

Having the empirical model $\widehat{P}$ estimated, QVI and PI rely on standard value iteration and policy iteration schemes to estimate the optimal action-value function: QVI iterates some action-value function Q_j, with the initial value of Q_0, through the empirical Bellman optimality operator $\widehat{\mathcal{T}}$ until Q_j admits some convergence criteria. PI, in contrast, relies on iterating some policy π_j with the initial value π_0: At each iteration $j > 0$, the algorithm solves the dynamic programming problem for a fixed policy π_j using the empirical model $\widehat{P}$. The next policy π_{j+1} is then determined as the greedy policy w.r.t. the action-value function $\widehat{Q}^{\pi_j}$, that is, $\pi_{j+1}(x) = \arg\max_{a \in \mathcal{A}} \widehat{Q}^{\pi_j}(x, a)$ for all $x \in \mathcal{X}$. Note that Q_k, as defined by PI and QVI are deferent, but nevertheless we use a same notation for both action-functions since we will show in the next section that they enjoy the same performance guarantees. The pseudo codes of both algorithms are provided in Algorithm 1 and Algorithm 2.

Algorithm 1 Model-based Q-value Iteration (QVI)

Require: reward function r, discount factor γ, initial action-value function Q_0, samples per state-action n, number of iterations k

$\widehat{P} = \text{ESTIMATEMODEL}(n)$ ▷ Estimate the model (defined in Algorithm 3)
for $j := 0, 1, \ldots, k - 1$ **do**
    for each $x \in \mathcal{X}$ **do**
        $\pi_j(x) = \arg\max_{a \in \mathcal{A}} Q_j(x, a)$ ▷ greedy policy w.r.t. the latest estimation of Q^*
        for each $a \in \mathcal{A}$ **do**
            $\widehat{\mathcal{T}}Q_j(x, a) = r(x, a) + \gamma(\widehat{P}^{\pi_j}Q_j)(x, a)$ ▷ empirical Bellman operator
            $Q_{j+1}(x, a) = \widehat{\mathcal{T}}Q_j(x, a)$ ▷ Iterate the action-value function Q_j
        end for
    end for
end for
return Q_k

Algorithm 2 Model-based Policy Iteration (PI)

Require: reward function r, discount factor γ, initial action-value function Q_0, samples per state-action n, number of iterations k

$\widehat{P} = \text{ESTIMATEMODEL}(n)$ ▷ Estimate the model (defined in Algorithm 3)
for $j := 0, 1, \ldots, k-1$ **do**
    for each $x \in \mathfrak{X}$ **do**
        $\pi_j(x) = \arg\max_{a \in \mathcal{A}} Q_j(x, a)$ ▷ greedy policy w.r.t. the latest estimation of Q^*
    end for
    $\widehat{Q}^{\pi_j} = \text{SOLVEDP}(\widehat{P}, \pi_j)$ ▷ Find the fixed point of the Bellman operator for the policy π_j
    $Q_{j+1} = \widehat{Q}^{\pi_j}$ ▷ Iterate the action-value function Q_j
end for
return Q_k

function $\text{SOLVEDP}(P, \pi)$
    $Q = (I - \gamma P^\pi)^{-1} r$
    return Q
end function

Algorithm 3 Function: ESTIMATEMODEL

Require: The generative model (simulator) P
    function $\text{ESTIMATEMODEL}(n)$ ▷ Estimating the transition model using n samples
        $\forall (y, z) \in \mathfrak{X} \times \mathfrak{Z} : m(y, z) = 0$ ▷ initialization
        for each $z \in \mathfrak{Z}$ **do**
            for $i := 1, 2, \ldots, n$ **do**
                $y \sim P(\cdot | z)$ ▷ Generate a state-transition sample
                $m(y, z) := m(y, z) + 1$ ▷ Count the transition samples
            end for
            $\forall y \in \mathfrak{X} : \widehat{P}(y | z) = \frac{m(y, z)}{n}$ ▷ Normalize by n
        end for
        return $\widehat{P}$ ▷ Return the empirical model
    end function

---

3 Main Results

Our main results are in the form of PAC (probably approximately correct) sample complexity bounds on the total number of samples required to attain a near-optimal estimate of the action-value function:

Theorem 1 (PAC-bound on $Q^* - Q_k$)
Let Assumption 1 hold and T be a positive integer. Then, there exist some constants c, c_0, d and d_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 $\|Q^* - Q_k\| \leq \varepsilon$, *w.p. at least* $1 - \delta$, *after* $k = \lceil d \log(d_0 \beta / \varepsilon) / \log(1/\gamma) \rceil$ *iteration of QVI or PI algorithm.*[5]

We also prove a similar bound on the sample-complexity of finding a near-optimal policy for small values of ε:

Theorem 2 (PAC-bound on $Q^* - Q^{\pi_k}$) *Let Assumption 1 hold and T be a positive integer. Define π_k as the greedy policy w.r.t. Q_k at iteration k of PI or QVI. Then, there exist some constants c', c'_0, c'_1, d' and d'_0 such that for all $\varepsilon \in (0, c'_1 \sqrt{\beta/(\gamma|\mathcal{X}|)})$ 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 $\|Q^* - Q^{\pi_k}\| \leq \varepsilon$, *w.p. at least* $1 - \delta$, *after* $k = d' \lceil \log(d'_0 \beta / \varepsilon) / \log(1/\gamma) \rceil$ *iteration of QVI or PI algorithm.*

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

Definition 1 ((ε, δ)-correct algorithm) *Let $Q^{\mathfrak{A}} : \mathcal{Z} \to \mathbb{R}$ be the output of some RL Algorithm $\mathfrak{A}$. We say that $\mathfrak{A}$ is (ε, δ)-correct on the class of MDPs $\mathbb{M} = \{M_1, M_2, \ldots, M_m\}$ if $\|Q^* - Q^{\mathfrak{A}}\| \leq \varepsilon$ with probability at least $1 - \delta$ for all $M \in \mathbb{M}$.*[6]

Theorem 3 (Lower bound on the sample complexity of RL)
Let Assumption 1 hold and T be a positive integer. There exist 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)-correct RL Algorithm $\mathfrak{A}$ on the class of MDPs $\mathbb{M}$ the total number of state-transition samples (sampling budget) 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 and PI, reported in Theorem 1 and Theorem 2, in Subsection 4.1. We then prove Theorem 3, a new RL lower bound, in Subsection 4.2.

