Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model

Gen Li* Tsinghua
Yuting Wei† CMU
Yuejie Chi‡ CMU
Yuantao Gu* Tsinghua
Yuxin Chen§ Princeton

May 2020; Revised: September 2020

Abstract

We investigate the sample efficiency of reinforcement learning in a γ-discounted infinite-horizon Markov decision process (MDP) with state space $\mathcal{S}$ and action space $\mathcal{A}$, assuming access to a generative model. Despite a number of prior work tackling this problem, a complete picture of the trade-offs between sample complexity and statistical accuracy is yet to be determined. In particular, prior results suffer from a sample size barrier, in the sense that their claimed statistical guarantees hold only when the sample size exceeds at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$ (up to some log factor). The current paper overcomes this barrier by certifying the minimax optimality of model-based reinforcement learning as soon as the sample size exceeds the order of $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ (modulo some log factor). More specifically, a *perturbed* model-based planning algorithm provably finds an ε-optimal policy with an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\log\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon}$ samples for any $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Along the way, we derive improved (instance-dependent) guarantees for model-based policy evaluation. To the best of our knowledge, this work provides the first minimax-optimal guarantee in a generative model that accommodates the entire range of sample sizes (beyond which finding a meaningful policy is information theoretically impossible).

Keywords: model-based reinforcement learning, minimaxity, planning, policy evaluation, instance-dependent guarantees, generative model

Contents

*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
†Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
§Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA.

1 Introduction

Reinforcement learning (RL) (Sutton and Barto, 2018; Szepesvári, 2010), which is frequently modeled as learning and decision making in a Markov decision process (MDP), is garnering growing interest in recent years due to its remarkable success in practice. A core objective of RL is to search for a policy — based on a collection of noisy data samples — that approximately maximizes expected rewards in an MDP, without direct access to a precise description of the underlying model.[1] In contemporary applications, it is increasingly more common to encounter environments with prohibitively large state and action space, thus exacerbating the challenge of collecting enough samples to learn the model. To enable faithful policy learning in the sample-starved regime (i.e. the regime where the model complexity overwhelms the sample size), it is crucial to obtain a quantitative picture of the trade-off between sample complexity and statistical accuracy, and to design efficient algorithms that provably achieve the optimal trade-off.

Broadly speaking, there are two common algorithmic approaches: a model-based approach and a model-free one. In the model-based approach, one first learns to describe the unknown model using the data samples in hand, and then leverages the fitted model to perform planning — a task that can be accomplished by resorting to Bellman's principle of optimality (Bellman, 1952; Puterman, 2014). An advantage of model-based algorithms is their flexibility: the learned model can be adapted to perform new ad-hoc tasks without revisiting the data samples. In comparison, the model-free approach attempts to compute the optimal policy (and the optimal value function) without learning the model explicitly, which lends itself to scenarios when a realistic model is difficult to construct or changes on the fly. Characterizing the sample complexities for both approaches has been the focal point of a large body of recent work, e.g. Agarwal et al. (2019); Azar et al. (2013); Jin et al. (2018); Kearns and Singh (1999); Sidford et al. (2018a,b); Tu and Recht (2018); Wainwright (2019a,b); Wang (2019).

In this paper, we pursue a comprehensive understanding of model-based RL, assuming access to a generative model — that is, a simulator that produces samples based on the transition kernel of the MDP for each state-action pair (Kakade, 2003; Kearns and Singh, 1999). To allow for more precise discussions, we focus our attention on an infinite-horizon discounted MDP with state space $\mathcal{S}$, action space $\mathcal{A}$ and discount factor $0 < \gamma < 1$. We obtain N samples per state-action pair by querying the generative model. For an *arbitrary* target accuracy level $\varepsilon > 0$, a desired model-based planning algorithm should return an ε-optimal policy with a minimal number of calls to the generative model. Particular emphasis is placed on the sub-linear sampling scenario, in which the total sample size is smaller than the total number $|\mathcal{S}|^2|\mathcal{A}|$ of model parameters (so that it might be infeasible to estimate the model accurately).

Motivation: a sample size barrier. Several prior work was dedicated to investigating model-based RL with a generative model, which uncovered the minimax optimality of this approach for an already wide

[1] Here and throughout, the "model" refers to the transition kernel and the rewards of the MDP taken collectively.

Algorithm	Sample size range	Sample complexity	ε-range
Phased Q-learning Kearns and Singh (1999)	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical QVI Azar et al. (2013)	$\big[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}]$
Sublinear randomized value iteration Sidford et al. (2018b)	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\big(0, \frac{1}{1-\gamma}\big]$
Variance-reduced QVI Sidford et al. (2018a)	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
Randomized primal-dual method Wang (2019)	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical MDP + planning Agarwal et al. (2019)	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$
Perturbed empirical MDP + planning **This paper**	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{1-\gamma}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$

Table 1: Comparisons with prior results (up to log factors) regarding finding an ε-optimal policy with a generative model. The sample size range and the ε-range stand for the range of sample size and optimality gap (e.g. ε-accuracy) for the claimed sample complexity to hold. Note that the results in (Kearns and Singh, 1999; Wang, 2019) only hold for a restricted family of MDPs satisfying certain ergodicity assumptions. In addition, Azar et al. (2013) (resp. Wainwright (2019b)) showed that empirical QVI (resp. variance-reduced Q-learning) finds an ε-optimal Q-function estimate with sample complexity $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$ ($\varepsilon \in (0,1]$) in a sample size range $[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}, \infty)$, which did not translate directly to an ε-optimal policy.

regime (Agarwal et al., 2019; Azar et al., 2013). However, the results therein often suffered from a sample complexity barrier that prevents us from obtaining a complete trade-off curve between sample complexity and statistical accuracy. For instance, the state-of-the-art result Agarwal et al. (2019) required the total sample size to at least exceed $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}$ (up to some log factor), thus restricting the validity of the theory for broader contexts. In truth, this is not merely an issue for model-based planning; the same barrier already showed up when analyzing the simpler task of model-based policy evaluation (Agarwal et al., 2019; Pananjady and Wainwright, 2019). Furthermore, a similar or even higher barrier emerged in prior theory for model-free methods; for instance, Sidford et al. (2018b); Wainwright (2019b) required the sample size to exceed $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}$ modulo some log factor. In stark contrast, however, no lower bounds developed thus far preclude us from attaining reasonable statistical accuracy when going below the aforementioned sample complexity barriers.

Our contributions. Following the model-based approach, we propose to perform planning based on an empirical MDP learned from samples with mild *reward perturbation*. The perturbed model-based algorithm we propose provably finds an ε-optimal policy with an order of $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$ samples (up to log factor), which matches the minimax lower bound (Azar et al., 2013). Our result accommodates the full range of accuracy level ε (namely, $\varepsilon \in (0, \frac{1}{1-\gamma}]$), thus unveiling the minimaxity of our algorithm as soon as the sample size exceeds $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{1-\gamma}$ (modulo some log factor); encouragingly, this covers the *full* range of sample sizes that enable one to find a policy strictly better than a random guess. See Table 1 for detailed comparisons with prior literature. Along the way, we also derive (instance-dependent) statistical guarantees for policy evaluation, which strengthens state-of-the-art results by broadening the sample size range.

Our theory is established upon a novel combination of several key ideas: (1) a high-order expansion of the estimation error for value functions, coupled with fine-grained analysis for each term in the expansion; (2) the construction of auxiliary leave-one-out type (state-action-absorbing) MDPs — motivated by Agarwal et al. (2019) — that help decouple the complicated statistical dependency between the empirically optimal policy (as opposed to value functions) and data samples; (3) a tie-breaking argument guaranteeing that the empirically optimal policy stands out from all other policies under reward perturbation.

2 Problem formulation

2.1 Models and background

Basics of Markov decision processes. Consider an infinite-horizon discounted MDP represented by a quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ denotes a finite set of states, $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$ is a finite set of actions, $\gamma \in (0, 1)$ stands for the discount factor, and $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ represents the reward function, namely, $r(s, a)$ is the immediate reward received upon executing action a while in state s (here and throughout, we consider the normalized setting where the rewards lie within $[0, 1]$). In addition, $P : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ represents the probability transition kernel of the MDP, where $P(s'|s, a)$ denotes the probability of transiting from state s to state s' when action a is executed, and $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$.

A deterministic policy is a mapping $\pi : \mathcal{S} \mapsto \mathcal{A}$ that maps a state to an action. The value function $V^\pi : \mathcal{S} \mapsto \mathbb{R}$ of a policy π is defined by

$$\forall s \in \mathcal{S} : \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s^t, a^t) \,\middle|\, s^0 = s\right], \tag{1}$$

which is the expected discounted total reward starting from the initial state $s^0 = s$, with the actions taken according to the policy π (namely, $a^t = \pi(s^t)$ for all $t \geq 0$) and the trajectory generated based on the transition kernel (namely, $s^{t+1} \sim P(\cdot|s^t, a^t)$). It is easily seen that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$. The corresponding action-value function (or Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ of a policy π is defined by

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \qquad Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s^t, a^t) \,\middle|\, s^0 = s, a^0 = a\right], \tag{2}$$

where the actions are taken according to the policy π after the initial action (i.e. $a^t = \pi(s^t)$ for all $t \geq 1$). It is well-known that there exists an optimal policy, denoted by $\pi^\star$, that simultaneously maximizes $V^\pi(s)$ (resp. $Q^\pi(s, a)$) for all states $s \in \mathcal{S}$ (resp. state-action pairs $(s, a) \in (\mathcal{S} \times \mathcal{A})$) (Sutton and Barto, 2018). The corresponding value function $V^\star := V^{\pi^\star}$ (resp. action-value function $Q^\star := Q^{\pi^\star}$) is called the optimal value function (resp. optimal action-value function).

A generative model and an empirical MDP. The current paper focuses on a stylized generative model (also called a simulator) as studied in Kakade (2003); Kearns et al. (2002). Assuming access to this generative model, we collect N independent samples $s^i_{s,a} \sim P(\cdot|s, a)$, $i = 1, \ldots, N$ for any state-action pair (s, a), which allows us to construct an empirical transition kernel $\widehat{P}$ given by

$$\forall s' \in \mathcal{S}, \qquad \widehat{P}(s'|s, a) = \frac{1}{N} \sum\nolimits_{i=1}^{N} \mathbb{1}\{s^i_{s,a} = s'\}, \tag{3}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. In words, $\widehat{P}(s'|s, a)$ counts the empirical frequency of transitions from (s, a) to state s'. The total sample size should be understood as $N^{\mathsf{total}} := N|\mathcal{S}||\mathcal{A}|$. This leads to an empirical MDP $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$ constructed from the data samples. We can define the value function and the action-value function of a policy π for $\widehat{\mathcal{M}}$ analogously, which we shall denote by $\widehat{V}^\pi$ and $\widehat{Q}^\pi$, respectively. The optimal policy of $\widehat{\mathcal{M}}$ is denoted by $\widehat{\pi}^\star$, with the optimal value function and Q-function denoted by $\widehat{V}^\star := \widehat{V}^{\widehat{\pi}^\star}$ and $\widehat{Q}^\star := \widehat{Q}^{\widehat{\pi}^\star}$, respectively.

Policy evaluation and planning. Given a few data samples in hand, the task of policy evaluation aims to compute or approximate the value function V^π under a given policy π. To be precise, for any target level $\varepsilon > 0$, the goal is to find an ε-accurate estimate V^π_{est} such that

$$\text{(policy evaluation)} \qquad \forall s \in \mathcal{S} : \qquad \left|V^\pi_{\mathsf{est}}(s) - V^\pi(s)\right| \leq \varepsilon. \tag{4}$$

In contrast, the task of planning seeks to identify a policy that (approximately) maximizes the expected discounted reward given the data samples. Specifically, for any target level $\varepsilon > 0$, the aim is to compute an ε-accurate policy π_{est} obeying

$$\text{(planning)} \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad V^{\pi_{\mathsf{est}}}(s) \geq V^{\star}(s) - \varepsilon, \quad Q^{\pi_{\mathsf{est}}}(s,a) \geq Q^{\star}(s,a) - \varepsilon. \tag{5}$$

Naturally, one would hope to accomplish these tasks with as few samples as possible. Recall that for the normalized reward setting with $0 \leq r \leq 1$, the value function and Q-function fall within the range $[0, \frac{1}{1-\gamma}]$; this means that the range of the target level ε should be set to $\varepsilon \in [0, \frac{1}{1-\gamma}]$. The model-based approach typically starts by constructing an empirical MDP $\widehat{\mathcal{M}}$ based on all collected samples, and then "plugs in" this empirical model directly into the Bellman recursion to perform policy evaluation or planning, with prominent examples including Q-value iteration (QVI) and policy iteration (PI) (Bertsekas, 2017).

