Mach Learn (2013) 91:325–349
DOI 10.1007/s10994-013-5368-1

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

Mohammad Gheshlaghi Azar · Rémi Munos · Hilbert J. Kappen

Received: 30 July 2012 / Accepted: 24 April 2013 / Published online: 14 May 2013

Abstract We consider the problems of learning the optimal action-value function and the optimal policy 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 $O(N \log(N/\delta)/((1-\gamma)^3\varepsilon^2))$ 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 $O(N \log(N/\delta)/((1-\gamma)^3\varepsilon^2))$ samples is required to find an ε-optimal policy w.p. $1-\delta$. We also prove a matching lower bound of $\Theta(N \log(N/\delta)/((1-\gamma)^3\varepsilon^2))$ on the sample complexity of estimating the optimal action-value function with ε accuracy. To the best of our knowledge, this is the first minimax result on the sample complexity of RL: the upper bounds match 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 $1/(1-\gamma)$.

Keywords Sample complexity · Markov decision processes · Reinforcement learning · Learning theory

Editor: Csaba Szepesvari.

M. Gheshlaghi Azar (✉) · H.J. Kappen
Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands
e-mail: m.azar@science.ru.nl

H.J. Kappen
e-mail: b.kappen@science.ru.nl

Present address:
M. Gheshlaghi Azar
School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

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

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 (Szepesvári 2010; Sutton and Barto 1998). To address this problem 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, when we have access to a generative model of the MDP, that is, a sampling device which can generate next-state samples for all state-action pairs 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 as well as model-based policy iteration (PI). These bounds 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 QVI and PI by an order of $1/(1-\gamma)$. We also present a new minimax lower bound of $\Theta(N\log(N/\delta)/((1-\gamma)^3\varepsilon^2))$, 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

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

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

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, p. 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 model estimation error may cause a large error in the estimate of the optimal value function, especially when the discount factor γ is close to 1.

We must emphasize that, in this work, we only consider the problem of estimating the optimal policy 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 (Szita and Szepesvári 2010; Strehl et al. 2009; Jaksch et al. 2010; Bartlett and Tewari 2009) in which the choice of exploration policy has an influence on the behavior of the learning algorithm and vise-versa. For that reason, we do not provide a detailed comparison of our results with those of online RL.

This paper extends on the results of Azar et al. (2012) by including a new sample complexity bound for finding an ε-optimal policy, whereas Azar et al. (2012) only prove bounds on the sample complexity of estimating the optimal action-value function. Also Azar et al. (2012) only consider the QVI algorithm. In this paper we prove bounds for PI as well as QVI.

The rest of the paper is organized as follows. After introducing the notations used in the paper in Sect. 2, we describe QVI and PI algorithms in Sect. 2.1. We then state our main theoretical results, which are in the form of PAC sample complexity bounds in Sect. 3. Section 4 contains the detailed proofs of the results of Sect. 3, that is, 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 Sect. 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 $(P_\pi V)(x) \triangleq \sum_{y \in \mathcal{X}} y \in X P_\pi(y|x) V(y)$ for all $x \in \mathcal{X}$, respectively. For any policy π, we also define the operator $(P^\pi)^k \cdot$ as

$$\left(P^\pi\right)^k Q(z) \triangleq \underbrace{P^\pi \cdots P^\pi}_{k} Q(z),$$

for all $k \geq 1$ and $z \in \mathcal{Z}$. Based on this definition we then define the operator $(I - \gamma P^\pi)^{-1} \cdot$ as $(I - \gamma P^\pi)^{-1} Q(z) \triangleq \sum_{i \geq 0} (\gamma P^\pi)^i Q(z)$ for all $z \in \mathcal{Z}$.

The optimal action-value function Q^* is the unique fixed-point of the *Bellman optimality operator* defined as

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

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, that is, 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,

[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 $(r_{\max} - r_{\min})^2$.

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.