4.1 Proofs of Theorem 1 and Theorem 2 - The Upper Bounds

We begin by introducing some new notation. For the stationary policy π, we define $\Sigma^\pi(z) \triangleq \mathbb{E}[|\sum_{t \geq 0} \gamma^t r(Z_t) - Q^\pi(z)|^2 | Z_0 = z]$ as the variance of the sum of discounted rewards starting from $z \in \mathcal{Z}$ under the policy π. We also make use of the

---

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

[6] Algorithm $\mathfrak{A}$, unlike QVI and PI, does not require a same number of transition samples for every state-action pair and can generate samples arbitrarily.

following definition of the variance of a function: For any real-valued function $f :$ $\mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is a finite set, we define $\mathbb{V}_{y\sim\rho}(f(y)) \triangleq \mathbb{E}_{y\sim\rho}|f(y)-\mathbb{E}_{y\sim\rho}(f(y))|^2$ as the variance of f under the probability distribution ρ, where $\mathcal{Y}$ is a finite set and ρ is a probability distribution on $\mathcal{Y}$. We then define $\sigma_{Q^\pi}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim P(\cdot|z)}[Q^\pi(y, \pi(y))]$ as the discounted variance of Q^π at $z \in \mathcal{Z}$. Also, we shall denote σ_{V^*} and σ_{V^*} as the discounted variance of the value function V^π and V^* defined as $\sigma_{V^\pi}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim P(\cdot|z)}[V^\pi(y)]$ and $\sigma_{V^*}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim P(\cdot|z)}[V^*(y)]$, for all $z \in \mathcal{Z}$, respectively. For each of these variances we define the corresponding empirical variance $\widehat{\sigma}_{Q^\pi}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim\widehat{P}(\cdot|z)}[Q^\pi(y, \pi(y))]$, $\widehat{\sigma}_{V^\pi}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim\widehat{P}(\cdot|z)}[V^\pi(y)]$ and $\widehat{\sigma}_{V^*}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim\widehat{P}(\cdot|z)}[V^*(y)]$, respectively, for all $z \in \mathcal{Z}$ under the model $\widehat{P}$. We also define $\widehat{\sigma}_{Q^*}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim\widehat{P}(\cdot|z)}[Q^*(y, \pi^*(y))]$. We also notice that σ_{Q^π} and σ_{V^π} can be written as follows: $\sigma_{Q^\pi}(z) = \gamma^2 P^\pi[|Q^\pi - P^\pi Q^\pi|^2](z)$ and $\sigma_{V^\pi}(z) = \gamma^2 P[|V^\pi - PV^\pi|^2](z)$ for all $z \in \mathcal{Z}$.

We now prove our first result which shows that Q_k, for both QVI and PI, 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 $\|Q^* - \widehat{Q}^*\|$ 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, for both QVI and PI, we have*

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

Proof

We begin by proving the result for QVI. For all $k \geq 0$, we have

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

Thus by an immediate recursion

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

In the case of PI, we notice that $Q_k = \widehat{Q}^{\pi_{k-1}} \geq \widehat{Q}^{\pi_{k-2}} = Q_{k-1}$, which implies that

$$
\begin{aligned}
0 \leq \widehat{Q}^* - Q_k &= \gamma\widehat{P}^{\widehat{\pi}^*}\widehat{Q}^* - \gamma\widehat{P}^{\pi_{k-1}}\widehat{Q}^{\pi_{k-1}} \leq \gamma(\widehat{P}^{\widehat{\pi}^*}\widehat{Q}^* - \widehat{P}^{\pi_{k-1}}\widehat{Q}^{\pi_{k-2}}) \\
&= \gamma(\widehat{P}^{\widehat{\pi}^*}\widehat{Q}^* - \widehat{P}^{\pi_{k-1}}Q_{k-1}) \leq \gamma\widehat{P}^{\widehat{\pi}^*}(\widehat{Q}^* - Q_{k-1}),
\end{aligned}
$$

where in the last line we rely on the fact hat π_{k-1} is the greedy policy w.r.t. Q_{k-1}. This implies the component-wise inequality $\widehat{P}^{\pi_{k-1}}Q_{k-1} \geq \widehat{P}^{\widehat{\pi}^*}Q_{k-1}$. The result then follows by taking the ℓ_∞-norm on both sides of the inequality and then recursively expand the resulted bound.
∎

One can easily prove the following corollary, which bounds the difference between $\widehat{Q}^*$ and $\widehat{Q}^{\pi_k}$, based on the result of Lemma 1 and the main result of Singh and Yee (1994). Corollary 1 is required for the proof of Theorem 2.

Corollary 1 *Let Assumption 1 hold and π_k be the greedy policy induced by the kth iterate of QVI and PI. Also, let $Q_0(z)$ takes value in the interval $[0, \beta]$ for all $z \in \mathcal{Z}$. Then we have*

$$\|\widehat{Q}^{\pi_k} - \widehat{Q}^*\| \le 2\gamma^k \beta^2, \qquad and \qquad \|\widehat{V}^{\pi_k} - \widehat{V}^*\| \le 2\gamma^k \beta^2.$$

We notice that the tight bound on $\|\widehat{Q}^{\pi_k} - \widehat{Q}^*\|$ for PI is of order $\gamma^{k+1}\beta$ since $\widehat{Q}^{\pi_k} = Q_{k+1}$. However, for ease of exposition we make use of the bound of Corollary 1 for both QVI and PI.

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 $\widetilde{O}(\beta^2/\sqrt{n})$ 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 and 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 our desired result of $\widetilde{O}(\beta^{1.5}/\sqrt{n})$: **(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 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 $\sigma_{Q^\pi}(z)$. Based on this result we prove an upper-bound of order $O(\beta^{1.5})$ on $(I - \gamma P^\pi)^{-1}\sqrt{\sigma_{Q^\pi}}$ for every policy π, which combined with the previous steps leads to an upper bound of $\widetilde{O}(\beta^{1.5}/\sqrt{n})$ on $\|Q^* - \widehat{Q}^*\|$. A similar approach leads to a bound of $\widetilde{O}(\beta^{1.5}/\sqrt{n})$ on $\|Q^* - Q^{\pi_k}\|$ under the assumption that there exist constants $c_1 > 0$ and $c_2 > 0$ such that $n > c_1\gamma^2\beta^2|\mathcal{X}|\log(c_2 N/\delta))$.