2.2 Notation

Let $\mathcal{X} := \big(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \frac{1}{\varepsilon}\big)$. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ means there exists a universal constant $C_1 > 0$ such that $f \leq C_1 g$, whereas the notation $f(\mathcal{X}) = \Omega(g(\mathcal{X}))$ means $g(\mathcal{X}) = O(f(\mathcal{X}))$. In addition, the notation $\widetilde{O}(\cdot)$ (resp. $\widetilde{\Omega}(\cdot)$) is defined in the same way as $O(\cdot)$ (resp. $\Omega(\cdot)$) except that it ignores logarithmic factors.

For any vector $\boldsymbol{a} = [a_i]_{1\leq i\leq n} \in \mathbb{R}^n$, we overload the notation $\sqrt{\cdot}$ and $|\cdot|$ in an entry-wise manner such that $\sqrt{\boldsymbol{a}} := [\sqrt{a_i}]_{1\leq i\leq n}$ and $|\boldsymbol{a}| := [|a_i|]_{1\leq i\leq n}$. For any vectors $\boldsymbol{a} = [a_i]_{1\leq i\leq n}$ and $\boldsymbol{b} = [b_i]_{1\leq i\leq n}$, the notation $\boldsymbol{a} \geq \boldsymbol{b}$ (resp. $\boldsymbol{a} \leq \boldsymbol{b}$) means $a_i \geq b_i$ (resp. $a_i \leq b_i$) for all $1 \leq i \leq n$, and we let $\boldsymbol{a} \circ \boldsymbol{b} := [a_i b_i]_{1\leq i\leq n}$ represent the Hadamard product. Additionally, we denote by $\mathbf{1}$ the all-one vector, and $\boldsymbol{I}$ the identity matrix. For any matrix $\boldsymbol{A}$, we define the norm $\|\boldsymbol{A}\|_1 := \max_i \sum_j |A_{i,j}|$.

3 Main results

3.1 Theoretical guarantees for model-based reinforcement learning

As summarized in Table 1, the theory of all prior work required the sample size per state-action pair to at least exceed $N \geq \Omega\big(\frac{1}{(1-\gamma)^2}\big)$. In order to break this sample size barrier, we propose to invoke the model-based planning approach applied to an empirical MDP with *perturbed rewards*.

Specifically, for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, let us randomly perturb the immediate reward to obtain

$$r_{\mathrm{p}}(s,a) = r(s,a) + \zeta(s,a), \qquad \zeta(s,a) \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(0,\xi), \tag{6}$$

where $\mathsf{Unif}(0,\xi)$ denotes the uniform distribution between 0 and some parameter $\xi > 0$ (to be specified momentarily).[2] For any policy π, we denote by $\widehat{V}_{\mathrm{p}}^{\pi}$ the corresponding value function of the perturbed empirical MDP $\widehat{\mathcal{M}}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r_{\mathrm{p}}, \gamma)$ with the probability transition kernel $\widehat{P}$ and the perturbed reward function r_{p}. Let $\widehat{\pi}_{\mathrm{p}}^{\star}$ represent the optimal policy of $\widehat{\mathcal{M}}_{\mathrm{p}}$, i.e.

$$\widehat{\pi}_{\mathrm{p}}^{\star} := \arg\max_{\pi} \widehat{V}_{\mathrm{p}}^{\pi}. \tag{7}$$

Encouragingly, this policy results in a value function $V^{\widehat{\pi}_{\mathrm{p}}^{\star}}$ (resp. Q-function $Q^{\widehat{\pi}_{\mathrm{p}}^{\star}}$) that well approximates the true optimal value function $V^{\star}$ (resp. optimal Q-function $Q^{\star}$), as asserted by the following theorem.

Theorem 1 (Perturbed model-based planning)**.** *There exist some universal constants $c_0, c_1 > 0$ such that: for any $\delta > 0$ and any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the policy $\widehat{\pi}_{\mathrm{p}}^{\star}$ defined in (7) obeys*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad V^{\widehat{\pi}_{\mathrm{p}}^{\star}}(s) \geq V^{\star}(s) - \varepsilon \qquad \text{and} \qquad Q^{\widehat{\pi}_{\mathrm{p}}^{\star}}(s,a) \geq Q^{\star}(s,a) - \gamma\varepsilon \tag{8}$$

with probability at least $1-\delta$, provided that the perturbation size is $\xi = \frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ and that the sample size per state-action pair exceeds

$$N \geq \frac{c_0 \log\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta}\big)}{(1-\gamma)^3\varepsilon^2}. \tag{9}$$

[2] Note that perturbation is only invoked when running the planning algorithms and does not require collecting new samples.

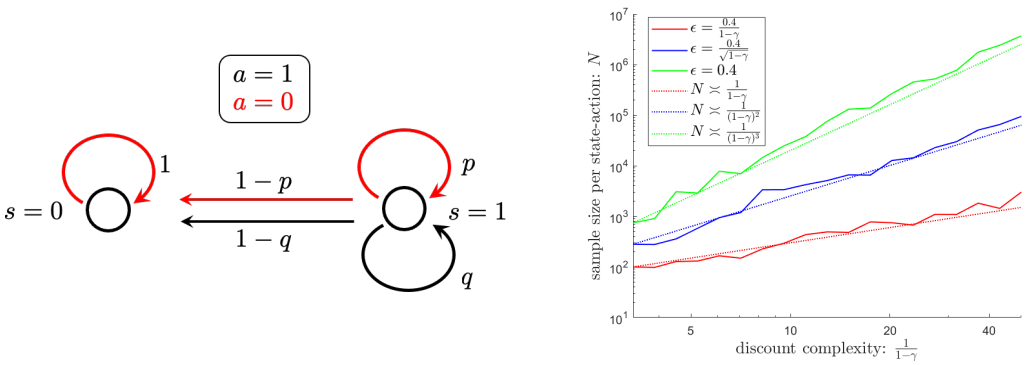

Figure 1: Numerical sample complexity per state-action pair N vs. discount complexity $\frac{1}{1-\gamma}$.

In addition, both the empirical QVI and PI algorithms w.r.t. $\widehat{\mathcal{M}}_{\mathrm{p}}$ *(cf. (Azar et al., 2013, Algorithms 1-2)) are able to recover* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *perfectly within* $O\big(\frac{1}{1-\gamma}\log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta})\big)$ *iterations.*

Remark 1. Theorem 1 holds unchanged if ξ is taken to be $\frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^{\alpha}|\mathcal{A}|^{\alpha}}$ for any $\alpha \geq 1$. The current paper picks the specific choice $\alpha = 5$ merely to convey that a very small level of perturbation suffices for our purpose.

Remark 2. Perturbation brings a side benefit: one can recover the optimal policy $\widehat{\pi}_{\mathrm{p}}^{\star}$ of the perturbed empirical MDP $\widehat{\mathcal{M}}_{\mathrm{p}}$ *exactly* in a small number of iterations without incurring further ptimization errors. To give a flavor of the overall computational complexity, let us take QVI for example Azar et al. (2013). Recall that each iteration of QVI takes time proportional to the time taken to read $\widehat{P}$ (which is a matrix with at most $N|\mathcal{S}||\mathcal{A}|$ nonzeros), hence the resulting computational complexity can be as low as $O\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\log^2(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta})\big)$.

The above theorem demonstrates that: the perturbed model-based planning approach finds an ε-optimal policy as soon as the total sample complexity exceeds the order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ (modulo some log factor). It is worth emphasizing that, compared to prior literature, our result imposes no restriction on the range of ε and, in particular, we allow the accuracy level ε to go all the way up to $\frac{1}{1-\gamma}$. Our result is particularly useful in the regime with small-to-moderate sample sizes, since its validity is guaranteed as long as

$$N \geq \widetilde{\Omega}\Big(\frac{1}{1-\gamma}\Big). \tag{10}$$

Tackling the sample-limited regime (in particular, the scenario when $N \in [\frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2}]$) requires us to develop new analysis frameworks beyond prior theory, which we shall discuss in detail momentarily.

We remark that the work Azar et al. (2013) established a minimax lower bound of the same order as (9) (up to some log factor) in the regime $\varepsilon = O(1)$. A closer inspection of their analysis, however, reveals that their argument and bound hold true as long as $\varepsilon = O(\frac{1}{1-\gamma})$. This in turn corroborates the *minimax optimality* of our perturbed model-based approach for the full ε-range (which is previously unavailable), and demonstrates the information-theoretic infeasibility to learn a policy strictly better than a random guess if $N \leq \widetilde{O}(\frac{1}{1-\gamma})$. Put another way, the condition (10) contains the full range of "meaningful" sample sizes.

To further demonstrate the effectiveness of our model-based approach, we conduct a series of numerical experiments (motivated by Azar et al. (2013)), focusing on the effect of the discount complexity $\frac{1}{1-\gamma}$ upon the sample complexity. More specifically, consider the following MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \{0, 1\}$, $\mathcal{A}_0 = \{0\}, \mathcal{A}_1 = \{0, 1\}$, $P(0|0, 0) = 1, P(1|1, 0) = p, P(1|1, 1) = q$ and $r(0, 0) = 0, r(1, 0) = r(1, 1) = 1$ (see Fig. 1 for an illustration), and we take the quantities p and q to be $p = \gamma + \frac{2\gamma(1-\gamma)^2\varepsilon}{(1+\gamma)^2}$ and $q = \gamma - \frac{2\gamma(1-\gamma)^2\varepsilon}{(1+\gamma)^2}$ as suggested by Azar et al. (2013). As depicted in Fig. 1, the numerical sample complexity per state-action pair N scales on the order of $\frac{1}{(1-\gamma)^3\varepsilon^2}$ for varying choices of ε, which is consistent with our theory.

Finally, we single out an intermediate result in the analysis of Theorem 1 concerning model-based policy evaluation, which might be of interest on its own. Specifically, for any fixed policy π independent of the data, this task concerns value function estimation via the plug-in estimate $\widehat{V}^{\pi}$ (i.e. the value function of

the empirical $\mathcal{M}$ under this policy). However simple as this might seem, existing theoretical underpinnings of this approach remain suboptimal, unless the sample size is already sufficiently large. Our result is the following, which does not require enforcing reward perturbation.

Theorem 2 (Model-based policy evaluation)**.** *Fix any policy π. There exists some universal constant $c_0 > 0$ such that: for any $0 < \delta < 1$ and any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, one has*

$$\forall s \in \mathcal{S}: \qquad \left|\widehat{V}^{\pi}(s) - V^{\pi}(s)\right| \leq \varepsilon \tag{11}$$

with probability at least $1-\delta$, provided that the sample size per state-action pair exceeds

$$N \geq c_0 \frac{\log\left(\frac{|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^3\varepsilon^2}. \tag{12}$$

In words, this theorem reveals that $\widehat{V}^{\pi}$ begins to outperform a random guess as soon as $N \geq \widetilde{\Omega}\big(\frac{1}{1-\gamma}\big)$. The sample complexity bound (12) enjoys *full* coverage of the ε-range $(0, \frac{1}{1-\gamma}]$, and matches the minimax lower bound derived in (Pananjady and Wainwright, 2019, Theorem 2(b)) up to only a $\log\log\frac{1}{1-\gamma}$ factor. In addition, a recent line of work investigated instance-dependent guarantees for policy evaluation (Khamaru et al. (2020); Pananjady and Wainwright (2019)). While this is not our focus, our analysis does uncover an instance-dependent bound with a broadened sample size range. See Lemma 1 and the discussion thereafter.

3.2 Comparisons with prior work and implications

In order to discuss the novelty of our results in context, we take a moment to compare them with prior theory. See Table 1 for a more complete list of comparisons.

Prior bounds for planning and policy learning. None of the prior results with a generative model (including both model-based or model-free approaches) was capable of efficiently finding the desired policy while accommodating the full sample size range (10). For instance, the state-of-the-art analysis for the model-based approach Agarwal et al. (2019) required the sample size to at least exceed

$$N \geq \widetilde{\Omega}\Big(\frac{1}{(1-\gamma)^2}\Big), \tag{13}$$

whereas the theory for the variance-reduced model-free approach Sidford et al. (2018a); Wainwright (2019b) imposed the sample size requirement

$$N \geq \widetilde{\Omega}\Big(\frac{1}{(1-\gamma)^3}\Big). \tag{14}$$

In fact, it was previously unknown what is achievable in the sample size range $N \in [\frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2}]$. In contrast, our results confirm the minimax-optimal statistical performance of the model-based approach with full coverage of the ε-range and the sample size range.