2.1 Algorithms

We begin by describing the model-estimation 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 (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 different, 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 Algorithms 1 and 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

```
P̂ = EstimateModel(n)                          ▷ Estimate the model (defined in Algorithm 3)
for j := 0, 1, ..., k − 1 do
    for each x ∈ X do
        π_j(x) = arg max_{a∈A} Q_j(x, a)       ▷ greedy policy w.r.t. the latest estimation of Q*
        for each a ∈ A do
            T̂Q_j(x, a) = r(x, a) + γ(P̂^{π_j} Q_j)(x, a)      ▷ empirical Bellman operator
            Q_{j+1}(x, a) = T̂Q_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 policy π_0, samples per state-action n, number of iterations k

$\widehat{P}$ =EstimateModel(n) ▷ Estimate the model (defined in Algorithm 3)
Q_0=SolveDP($\widehat{P}, \pi_0$)
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^*
  end for
  $\widehat{Q}^{\pi_j}$=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 SolveDP(P, π)
  $Q = (I - \gamma P^{\pi})^{-1} r$
  return Q
end function

Algorithm 3 Function: EstimateModel

Require: The generative model (simulator) of P
function EstimateModel(n) ▷ Estimating the transition model using n samples
  $\forall (y, z) \in \mathcal{X} \times \mathcal{Z} : m(y, z) = 0$ ▷ initialization
  for each $z \in \mathcal{Z}$ **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
    $\forall y \in \mathcal{X} : \widehat{P}(y|z) = \frac{m(y,z)}{n}$ ▷ 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. 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*

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

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. 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*

$$T = \left\lceil \frac{c'\beta^3 N}{\varepsilon^2} \log \frac{c'_0 N}{\delta} \right\rceil,$$

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 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)-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 $M \in \mathbb{M}$.

Theorem 3 (Lower bound on the sample complexity of RL) *Let Assumption 1 hold. There exist some constants* ε_0, δ_0, c_1, c_2, *and a class of MDPs* $\mathbb{M}$, *such that for all* $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0)$, *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*

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

We note that the result of Theorem 3 is rather general and algorithm independent: this result provides a tight lower-bound for every RL algorithm regardless of whether it makes use of the generative or it is an online approach.

4 Analysis

In this section, we first provide the full proof of the finite-time PAC bound of QVI and PI, reported in Theorems 1 and 2, in Sect. 4.1. We then prove Theorem 3, a new RL lower bound, in Sect. 4.2.

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

4.1 Proofs of Theorems 1 and 2—the upper bounds

We begin by introducing some new notation required for the analysis. For any 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 for the sequence of state-action pairs $\{Z_0, Z_1, \ldots\}$ starting from $z \in \mathcal{Z}$ under the policy π. We also make use of the 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 ρ is a probability distribution on $\mathcal{Y}$. shall denote σ_{V^π} 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 $\widehat{\sigma}_{V^\pi}(z) \triangleq \gamma^2 \mathbb{V}_{y\sim\widehat{P}(\cdot|z)}[V^\pi(y)]$ 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 notice that for any policy π and for all $z \in \mathcal{Z}$, σ_{V^π} can be written as

$$\sigma_{V^\pi}(z) = \gamma^2 P\big[\big|V^\pi - PV^\pi\big|^2\big](z) = \gamma^2 P^\pi\big[\big|Q^\pi - P^\pi Q^\pi\big|^2\big](z).$$

In this subsection, we focus on proving high probability bounds on $\|Q^* - Q_k\|$ and $\|Q^* - Q^{\pi_k}\|$ for both QVI and PI. These high probability bounds imply the sample complexity bounds of Theorems 1 and 2. One can easily show 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 a bound on $\|Q^* - Q_k\|$, we only need to bound $\|Q^* - \widehat{Q}^*\|$ in high probability. One can prove a crude bound of $\widetilde{O}(\beta^2/\sqrt{n})$ on $\|Q^* - \widehat{Q}^*\|$ by first proving that $\|Q^* - \widehat{Q}^*\| \leq \beta\|(P - \widehat{P})V^*\|$ and then using the Hoeffding's tail inequality (Cesa-Bianchi and Lugosi 2006, Appendix, p. 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 3 component-wise upper and lower bounds on the error $Q^* - \widehat{Q}^*$, which are expressed in terms of $(I - \gamma\widehat{P}^{\pi^*})^{-1}[P - \widehat{P}]V^*$ and $(I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}[P - \widehat{P}]V^*$, respectively. (ii) We make use of Bernstein's inequality to bound $[P - \widehat{P}]V^*$ in terms of the squared root of the variance of V^* in high probability. (iii) We prove the key result of this subsection (Lemma 7), 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 $n > c_1\gamma^2\beta^2|\mathcal{X}|\log(c_2 N/\delta)$.