The following component-wise results bound $Q^* - \widehat{Q}^*$ from above and below:

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

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

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

Proof

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

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

In the case of Ineq. (2) we have

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

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

$$\widehat{P}^{\pi^*} Q^* \geq \widehat{P}^{\widehat{\pi}^*} Q^*, \qquad \text{and} \qquad (I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1} = \sum_{i \geq 0} \left(\gamma \widehat{P}^{\widehat{\pi}^*} \right)^i \geq \mathbf{0},$$

where $\mathbf{0}$ is a function which assigns 0 to all $(z_1, z_2) \in \mathcal{Z} \times \mathcal{Z}$.

∎

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

Lemma 3 *Let Assumption 1 hold. Then, for any $0 < \delta < 1$ w.p at least $1 - \delta$*

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

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

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

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

By solving this inequality w.r.t. $\|V^* - \widehat{V}^{\pi^*}\|$ we deduce

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

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

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

We then make use of Hoeffding's inequality (Cesa-Bianchi and Lugosi, 2006, Appendix A, pg. 359) to bound $|(P_{\pi^*} - \widehat{P}_{\pi^*}) V^*(x)|$ for all $x \in \mathcal{X}$ in high probability:

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

By applying the union bound we deduce

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

We then define the probability of failure δ as

$$\delta \triangleq 2|\mathcal{X}| \exp\left(\frac{-n\varepsilon^2}{2\beta^2}\right). \tag{6}$$

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). ∎

We now state Lemma 4 which relates σ_{V^*} to $\widehat{\sigma}_{\widehat{Q}^{\pi^*}}$ and $\widehat{\sigma}_{\widehat{Q}^*}$. Later, we make use of this result in the proof of Lemma 5.

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

$$\sigma_{V^*} \leq \widehat{\sigma}_{\widehat{Q}^{\pi^*}} + b_v \mathbf{1}, \tag{8}$$

$$\sigma_{V^*} \leq \widehat{\sigma}_{\widehat{Q}^*} + b_v \mathbf{1}, \tag{9}$$

where b_v is defined as

$$b_v \triangleq \sqrt{\frac{18\gamma^4\beta^4\log\frac{3N}{\delta}}{n}} + \frac{4\gamma^2\beta^4\log\frac{6N}{\delta}}{n},$$

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.

$$\sigma_{V^*}(z) = \sigma_{V^*}(z) - \gamma^2 \mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y)) + \gamma^2 \mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y))$$
$$\leq \gamma^2\left((P - \widehat{P})V^{*2}\right)(z) - \gamma^2[(PV^*)^2(z) - (\widehat{P}V^*)^2(z)]$$
$$+ \gamma^2 \mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(V^*(Y) - \widehat{V}^{\pi^*}(Y)) + \gamma^2 \mathbb{V}_{Y\sim\widehat{P}(\cdot|z)}(\widehat{V}^{\pi^*}(Y)).$$

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

$$\sigma_{V^*}(z) \leq \gamma^2[P - \widehat{P}]V^{*2}(z) - \gamma^2[(P - \widehat{P})V^*][(P + \widehat{P})V^*](z)$$
$$+ \gamma^2\|V^* - \widehat{V}^{\pi^*}\|^2 + \widehat{\sigma}_{\widehat{V}^{\pi^*}}(z).$$

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

$$\sigma_{V^*}(z) \leq \widehat{\sigma}_{\widehat{V}^{\pi^*}}(z) + \gamma^2\left[3\beta^2\sqrt{2\frac{\log\frac{3}{\delta}}{n}} + \frac{2\beta^4\log\frac{6N}{\delta}}{n}\right], \tag{10}$$

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

■

The following result proves a bound on $\gamma(P - \widehat{P})V^*$, for which we make use of the Bernstein's inequality (Cesa-Bianchi and 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*

$$b_{pv} \triangleq \left(\frac{5(\gamma\beta)^{4/3}\log\frac{6N}{\delta}}{n}\right)^{3/4} + \frac{3\beta^2\log\frac{12N}{\delta}}{n}.$$

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

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

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

Proof

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

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

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

We deduce (using a union bound)

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

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

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 Σ^π satisfies the following Bellman-like recursion. Later, we use this result, in Lemma 7, to bound $(I - \gamma P^\pi)^{-1}\sigma_{Q^\pi}$.

Lemma 6 Σ^π *satisfies the Bellman equation*

$$\Sigma^\pi = \sigma_{Q^\pi} + \gamma^2 P^\pi \Sigma^\pi. \tag{15}$$

Proof

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

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

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 discounted variance.

Lemma 7

$$
\|(I - \gamma^2 P^\pi)^{-1}\sigma_{Q^\pi}\| = \|\Sigma^\pi\| \leq \beta^2, \tag{16}
$$

$$
\|(I - \gamma P^\pi)^{-1}\sqrt{\sigma_{Q^\pi}}\| \leq 2\log(2)\|\sqrt{\beta \Sigma^\pi}\| \leq 2\log(2)\beta^{1.5}. \tag{17}
$$

Proof

The first inequality follows from Lemma 6 by solving (15) in terms of Σ^π and taking the sup-norm over both sides of the resulted equation. In the case of Eq. (17) we have

$$
\begin{aligned}
\|(I - \gamma P^\pi)^{-1}\sqrt{\sigma_{Q^\pi}}\| &= \left\|\sum_{k \geq 0}(\gamma P^\pi)^k \sqrt{\sigma_{Q^\pi}}\right\| = \left\|\sum_{l \geq 0}(\gamma P^\pi)^{tl} \sum_{j=0}^{t-1}(\gamma P^\pi)^j \sqrt{\sigma_{Q^\pi}}\right\| \\
&\leq \sum_{l \geq 0}(\gamma^t)^l \left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j \sqrt{\sigma_{Q^\pi}}\right\| = \frac{1}{1 - \gamma^t}\left\|\sum_{j=0}^{t-1}(\gamma P^\pi)^j \sqrt{\sigma_{Q^\pi}}\right\|,
\end{aligned}
\tag{18}
$$