Remark 3. We briefly point out why the sample size barrier (13) appeared in the analysis of Agarwal et al. (2019). Take (Agarwal et al., 2019, Section 4.3) for example: the contraction factor $\gamma\sqrt{\frac{8\log(|\mathcal{S}||\mathcal{A}|/(1-\gamma)\delta)}{N}}\frac{1}{1-\gamma}$ therein needs to be smaller than 1, thereby requiring $N \geq \widetilde{\Omega}\big((1-\gamma)^{-2}\big)$.

Prior bounds for policy evaluation. Regarding value function estimation for any fixed policy π, the prior results Agarwal et al. (2019); Azar et al. (2013); Pananjady and Wainwright (2019) for the plug-in approach all operated under the assumption that $N \geq \widetilde{\Omega}\big(\frac{1}{(1-\gamma)^2}\big)$, which is more stringent than our result by a factor of at least $\frac{1}{1-\gamma}$. In addition, our sample complexity matches the state-of-the-art guarantees in the regime where $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (Agarwal et al., 2019; Pananjady and Wainwright, 2019), while extending them to the range $\varepsilon \in \big[\frac{1}{\sqrt{1-\gamma}}, \frac{1}{1-\gamma}\big]$ uncovered in these previous papers.

4 Other related work

Classical analyses of reinforcement learning algorithms have largely focused on asymptotic performance (e.g. Jaakkola et al. (1994); Szepesvári (1998); Tsitsiklis and Van Roy (1997); Tsitsiklis (1994)). Leveraging the toolkit of concentration inequalities, a number of recent papers have shifted attention towards understanding the performance in the non-asymptotic and finite-time settings. A highly incomplete list includes Azar et al. (2017); Beck and Srikant (2012); Bhandari et al. (2018); Bradtke and Barto (1996); Cai et al. (2019); Chen et al. (2020); Dalal et al. (2018); Even-Dar and Mansour (2003); Fan et al. (2019); Gupta et al. (2019); Jin et al. (2018); Kaledin et al. (2020); Kearns and Singh (1999); Khamaru et al. (2020); Lakshminarayanan and Szepesvari (2018); Li et al. (2020); Mou et al. (2020); Qu and Wierman (2020); Shah and Xie (2018); Sidford et al. (2018a); Srikant and Ying (2019); Strehl et al. (2006); Wainwright (2019b); Xu and Gu (2020); Xu et al. (2019), a large fraction of which is concerned with model-free algorithms.

The generative model (or simulator) adopted in this paper was first proposed in Kearns and Singh (1999), which has been invoked in Agarwal et al. (2019); Azar et al. (2012, 2013); Kakade (2003); Kearns et al. (2002); Kearns and Singh (1999); Khamaru et al. (2020); Lattimore and Hutter (2012); Pananjady and Wainwright (2019); Sidford et al. (2018a,b); Wainwright (2019b); Wang (2019); Yang and Wang (2019), to name just a few. In particular, Azar et al. (2013) developed the minimax lower bound on the sample complexity $N = \Omega\big(\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^3\varepsilon^2}\big)$ necessary for finding an ε-optimal policy, and showed that, for any $\varepsilon \in (0,1)$, a model-based approach (e.g. applying QVI or PI to the empirical MDP) can estimate the optimal Q-function to within an ε-accuracy given near-minimal samples. Note, however, that directly translating this result to the policy guarantees leads to an additional factor of $\frac{1}{1-\gamma}$ in estimation accuracy and of $\frac{1}{(1-\gamma)^2}$ in sample complexity. In light of this, Azar et al. (2013) further showed that a near-optimal sample complexity is possible for policy learning if the sample size is at least on the order of $\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^2}$ which, however, is no longer sub-linear in the model complexity. A recent breakthrough Agarwal et al. (2019) substantially improved the model-based guarantee with the aid of auxiliary state-absorbing MDPs, extending the range of sample complexity to $[\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2}, \infty)$. Our analysis is motivated in part by Agarwal et al. (2019), but also relies on several other novel techniques to complete the picture.

Finally, we remark that the construction of state-absorbing MDPs or state-action-absorbing MDPs falls under the category of "leave-one-out" type analysis, which is particularly effective in decoupling complicated statistical dependency in various statistical estimation problems Agarwal et al. (2019); Chen et al. (2019); El Karoui (2015); Ma et al. (2020); Pananjady and Wainwright (2019). The application of such an analysis framework to MDPs should be attributed to Agarwal et al. (2019). Other applications to Markov chains include Chen et al. (2019); Pananjady and Wainwright (2019).

5 Analysis

This section presents the key ideas for proving our main results, following an introduction of some convenient matrix notation.

5.1 Matrix notation and Bellman equations

It is convenient to present our proof based on some matrix notation for MDPs. Denoting by $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$ the standard basis vectors, we can define:

- $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$: a vector representing the reward function r (so that $r_{(s,a)} = r(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$).
- $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$: a vector representing the value function V^π (so that $V^\pi_s = V^\pi(s)$ for all $s \in \mathcal{S}$).
- $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$: a vector representing the Q-function Q^π (so that $Q^\pi_{(s,a)} = Q^\pi(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$).
- $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$ and $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$: representing the optimal value function $V^\star$ and optimal Q-function $Q^\star$.
- $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$: a matrix representing the probability transition kernel P, where the (s,a)-th row of $\boldsymbol{P}$ is a probability vector representing $P(\cdot|s,a)$. Denote $\boldsymbol{P}_{s,a}$ as the (s,a)-th row of the transition matrix $\boldsymbol{P}$.

- $\boldsymbol{\Pi}^\pi \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$: a projection matrix associated with a given policy π taking the following form

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

- $\boldsymbol{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$: two *square* probability transition matrices induced by the policy π over the state-action pairs and the states respectively, defined by

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \qquad \text{and} \qquad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}. \tag{16}$$

- $\boldsymbol{r}_\pi \in \mathbb{R}^{|\mathcal{S}|}$: a reward vector restricted to the actions chosen by the policy π, namely, $r_\pi(s) = r(s,\pi(s))$ for all $s \in \mathcal{S}$ (or simply, $\boldsymbol{r}_\pi = \boldsymbol{\Pi}^\pi \boldsymbol{r}$).

Armed with the above matrix notation, we can write, for any policy π, the *Bellman consistency equation* as

$$\boldsymbol{Q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}^\pi \boldsymbol{Q}^\pi, \tag{17}$$

which implies that

$$\boldsymbol{Q}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}^\pi)^{-1}\boldsymbol{r}; \tag{18}$$

$$\boldsymbol{V}^\pi = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi \boldsymbol{V}^\pi \qquad \text{and} \qquad \boldsymbol{V}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi. \tag{19}$$

For a vector $\boldsymbol{V} = [V_i]_{1\le i\le|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$, we define the vector $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ whose entries are given by

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \qquad \big[\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V})\big]_{(s,a)} := \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{s'}^2 - \Big(\sum_{s'\in\mathcal{S}} P(s'|s,a)V_{s'}\Big)^2,$$

i.e. the variance of $\boldsymbol{V}$ w.r.t. $P(\cdot|s,a)$. This can be expressed using our matrix notation as follows

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

Similarly, for any given policy π we define

$$\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V}) = \boldsymbol{\Pi}^\pi \mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_\pi\boldsymbol{V})\circ(\boldsymbol{P}_\pi\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}|}.$$

We shall also define $\widehat{\boldsymbol{V}}^\pi$, $\widehat{\boldsymbol{Q}}^\pi$, $\widehat{\boldsymbol{V}}^\star$, $\widehat{\boldsymbol{Q}}^\star$, $\widehat{\boldsymbol{P}}$, $\widehat{\boldsymbol{P}}^\pi$, $\widehat{\boldsymbol{P}}_\pi$, $\mathsf{Var}_{\widehat{\boldsymbol{P}}}(\boldsymbol{V})$, $\mathsf{Var}_{\widehat{\boldsymbol{P}}_\pi}(\boldsymbol{V})$ w.r.t. the empirical MDP $\widehat{\mathcal{M}}$ in an analogous fashion.

5.2 Analysis: model-based policy evaluation

We start with the simpler task of policy evaluation, which also plays a crucial role in the analysis of planning. To establish our guarantees in Theorem 2, we aim to prove the following result. Here, we recall that the true value function under a policy π and the model-based empirical estimate are given respectively by

$$\boldsymbol{V}^\pi = (\boldsymbol{I} - \gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi \qquad \text{and} \qquad \widehat{\boldsymbol{V}}^\pi = (\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}_\pi. \tag{20}$$

Lemma 1. *Fix any policy π. Consider any $0<\delta<1$, and suppose $N \ge \frac{32e^2}{1-\gamma}\log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)$. Then with probability at least $1-\delta$, the vectors defined in (20) obey*

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\big\|_\infty &\le 4\gamma\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)}{N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^\pi\big]}\Big\|_\infty + \frac{2\gamma\log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)}{(1-\gamma)N}\big\|\boldsymbol{V}^\pi\big\|_\infty \\ &\le 6\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)}{N(1-\gamma)^3}}.\end{aligned} \tag{21}$$

Proof. The key proof idea is to resort to a high-order successive expansion of $\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}$, followed by fine-grained analysis of each term up to a certain logarithmic order. See Appendix B.1. □

Clearly, Theorem 2 is a straightforward consequence of Lemma 1. Further, we strengthen the result by providing an additional instance-dependent bound (see the first line of (21) that depends on the true instance $\boldsymbol{P}_{\pi}, \boldsymbol{V}^{\pi}$), which is often tighter than the worst-case bound stated in the second line of (21). Our contribution can be better understood when compared with Pananjady and Wainwright (2019). Assuming that there is no noise in the rewards, our instance-dependent guarantee matches Pananjady and Wainwright (2019, Theorem 1(a)) up to some $\log\log\frac{1}{1-\gamma}$ factor, while being capable of covering the full sample size range $N \geq \widetilde{\Omega}(\frac{1}{1-\gamma})$. In contrast, Pananjady and Wainwright (2019, Theorem 1) is only valid when $N \geq \widetilde{\Omega}(\frac{1}{(1-\gamma)^2})$.

Proof ideas. We now briefly and informally describe the key proof ideas. As a starting point, the elementary identities (20) allow us to obtain

$$\begin{aligned}\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi} &= \big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\boldsymbol{r}_{\pi} - \boldsymbol{V}^{\pi} \\ &= \big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\big(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi} - \big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)\boldsymbol{V}^{\pi} \\ &= \gamma\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}. \end{aligned} \tag{22}$$

Due to the complicated dependency between $(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}$ and $\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}$, a natural strategy is to control these two terms separately and then to combine bounds; see Agarwal et al. (2019, Lemma 5) for an introduction. This simple approach, however, leads to sub-optimal statistical guarantees.

In order to refine the statistical analysis, we propose to further expand (22) in a similar way to deduce

$$\begin{aligned}(22) &= \gamma\big(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi} + \gamma\left\{\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1} - \big(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi}\big)^{-1}\right\}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi} \\ &= \gamma\big(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi} + \gamma^2\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\big(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}, \end{aligned} \tag{23}$$

where the last line holds due to the same reason as (22) (basically it can be seen by replacing $\boldsymbol{r}_{\pi}$ with $\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}$ in (22)). This can be viewed as a "second-order" expansion, with (22) being a "first-order" counterpart. The advantage is that: the first term in (23) becomes easier to cope with than its counterpart (22), owing to the independence between $(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi})^{-1}$ and $\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}$. However, the second term in (23) remains difficult to control optimally. To remedy this issue, we shall continue to expand it to higher order (up to some logarithmic order), which eventually allows for optimal control of the estimation error.

Another crucial issue is that: in order to obtain fine-grained analyses on each term in the expansion (except for the first-order term), a common approach is to combine the Bernstein inequality with a classical entrywise bound on a quantity taking the form $(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}(\boldsymbol{V})}$ (which dates back to Azar et al. (2013)). Such a classical bound in prior literature, however, is not sufficiently tight for our purpose, which calls for refinement; see Lemma 8. Details are deferred to Appendix B.1.

5.3 Analysis: model-based planning

This subsection moves on to establishing our theory for model-based planning (cf. Theorem 1) and outlines the key ideas. In what follows, we shall start by analyzing the unperturbed version, which will elucidate the role of reward perturbation in our analysis.

We first make note of the following elementary decomposition:

$$\begin{aligned}\boldsymbol{V}^{\star} - \boldsymbol{V}^{\widehat{\pi}^{\star}} &= \big(\widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}} - \boldsymbol{V}^{\widehat{\pi}^{\star}}\big) + \big(\widehat{\boldsymbol{V}}^{\pi^{\star}} - \widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}\big) + \big(\boldsymbol{V}^{\star} - \widehat{\boldsymbol{V}}^{\pi^{\star}}\big) \\ &\leq \big(\widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}} - \boldsymbol{V}^{\widehat{\pi}^{\star}}\big) + \big(\boldsymbol{V}^{\pi^{\star}} - \widehat{\boldsymbol{V}}^{\pi^{\star}}\big), \end{aligned} \tag{24}$$

where the inequality follows from the optimality of $\widehat{\pi}^{\star}$ w.r.t. $\widehat{\boldsymbol{V}}$ (so that $\widehat{\boldsymbol{V}}^{\pi^{\star}} \leq \widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}$) and the definition $\boldsymbol{V}^{\star} = \boldsymbol{V}^{\pi^{\star}}$. This leaves us with two terms to control.