We begin by Lemma 1, which bounds $\|Q_k - \widehat{Q}^*\|$, for both QVI and PI.[6]

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*

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

[6] There exist similar results in the literature regarding the convergence rate of the iterates of the Bellman recursion to its fixed point (see, e.g., Puterman 1994, Chap. 6). However, those results are often expressed in terms of the state-value function, whereas here we consider the action-value function.

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

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

Thus by an immediate recursion

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

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

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

where in the last line we rely on the fact that π_{k-1} is the greedy policy w.r.t. Q_{k-1}, which implies the component-wise inequality $\widehat{P}^{\pi_{k-1}} Q_{k-1} \geq \widehat{P}^{\widehat{\pi}^*} Q_{k-1}$. The result follows from Eq. 1 by taking the ℓ_∞-norm on both sides of the inequality and then recursively expanding the resulting bound. □

One can easily prove the following lemma, 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). Lemma 2 is required for the proof of Theorem 2.

Lemma 2 *Let Assumption 1 hold and π_k be the greedy policy induced by the k^{th} 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*

$$\left\| \widehat{Q}^{\pi_k} - \widehat{Q}^* \right\| \leq 2\gamma^{k+1}\beta^2, \quad \text{and} \quad \left\| \widehat{V}^{\pi_k} - \widehat{V}^* \right\| \leq 2\gamma^{k+1}\beta^2.$$

Proof Based on the main theorem of Singh and Yee (1994) we have, for both QVI and PI:

$$\begin{aligned} \left\| \widehat{V}^{\pi_k} - \widehat{V}^* \right\| \leq \left\| \widehat{Q}^{\pi_k} - \widehat{Q}^* \right\| &\leq 2\gamma\beta \left\| Q_k - \widehat{Q}^* \right\| \\ &\leq 2\gamma^{k+1}\beta^2, \end{aligned}$$

where in the last line we make use of the result of Lemma 1. □

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 2 for both QVI and PI.

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

Lemma 3 (Component-wise bounds on $Q^* - \widehat{Q}^*$)

$$Q^* - \widehat{Q}^* \leq \gamma \left(I - \gamma \widehat{P}^{\pi^*} \right)^{-1} [P - \widehat{P}] V^*, \tag{2}$$

$$Q^* - \widehat{Q}^* \geq \gamma \left(I - \gamma \widehat{P}^{\widehat{\pi}^*} \right)^{-1} [P - \widehat{P}] V^*. \tag{3}$$

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

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

In the case of Eq. 3 we have

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

where in the last line we take the following steps:

$$\begin{aligned} -\left(I - \gamma \widehat{P}^{\widehat{\pi}^*}\right)^{-1} \widehat{P}^{\widehat{\pi}^*} Q^* &= -\sum_{i \geq 0} \left(\gamma \widehat{P}^{\widehat{\pi}^*}\right)^i \widehat{P}^{\widehat{\pi}^*} Q^* \\ &\geq -\sum_{i \geq 0} \left(\gamma \widehat{P}^{\widehat{\pi}^*}\right)^i \widehat{P}^{\pi^*} Q^* = -\left(I - \gamma \widehat{P}^{\widehat{\pi}^*}\right)^{-1} \widehat{P} V^*. \end{aligned}$$

The inequality $-\widehat{P}^{\widehat{\pi}^*} Q^* \geq -\widehat{P}^{\pi^*} Q^*$ holds since π^* is the greedy policy w.r.t. Q^*. □

We now concentrate on bounding the right hand sides (RHS) of Eqs. 2 and 3 in high probability, for that we need the following technical lemmas (Lemmas 4 and 5).

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

$$\left\| V^* - \widehat{V}^{\pi^*} \right\| \leq c_v, \quad \textit{and} \quad \left\| V^* - \widehat{V}^* \right\| \leq c_v,$$

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

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

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

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

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

Now we focus on bounding $\| V^* - \widehat{V}^* \|$:

$$\begin{aligned} \left\| V^* - \widehat{V}^* \right\| &\leq \left\| Q^* - \widehat{Q}^* \right\| = \left\| \mathcal{T} Q^* - \widehat{\mathcal{T}} \widehat{Q}^* \right\| \\ &\leq \left\| \mathcal{T} Q^* - \widehat{\mathcal{T}}^{\pi^*} Q^* \right\| + \left\| \widehat{\mathcal{T}}^{\pi^*} Q^* - \widehat{\mathcal{T}} \widehat{Q}^* \right\| \\ &= \gamma \left\| P^{\pi^*} Q^* - \widehat{P}^{\pi^*} Q^* \right\| + \gamma \left\| \widehat{P}^{\pi^*} Q^* - \widehat{P}^{\widehat{\pi}^*} \widehat{Q}^* \right\| \\ &= \gamma \left\| (P - \widehat{P}) V^* \right\| + \gamma \left\| \widehat{P} \left(V^* - \widehat{V}^*\right) \right\| \\ &\leq \gamma \left\| (P - \widehat{P}) V^* \right\| + \gamma \left\| V^* - \widehat{V}^* \right\|. \end{aligned} \tag{5}$$