in which we write $k = tl + j$ with t any positive integer.[7] We now prove a bound on $\left\| \sum_{j=0}^{t-1} (\gamma P^\pi)^j \sqrt{\sigma_{Q^\pi}} \right\|$ by making use of Jensen's inequality, Cauchy-Schwarz inequality and Eq. 16:

$$
\left\| \sum_{j=0}^{t-1} (\gamma P^\pi)^j \sqrt{\sigma_{Q^\pi}} \right\| \le \left\| \sum_{j=0}^{t-1} \gamma^j \sqrt{(P^\pi)^j \sigma_{Q^\pi}} \right\| \le \sqrt{t} \left\| \sqrt{\sum_{j=0}^{t-1} (\gamma^2 P^\pi)^j \sigma_{Q^\pi}} \right\|
$$
$$
\le \sqrt{t} \left\| \sqrt{(I - \gamma^2 P^\pi)^{-1} \sigma_{Q^\pi}} \right\| = \| \sqrt{t \Sigma^\pi} \|.
$$
(19)

The result then follows by plugging (19) into (18) 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 $0 < \delta < 1$:*

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

w.p. at least $1 - \delta$, where ε' is defined as

$$
\varepsilon' \triangleq \sqrt{\frac{4\beta^3 \log \frac{4N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{12N}{\delta}}{n} \right)^{3/4} + \frac{3\beta^3 \log \frac{24N}{\delta}}{n}.
$$
(20)

Proof

By incorporating the result of Lemma 5 and Lemma 7 into Lemma 2 and taking in to account that $(I - \gamma \widehat{P}^{\pi^*})^{-1} \mathbf{1} = \beta \mathbf{1}$, we deduce

$$
\begin{aligned}
Q^* - \widehat{Q}^* &\le b\mathbf{1}, \\
Q^* - \widehat{Q}^* &\ge -b\mathbf{1},
\end{aligned}
$$
(21)

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

$$
b \triangleq \sqrt{\frac{4\beta^3 \log \frac{2N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{6N}{\delta}}{n} \right)^{3/4} + \frac{3\beta^3 \log \frac{12N}{\delta}}{n}.
$$
(22)

The result then follows by combining these two bounds using a union bound and taking the ℓ_∞ norm. ∎

---

[7] For any real-valued function f, $\sqrt{f}$ is defined as a component wise squared-root operator on f. Also, for any policy π and $k \ge 1$: $(P^\pi)^k(\cdot) \triangleq \underbrace{P^\pi \cdots P^\pi}_{k}(\cdot)$.

Proof of Theorem 1

We define the total error $\varepsilon = \varepsilon' + \gamma^k \beta$ which bounds $\|Q^* - Q_k\| \le \|Q^* - \widehat{Q}^*\| + \|\widehat{Q}^* - Q_k\|$ in high probability (ε' is defined in Lemma 8). The results then follows by solving this bound w.r.t. n and k and then quantifying the total number of samples by $T = nk$.

$\blacksquare$

We now draw our attention to the proof of Theorem 2, for which we need the following component-wise bound on $Q^* - Q^{\pi_k}$.

Lemma 9 *Let Assumption 1 hold. Then w.p. at least $1 - \delta$*

$$Q^* - Q^{\pi_k} \le \widehat{Q}^{\pi_k} - Q^{\pi_k} + (b + 2\gamma^k \beta^2)\mathbf{1},$$

where b is defined by (22).

Proof

We make use of Corollary 1 and Lemma 8 to prove the result:

$$
\begin{aligned}
Q^* - Q^{\pi_k} &= Q^* - \widehat{Q}^* + \widehat{Q}^* - \widehat{Q}^{\pi_k} + \widehat{Q}^{\pi_k} - Q^{\pi_k} \\
&\le b\mathbf{1} + \widehat{Q}^* - \widehat{Q}^{\pi_k} + \widehat{Q}^{\pi_k} - Q^{\pi_k} && \text{by Eq. 21} \\
&\le (b + 2\gamma^k \beta^2)\mathbf{1} + \widehat{Q}^{\pi_k} - Q^{\pi_k} && \text{by Corollary 1.}
\end{aligned}
$$

$\blacksquare$

Lemma 9 states that w.h.p. $Q^* - Q^{\pi_k}$ is close to $\widehat{Q}^{\pi_k} - Q^{\pi_k}$ for large values of k and n. Therefore, to prove the result of Theorem 2 we only need to bound $\widehat{Q}^{\pi_k} - Q^{\pi_k}$ in high probability:

Lemma 10 (Component-wise upper bound on $\widehat{Q}^{\pi_k} - Q^{\pi_k}$)

$$\widehat{Q}^{\pi_k} - Q^{\pi_k} \le \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^* + \gamma\beta\|(P - \widehat{P})(V^* - V^{\pi_k})\|\mathbf{1}. \quad (23)$$

Proof

We prove this result using a similar argument as in the proof of Lemma 2:

$$
\begin{aligned}
\widehat{Q}^{\pi_k} - Q^{\pi_k} &= (I - \gamma\widehat{P}^{\pi_k})^{-1}r - (I - \gamma P^{\pi_k})^{-1}r = \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P^{\pi_k} - \widehat{P}^{\pi_k})Q^{\pi_k} \\
&= \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^{\pi_k} \\
&= \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^* + \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})(V^{\pi_k} - V^*) \\
&\le \gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^* + \gamma\beta\|(P - \widehat{P})(V^* - V^{\pi_k})\|\mathbf{1}.
\end{aligned}
$$

$\blacksquare$

Now we bound the terms in the RHS of Eq. (23) in high probability. We begin by bounding $\gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^*$:

Lemma 11 *Let Assumption 1 hold. Then, w.p. at least $1 - \delta$ we have*

$$\gamma(I - \gamma\widehat{P}^{\pi_k})^{-1}(P - \widehat{P})V^* \leq \left(\sqrt{\frac{4\beta^3 \log\frac{2N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log\frac{6N}{\delta}}{n} \right)^{3/4} \right) \mathbf{1}$$

$$+ \left(\frac{3\beta^3 \log\frac{12N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6 \log\frac{2N}{\delta}}{n}} \right) \mathbf{1}.$$