Step 1: bounding $\|\boldsymbol{V}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\pi^\star}\|_\infty$. Given that $\pi^\star$ is independent of the data, we can carry out this step using Lemma 1. Specifically, taking $\pi = \pi^\star$ in Lemma 1 yields that, with probability at least $1-\delta$,

$$\left\|\widehat{\boldsymbol{V}}^{\pi^\star} - \boldsymbol{V}^{\pi^\star}\right\|_\infty \le 6\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{N(1-\gamma)^3}}. \tag{25}$$

Step 2: bounding $\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\|_\infty$. Extending the result in Step 1 to $\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\|_\infty$ is considerably more challenging, primarily due to the complicated statistical dependency between $(\boldsymbol{V}^{\widehat{\pi}^\star}, \widehat{\boldsymbol{V}}^{\widehat{\pi}^\star})$ and the data matrix $\widehat{\boldsymbol{P}}$. The recent work Agarwal et al. (2019) developed a clever "leave-one-out" type argument by constructing some auxiliary state-absorbing MDPs to decouple the statistical dependency when $\varepsilon < 1/\sqrt{1-\gamma}$. However, their argument falls short of accommodating the full range of ε. To address this challenge, our analysis consists of the following two steps, both of which require new ideas beyond Agarwal et al. (2019).

- *Decoupling statistical dependency between $\widehat{\pi}^\star$ and $\widehat{\boldsymbol{P}}$.* Instead of attempting to decouple the statistical dependency between $\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star}$ and $\widehat{\boldsymbol{P}}$ as in Agarwal et al. (2019), we focus on decoupling the statistical dependency between the policy $\widehat{\pi}^\star$ and $\widehat{\boldsymbol{P}}$. If this can be achieved, then the proof strategy adopted in Step 1 for a fixed policy becomes applicable (see Section 5.3.1). A key ingredient of this step lies in the construction of a collection of auxiliary state-action-absorbing MDPs (motivated by Agarwal et al. (2019)), which allows us to get hold of $\|\boldsymbol{V}^{\widehat{\pi}^\star} - \widehat{\boldsymbol{V}}^{\widehat{\pi}^\star}\|_\infty$. See Section 5.3.2 for details, with a formal bound delivered in Lemma 5.
- *Tie-breaking via reward perturbation.* A shortcoming of the above-mentioned approach, however, is that it relies crucially on the separability of $\widehat{\pi}^\star$ from other policies; in other words, the proof might fail if $\widehat{\pi}^\star$ is non-unique or not sufficiently distinguishable from others. Consequently, it remains to ensure that the optimal policy $\widehat{\pi}^\star$ stands out from all the rest for all MDPs of interest. As it turns out, this can be guaranteed with high probability by slightly perturbing the reward function so as to break the ties. See Section 5.3.3 for details.

In the sequel, we shall flesh out these key ideas.

5.3.1 Value function estimation for a policy obeying Bernstein-type conditions

Before discussing how to decouple statistical dependency, we record a useful result that plays an important role in the analysis. Specifically, Lemma 1 can be generalized beyond the family of fixed policies (namely, those independent of $\widehat{\boldsymbol{P}}$), as long as a certain Bernstein-type condition — to be formalized in (27) — is satisfied. To make it precise, we need to introduce a set of auxiliary vectors as follows

$$\begin{aligned} \boldsymbol{r}^{(0)} &:= \boldsymbol{r}_\pi, & \boldsymbol{V}^{(0)} &:= (\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(0)}, \\ \boldsymbol{r}^{(l)} &:= \sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(l-1)}\right]}, & \boldsymbol{V}^{(l)} &:= (\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(l)}, \qquad l \ge 1. \end{aligned} \tag{26}$$

Our generalization of Lemma 1 is as follows, which does *not* require statistical independence between the policy π and the data $\widehat{\boldsymbol{P}}$. Here, we remind the reader of the notation $|\boldsymbol{z}| := [|z_1|, \cdots, |z_n|]^\top$ and $\sqrt{\boldsymbol{z}} := [\sqrt{z_1}, \cdots, \sqrt{z_n}]^\top$ for any vector $\boldsymbol{z} \in \mathbb{R}^n$.

Lemma 2. *Suppose that there exists some quantity $\beta_1 > 0$ such that $\{\boldsymbol{V}^{(l)}\}$ (cf. (26)) obeys*

$$\left|(\widehat{\boldsymbol{P}}_\pi - \boldsymbol{P}_\pi)\boldsymbol{V}^{(l)}\right| \le \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(l)}\right]} + \frac{\beta_1\left\|\boldsymbol{V}^{(l)}\right\|_\infty}{N}\boldsymbol{1}, \qquad \text{for all } 0 \le l \le \log\left(\frac{e}{1-\gamma}\right). \tag{27}$$

Suppose that $N > \frac{16e^2}{1-\gamma}\beta_1$. Then the vectors $\boldsymbol{V}^\pi = (\boldsymbol{I} - \gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi$ and $\widehat{\boldsymbol{V}}^\pi = (\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}_\pi$ satisfy

$$\left\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\right\|_\infty \le \frac{6}{1-\gamma}\sqrt{\frac{\beta_1}{N(1-\gamma)}}. \tag{28}$$

While the Bernstein-type condition (27) clearly holds for some reasonably small β_1 if π is independent of $\widehat{\boldsymbol{P}}$, it might remain valid if π exhibits fairly "weak" statistical dependency on the data samples. This is a key step that paves the way for our subsequent analysis of $\widehat{\pi}^\star$.

5.3.2 Decoupling statistical dependency via (s,a)-absorbing MDPs

We are now positioned to demonstrate how to control $\big\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\big\|_\infty$ w.r.t. the optimal policy $\widehat{\pi}^\star$ to $\widehat{\boldsymbol{V}}$. A crucial technical challenge lies in how to decouple the complicated statistical dependency between the optimal policy $\widehat{\pi}^\star$ and the $\widehat{\boldsymbol{V}}^\star$ (which heavily relies on the data samples). Towards this, we resort to a leave-one-row-out argument built upon a collection of auxiliary MDPs, largely motivated by the novel construction in (Agarwal et al., 2019, Section 4.2). In comparison to Agarwal et al. (2019) that introduces state-absorbing MDPs (so that a state s is absorbing regardless of the subsequent actions chosen), our construction is a set of state-action-absorbing MDPs, in which a state s is absorbing only when a designated action a is always executed at the state s.

Construction of (s,a)-absorbing MDPs. For each state-action pair (s,a) and each scalar u with $|u| \le 1/(1-\gamma)$, we construct an auxiliary MDP $\mathcal{M}_{s,a,u}$ — it is identical to the original $\mathcal{M}$ except that it is absorbing in state s if we always choose action a in state s. More specifically, the probability transition kernel associated with $\mathcal{M}_{s,a,u}$ (denoted by $P_{\mathcal{M}_{s,a,u}}$) can be specified by

$$\begin{aligned} P_{\mathcal{M}_{s,a,u}}(s \mid s,a) &= 1, \\ P_{\mathcal{M}_{s,a,u}}(s' \mid s,a) &= 0, && \text{for all } s' \neq s, \\ P_{\mathcal{M}_{s,a,u}}(\cdot \mid s',a') &= P_{\mathcal{M}}(\cdot \mid s',a'), && \text{for all } (s',a') \neq (s,a), \end{aligned} \tag{29}$$

where $P_{\mathcal{M}}$ is the probability transition kernel w.r.t. the original $\mathcal{M}$. Meanwhile, the instant reward received at (s,a) in $\mathcal{M}_{s,a,u}$ is set to be u, while the rewards at all other state-action pairs stay unchanged. We can define $\widehat{\mathcal{M}}_{s,a,u}$ analogously (so that its probability transition matrix is identical to $\widehat{\boldsymbol{P}}$ except that the (s,a)-th row becomes absorbing). The main advantage of this construction is that: for any fixed u, the MDP $\widehat{\mathcal{M}}_{s,a,u}$ is statistically independent of $\widehat{\boldsymbol{P}}_{s,a}$ (the row of $\widehat{\boldsymbol{P}}$ corresponding to the state-action pair (s,a), determined by the samples collected for the (s,a) pair).

To streamline notation, we let $\boldsymbol{Q}^\pi_{s,a,u}$ represent the Q-function of $\mathcal{M}_{s,a,u}$ under a policy π, denote by $\pi^\star_{s,a,u}$ the optimal policy associated with $\mathcal{M}_{s,a,u}$, and let $\boldsymbol{Q}^\star_{s,a,u}$ be the Q-function under this optimal policy $\pi^\star_{s,a,u}$. The notation $\boldsymbol{V}^\pi_{s,a,u}$ and $\boldsymbol{V}^\star_{s,a,u}$ regarding value functions, as well as their counterparts (i.e. $\widehat{\boldsymbol{Q}}^\pi_{s,a,u}$, $\widehat{\boldsymbol{Q}}^\star_{s,a,u}$, $\widehat{\boldsymbol{V}}^\pi_{s,a,u}$, $\widehat{\boldsymbol{V}}^\star_{s,a,u}$, $\widehat{\pi}^\star_{s,a,u}$) in the empirical MDP $\widehat{\mathcal{M}}$, can be defined in an analogous fashion.

Remark 4. The careful reader will remark that the instant reward u is constrained to reside within $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$ rather than the usual range $[0,1]$. Fortunately, none of the subsequent steps that involve u requires u to lie within $[0,1]$.

Intimate connections between the auxiliary MDPs and the original MDP. In the following, we introduce a result that connects the Q-function and the value function of the absorbing MDP with those of the original MDP. The idea is motivated by Agarwal et al. (2019, Lemma 7).

Lemma 3. *Setting* $u^\star := r(s,a) + \gamma(\boldsymbol{P}\boldsymbol{V}^\star)_{s,a} - \gamma V^\star(s)$, *one has*

$$\boldsymbol{Q}^\star_{s,a,u^\star} = \boldsymbol{Q}^\star \qquad \text{and} \qquad \boldsymbol{V}^\star_{s,a,u^\star} = \boldsymbol{V}^\star. \tag{30}$$

Proof. See Appendix B.2. □

Remark 5. Lemma 3 does not rely on the particular form of $\boldsymbol{P}$, and can be directly generalized to the empirical model $\widehat{\boldsymbol{P}}$ and the auxiliary MDPs built upon $\widehat{\boldsymbol{P}}$.

In words, by properly setting the instant reward $u = u^\star$ (which can be easily shown to reside within $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$), one guarantees that the (s,a)-absorbing MDP and the original MDP have the same Q-function and value function under the respective optimal policies.

Representing $\widehat{\pi}^\star$ via a small set of policies independent of $\widehat{\boldsymbol{P}}_{s,a}$. With Lemma 3 in place, it is tempting to use $\widehat{\mathcal{M}}_{s,a,\widehat{u}^\star}$ with $\widehat{u}^\star := r(s,a) + \gamma(\widehat{\boldsymbol{P}}\widehat{\boldsymbol{V}}^\star)_{s,a} - \gamma\widehat{V}^\star(s)$ to replace the original $\widehat{\mathcal{M}}$. The rationale is simple: given that the probability transition matrix of $\widehat{\mathcal{M}}_{s,a,\widehat{u}^\star}$ does not rely upon $\widehat{\boldsymbol{P}}_{s,a}$, the statistical

dependency between $\widehat{\mathcal{M}}_{s,a,\widehat{u}^\star}$ and $\widehat{\boldsymbol{P}}_{s,a}$ is now fully embedded into a single parameter $\widehat{u}^\star$. This motivates us to decouple the statistical dependency effectively by constructing an epsilon-net (see, e.g., Vershynin (2018)) w.r.t. this single parameter. The aim is to locate a point u_0 over a small fixed set such that *(i)* it is close to $\widehat{u}^\star$, and *(ii)* its associated optimal policy is identical to the original optimal policy $\widehat{\pi}^\star$.