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

$$\left\| V^* - \widehat{V}^* \right\| \leq \gamma \beta \left\| (P - \widehat{P}) V^* \right\|. \tag{6}$$

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

$$\mathbb{P}\Big(\big|\big((P-\widehat{P})V^*\big)(z)\big| \geq \varepsilon\Big) \leq 2\exp\left(\frac{-n\varepsilon^2}{2\beta^2}\right).$$

By applying the union bound we deduce

$$\mathbb{P}\Big(\big\|(P-\widehat{P})V^*\big\| \geq \varepsilon\Big) \leq 2|\mathcal{Z}|\exp\left(\frac{-n\varepsilon^2}{2\beta^2}\right). \tag{7}$$

We then define the probability of failure δ as

$$\delta \triangleq 2N\exp\left(\frac{-n\varepsilon^2}{2\beta^2}\right). \tag{8}$$

By plugging Eq. 8 into Eq. 7 we deduce

$$\mathbb{P}\Big[\big\|(P-\widehat{P})V^*\big\| < \beta\sqrt{2\log(2N/\delta)/n}\,\Big] \geq 1-\delta. \tag{9}$$

The results then follow by plugging Eq. 9 into Eqs. 6 and 5. □

We now state Lemma 5 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 6.

Lemma 5 *Let Assumption 1 hold and* $0 < \delta < 1$. *Then, w.p. at least* $1-\delta$:

$$\sqrt{\sigma_{V^*}} \leq \sqrt{\widehat{\sigma}_{\widehat{V}^{\pi^*}}} + b_v\mathbf{1}, \tag{10}$$

$$\sqrt{\sigma_{V^*}} \leq \sqrt{\widehat{\sigma}_{\widehat{V}^*}} + b_v\mathbf{1}, \tag{11}$$

where b_v *is defined as*

$$b_v \triangleq \left(\frac{18\gamma^4\beta^4\log\frac{3N}{\delta}}{n}\right)^{1/4} + \sqrt{\frac{4\gamma^2\beta^4\log\frac{6N}{\delta}}{n}},$$

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

Proof Here, we only prove Eq. 10. One can prove Eq. 11 following similar lines:

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

where in the last line we rely on a triangle inequality argument. 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

$$\begin{aligned}
\sigma_{V^*}(z) &\leq \gamma^2[P-\widehat{P}]V^{*2}(z) - \gamma^2\big[(P-\widehat{P})V^*\big]\big[(P+\widehat{P})V^*\big](z) \\
&\quad + \big(\gamma\,\|V^* - \widehat{V}^{\pi^*}\| + \sqrt{\widehat{\sigma}_{\widehat{V}^{\pi^*}}(z)}\big)^2.
\end{aligned}$$

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

$$\sigma_{V^*}(z) \leq \left[\sqrt{\widehat{\sigma}_{\widehat{V}^{\pi^*}}(z)} + \sqrt{\frac{4\gamma^2\beta^4 \log\frac{6N}{\delta}}{n}}\right]^2 + 3\gamma^2\beta^2\sqrt{\frac{2\log\frac{3}{\delta}}{n}}, \tag{12}$$

in which we make use of Hoeffding's inequality as well as Lemma 4 and a union bound to prove the bound on σ_{V^*} in high probability. The result follows from Eq. 12 by taking the square root on both sides of the inequality as well as applying union bound on all state-action pairs. □

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, p. 361) as well as Lemma 5.

Lemma 6 *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*

$$\gamma(P-\widehat{P})V^* \leq \sqrt{\frac{c_{pv}\widehat{\sigma}_{\widehat{V}^{\pi^*}}}{n}} + b_{pv}\mathbf{1}, \tag{13}$$

$$\gamma(P-\widehat{P})V^* \geq -\sqrt{\frac{c_{pv}\widehat{\sigma}_{\widehat{V}^*}}{n}} - b_{pv}\mathbf{1}. \tag{14}$$

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)

$$\gamma(P-\widehat{P})V^* \leq \sqrt{c'_{pv}\frac{\sigma_{V^*}}{n}} + b'_{pv}\mathbf{1}, \tag{15}$$

$$\gamma(P-\widehat{P})V^* \geq -\sqrt{c'_{pv}\frac{\sigma_{V^*}}{n}} - b'_{pv}\mathbf{1}, \tag{16}$$

where $c'_{pv} \triangleq 2\log(N/\delta)$ and $b'_{pv} \triangleq 2\gamma\beta\log(N/\delta)/3n$. Plugging Eqs. 10 and 11 into Eqs. 15 and 16, respectively, and then taking a union bound conclude the proof. □

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. We note that this result is similar to those of Munos and Moore (1999), Sobel (1982) in the sense that, like those previous results, it shows that the variance Σ^π satisfies a Bellman-like equation. The difference is that, here, we consider the total of variance of the sum of rewards for every state-action

pair, whereas Munos and Moore (1999), Sobel (1982) express their results in terms of the variance of the sum of rewards of every state. Later, we use Lemma 7, in Lemma 8, to bound $(I-\gamma P^{\pi})^{-1}\sigma_{V^{\pi}}$.

Lemma 7 *Σ^{π} satisfies the Bellman equation*

$$\Sigma^{\pi} = \sigma_{V^{\pi}} + \gamma^2 P^{\pi}\Sigma^{\pi}. \tag{17}$$

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

$$\begin{aligned}
\Sigma^{\pi}(z) &= \mathbb{E}\left[\left|\sum_{t\geq 0}\gamma^t r(Z_t) - Q^{\pi}(z)\right|^2\right] \\
&= \mathbb{E}_{Z_1\sim P^{\pi}(\cdot|z)}\mathbb{E}\left[\left|\sum_{t\geq 1}\gamma^t r(Z_t) - \gamma Q^{\pi}(Z_1) - \left(Q^{\pi}(z) - r(z) - \gamma Q^{\pi}(Z_1)\right)\right|^2\right] \\
&= \gamma^2\mathbb{E}_{Z_1\sim P^{\pi}(\cdot|z)}\mathbb{E}\left[\left|\sum_{t\geq 1}\gamma^{t-1} r(Z_t) - Q^{\pi}(Z_1)\right|^2\right] \\
&\quad - 2\mathbb{E}_{Z_1\sim P^{\pi}(\cdot|z)}\left[\left(Q^{\pi}(z) - r(z) - \gamma Q^{\pi}(Z_1)\right)\mathbb{E}\left(\sum_{t\geq 1}\gamma^t r(Z_t) - \gamma Q^{\pi}(Z_1)\,\middle|\, Z_1\right)\right] \\
&\quad + \mathbb{E}_{Z_1\sim P^{\pi}(\cdot|z)}\left(\left|Q^{\pi}(z) - r(z) - \gamma Q^{\pi}(Z_1)\right|^2\right) \\