Proof

From Lemma 5, w.p. at least $1 - \delta$, we have

$$
\begin{aligned}
\gamma(P - \widehat{P})V^* &\leq \sqrt{\frac{2 \log\frac{2N}{\delta}\widehat{\sigma}_{\widehat{Q}^*}}{n}} + b_{pv}\mathbf{1} \\
&\leq \sqrt{\frac{2 \log\frac{2N}{\delta}\left(\widehat{\sigma}_{\widehat{Q}^{\pi_k}} + \gamma^2\|\widehat{Q}^{\pi_k} - \widehat{Q}^*\|^2 \right)}{n}} + b_{pv}\mathbf{1} \qquad (24) \\
&\leq \sqrt{\frac{2 \log\frac{2N}{\delta}\widehat{\sigma}_{\widehat{Q}^{\pi_k}}}{n}} + \left(b_{pv} + \sqrt{\frac{8\gamma^{2k+2}\beta^4 \log\frac{2N}{\delta}}{n}} \right) \mathbf{1}.
\end{aligned}
$$

where in the last line we rely on Corollary 1. The result then follows by combining (24) with the result of Lemma 7. $\blacksquare$

We now prove bound on $\|(P - \widehat{P})(V^* - \widehat{V}^{\pi_k})\|$ in high probability, for which we require the following technical result:

Lemma 12 (Weissman et. al. 2003)

Let ρ be a probability distribution on the finite set $\mathcal{X}$. Let $\{X_1, X_2, \cdots, X_n\}$ be a set of i.i.d. samples distributed according to ρ and $\widehat{\rho}$ be the empirical estimation of ρ using this set of samples. Define $\pi_\rho \triangleq \max_{X \subseteq \mathcal{X}} \min(\mathbb{P}_\rho(X), 1 - \mathbb{P}_\rho(X))$, where $P_\rho(X)$ is the probability of X under the distribution ρ and $\varphi(p) \triangleq 1/(1-2p) \log((1-p)/p)$ for all $p \in [0, 1/2)$ with the convention $\varphi(1/2) = 2$, then w.p. at least $1 - \delta$ we have

$$\|\rho - \widehat{\rho}\|_1 \leq \sqrt{\frac{2 \log\frac{2^{|\mathcal{X}|-2}}{\delta}}{n\varphi(\pi_\rho)}} \leq \sqrt{\frac{2|\mathcal{X}| \log\frac{2}{\delta}}{n}}.$$

Lemma 13 *Let Assumption 1 hold. Then, w.p. at least $1 - \delta$ we have*

$$\gamma\|(P - \widehat{P})(V^* - V^{\pi_k})\| \leq \sqrt{\frac{2\gamma^2|\mathcal{X}| \log\frac{2N}{\delta}}{n}}\|Q^* - Q^{\pi_k}\|.$$

Proof

From the Hölder's inequality for all $z \in \mathcal{Z}$ we have

$$\gamma |(P - \widehat{P})(V^* - V^{\pi_k})(z)| \leq \gamma \|P(\cdot|z) - \widehat{P}(\cdot|z)\|_1 \|V^* - V^{\pi_k}\|$$
$$\leq \gamma \|P(\cdot|z) - \widehat{P}(\cdot|z)\|_1 \|Q^* - Q^{\pi_k}\|.$$

This combined with Lemma 12 implies that

$$\gamma |(P - \widehat{P})(V^* - V^{\pi_k})(z)| \leq \sqrt{\frac{2\gamma^2 |\mathcal{X}| \log \frac{2}{\delta}}{n}} \|Q^* - Q^{\pi_k}\|.$$

The result then follows by taking union bound on all $z \in \mathcal{Z}$. ■

We now make use of the results of Lemma 13 and Lemma 11 to bound $\|Q^* - Q^{\pi_k}\|$ in high probability:

Lemma 14 *Let Assumption 1 hold. Assume that*

$$n \geq 8\gamma^2 \beta^2 |\mathcal{X}| \log \frac{4N}{\delta}. \tag{25}$$

Then, w.p. at least $1 - \delta$ we have

$$\|Q^* - Q^{\pi_k}\| \leq 2 \left[\varepsilon' + 2\gamma^k \beta^2 + \sqrt{\frac{4\beta^3 \log \frac{4N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{12N}{\delta}}{n} \right)^{3/4} \right.$$
$$\left. + \frac{4\beta^3 \log \frac{24N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6 \log \frac{4N}{\delta}}{n}} \right].$$

where ε' is defined by Eq. (20).

Proof

By incorporating the result of Lemma 13 and Lemma 11 into Lemma 10 we deduce

$$\widehat{Q}^{\pi_k} - Q^{\pi_k} \leq \sqrt{\frac{2\beta^2 \gamma^2 |\mathcal{X}| \log \frac{2N}{\delta}}{n}} \|Q^* - Q^{\pi_k}\| \mathbf{1}$$
$$+ \left(\sqrt{\frac{4\beta^3 \log \frac{2N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{6N}{\delta}}{n} \right)^{3/4} \right) \mathbf{1} \tag{26}$$
$$+ \left(\frac{3\beta^3 \log \frac{12N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6 \log \frac{2N}{\delta}}{n}} \right) \mathbf{1}.$$

w.p. $1 - \delta$. Eq. (26) combined with the result of Lemma 9 and a union bound implies that

$$Q^* - Q^{\pi_k} \le (\varepsilon' + 2\gamma^k\beta^2)\mathbf{1} + \sqrt{\frac{2\beta^2\gamma^2|\mathfrak{X}|\log\frac{4N}{\delta}}{n}}\|Q^* - Q^{\pi_k}\|\mathbf{1}+$$

$$+ \left(\sqrt{\frac{4\beta^3\log\frac{2N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{12N}{\delta}}{n}\right)^{3/4}\right)\mathbf{1}$$

$$+ \left(\frac{3\gamma\beta^3\log\frac{24N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{4N}{\delta}}{n}}\right)\mathbf{1},$$

By taking the ℓ_∞-norm and solving the resulted bound in terms of $\|Q^* - Q^{\pi_k}\|$ we deduce

$$\|Q^* - Q^{\pi_k}\| \le \frac{1}{1 - \sqrt{\frac{2\beta^2\gamma^2|\mathfrak{X}|\log\frac{4N}{\delta}}{n}}}\left[\varepsilon' + 2\gamma^k\beta^2\right.$$

$$+ \sqrt{\frac{4\beta^3\log\frac{4N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{12N}{\delta}}{n}\right)^{3/4}$$

$$\left.+ \frac{3\beta^3\log\frac{24N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{4N}{\delta}}{n}}\right].$$