It turns out that this aim can be accomplished as long as the original Q-function $\widehat{\boldsymbol{Q}}^\star$ satisfies a sort of separation condition (which indicates that there is no tie when it comes to the optimal policy). To make it precise, given any $0 < \omega < 1$, our separation condition is characterized through the following event

$$\mathcal{B}_\omega := \Big\{ \widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a: a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s,a) \geq \omega \quad \text{for all } s \in \mathcal{S} \Big\}. \tag{31}$$

Clearly, on the event $\mathcal{B}_\omega$, the optimal policy $\widehat{\pi}^\star$ is unique, since for each s the action $\widehat{\pi}^\star(s)$ results in a strictly higher Q-value compared to any other action. With this separation condition in mind, our result is stated below. Here and throughout, we define an epsilon-net of the interval $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$ as follows

$$\mathcal{N}_\epsilon := \big\{ -n_\epsilon \epsilon, \ldots, -\epsilon, 0, \epsilon, \ldots, n_\epsilon \epsilon \big\}, \qquad \text{for the largest integer } n_\epsilon \text{ obeying } n_\epsilon \epsilon < \frac{1}{1-\gamma}, \tag{32}$$

which has cardinality at most $\frac{2}{(1-\gamma)\epsilon}$.

Lemma 4. *Consider any $\omega > 0$, and suppose the event $\mathcal{B}_\omega$ (cf. (31)) holds. Then for any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$, such that*

$$\widehat{\pi}^\star = \widehat{\pi}^\star_{s,a,u_0}. \tag{33}$$

Proof. See Appendix B.3. □

Deriving an optimal error bound under the separation condition. Armed with the above bounds, we are ready to derive the desired error bound by combining Lemma 2 and Lemma 4.

Lemma 5. *Given $0 < \omega < 1$ and $\delta > 0$, suppose that $\mathcal{B}_\omega$ (defined in (31)) occurs with probability at least $1 - \delta$. Then with probability at least $1 - 3\delta$,*

$$\big\| \widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star} \big\|_\infty \leq 6 \sqrt{\frac{2 \log\big(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \omega \delta}\big)}{N(1-\gamma)^3}} \quad \text{and} \quad \boldsymbol{V}^\star - \boldsymbol{V}^{\widehat{\pi}^\star} \leq 12 \sqrt{\frac{2 \log\big(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \omega \delta}\big)}{N(1-\gamma)^3}}\, \mathbf{1}, \tag{34}$$

provided that $N \geq \frac{c_0 \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\omega}\right)}{1-\gamma}$ for some sufficiently large constant $c_0 > 0$.

Proof. See Appendix B.4. □

5.3.3 A tie-breaking argument

Unfortunately, the separation condition specified in $\mathcal{B}_\omega$ (cf. (31)) does not always hold. In order to accommodate all possible MDPs of interest without imposing such a special separation condition, we put forward a perturbation argument allowing one to generate a new MDP that *(i)* satisfies the separation condition, and that *(ii)* is sufficiently close to the original MDP.

Specifically, let us represent the proposed reward perturbation (6) in a vector form as follows

$$\boldsymbol{r}_{\mathrm{p}} := \boldsymbol{r} + \boldsymbol{\zeta}, \tag{35}$$

where $\boldsymbol{\zeta} = [\zeta(s,a)]_{(s,a) \in |\mathcal{S}| \times |\mathcal{A}|}$ is an $|\mathcal{S}||\mathcal{A}|$-dimensional vector composed of independent entries with each $\zeta(s,a) \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(0, \xi)$. We aim to show that: by randomly perturbing the reward function, we can "break the tie" in the Q-function and ensure sufficient separation of Q-values associated with different actions.

To formalize our result, we find it convenient to introduce additional notation. Denote by $\pi^\star_{\mathrm{p}}$ the optimal policy of the MDP $\mathcal{M}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \boldsymbol{r}_{\mathrm{p}}, \gamma)$, and $Q^\star_{\mathrm{p}}$ its optimal state-action value function. We can define $\widehat{Q}^\star_{\mathrm{p}}$ and $\widehat{\pi}^\star_{\mathrm{p}}$ analogously for the MDP $\widehat{\mathcal{M}}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \widehat{\boldsymbol{P}}, \boldsymbol{r}_{\mathrm{p}}, \gamma)$. Our result is phrased as follows.

Lemma 6. *Consider the perturbed reward vector defined in expression (35). With probability at least* $1-\delta$,

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A} \text{ with } a\neq\pi_{\mathrm{p}}^{\star}(s): \qquad Q_{\mathrm{p}}^{\star}(s,\pi_{\mathrm{p}}^{\star}(s))-Q_{\mathrm{p}}^{\star}(s,a)>\frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2}. \tag{36}$$

This result holds unchanged if $(Q_{\mathrm{p}}^{\star},\pi_{\mathrm{p}}^{\star})$ *is replaced by* $(\widehat{Q}_{\mathrm{p}}^{\star},\widehat{\pi}_{\mathrm{p}}^{\star})$.

Proof. See Appendix B.5. □

Lemma 6 reveals that at least a polynomially small degree of separation ($\omega=\frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2}$) arises upon random perturbation (with size ξ) of the reward function. As we shall see momentarily, this level of separation suffices for our purpose.

5.3.4 Proof of Theorem 1

Let us consider the randomly perturbed reward function as in (35). For any policy π, we denote by $\boldsymbol{V}_{\mathrm{p}}^{\pi}$ (resp. $\widehat{\boldsymbol{V}}_{\mathrm{p}}^{\pi}$) the corresponding value function vector in the MDP with probability transition matrix $\boldsymbol{P}$ (resp. $\widehat{\boldsymbol{P}}$) and reward vector $\boldsymbol{r}_{\mathrm{p}}$. Note that $\pi_{\mathrm{p}}^{\star}$ (resp. $\widehat{\pi}_{\mathrm{p}}^{\star}$) denotes the optimal policy that maximizes $\boldsymbol{V}_{\mathrm{p}}^{\pi}$ (resp. $\widehat{\boldsymbol{V}}_{\mathrm{p}}^{\pi}$).

In view of Lemma 6, with probability at least $1-\delta$ one has the separation

$$\left|\widehat{Q}_{\mathrm{p}}^{\star}(s,\widehat{\pi}_{\mathrm{p}}^{\star}(s))-\widehat{Q}_{\mathrm{p}}^{\star}(s,a)\right|>\frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2} \tag{37}$$

uniformly over all s and $a\neq\widehat{\pi}_{\mathrm{p}}^{\star}(s)$. With this separation in place, taking $\omega:=\frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2}$ in Lemma 5 yields

$$\left\|\boldsymbol{V}_{\mathrm{p}}^{\pi_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}\right\|_{\infty}\leq 12\sqrt{\frac{2\log\left(\frac{96|\mathcal{S}|^2|\mathcal{A}|^3}{(1-\gamma)^4\xi\delta^2}\right)}{N(1-\gamma)^3}}. \tag{38}$$

In addition, the value functions under any policy π obeys

$$\boldsymbol{V}^{\pi}-\boldsymbol{V}_{\mathrm{p}}^{\pi}=\boldsymbol{\Pi}^{\pi}\Big((\boldsymbol{I}-\boldsymbol{P}^{\pi})^{-1}\boldsymbol{r}-(\boldsymbol{I}-\boldsymbol{P}^{\pi})^{-1}\boldsymbol{r}_{\mathrm{p}}\Big),$$

which taken collectively with the facts $\|\boldsymbol{r}-\boldsymbol{r}_{\mathrm{p}}\|_{\infty}\leq\xi$ and $\|(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi})^{-1}\|_1\leq\frac{1}{1-\gamma}$ gives

$$\left\|\boldsymbol{V}^{\pi}-\boldsymbol{V}_{\mathrm{p}}^{\pi}\right\|_{\infty}\leq\|(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi})^{-1}\|_1\|\boldsymbol{r}-\boldsymbol{r}_{\mathrm{p}}\|_{\infty}\leq\frac{1}{1-\gamma}\xi.$$

Specializing the above relation to $\pi^{\star}$ and $\widehat{\pi}_{\mathrm{p}}^{\star}$ gives

$$\left\|\boldsymbol{V}^{\pi^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\pi^{\star}}\right\|_{\infty}\leq\frac{1}{1-\gamma}\xi \qquad\text{and}\qquad \left\|\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}\right\|_{\infty}\leq\frac{1}{1-\gamma}\xi. \tag{39}$$

Now let us consider the following decomposition

$$\begin{aligned}\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}^{\star}&=\big(\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}\big)+\big(\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\pi_{\mathrm{p}}^{\star}}\big)+\big(\boldsymbol{V}_{\mathrm{p}}^{\pi_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\pi^{\star}}\big)+\big(\boldsymbol{V}_{\mathrm{p}}^{\pi^{\star}}-\boldsymbol{V}^{\star}\big)\\&\geq\big(\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}\big)+\big(\boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}_{\mathrm{p}}^{\pi_{\mathrm{p}}^{\star}}\big)+\big(\boldsymbol{V}_{\mathrm{p}}^{\pi^{\star}}-\boldsymbol{V}^{\star}\big),\end{aligned}$$

where the last step follows from the optimality of $\pi_{\mathrm{p}}^{\star}$ w.r.t. $\boldsymbol{V}_{\mathrm{p}}$. Taking this collectively with the inequalities (38) and (39), one shows that with probability greater than $1-3\delta$,

$$\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^{\star}}-\boldsymbol{V}^{\star}\geq-\left(\frac{2}{1-\gamma}\xi+12\sqrt{\frac{2\log\left(\frac{96|\mathcal{S}|^2|\mathcal{A}|^3}{\xi(1-\gamma)^4\delta^2}\right)}{N(1-\gamma)^3}}\right)\mathbf{1}$$

By taking $\xi = \frac{(1-\gamma)\varepsilon}{3|\mathcal{S}|^5|\mathcal{A}|^5}$ and $N \geq \frac{c_0 \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\varepsilon}\right)}{(1-\gamma)^3\varepsilon^2}$ for some constant $c_0 > 0$ large enough, we can ensure that $\mathbf{0} \geq \boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{V}^\star \geq -\varepsilon\mathbf{1}$ as claimed. Regarding the Q-functions, the Bellman equation gives

$$\boldsymbol{Q}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^\star} - \left(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star\right) = \gamma \boldsymbol{P}(\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{V}^\star).$$

Consequently, one has

$$\boldsymbol{Q}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{Q}^\star \geq -\left(\gamma \|\boldsymbol{P}\|_1 \|\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{V}^\star\|_\infty\right)\mathbf{1} \geq -\gamma\varepsilon\mathbf{1}.$$

Finally, we demonstrate that both the empirical QVI and PI w.r.t. $\widehat{\mathcal{M}}_{\mathrm{p}}$ are guaranteed to find $\widehat{\pi}_{\mathrm{p}}^\star$ in a few iterations. Suppose for the moment that we can obtain a policy π_k obeying

$$\|\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k} - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\|_\infty < \frac{\xi\delta(1-\gamma)}{8|\mathcal{S}||\mathcal{A}|^2}. \tag{40}$$

Then for any $s \in \mathcal{S}$ and any action $a \neq \widehat{\pi}_{\mathrm{p}}^\star(s)$ one has

$$\begin{aligned}
&\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k}\left(s, \widehat{\pi}_{\mathrm{p}}^\star(s)\right) - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k}(s,a) \\
&\qquad = \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\left(s, \widehat{\pi}_{\mathrm{p}}^\star(s)\right) - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star(s,a) + \left(\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k}\left(s, \widehat{\pi}_{\mathrm{p}}^\star(s)\right) - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\left(s, \widehat{\pi}_{\mathrm{p}}^\star(s)\right)\right) - \left(\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k}(s,a) - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star(s,a)\right) \\
&\qquad \geq \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\left(s, \widehat{\pi}_{\mathrm{p}}^\star(s)\right) - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star(s,a) - 2\left\|\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k} - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\right\|_\infty \\
&\qquad > \frac{\xi\delta(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2} - 2 \cdot \frac{\xi\delta(1-\gamma)}{8|\mathcal{S}||\mathcal{A}|^2} = 0,
\end{aligned}$$

where the last line results from (37) and (40). In other words, we can perfectly recover the policy $\widehat{\pi}_{\mathrm{p}}^\star$ from the estimate $\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k}$, provided that (40) is satisfied. In addition, it has been shown that (Azar et al., 2013, Lemma 2) the greedy policy induced by k-th iteration of both algorithms — denoted by π_k — satisfies $\left\|\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k} - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\right\|_\infty \leq \frac{2\gamma^{k+1}}{(1-\gamma)^2}$. Taking $\xi = \frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ and $k = \frac{c_2}{1-\gamma}\log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta})$ for some constant $c_2 > 0$ large enough, one guarantees that π_k satisfies (40), which in turn ensures perfect recovery of $\widehat{\pi}_{\mathrm{p}}^\star$.

6 Discussion

This paper demonstrates that a *perturbed* model-based planning algorithm achieves the minimax sample complexity in a generative model, as soon as the sample size exceeds the order of $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ (modulo some log factor). Compared to prior literature, our result considerably broadens the sample size range, allowing us to pin down a complete trade-off curve between sample complexities and statistical accuracy.