&= \gamma^2\mathbb{E}_{Z_1\sim P^{\pi}(\cdot|z)}\mathbb{E}\left[\left|\sum_{t\geq 1}\gamma^{t-1} r(Z_t) - Q^{\pi}(Z_1)\right|^2\right] + \gamma^2\mathbb{V}_{Y_1\sim P(\cdot|z)}\left(Q^{\pi}\left(Y_1,\pi(Y_1)\right)\right) \\
&= \gamma^2\left[P^{\pi}\Sigma^{\pi}\right](z) + \sigma_{V^{\pi}}(z),
\end{aligned}$$

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 7, one can prove the following result on the discounted variance.

Lemma 8

$$\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1}\sigma_{V^{\pi}}\right\| = \left\|\Sigma^{\pi}\right\| \leq \beta^2, \tag{18}$$

$$\left\|\left(I-\gamma P^{\pi}\right)^{-1}\sqrt{\sigma_{V^{\pi}}}\right\| \leq 2\log(2)\left\|\sqrt{\beta\Sigma^{\pi}}\right\| \leq 2\log(2)\beta^{1.5}. \tag{19}$$

Proof The first inequality follows from Lemma 7 by solving Eq. 17 in terms of Σ^{π} and taking the sup-norm over both sides of the resulting equation. In the case of Eq. 19 we have

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

$$\leq \sum_{l\geq 0}\left(\gamma^t\right)^l \left\| \sum_{j=0}^{t-1} \left(\gamma P^\pi\right)^j \sqrt{\sigma_{V^\pi}} \right\|$$

$$= \frac{1}{1-\gamma^t} \left\| \sum_{j=0}^{t-1} \left(\gamma P^\pi\right)^j \sqrt{\sigma_{V^\pi}} \right\|, \tag{20}$$

in which we write $k = tl + j$ with t any positive integer.[7] We now prove a bound on $\| \sum_{j=0}^{t-1} (\gamma P^\pi)^j \sqrt{\sigma_{V^\pi}} \|$ by making use of Jensen's inequality, Cauchy-Schwarz inequality and Eq. 18:

$$\left\| \sum_{j=0}^{t-1} \left(\gamma P^\pi\right)^j \sqrt{\sigma_{V^\pi}} \right\| \leq \left\| \sum_{j=0}^{t-1} \gamma^j \sqrt{\left(P^\pi\right)^j \sigma_{V^\pi}} \right\| \leq \sqrt{t} \left\| \sqrt{\sum_{j=0}^{t-1} \left(\gamma^2 P^\pi\right)^j \sigma_{V^\pi}} \right\|$$

$$\leq \sqrt{t} \left\| \sqrt{\left(I - \gamma^2 P^\pi\right)^{-1} \sigma_{V^\pi}} \right\| = \left\| \sqrt{t \Sigma^\pi} \right\|. \tag{21}$$

The result then follows by plugging Eq. 21 into Eq. 20 and optimizing the bound in terms of t to achieve the best dependency on β. □

Now, we make use of Lemmas 8 and 6 to bound $\| Q^* - \widehat{Q}^* \|$ in high probability.

Lemma 9 *Let Assumption 1 hold. Then, for any* $0 < \delta < 1$:

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

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

$$\varepsilon' \triangleq \sqrt{\frac{4\beta^3 \log \frac{4N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{12N}{\delta}}{n} \right)^{3/4} + \frac{3\beta^3 \log \frac{24N}{\delta}}{n}. \tag{22}$$

Proof By incorporating the result of Lemmas 6 and 8 into Lemma 3 and taking in to account that $(I - \gamma \widehat{P}^{\pi^*})^{-1} \mathbf{1} = \beta \mathbf{1}$, we deduce[8]

$$\begin{aligned} Q^* - \widehat{Q}^* &\leq b\mathbf{1}, \\ Q^* - \widehat{Q}^* &\geq -b\mathbf{1}, \end{aligned} \tag{23}$$

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

$$b \triangleq \sqrt{\frac{4\beta^3 \log \frac{2N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3} \log \frac{6N}{\delta}}{n} \right)^{3/4} + \frac{3\beta^3 \log \frac{12N}{\delta}}{n}. \tag{24}$$

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.

[8] Note that, for any policy π, Lemma 8 implies the component-wise inequality $(I - \gamma P^\pi)^{-1} \sqrt{\sigma_{V^\pi}} \leq 2\log(2)\beta^{1.5}\mathbf{1}$.

Proof of Theorem 1 We define the total error $\varepsilon \triangleq \varepsilon' + \gamma^k \beta$, which bounds $\|Q^* - Q_k\| \leq \|Q^* - \widehat{Q}^*\| + \|\widehat{Q}^* - Q_k\|$ in high probability (ε' is defined in Lemma 9). The results then follows by solving this bound w.r.t. n and k and then quantifying the total number of samples by $T = Nn$. □

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

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

where b is defined by Eq. 24.

Proof We make use of Lemmas 2 and 9 to prove the result:

$$\begin{aligned} Q^* - Q^{\pi_k} &= Q^* - \widehat{Q}^* + \widehat{Q}^* - \widehat{Q}^{\pi_k} + \widehat{Q}^{\pi_k} - Q^{\pi_k} \\ &\leq b\mathbf{1} + \widehat{Q}^* - \widehat{Q}^{\pi_k} + \widehat{Q}^{\pi_k} - Q^{\pi_k} && \text{by Eq. 23} \\ &\leq \big(b + 2\gamma^k \beta^2\big)\mathbf{1} + \widehat{Q}^{\pi_k} - Q^{\pi_k} && \text{by Lemma 2.} \end{aligned}$$

□

Lemma 10 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, for which we make use of the following lemma:

Lemma 11 (Component-wise upper bound on $\widehat{Q}^{\pi_k} - Q^{\pi_k}$)

$$\widehat{Q}^{\pi_k} - Q^{\pi_k} \leq \gamma \big(I - \gamma \widehat{P}^{\pi_k}\big)^{-1} (P - \widehat{P}) V^* + \gamma\beta \big\| (P - \widehat{P})\big(V^* - V^{\pi_k}\big)\big\| \mathbf{1}. \tag{25}$$

Proof We prove this result using a similar argument as in the proof of Lemma 3:

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

where in the last line, we rely on the following:

$$\big(I - \gamma \widehat{P}^{\pi_k}\big)^{-1} \mathbf{1} = \left[I + \sum_{i>0} \big(\gamma \widehat{P}^{\pi_k}\big)^i \right] \mathbf{1} = \left(1 + \sum_{i>0} \gamma^i \right) \mathbf{1} = \beta \mathbf{1}.$$

□

Now we bound the terms in the RHS of Eq. 25 in high probability. We begin by bounding $\gamma (I - \gamma \widehat{P}^{\pi_k})^{-1} (P - \widehat{P}) V^*$:

Lemma 12 *Let Assumption 1 hold. Then, w.p. at least* $1-\delta$ *we have*

$$\gamma\left(I-\gamma\widehat{P}^{\pi_k}\right)^{-1}(P-\widehat{P})V^* \le \left(\sqrt{\frac{4\beta^3\log\frac{2N}{\delta}}{n}}+\left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{6N}{\delta}}{n}\right)^{3/4}\right)\mathbf{1} + \left(\frac{3\beta^3\log\frac{12N}{\delta}}{n}+\sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{2N}{\delta}}{n}}\right)\mathbf{1}.$$

Proof From Lemma 6, w.p. at least $1-\delta$, we have

$$\begin{aligned}\gamma(P-\widehat{P})V^* &\le \sqrt{\frac{2\log\frac{2N}{\delta}\widehat{\sigma}_{\widehat{V}^*}}{n}}+b_{pv}\mathbf{1}\\ &\le \sqrt{\frac{2\log\frac{2N}{\delta}(\sqrt{\widehat{\sigma}_{\widehat{V}^{\pi_k}}}+\gamma\|\widehat{Q}^{\pi_k}-\widehat{Q}^*\|)^2}{n}}+b_{pv}\mathbf{1}\\ &\le \sqrt{\frac{2\log\frac{2N}{\delta}\widehat{\sigma}_{\widehat{V}^{\pi_k}}}{n}}+\left(b_{pv}+\sqrt{\frac{8\gamma^{2k+2}\beta^4\log\frac{2N}{\delta}}{n}}\right)\mathbf{1}, \end{aligned} \tag{26}$$

where in the last line we rely on Lemma 2. The result then follows by combining Eq. 26 with the result of Lemma 8. □

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 13 (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*

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

Lemma 14 *Let Assumption 1 hold. Then, w.p. at least* $1-\delta$ *we have*

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

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

$$\begin{aligned}\gamma\left|(P-\widehat{P})\left(V^*-V^{\pi_k}\right)(z)\right| &\le \gamma\left\|P(\cdot|z)-\widehat{P}(\cdot|z)\right\|_1\left\|V^*-V^{\pi_k}\right\|\\ &\le \gamma\left\|P(\cdot|z)-\widehat{P}(\cdot|z)\right\|_1\left\|Q^*-Q^{\pi_k}\right\|.\end{aligned}$$

This combined with Lemma 13 implies that

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

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

We now make use of the results of Lemma 14 and Lemma 12 to bound $\|Q^*-Q^{\pi_k}\|$ in high probability:

Lemma 15 *Let Assumption 1 hold. Assume that*

$$n \ge 8\gamma^2\beta^2|\mathcal{X}|\log\frac{4N}{\delta}. \tag{27}$$

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

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

where ε' *is defined by Eq. 22.*

Proof By incorporating the result of Lemmas 14 and 12 into Lemma 11 we deduce

$$\begin{aligned}\widehat{Q}^{\pi_k}-Q^{\pi_k} \le{}& \sqrt{\frac{2\beta^2\gamma^2|\mathcal{X}|\log\frac{2N}{\delta}}{n}}\left\|Q^*-Q^{\pi_k}\right\|\mathbf{1} \\ &+\left(\sqrt{\frac{4\beta^3\log\frac{2N}{\delta}}{n}}+\left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{6N}{\delta}}{n}\right)^{3/4}\right)\mathbf{1} \\ &+\left(\frac{3\beta^3\log\frac{12N}{\delta}}{n}+\sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{2N}{\delta}}{n}}\right)\mathbf{1},\end{aligned} \tag{28}$$

w.p. $1-\delta$. Equation 28 combined with the result of Lemma 10 and a union bound implies that

$$\begin{aligned}Q^*-Q^{\pi_k} \le{}& \big(\varepsilon'+2\gamma^k\beta^2\big)\mathbf{1}+\sqrt{\frac{2\beta^2\gamma^2|\mathcal{X}|\log\frac{4N}{\delta}}{n}}\left\|Q^*-Q^{\pi_k}\right\|\mathbf{1} \\ &+\left(\sqrt{\frac{4\beta^3\log\frac{2N}{\delta}}{n}}+\left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{12N}{\delta}}{n}\right)^{3/4}\right)\mathbf{1} \\ &+\left(\frac{3\gamma\beta^3\log\frac{24N}{\delta}}{n}+\sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{4N}{\delta}}{n}}\right)\mathbf{1}.\end{aligned}$$

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

$$\|Q^* - Q^{\pi_k}\| \leq \frac{1}{1 - \sqrt{\frac{2\beta^2\gamma^2|\mathcal{X}|\log\frac{4N}{\delta}}{n}}} \Bigg[\varepsilon' + 2\gamma^k\beta^2 + \sqrt{\frac{4\beta^3\log\frac{4N}{\delta}}{n}} + \left(\frac{5(\gamma\beta^2)^{4/3}\log\frac{12N}{\delta}}{n}\right)^{3/4} + \frac{3\beta^3\log\frac{24N}{\delta}}{n} + \sqrt{\frac{8\gamma^{2k+2}\beta^6\log\frac{4N}{\delta}}{n}} \Bigg].$$

The choice of $n > 8\beta^2\gamma^2|\mathcal{X}|\log\frac{4N}{\delta}$ completes the proof. □

Proof of Theorem 2 The result follows by solving the bound of Lemma 15 w.r.t. n and k, in that we also need to assume that $\varepsilon \leq c\sqrt{\frac{\beta}{\gamma|\mathcal{X}|}}$ for some $c > 0$ in order to reconcile the bound of Theorem 2 with Eq. 27. □

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 Fig. 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 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 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 $\mathcal{S}$, that is, there is no overlapping path in the MDP. Further, for all $M \in \mathbb{M}$, $\mathcal{Y}^2$ consists of only absorbing states, that is, for all $y \in \mathcal{Y}^2$, $P(y|y) = 1$. The instant reward