The choice of $n > 8\beta^2\gamma^2|\mathfrak{X}|\log\frac{4N}{\delta}$ deduce the result. ∎

Proof of Theorem 2

The result follows by solving the bound of Lemma 14 w.r.t. n and k, in that we also need to assume that $\varepsilon \le c\sqrt{\frac{\beta}{\gamma|\mathfrak{X}|}}$ for some $c > 0$ in order to reconcile the bound of Theorem 2 with Eq. (25). ∎

4.2 Proof of Theorem 3 - The Lower-Bound

In this section, we provide the proof of Theorem 3. 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 and 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 $\mathfrak{X}$ consists of three smaller subsets $\mathcal{S}$, $\mathcal{Y}_1$ 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
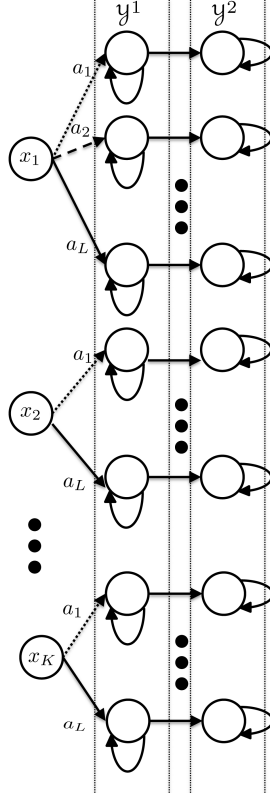
Fig. 1 The class of MDPs considered in the proof of Theorem 3. Nodes represent states and arrows show transitions between the states (see the text for details).

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 $y(z) \in \mathcal{Y}^2$. We notice that 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. Further, for all $M \in \mathbb{M}$, $\mathcal{Y}^2$ consists of only absorbing states, i.e., for all $y \in \mathcal{Y}^2$, $P(y|y) = 1$. 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_M^* can be solved in closed form from the Bellman equation. For all $M \in \mathbb{M}$

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

Now, let us consider two MDPs M_0 and 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 $0 < p < p + \alpha \leq 1$, to be quantified later in this section. We denote the set $\{M_0, M_1\} \subset \mathbb{M}$ with $\mathbb{M}^*$.

In the rest of this section, we concentrate on proving the lower bound on $\|Q_M^* - Q_T^{\mathfrak{A}}\|$ for all $M \in \mathbb{M}^*$, where $Q_T^{\mathfrak{A}}$ is the output of Algorithm $\mathfrak{A}$ after observing T state-transition samples. It turns out that a lower-bound on the sample complexity of $\mathbb{M}^*$ also bounds the sample complexity of $\mathbb{M}$ from below. In the sequel, we make use of the notation $\mathbb{E}_m$ ad $\mathbb{P}_m$ for the expectation and the probability under the model $M_m : m \in \{0, 1\}$, respectively.

We follow the following steps in the proof: **(i)** we prove a lower bound on the sample-complexity of learning the action-value function for every state-action pair $z \in \mathcal{S} \times \mathcal{A}$ 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 the bounds for all $z \in \mathcal{S} \times \mathcal{A}$ and prove the tight result of Theorem 3.

We begin our analysis of the lower bound by proving a lower-bound on the probability of failure of any RL algorithm to estimate a near-optimal action-value function for every state-action pair $z \in \mathcal{S} \times \mathcal{A}$. In order to prove this result (Lemma 16) we need to introduce some new notation: We define $Q_t^{\mathfrak{A}}(z)$ as the output of Algorithm $\mathfrak{A}$ using $t > 0$ transition samples from the state $y(z) \in \mathcal{Y}^1$ for all $z \in \mathcal{S} \times \mathcal{A}$. We also define the event $\mathcal{E}_1(z) \triangleq \{|Q_{M_0}^*(z) - Q_t^{\mathfrak{A}}(z)| \leq \varepsilon\}$ for all $z \in \mathcal{S} \times \mathcal{A}$. 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

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

where we have defined $\theta \triangleq \exp\left(-c_1'\alpha^2 t/(p(1-p))\right)$. Further, we define $\mathcal{E}(z) \triangleq \mathcal{E}_1(z) \cap \mathcal{E}_2(z)$.

We also make use of the following technical lemma which bounds the probability of the event $\mathcal{E}_2(z)$ from below:

Lemma 15 *For all $p > \frac{1}{2}$ and every $z \in \mathcal{S} \times \mathcal{A}$, we have*

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

Proof

We make use of the Chernoff-Hoeffding bound for Bernoulli's (Hagerup and Rüb, 1990) to prove the result: For $p > \frac{1}{2}$, define $\varepsilon = \sqrt{2p(1-p)t \log \frac{c_2'}{2\theta}}$, we then have

$$
\begin{aligned}
\mathbb{P}_0(\mathcal{E}_2(z)) &> -\exp\left(-\frac{\mathrm{KL}(p + \varepsilon \| p)}{t}\right) \\
&\geq 1 - \exp\left(-\frac{\varepsilon^2}{2tp(1-p)}\right) \\
&= 1 - \exp\left(-\frac{2tp(1-p)\log\frac{c_2'}{2\theta}}{2tp(1-p)}\right), \qquad \forall z \in \mathcal{S} \times \mathcal{A}, \\
&= 1 - \exp\left(-\log\frac{c_2'}{2\theta}\right) = 1 - \frac{2\theta}{c_2'}
\end{aligned}
$$

where $\mathrm{KL}(p\|q) \triangleq p \log(p/q) + (1 - p)\log((1 - p)/(1 - q))$ denotes the Kullback-Leibler divergence between p and q.

■

We now state the key result of this section:

Lemma 16 *For every RL Algorithm $\mathfrak{A}$ and every $z \in \mathcal{S} \times \mathcal{A}$, there exists an MDP $M_m \in \mathbb{M}^*$ and constants $c_1' > 0$ and $c_2' > 0$ such that*

$$\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) > \varepsilon) > \frac{\theta}{c_2'}, \tag{27}$$

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:

$$\mathbb{P}_m((|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) > \varepsilon) \le \frac{\theta}{c_2'}, \quad \text{or} \quad \mathbb{P}_m((|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) \le \varepsilon) \ge 1 - \frac{\theta}{c_2'}, \tag{28}$$

for all $M_m \in \mathbb{M}^*$ and show that this assumption leads to a contradiction.