The present work opens up several directions for future investigation, which we discuss in passing below.

- *Is perturbation necessary?* The planning algorithm analyzed here is applied to a perturbed variant of the original empirical MDP. This, however, gives rise to a natural question regarding the necessity of perturbation: can we achieve optimal planning performance directly using the original empirical MDP without perturbation? While we conjecture that the answer is affirmative, settling this conjecture requires developing new analysis techniques beyond the scope of this paper.

- *Improved analysis for model-free algorithms.* As mentioned previously, sample complexity barriers exist in prior theory for model-free approaches (e.g. Sidford et al. (2018a); Wainwright (2019b)). Our analysis might shed light on how to break such barriers for model-free approaches.

- *Episodic finite-horizon MDPs.* This work concentrates on infinite-horizon discounted MDPs, and we expect that the insights analysis can be extended to study the finite-horizon setting as well, where the samples are drawn in an episodic fashion (Dann and Brunskill (2015); Jiang and Agarwal (2018); Wang et al. (2020)). We leave this extension to future work.

- *Beyond the tabular setting.* This current paper focuses on the tabular setting with finite state and action spaces. While we improve the sample size range, the sample complexities might still be prohibitive when $|\mathcal{S}|$ and $|\mathcal{A}|$ are enormous. Therefore, it is desirable to further investigate settings where function approximation is employed to improve the efficiency (e.g. Duan and Wang (2020); Jin et al. (2019)).

Acknowledgements

Y. Wei is supported in part by the grants NSF CCF-2007911 and DMS-2015447. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, and NSF CCF-1806154 and CCF-2007911. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, DMS-2014279 and IIS-1900140, and the Princeton SEAS Innovation Award. We thank Qiwen Cui for pointing out an issue in Appendix B.5 in an early version of this paper, and thank Shicong Cen, Chen Cheng and Cong Ma for numerous discussions about reinforcement learning.

A Preliminary facts

We begin by recording a few elementary facts about $\boldsymbol{P}^{\pi}$ and $\boldsymbol{P}_{\pi}$ (see definitions in (16)). These are standard results and we omit the proofs for brevity.

Lemma 7. *For any policy π, any probability transition matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ and any $0<\gamma<1$, one has*

(a) $(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}=\sum_{i=0}^{\infty}(\gamma\boldsymbol{P}_{\pi})^{i}$;

(b) All entries of the matrix $(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}$ are non-negative;

(c) $\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\|_{1}\le 1/(1-\gamma)$;

(d) $(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\mathbf{1}=\mathbf{1}$;

(e) For any non-negative vectors $\mathbf{0}\le\boldsymbol{r}_1\le\boldsymbol{r}_2$ of compatible dimension, one has $\mathbf{0}\le(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_1\le(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_2$.

The above results continue to hold if $\boldsymbol{P}_{\pi}$ is replaced by $\boldsymbol{P}^{\pi}$.

B Proofs of auxiliary lemmas

B.1 Proofs of Lemma 1 and Lemma 2

Auxiliary notation and preliminaries. Before proceeding, we define several $|\mathcal{S}|$-dimensional auxiliary vectors $\boldsymbol{r}^{(i)},\boldsymbol{V}^{(i)},\widehat{\boldsymbol{V}}^{(i)}$ $(1\le i\le m)$ recursively as follows

$$\begin{array}{lll}
\boldsymbol{r}^{(0)}:=\boldsymbol{r}_{\pi}, & \boldsymbol{V}^{(0)}:=(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}^{(0)}, & \widehat{\boldsymbol{V}}^{(0)}:=(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}^{(0)},\\
\boldsymbol{r}^{(1)}:=\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\big[\boldsymbol{V}^{(0)}\big]}, & \boldsymbol{V}^{(1)}:=(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}^{(1)}, & \widehat{\boldsymbol{V}}^{(2)}:=(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}^{(1)},\\
\quad\vdots & \quad\vdots & \quad\vdots\\
\boldsymbol{r}^{(m)}:=\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\big[\boldsymbol{V}^{(m-1)}\big]}, & \boldsymbol{V}^{(m)}:=(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}^{(m)}, & \widehat{\boldsymbol{V}}^{(m)}:=(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}^{(m)},
\end{array}\tag{41}$$

where m will be specified momentarily.

A crucial quantity that appears repeatedly in analyzing the above terms is $\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}(\boldsymbol{V})}\|_{\infty}$, whose importance was already made apparent in the work Azar et al. (2013). A widely used upper bound on this quantity, originally due to (Azar et al., 2013, Lemma 8), is given by

$$\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}(\boldsymbol{V})}\Big\|_{\infty}\le\frac{2\log 2}{\gamma(1-\gamma)^{1.5}}\,\|\boldsymbol{r}\|_{\infty}\,.\tag{42}$$

This bound turns out to be loose for our purpose, and we develop an improved bound as follows, whose proof is deferred to Appendix B.1.1.

Lemma 8. *Consider any policy π and any probability transition matrix $\boldsymbol{P}\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$. Let $\boldsymbol{V}$ be a vector obeying $\boldsymbol{V}=(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_{\pi}$ for some $|\mathcal{S}|$-dimensional vector $\boldsymbol{r}_{\pi}\ge\mathbf{0}$. For any $0<\gamma<1$, one has*

$$\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}(\boldsymbol{V})}\Big\|_{\infty}\le\frac{4}{\gamma\sqrt{1-\gamma}}\,\|\boldsymbol{V}\|_{\infty}\,.\tag{43}$$

Remark 6. In comparison to the bound (42) derived in (Azar et al., 2013, Lemma 8), Lemma 8 offers an improved upper bound stated directly in terms of the properties of $\boldsymbol{V}$ rather than those of $\boldsymbol{r}$.

As it turns out, Lemma 8 allows us to obtain an entrywise bound for $\boldsymbol{V}^{(l)}$ $(1 \le l \le m)$. To begin with, the first term $\boldsymbol{V}^{(1)}$ satisfies

$$\big\|\boldsymbol{V}^{(1)}\big\|_\infty = \Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(0)}\big]}\Big\|_\infty \tag{44}$$

since $\boldsymbol{r}^{(1)} = \sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(0)}\big]}$. Next, for any $l > 1$ one has

$$\begin{aligned}\big\|\boldsymbol{V}^{(l)}\big\|_\infty &= \big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(l)}\big\|_\infty = \Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(l-1)}\big]}\Big\|_\infty \\ &\le \frac{4}{\gamma\sqrt{1-\gamma}}\big\|\boldsymbol{V}^{(l-1)}\big\|_\infty,\end{aligned}$$

where the second identity results from the definition of $\boldsymbol{r}^{(l)}$, and the last inequality comes from Lemma 8. As a consequence, applying this inequality recursively gives

$$\big\|\boldsymbol{V}^{(l)}\big\|_\infty \le \Big(\frac{4}{\gamma\sqrt{1-\gamma}}\Big)^{l-1}\big\|\boldsymbol{V}^{(1)}\big\|_\infty. \tag{45}$$

Main proof. Equipped with the above facts, we are now in a position to prove the lemmas, for which we start with the more general one — Lemma 2. Consider any $0 \le l \le m$. We first observe that

$$\begin{aligned}\widehat{\boldsymbol{V}}^{(l)} - \boldsymbol{V}^{(l)} &= (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}^{(l)} - \boldsymbol{V}^{(l)} \\ &= \big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)} - \big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)\boldsymbol{V}^{(l)} \\ &= \gamma\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)},\end{aligned} \tag{46}$$

where the second line follows since, by definition, $\big(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)} = \boldsymbol{r}^{(l)}$. Suppose that there exists some quantity $\beta_1 > 0$ such that the following condition

$$\Big|\big(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)}\Big| \le \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(l)}\big]} + \frac{\big\|\boldsymbol{V}^{(l)}\big\|_\infty\beta_1}{N}\mathbf{1} \tag{47}$$

holds uniformly for all $0 \le l \le m$. Then this combined with (46) gives

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(l)} - \boldsymbol{V}^{(l)}\big\|_\infty &= \gamma\Big\|\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)}\Big\|_\infty \\ &\overset{\text{(i)}}{\le} \gamma\Big\|\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\Big|\big(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\big)\boldsymbol{V}^{(l)}\Big|\Big\|_\infty \\ &\overset{\text{(ii)}}{\le} \gamma\sqrt{\frac{\beta_1}{N}}\Big\|\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(l)}\big]}\Big\|_\infty + \frac{\gamma\big\|\boldsymbol{V}^{(l)}\big\|_\infty\beta_1}{N}\Big\|\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\mathbf{1}\Big\|_\infty.\end{aligned}$$

Here, (i) follows since $\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}$ is a non-negative matrix, (ii) comes from (47) and the triangle inequality. Recalling the definition of $\boldsymbol{r}^{(l)}$ and $\widehat{\boldsymbol{V}}^{(l)}$ and invoking Lemma 7(d), we can further bound the above as

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(l)} - \boldsymbol{V}^{(l)}\big\|_\infty &\le \gamma\sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}^{(l+1)}\big\|_\infty + \frac{\gamma\big\|\boldsymbol{V}^{(l)}\big\|_\infty\beta_1}{(1-\gamma)N} \\ &\le \gamma\sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}^{(l+1)} - \boldsymbol{V}^{(l+1)}\big\|_\infty + \gamma\sqrt{\frac{\beta_1}{N}}\big\|\boldsymbol{V}^{(l+1)}\big\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(l)}\big\|_\infty,\end{aligned} \tag{48}$$

where the last inequality is a consequence of the triangle inequality.

The above inequality (48) provides a recursive relation that in turn allows for an effective upper bound. Specifically, combining the inequalities (45) and (48) leads to

$$\big\|\widehat{\boldsymbol{V}}^{(0)} - \boldsymbol{V}^{(0)}\big\|_\infty \le \gamma\sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}^{(1)} - \boldsymbol{V}^{(1)}\big\|_\infty + \gamma\sqrt{\frac{\beta_1}{N}}\big\|\boldsymbol{V}^{(1)}\big\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty$$

$$=: b_1\big\|\widehat{\boldsymbol{V}}^{(1)}-\boldsymbol{V}^{(1)}\big\|_\infty+b_1\big\|\boldsymbol{V}^{(1)}\big\|_\infty+\frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty,$$

and for $l\geq 1$,

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(l)}-\boldsymbol{V}^{(l)}\big\|_\infty &\leq \gamma\sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}^{(l+1)}-\boldsymbol{V}^{(l+1)}\big\|_\infty+\left(4\sqrt{\frac{\beta_1}{(1-\gamma)N}}+\frac{\gamma\beta_1}{(1-\gamma)N}\right)\left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^{l-1}\|\boldsymbol{V}^{(1)}\|_\infty\\ &=: b_1\big\|\widehat{\boldsymbol{V}}^{(l+1)}-\boldsymbol{V}^{(l+1)}\big\|_\infty+b_2b_3^{l-1}\|\boldsymbol{V}^{(1)}\|_\infty.\end{aligned}$$

Here for notational simplicity, we introduce

$$b_1:=\gamma\sqrt{\frac{\beta_1}{N}},\qquad b_2:=4\sqrt{\frac{\beta_1}{(1-\gamma)N}}+\frac{\gamma\beta_1}{(1-\gamma)N},\qquad b_3:=\frac{4}{\gamma\sqrt{1-\gamma}}.$$

Invoking the above recursive relation, we can arrange terms to reach

$$\big\|\widehat{\boldsymbol{V}}^{(0)}-\boldsymbol{V}^{(0)}\big\|_\infty\leq\underbrace{b_1^m\big\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\big\|_\infty}_{=:\alpha_1}+\underbrace{\left(b_1b_2\sum_{l=0}^{m-2}(b_1b_3)^l+b_1\right)}_{=:\alpha_2}\big\|\boldsymbol{V}^{(1)}\big\|_\infty+\frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty. \tag{49}$$

Controlling the quantity α_2. Now it suffices to control the two terms on the right-hand side of the inequality (49) separately, towards which we shall start with the quantity α_2. Assuming that $N\geq 64\beta_1/(1-\gamma)$, one can easily verify $b_1b_3\leq 1/2$. The summation of the geometric sequence thus gives

$$\begin{aligned}\alpha_2:=b_1b_2\sum_{l=0}^{m-2}(b_1b_3)^l+b_1&\leq\frac{b_1b_2}{1-b_1b_3}+b_1\leq 2b_1b_2+b_1\\ &=\gamma\sqrt{\frac{\beta_1}{N}}\left\{1+8\sqrt{\frac{\beta_1}{(1-\gamma)N}}+\frac{2\gamma\beta_1}{(1-\gamma)N}\right\}\leq 3\gamma\sqrt{\frac{\beta_1}{N}},\end{aligned} \tag{50}$$

where the last step holds with the assumption $N\geq\frac{64\beta_1}{1-\gamma}$.