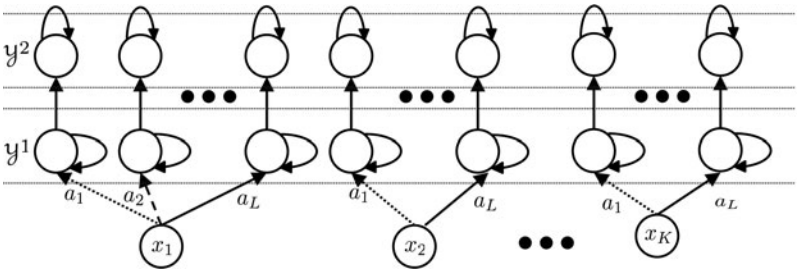

Fig. 1 The class of MDPs considered in the proof of Theorem 3. *Nodes* represent states and *arrows* show transitions between the states (see the text for details)

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 $M \in \mathbb{M}$

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

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 $0 < p < p+\alpha \leq 1$. The exact values of p and α 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 $\|Q^*_M - Q^{\mathfrak{A}}_T\|$ for all $M \in \mathbb{M}^*$, where $Q^{\mathfrak{A}}_T$ 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 other 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 achieve an ε-close estimate of the optimal action-value function for every state-action pair $z \in \mathcal{S} \times \mathcal{A}$. In order to prove this result (Lemma 17) we need to introduce some new notation: We define $Q^{\mathfrak{A}}_t(z)$ 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 $\mathcal{E}_1(z) \triangleq \{|Q^*_{M_0}(z) - Q^{\mathfrak{A}}_t(z)| \leq \varepsilon\}$ 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 $\theta \triangleq \exp(-c'_2\alpha^2 t/(p(1-p)))$. 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 16 *For all $p > \frac{1}{2}$ and every $z \in \mathcal{S} \times \mathcal{A}$, we have*

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

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

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

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 17 *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*

$$\mathbb{P}_m\big(\big|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)\big|\big) > \varepsilon) > \frac{\theta}{c_2'}, \tag{29}$$

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, that is, we assume that there exists an algorithm $\mathfrak{A}$ for which

$$\mathbb{P}_m\big(\big(\big|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)\big|\big) > \varepsilon\big) \leq \frac{\theta}{c_2'}, \tag{30}$$

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

By the assumption that $\mathbb{P}_m(|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)|) > \varepsilon) \leq \theta/c_2'$ for all $M_m \in \mathbb{M}^*$, we have $\mathbb{P}_0(\mathcal{E}_1(z)) \geq 1-\theta/c_2' \geq 1-1/c_2'$. This combined with Lemma 16 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 $\log(1-u) \geq -u-u^2$ for $0 \leq u \leq 1/2$, and $\exp(-u) \geq 1-u$ for $0 \leq u \leq 1$, we have

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

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 $\alpha^2 \leq p(1-p)$,

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

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

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

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 $\alpha \leq (1-p)/2$:

$$\left(1-\frac{\alpha}{1-p}\right)^{t-\frac{k}{p}} \geq \left(1-\frac{\alpha}{1-p}\right)^{\sqrt{2\frac{1-p}{p}t\log(c_2'/2\theta)}}$$
$$\geq \exp\left(-\sqrt{\frac{2\alpha^2}{p(1-p)}t\log\left(c_2'/(2\theta)\right)}\right)$$
$$\geq \exp\left(-\sqrt{2/c_1'}\log\left(c_2'/\theta\right)\right) = \left(2\theta/c_2'\right)^{\sqrt{2/c_1'}}.$$

We then deduce that

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

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

$$\begin{aligned}\mathbb{P}_1\big(\mathcal{E}_1(z)\big) &\geq \mathbb{P}_1\big(\mathcal{E}(z)\big) = \mathbb{E}_1[\mathbb{1}_{\mathcal{E}}] = \mathbb{E}_0\left(\frac{L_1(W)}{L_0(W)}\mathbb{1}_{\mathcal{E}}\right)\\ &\geq \mathbb{E}_0\big[2\theta/c_2'\mathbb{1}_{\mathcal{E}}\big] = 2\theta/c_2'\mathbb{P}_0\big(\mathcal{E}(z)\big) > \theta/c_2', \end{aligned} \tag{31}$$

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

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. 30, 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. 31, 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. 30. Therefore, the lower bound of Eq. 29 shall hold. □

Based on the result of Lemma 17 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

$$\mathbb{P}_m\big(\big|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)\big| > \varepsilon\big) > \frac{1}{c_2'}\exp\left(\frac{-c_1 t\varepsilon^2}{6\beta^3}\right).$$

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

$$\mathbb{P}_m\big(\big|Q^*_{M_m}(z) - Q^{\mathfrak{A}}_t(z)\big| > \varepsilon\big) > \delta, \tag{32}$$

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 for which $\|Q^*_{M_m} - Q^{\mathfrak{A}}_T\| > \varepsilon$ on either M_0 or M_1:

Lemma 18 *For any $\delta' \in (0, 1/2)$ and any Algorithm $\mathfrak{A}$ using a total number of transition samples less than $T = \frac{N}{6}\xi(\varepsilon, \frac{12\delta'}{N})$, there exists an MDP $M_m \in \mathbb{M}^*$ such that*