By the assumption that $\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) > \varepsilon) \le \theta/c_2'$ 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 15 and by the choice of $c_2' = 6$ implies that, for all $z \in \mathcal{S} \times \mathcal{A}$, $\mathbb{P}_0(\mathcal{E}(z)) > 1/2$. Based on this result we now prove a bound from below on $\mathbb{P}_1(\mathcal{E}_1(z))$.

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 path of the random process (the sequence of observed transitions). The likelihood ratio of the event W between two MDPs M_1 and M_0 can then be written as

$$\frac{L_1(W)}{L_0(W)} = \frac{(p + \alpha)^k (1 - p - \alpha)^{t-k}}{p^k (1 - p)^{t-k}} = \left(1 + \frac{\alpha}{p}\right)^k \left(1 - \frac{\alpha}{1 - p}\right)^{t-k}$$

$$= \left(1 + \frac{\alpha}{p}\right)^k \left(1 - \frac{\alpha}{1 - p}\right)^{k\frac{1-p}{p}} \left(1 - \frac{\alpha}{1 - p}\right)^{t - \frac{k}{p}}.$$

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

$$\left(1 - \frac{\alpha}{1 - p}\right)^{(1-p)/p} \ge \exp\left(\frac{1 - p}{p}\left(-\frac{\alpha}{1 - p} - \left(\frac{\alpha}{1 - p}\right)^2\right)\right)$$

$$\ge \left(1 - \frac{\alpha}{p}\right)\left(1 - \frac{\alpha^2}{p(1 - p)}\right),$$

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

$$\frac{L_1(W)}{L_0(W)} \geq \left(1 - \frac{\alpha^2}{p^2}\right)^k \left(1 - \frac{\alpha^2}{p(1-p)}\right)^k \left(1 - \frac{\alpha}{1-p}\right)^{t-\frac{k}{p}}$$

$$= \geq \left(1 - \frac{\alpha^2}{p^2}\right)^t \left(1 - \frac{\alpha^2}{p(1-p)}\right)^t \left(1 - \frac{\alpha}{1-p}\right)^{t-\frac{k}{p}},$$

since $k \leq t$.

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

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

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

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

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

$$\left(1 - \frac{\alpha}{1-p}\right)^{t-\frac{k}{p}} \geq \left(1 - \frac{\alpha}{1-p}\right)^{\sqrt{2\frac{1-p}{p}t\log(c_2'/2\theta)}}$$

$$\geq \exp\left(-\sqrt{2\frac{\alpha^2}{p(1-p)}t\log(c_2'/(2\theta))}\right)$$

$$\geq \exp\left(-\sqrt{2/c1}\log(c_2'/\theta)\right) = (2\theta/c_2')^{\sqrt{2/c_1'}}.$$

We then 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

$$\mathbb{P}_1(\mathcal{E}_1(z)) \geq \mathbb{P}_1(\mathcal{E}(z)) = \mathbb{E}_1[\mathbb{1}_{\mathcal{E}}] = \mathbb{E}_0\left(\frac{L_1(W)}{L_0(W)}\mathbb{1}_{\mathcal{E}}\right)$$

$$\geq \mathbb{E}_0\left[2\theta/c_2'\mathbb{1}_{\mathcal{E}}\right] = 2\theta/c_2'\mathbb{P}_0(\mathcal{E}(z)) > \theta/c_2', \tag{29}$$

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

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 $Q^*_{M_1}(z) - Q^*_{M_0}(z) = \frac{\gamma}{1-\gamma(p+\alpha)} - \frac{\gamma}{1-\gamma p} > 2\varepsilon$, thus $Q^*_{M_0}(z) + \varepsilon < Q^*_{M_1}(z) - \varepsilon$. In words, the random event $\{|Q^*_{M_0}(z) - Q(z)| \leq \varepsilon\}$ does not overlap with the event $\{|Q^*_{M_1}(z) - Q(z)| \leq \varepsilon\}$.

Now let us return to the assumption of Eq. (28), which states that for all $M_m \in \mathbb{M}^*$, $\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) \leq \varepsilon) \geq 1 - \theta/c_2'$ under Algorithm $\mathfrak{A}$. Based

on Eq. (29), we have $\mathbb{P}_1(|Q^*_{M_0}(z) - Q^{\mathfrak{A}}_t(z)| \leq \varepsilon) > \theta/c'_2$. This combined with the fact that $\{|Q^*_{M_0}(z) - Q^{\mathfrak{A}}_t(z)|\}$ and $\{|Q^*_{M_1}(z) - Q^{\mathfrak{A}}_t(z)|\}$ do not overlap implies that $\mathbb{P}_1(|Q^*_{M_1}(z) - Q^{\mathfrak{A}}_t(z)|) \leq \varepsilon) \leq 1 - \theta/c'_2$, which violates the assumption of Eq. (28). Therefore, the lower bound of Eq. (27) shall hold.   ∎

Based on the result of Lemma 16 and 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

$$\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)) > \varepsilon) > \frac{1}{c'_2} \exp\left(\frac{-c_1 t \varepsilon^2}{6\beta^3}\right),$$

This result implies that for any state-action pair $z \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)| > \varepsilon) > \delta, \tag{30}$$

on M_0 or M_1 whenever the number of transition samples t is less than $\xi(\varepsilon, \delta) \triangleq \frac{6\beta^3}{c_1 \varepsilon^2} \log \frac{1}{c'_2 \delta}$.

Based on this result, we prove a lower bound on the number of samples T foe which $\|Q^*_{M_m} - Q^{\mathfrak{A}}_T\| > \varepsilon$ on either M_0 or M_1:

Lemma 17 *For any $\delta' \in (0, 1/2)$ and any Algorithm $\mathfrak{A}$ using a total number of transition samples less than $T = \frac{N}{6}\xi\left(\varepsilon, \frac{12\delta'}{N}\right)$, there exists an MDP $M_m \in \mathbb{M}^*$ such that*

$$\mathbb{P}_m\left(\|Q^*_{M_m} - Q^{\mathfrak{A}}_T\| > \varepsilon\right) > \delta'. \tag{31}$$

Proof

First, we 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 $z_{(1)}, \ldots, z_{(N/6)}$.