Controlling the quantity α_1. Next, we proceed to the quantity α_1, which requires the control of $\big\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\big\|_\infty$. In view of the identity (46), we obtain

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\big\|_\infty&=\gamma\big\|(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}(\boldsymbol{P}_\pi-\widehat{\boldsymbol{P}}_\pi)\boldsymbol{V}^{(m)}\big\|_\infty\\ &\leq\gamma\left\|(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\left(\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]}+\frac{\big\|\boldsymbol{V}^{(m)}\big\|_\infty\beta_1}{N}\mathbf{1}\right)\right\|_\infty,\end{aligned}$$

where the last inequality follows from the Bernstein-type condition (47) and the fact that $\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}$ has non-negative entries. By virtue of the simple relation $\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]}\leq\big\|\boldsymbol{V}^{(m)}\big\|_\infty$ and the fact that $\big\|\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big\|_1\leq\frac{1}{1-\gamma}$ (cf. Lemma 7(c)), it is further guaranteed that

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\big\|_\infty&\leq\gamma\|(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\|_1\cdot\left\|\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]}+\frac{\big\|\boldsymbol{V}^{(m)}\big\|_\infty\beta_1}{N}\mathbf{1}\right\|_\infty,\\ &\leq\frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}}+\frac{\beta_1}{N}\right)\big\|\boldsymbol{V}^{(m)}\big\|_\infty,\end{aligned}$$

which combined with the bound (45) yields

$$\big\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\big\|_\infty\leq\frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}}+\frac{\beta_1}{N}\right)\left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^{m-1}\|\boldsymbol{V}^{(1)}\|_\infty.$$

Putting the above bounds together yields

$$\begin{aligned}\alpha_1 := b_1^m \big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty &\le \left(\gamma\sqrt{\frac{\beta_1}{N}}\right)^m \frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right)\Big(\frac{4}{\gamma\sqrt{1-\gamma}}\Big)^{m-1}\|\boldsymbol{V}^{(1)}\|_\infty \\ &= \left(\sqrt{\frac{16\beta_1}{N(1-\gamma)}}\right)^{m-1}\left(\sqrt{\frac{\beta_1}{N}} + 1\right)\frac{\gamma^2\beta_1}{(1-\gamma)N}\|\boldsymbol{V}^{(1)}\|_\infty \\ &\le \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\|\boldsymbol{V}^{(1)}\|_\infty, \end{aligned} \tag{51}$$

where the last inequality holds provided that $N > \frac{16e^2}{1-\gamma}\beta_1$.

Putting all this together. Combining the inequalities (49), (50) and (51) gives

$$\big\|\widehat{\boldsymbol{V}}^{(0)} - \boldsymbol{V}^{(0)}\big\|_\infty \le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty + \left\{3\gamma\sqrt{\frac{\beta_1}{N}} + \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\right\}\|\boldsymbol{V}^{(1)}\|_\infty.$$

To finish up, set $m = \log(\frac{e}{1-\gamma})$ and assume that $N > \frac{16e^2}{1-\gamma}\beta_1$. Recognizing that $\boldsymbol{V}^{(0)} = \boldsymbol{V}^\pi$ and $\widehat{\boldsymbol{V}}^{(0)} = \widehat{\boldsymbol{V}}^\pi$, we arrive at

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\big\|_\infty = \big\|\widehat{\boldsymbol{V}}^{(0)} - \boldsymbol{V}^{(0)}\big\|_\infty &\le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty + \left\{3\gamma\sqrt{\frac{\beta_1}{N}} + \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\right\}\|\boldsymbol{V}^{(1)}\|_\infty \\ &\le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^\pi\big\|_\infty + 4\gamma\sqrt{\frac{\beta_1}{N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^\pi\big]}\Big\|_\infty, \end{aligned} \tag{52}$$

provided that $N \ge \frac{16e^2\beta_1}{1-\gamma}$.

Proof of Lemma 2. Invoking the inequality (42) to bound the second term of (52), we reach

$$\big\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\big\|_\infty \le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^\pi\big\|_\infty + 8\log 2\sqrt{\frac{\beta_1}{N(1-\gamma)^3}}\|\boldsymbol{r}\|_\infty \le 6\sqrt{\frac{\beta_1}{N(1-\gamma)^3}}\|\boldsymbol{r}\|_\infty, \tag{53}$$

where the last inequality uses the elementary fact that $\big\|\boldsymbol{V}^\pi\big\|_\infty \le \frac{1}{1-\gamma}\|\boldsymbol{r}\|_\infty$ and the assumption that $N > \frac{16e^2}{1-\gamma}\beta_1$. We complete the proof of Lemma 2.

Proof of Lemma 1. Finally, to establish Lemma 1, we observe that: for any fixed policy π, the vector $\boldsymbol{V}^{(l)}$ is independent of $\widehat{\boldsymbol{P}}_\pi$. The Bernstein inequality (e.g. (Agarwal et al., 2019, Lemma 6)) then reveals that with probability at least $1-\delta$,

$$\Big|(\widehat{\boldsymbol{P}}_\pi - \boldsymbol{P}_\pi)\boldsymbol{V}^{(l)}\Big| \le \sqrt{\frac{2\log\big(\frac{4m|\mathcal{S}|}{\delta}\big)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(l)}\big]} + \frac{\big\|\boldsymbol{V}^{(l)}\big\|_\infty\log\big(\frac{4m|\mathcal{S}|}{\delta}\big)}{N}\mathbf{1} \tag{54}$$

holds uniformly for all $0 \le l \le m$. This means that we can take $\beta_1 := 2\log\big(\frac{4m|\mathcal{S}|}{\delta}\big)$ with $m = \log(\frac{e}{1-\gamma})$ for this case. Combining this with the inequality (52), we derive the advertised instance-dependent bound

$$\big\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\big\|_\infty \le \frac{2\gamma\log\big(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\big)}{(1-\gamma)N}\big\|\boldsymbol{V}^\pi\big\|_\infty + 4\gamma\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\big)}{(1-\gamma)N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^\pi\big]}\Big\|_\infty.$$

Further, this taken collectively with (42) and the crude bound $\|\boldsymbol{V}^\pi\|_\infty \le \frac{1}{1-\gamma}\|\boldsymbol{r}\|_\infty$ gives

$$\big\|\widehat{\boldsymbol{V}}^\pi - \boldsymbol{V}^\pi\big\|_\infty \le \frac{2\gamma\log\big(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\big)}{(1-\gamma)^2N}\big\|\boldsymbol{r}\big\|_\infty + 8\log 2\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\big)}{(1-\gamma)^3N}}\|\boldsymbol{r}\|_\infty$$

$$\leq 6\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^3 N}}\,\|\boldsymbol{r}\|_\infty,$$

with the proviso that $N \geq \frac{32e^2}{1-\gamma}\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)$.

B.1.1 Proof of Lemma 8

To begin with, we make the observation that

$$\begin{aligned}
\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V}) &= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_\pi\boldsymbol{V})\circ(\boldsymbol{P}_\pi\boldsymbol{V}) \\
&= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - \frac{1}{\gamma^2}(\boldsymbol{V}-\boldsymbol{r}_\pi)\circ(\boldsymbol{V}-\boldsymbol{r}_\pi) \qquad (55)\\
&= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - \frac{1}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{V} + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi - \frac{1}{\gamma^2}\boldsymbol{r}_\pi\circ\boldsymbol{r}_\pi \\
&\leq \frac{1}{\gamma^2}(\gamma^2\boldsymbol{P}_\pi - \boldsymbol{I})(\boldsymbol{V}\circ\boldsymbol{V}) + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi, \qquad (56)
\end{aligned}$$

where the identity (55) makes use of the relation $\boldsymbol{V} = \boldsymbol{r}_\pi + \gamma\boldsymbol{P}_\pi\boldsymbol{V}$. In addition, one can deduce that

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty &= \frac{1}{1-\gamma}\left\|(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \\
&\overset{\text{(i)}}{\leq} \frac{1}{\sqrt{1-\gamma}}\left\|\sqrt{(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \\
&\overset{\text{(ii)}}{\leq} \frac{1}{\sqrt{1-\gamma}}\sqrt{\|2(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})\|_\infty}.
\end{aligned}$$

Here, (i) comes from Jensen's inequality (so that $\mathbb{E}[\sqrt{v}] \leq \sqrt{\mathbb{E}[v]}$) recognizing that each row of $(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$ is a probability distribution, and Lemma 7(d), (ii) is an elementary fact established in (Agarwal et al., 2019, Lemma 4). Combining Lemma 7(e) and the inequality (56) further yields

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty &\leq \frac{1}{\sqrt{1-\gamma}}\sqrt{\left\|2(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}\left(\frac{1}{\gamma^2}\left(\gamma^2\boldsymbol{P}_\pi-\boldsymbol{I}\right)(\boldsymbol{V}\circ\boldsymbol{V}) + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi\right)\right\|_\infty} \\
&\leq \frac{1}{\gamma\sqrt{1-\gamma}}\sqrt{2\,\|\boldsymbol{V}\circ\boldsymbol{V}\|_\infty} + \frac{2}{\gamma\sqrt{1-\gamma}}\sqrt{\left\|(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\right\|_\infty}. \qquad (57)
\end{aligned}$$

where the last step arises from the triangle inequality and $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$. This leaves us with two terms to deal with.

Regarding the first term of (57), we observe that $\|(\boldsymbol{V}\circ\boldsymbol{V})\|_\infty = \|\boldsymbol{V}\|_\infty^2$. When it comes to the second term of (57), it is seen that

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\right\|_\infty &\leq \left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\right\|_\infty \\
&\leq \|\boldsymbol{V}\|_\infty\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi\right\|_\infty \\
&= \|\boldsymbol{V}\|_\infty^2.
\end{aligned}$$

Here, the first inequality holds true since $(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1} = \sum_{i=0}^\infty \gamma^{2i}\boldsymbol{P}_\pi^i \leq \sum_{i=0}^\infty \gamma^i\boldsymbol{P}_\pi^i = (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$, while the second line holds since $\boldsymbol{V}$, $\boldsymbol{r}$ and $(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$ are all non-negative. Substitution into (57) thus yields

$$\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \leq \frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty + \frac{2}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty \leq \frac{4}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty$$

as claimed.

B.2 Proof of Lemma 3

To establish Lemma 3, it suffices to check that $\boldsymbol{V}^\star$ and $\boldsymbol{Q}^\star$ satisfy the Bellman optimality equations underlying $\mathcal{M}_{s,a,u^\star}$. Towards this end, we study the absorbing state-action pair (s,a) and other pairs separately. For notational simplicity, we shall let $\boldsymbol{P}^{\mathsf{abs}}$ and $r^{\mathsf{abs}}(\cdot,\cdot)$ denote respectively the probability transition matrix and the reward function associated with $\mathcal{M}_{s,a,u^\star}$.

First, we observe that, by construction,

$$r^{\mathsf{abs}}(s,a) + \gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s,a} = u^\star + \gamma V^\star(s).$$

Recall that $\boldsymbol{V}^\star$ satisfies the Bellman optimality equation w.r.t. the original MDP, namely, $Q^\star(s,a) = r(s,a) + \gamma(\boldsymbol{P}\boldsymbol{V}^\star)_{s,a}$. This together with our choice of $u^\star$ gives

$$u^\star + \gamma V^\star(s) = Q^\star(s,a) - \gamma V^\star(s) + \gamma V^\star(s) = Q^\star(s,a).$$

Putting the above identities together, we arrive at

$$Q^\star(s,a) = r^{\mathsf{abs}}(s,a) + \gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s,a}. \tag{58}$$

Next, consider any state-action pair $(s',a') \neq (s,a)$. Recalling again the properties of $\mathcal{M}_{s,a,u^\star}$, we reach

$$r^{\mathsf{abs}}(s',a') + \gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s',a'} = r(s',a') + \gamma(\boldsymbol{P}\boldsymbol{V}^\star)_{s',a'} = Q^\star(s',a'). \tag{59}$$

Here the last identity is due to the Bellman equation w.r.t. the original MDP. Combining (58) and (59) implies that $\boldsymbol{V}^\star$ and $\boldsymbol{Q}^\star$ satisfy Bellman's optimality equations in $\mathcal{M}_{s,a,u^\star}$, thus concluding the proof.