$$\mathbb{P}_m\big(\big\|Q^*_{M_m} - Q^{\mathfrak{A}}_T\big\| > \varepsilon\big) > \delta'. \tag{33}$$

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 Eq. 33 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.[9] 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

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

from Eq. 32, thus

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

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

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

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

Lemma 18 implies that if the total number of samples T is less than $\beta^3 N/(c_1\varepsilon^2)\log(N/(c_2\delta))$ then, with the choice of $c_1 = 8100$ and $c_2 = 72$, the probability of $\|Q^*_M - Q^{\mathfrak{A}}_T\| \le \varepsilon$ 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 $T \ge \beta^3 N/(c_1\varepsilon^2)\log(N/(c_2\delta))$, 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

[9] We let T_z to be random.

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 the proof of Theorem 2, we rely on the restrictive assumption that $\varepsilon \leq c\sqrt{\beta/(\gamma|\mathcal{X}|)}$ for some $c > 0$. This assumption restricts the applicability of Theorem 2 in problems with very large number of states. We are not sure whether the above assumption is essential for the result of Theorem 2 or it can be avoided by using a better proof technique. Improving this result, such that the above assumption is not required anymore, can be a subject for future work.

Another direction for future work would be to improve on the state-of-the-art in PAC-MDP, based on the results of this paper. 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 (2012a) have independently proven a similar upper-bound of order $\widetilde{O}(N/(\varepsilon^2(1-\gamma)^3))$ for γ-UCRL algorithm (which is a discounted version of the UCRL algorithm) under the assumption that only two states are accessible from every state-action pair.[10] 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. The difference is that Lattimore and Hutter (2012a) consider the online setting, whereas, in this paper, we assume that a generative model of the MDP is available.

References

Azar, M. G., Munos, R., Ghavamzadeh, M., & Kappen, H. J. (2011a). Reinforcement learning with a near optimal rate of convergence. Tech. rep. http://hal.inria.fr/inria-00636615.

Azar, M. G., Munos, R., Ghavamzadeh, M., & Kappen, H. J. (2011b). Speedy Q-learning. In *Advances in neural information processing systems* (Vol. 24, pp. 2411–2419).

Azar, M. G., Munos, R., Kappen, H. J. (2012). On the sample complexity of reinforcement learning with a generative model. In *ICML*. Omnipress.

Bartlett, P. L., & 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* (pp. 35–42).

Bertsekas, D. P. (2007). *Dynamic programming and optimal control* (Vol. II, 3rd edn.). Belmount: Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont: Athena Scientific.

[10] In the case that more than two states are accessible from every state-action pair, the result of Lattimore and Hutter (2012a) translates to an upper bound of $\widetilde{O}(|\mathcal{X}|^2|\mathcal{A}|/(\varepsilon^2(1-\gamma)^3))$ which has a quadratic dependency on the size of state space $|\mathcal{X}|$, whereas our bounds, at least for small values of ε, scale linearly with $|\mathcal{X}|$ (see also Lattimore and Hutter 2012b).

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York: Cambridge University Press.

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, S. M. (2004). On the sample complexity of reinforcement learning. Ph.D. 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* (Vol. 12, pp. 996–1002). Cambridge: MIT Press.

Lattimore, T., & Hutter, M. (2012a). PAC bounds for discounted MDPs. CoRR arXiv:1202.3890.

Lattimore, T., & Hutter, M. (2012b). PAC bounds for discounted MDPs. In *Algorithmic learning theory* (pp. 320–334). Berlin: Springer.

Mannor, S., & Tsitsiklis, J. N. (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*.

Puterman, M. L. (1994). *Markov decision processes, discrete stochastic dynamic programming*. New York: Wiley.

Singh, S. P., & Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, *16*(3), 227–233.

Sobel, M. J. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability*, *19*, 794–802.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.

Szepesvári, C. (2010). *Algorithms for reinforcement learning*. *Synthesis lectures on artificial intelligence and machine learning*. Morgan & Claypool.

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* (pp. 1031–1038). Omnipress.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., & Weinberger, M. J. (2003). Inequalities for the L1 deviation of the empirical distribution. Tech. rep.

Wiering, M., & van Otterlo, M. (2012). *Reinforcement learning: State-of-the-Art* (pp. 3–39). Berlin: Springer. Chap. 1.