In order to prove that (31) holds for every RL algorithm, it is sufficient to prove it for the class of algorithms that return an estimate $Q^{\mathfrak{A}}_{T_z}(z)$, where T_z is the number of samples collected from z, for each state-action z based on the transition samples observed from z only.[8] This is due to the fact that the samples from z and z' are independent. Therefore, the samples collected from z' do not bring more information about $Q^*_M(z)$ than the information brought by the samples collected from z. Thus, by defining $\mathcal{Q}(z) \triangleq \{|Q^*_M(z) - Q^{\mathfrak{A}}_{T_z}(z)| > \varepsilon\}$ for all $M \in \mathbb{M}^*$ 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 exists an MDP $M_m \in \mathbb{M}^*$ such that

---

[8] We let T_z to be random.

$$\mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \cap \{T_{z(i)} \leq \xi(\varepsilon,\delta)\}_{1\leq i\leq N/6}\Big)$$

$$= \sum_{t_1=0}^{\xi(\varepsilon,\delta)} \cdots \sum_{t_{N/6}=0}^{\xi(\varepsilon,\delta)} \mathbb{P}_m\Big(\{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big)$$

$$\mathbb{P}_m\Big(\{\mathcal{Q}(z_{(i)})^c\}_{1\leq i\leq N/6} \cap \{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big)$$

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

$$\leq \sum_{t_1=0}^{\xi(\varepsilon,\delta)} \cdots \sum_{t_{N/6}=0}^{\xi(\varepsilon,\delta)} \mathbb{P}_m\Big(\{T_{z(i)} = t_i\}_{1\leq i\leq N/6}\Big)(1-\delta)^{N/6},$$

from Eq. (30), thus

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

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

$$\mathbb{P}_m(\|Q^*_{M_m} - Q^{\mathfrak{A}}_T\| > \varepsilon) \geq \mathbb{P}_m\Big(\bigcup_{z\in\mathcal{S}\times\mathcal{A}} \mathcal{Q}(z)\Big)$$

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

$$\geq 1 - (1-\delta)^{N/6} \geq \frac{\delta N}{12},$$

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

Lemma 17 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 $\|Q^*_M - Q^{\mathfrak{A}}_T\| \leq \varepsilon$ is at maximum $1 - \delta$ on either M_0 or M_1. This is equivalent to the argument that for every RL algorithm $\mathfrak{A}$ to be (ε,δ)-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 $T \geq \beta^3 N/(c_1\varepsilon^2)\log(N/(c_2\delta))$, which concludes the proof of Theorem 3.

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 both model-based Q-value iteration (QVI) and model-based policy iteration (PI), in the presence of the generative model of the MDP,

are optimal in the sense that the dependency of their performances on $1/\varepsilon$, N, δ and $1/(1-\gamma)$ matches the lower bound of RL. Also, our results have significantly improved on the state-of-the-art in terms of dependency on $1/(1-\gamma)$.

Overall, we conclude that both QVI and PI are efficient RL algorithms in terms of the number of samples required to attain a near optimal solution as the upper bounds on the performance loss of both algorithms completely match the lower bound of RL up to a multiplicative factor.

In this work, we only consider the problem of estimating the optimal action-value function when a generative model of the MDP is available. This allows us to make an accurate estimate of the state-transition distribution for all state-action pairs and then estimate the optimal control policy based on this empirical model. This is in contrast to the online RL setup in which the choice of the exploration policy has an influence on the behavior of the learning algorithm and vise-versa. Therefore, we do not compare our results with those of online RL algorithms such as PAC-MDP (Szita and Szepesvári, 2010; Strehl et al, 2009), upper-confidence-bound reinforcement learning (UCRL) (Jaksch et al, 2010) and REGAL of Bartlett and Tewari (2009). 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. However, those results bound the estimation error in terms of $V_{\max}$ rather than the total variance of discounted reward which leads to a non-tight sample complexity bound. 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 to bound the estimation error in terms of the variance of sum of discounted rewards. It must be pointed out that, almost contemporaneously to our work, Lattimore and Hutter (2012) have independently proven a similar upper-bound of order $\widetilde{O}(N/(\varepsilon^2(1-\gamma)^3))$ for UCRL algorithm under the assumption that only two states are accessible form any state-action pair. Their work also includes a similar lower bound of $\widetilde{\Omega}(N/(\varepsilon^2(1-\gamma)^3))$ for any RL algorithm which matches, up to a logarithmic factor, the result of Theorem 3.

References

Azar MG, Munos R, Ghavamzadeh M, Kappen HJ (2011a) Reinforcement Learning with a Near Optimal Rate of Convergence. Tech. rep., URL http://hal.inria.fr/inria-00636615

Azar MG, Munos R, Ghavamzadeh M, Kappen HJ (2011b) Speedy q-learning. In: Advances in Neural Information Processing Systems 24, pp 2411–2419

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

Bertsekas DP (2007) Dynamic Programming and Optimal Control, vol II, 3rd edn. Athena Scientific, Belmount, Massachusetts

Bertsekas DP, Tsitsiklis JN (1996) Neuro-Dynamic Programming. Athena Scientific, Belmont, Massachusetts

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

Even-Dar E, Mannor S, Mansour Y (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of Machine Learning Research 7:1079–1105

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

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

Kakade SM (2004) On the sample complexity of reinforcement learning. PhD thesis, Gatsby Computational Neuroscience Unit

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

Lattimore T, Hutter M (2012) Pac bounds for discounted mdps. CoRR abs/1202.3890

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. Journal of Machine Learning Research 5:623–648

Munos R, Moore A (1999) 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

Singh SP, Yee RC (1994) An upper bound on the loss from approximate optimal-value functions. Machine Learning 16(3):227–233

Sobel MJ (1982) The variance of discounted markov decision processes. Journal of Applied Probability 19:794–802

Strehl AL, Li L, Littman ML (2009) Reinforcement learning in finite MDPs: PAC analysis. Journal of Machine Learning Research 10:2413–2444

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. MIT Press, Cambridge, Massachusetts

Szepesvári C (2010) Algorithms for Reinforcement Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers

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

Wiering M, van Otterlo M (2012) Reinforcement Learning: State-of-the-Art, Springer, chap 1, pp 3–39