B.3 Proof of Lemma 4

Our first observation is that $\widehat{\boldsymbol{Q}}^\star_{s,a,u}$ satisfies Lipschitz continuity w.r.t. u in the sense that

$$\big\|\widehat{\boldsymbol{Q}}^\star_{s,a,u} - \widehat{\boldsymbol{Q}}^\star_{s,a,u'}\big\|_\infty \le \frac{1}{1-\gamma}|u-u'|. \tag{60}$$

The proof of this relation is identical to that of (Agarwal et al., 2019, Lemma 8); we omit here for brevity. In view of Lemma 3, if we set $\widehat{u}^\star := r(s,a) + \gamma(\widehat{\boldsymbol{P}}\widehat{\boldsymbol{V}}^\star)_{s,a} - \gamma\widehat{V}^\star(s)$, then one has

$$\widehat{\boldsymbol{Q}}^\star_{s,a,\widehat{u}^\star} = \widehat{\boldsymbol{Q}}^\star \qquad \text{and} \quad \widehat{\boldsymbol{V}}^\star_{s,a,\widehat{u}^\star} = \widehat{\boldsymbol{V}}^\star.$$

In addition, there exists a point u_0 in the epsilon-net $\mathcal{N}_{(1-\gamma)\omega/4}$ such that $|\widehat{u}^\star - u_0| \le (1-\gamma)\omega/4$, which combined with the Lipschitz continuity property (60) gives

$$\big\|\widehat{\boldsymbol{Q}}^\star - \widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty = \big\|\widehat{\boldsymbol{Q}}^\star_{s,a,\widehat{u}^\star} - \widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty \le \frac{1}{1-\gamma}|\widehat{u}^\star - u_0| \le \frac{\omega}{4}. \tag{61}$$

Additionally, for any $s' \in \mathcal{S}$ and any $a_1, a_2 \in \mathcal{A}$ with $a_1 \neq a_2$, we have the following decomposition

$$\begin{aligned}
&\widehat{Q}^\star_{s,a,u_0}(s',a_1) - \widehat{Q}^\star_{s,a,u_0}(s',a_2) \\
&\quad= \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) + \widehat{Q}^\star_{s,a,u_0}(s',a_1) - \widehat{Q}^\star(s',a_1) - \Big(\widehat{Q}^\star_{s,a,u_0}(s',a_2) - \widehat{Q}^\star(s',a_2)\Big) \\
&\quad\ge \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) - 2\big\|\widehat{\boldsymbol{Q}}^\star - \widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty \\
&\quad\ge \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) - \frac{\omega}{2},
\end{aligned} \tag{62}$$

where the last inequality invokes the inequality (61). Moreover, our separation condition defined in (31) requires that: for any $s' \in \mathcal{S}$ and any $a_2 \neq \widehat{\pi}^\star(s')$, one has $\widehat{Q}^\star(s',\widehat{\pi}^\star(s')) - \widehat{Q}^\star(s',a_2) \ge \omega$, which together with (62) reveals that

$$\widehat{Q}^\star_{s,a,u_0}(s',\widehat{\pi}^\star(s')) - \widehat{Q}^\star_{s,a,u_0}(s',a_2) \ge \widehat{Q}^\star(s',\widehat{\pi}^\star(s')) - \widehat{Q}^\star(s',a_2) - \frac{\omega}{2} \ge \frac{\omega}{2}. \tag{63}$$

Given that $\widehat{\pi}^\star_{s,a,u_0}(s') := \arg\max_{a'} \widehat{Q}^\star_{s,a,u_0}(s',a')$, it is seen from (63) that

$$\widehat{\pi}^\star_{s,a,u_0}(s') = \widehat{\pi}^\star(s'),$$

which holds true for all $s' \in \mathcal{S}$. This concludes the proof.

B.4 Proof of Lemma 5

We start by bounding $\big\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\big\|_\infty$. Recall the definition of the series $\{\boldsymbol{V}^{(l)}\}$ in (41). Throughout this proof, we shall write $\boldsymbol{V}_\pi^{(l)}$ instead in order to make apparent the dependency on the policy π.

For each state-action pair (s,a), let us construct the epsilon-net $\mathcal{N}_{(1-\gamma)\omega/4}$ as in the expression (32). For every $u \in \mathcal{N}_{(1-\gamma)\omega/4}$, recall that $\widehat{\pi}^\star_{s,a,u}$ is defined as the optimal policy with respect to the (s,a)-absorbing MDP $\widehat{\mathcal{M}}_{s,a,u}$. By construction, the set of policies $\widehat{\pi}^\star_{s,a,u}$ $(u \in \mathcal{N}_{(1-\gamma)\omega/4})$ is independent of $\widehat{\boldsymbol{P}}_{s,a}$. The Bernstein inequality (e.g. (Agarwal et al., 2019, Lemma 6)) taken together with the union bound thus guarantees that with probability at least $1-\delta$,

$$\Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a} \boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\Big| \le \sqrt{\frac{\beta_1}{N}}\sqrt{\big(\mathsf{Var}_{\boldsymbol{P}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\big]\big)_{s,a}} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\big\|_\infty \beta_1}{N} \tag{64}$$

holds uniformly over all $0 \le l \le \log\frac{e}{1-\gamma}$, $u \in \mathcal{N}_{(1-\gamma)\omega/4}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$. Here, β_1 is given by

$$\beta_1 := 2\log\Big(\frac{4\log\big(\frac{e}{1-\gamma}\big)|\mathcal{N}_{(1-\gamma)\omega/4}||\mathcal{S}||\mathcal{A}|}{\delta}\Big) \le 2\log\Big(\frac{32}{(1-\gamma)^2\omega\delta}|\mathcal{S}||\mathcal{A}|\log\big(\frac{e}{1-\gamma}\big)\Big),$$

where we have used the fact $|\mathcal{N}_{(1-\gamma)\omega/4}| \le \frac{8}{(1-\gamma)^2\omega}$. In addition, for any $0<\omega<1$, Lemma 4 guarantees that for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$ such that $\widehat{\pi}^\star = \widehat{\pi}^\star_{s,a,u_0}$. Invoking this important fact, we obtain

$$\begin{aligned}
\Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a}\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\Big| &= \Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a}\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\Big| \\
&\le \sqrt{\frac{\beta_1}{N}}\sqrt{\big(\mathsf{Var}_{\boldsymbol{P}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\big]\big)_{s,a}} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\big\|_\infty\beta_1}{N} \\
&= \sqrt{\frac{\beta_1}{N}}\sqrt{\big(\mathsf{Var}_{\boldsymbol{P}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big]\big)_{s,a}} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big\|_\infty\beta_1}{N}.
\end{aligned}$$

The above inequality further allows us to deduce that, with probability $1-\delta$,

$$\begin{aligned}
\Big|\big(\widehat{\boldsymbol{P}}_{\widehat{\pi}^\star} - \boldsymbol{P}_{\widehat{\pi}^\star}\big)\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\Big| &= \Big|\boldsymbol{\Pi}^{\widehat{\pi}^\star}(\widehat{\boldsymbol{P}} - \boldsymbol{P})\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\Big| \\
&\le \sqrt{\frac{\beta_1}{N}}\sqrt{\boldsymbol{\Pi}^{\widehat{\pi}^\star}\mathsf{Var}_{\boldsymbol{P}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big]} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big\|_\infty\beta_1}{N}\mathbf{1} \\
&= \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\widehat{\pi}^\star}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big]} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\big\|_\infty\beta_1}{N}\mathbf{1}.
\end{aligned}$$

The above derivation validates the assumption required for Lemma 2. As a result, if $N > \frac{16e^2}{1-\gamma}\beta_1$ and $0<\omega<1$, then Lemma 2 leads to the advertised bound

$$\begin{aligned}
\big\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\big\|_\infty &\le \frac{6}{1-\gamma}\sqrt{\frac{\beta_1}{N(1-\gamma)}} \le \frac{6}{1-\gamma}\sqrt{\frac{2\log\Big(\frac{32}{(1-\gamma)^2\omega\delta}|\mathcal{S}||\mathcal{A}|\log\big(\frac{e}{1-\gamma}\big)\Big)}{N(1-\gamma)}} \\
&\le 6\sqrt{\frac{2\log\Big(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\omega\delta}\Big)}{N(1-\gamma)^3}}.
\end{aligned} \tag{65}$$

Finally, we move on to the term $\boldsymbol{V}^\star - \widehat{\boldsymbol{V}}^{\pi^\star}$. Given that $\pi^\star$ is independent of $\widehat{\boldsymbol{P}}$, invoke Lemma 1 to reach

$$\Big\|\widehat{\boldsymbol{V}}^{\pi^\star} - \boldsymbol{V}^\star\Big\|_\infty = \big\|\widehat{\boldsymbol{V}}^{\pi^\star} - \boldsymbol{V}^{\pi^\star}\big\|_\infty \le 6\sqrt{2}\sqrt{\frac{\log\big(\frac{4|\mathcal{S}|}{\delta}\big) + \log\log\big(\frac{e}{1-\gamma}\big)}{N(1-\gamma)^3}}$$

$$\leq 6\sqrt{\frac{2\log\left(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2\omega\delta}\right)}{N(1-\gamma)^3}} \tag{66}$$

with probability at least $1-\delta$. This together with (24) and (65) immediately establishes Lemma 5.

B.5 Proof of Lemma 6

The proofs for $Q_{\mathrm{p}}^{\star}$ and $\widehat{Q}_{\mathrm{p}}^{\star}$ are exactly the same; for the sake of conciseness, we shall only provide the proof for $Q_{\mathrm{p}}^{\star}$. Here we aim to prove a more general result than Lemma 6, namely, with probability at least $1-\delta$,

$$\forall s\in\mathcal{S} \text{ and } a_1,a_2\in\mathcal{A} \text{ with } a_1\neq a_2: \qquad \left|Q_{\mathrm{p}}^{\star}(s,a_1)-Q_{\mathrm{p}}^{\star}(s,a_2)\right| > \frac{\xi\delta(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2}.$$

Consider any state s and any actions $a_1\neq a_2$. In what follows, we allow $r_{\mathrm{p}}(s,a_1)=\tau$ to vary, while *freezing* the values of all other rewards $\{r_{\mathrm{p}}(\tilde{s},a)\mid(\tilde{s},a)\neq(s,a_1)\}$. To streamline notation, we define

- $\boldsymbol{r}_\tau=[r_\tau(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$: the reward vector obeying

$$r_\tau(s,a_1)=\tau \qquad \text{and} \qquad r_\tau(\tilde{s},a)=r_{\mathrm{p}}(\tilde{s},a) \quad \text{for all } (\tilde{s},a)\neq(s,a_1);$$

- $Q_\tau^{\star}$: the optimal Q-function when the reward vector is $\boldsymbol{r}_\tau$;
- $V_\tau^{\star}$: the optimal value function when the reward vector is $\boldsymbol{r}_\tau$;
- $\pi_\tau^{\star}$: the optimal policy when the reward vector is $\boldsymbol{r}_\tau$.

Additionally, we claim for the moment that there exists a phase transition boundary τ_{th} such that

$$\pi_\tau^{\star}(s)\neq a_1, \quad \text{for all } \tau<\tau_{\mathrm{th}}; \tag{67a}$$
$$\pi_\tau^{\star}(s)=a_1, \quad \text{for all } \tau>\tau_{\mathrm{th}}. \tag{67b}$$

The proof of this claim is deferred to the end of this section. To establish Lemma 6, the idea is to control the size of the set

$$\mathcal{I}_{0,\omega}:=\left\{\tau\mid\left|Q_\tau^{\star}(s,a_1)-Q_\tau^{\star}(s,a_2)\right|<\omega\right\} \tag{68}$$

for some $\omega>0$ to be specified shortly. As motivated by (67), we further break down this set into two parts $\mathcal{I}_{0,\omega}=\mathcal{I}_{1,\omega}\cup\mathcal{I}_{2,\omega}$, where

$$\mathcal{I}_{1,\omega}:=\left\{\tau\mid\tau<\tau_{\mathrm{th}},\ \left|Q_\tau^{\star}(s,a_1)-Q_\tau^{\star}(s,a_2)\right|<\omega\right\}, \tag{69a}$$
$$\mathcal{I}_{2,\omega}:=\left\{\tau\mid\tau\geq\tau_{\mathrm{th}},\ \left|Q_\tau^{\star}(s,a_1)-Q_\tau^{\star}(s,a_2)\right|<\omega\right\}. \tag{69b}$$

In what follows, we first control the size of each set separately, and then demonstrate that the probability of these events happening is very small.

Step 1. We begin with $\mathcal{I}_{1,\omega}$ associated with the range $\tau<\tau_{\mathrm{th}}$. In this case, the value function $V_\tau^{\star}$ does not vary with τ, since the reward $r_\tau(s,a_1)=\tau$ is never active when calculating $V_\tau^{\star}$ (by virtue of (67a)). Thus, the Bellman equation allows us to write

$$Q_\tau^{\star}(s,a_1)=\tau+B_1 \qquad \text{and} \qquad Q_\tau^{\star}(s,a_2)=B_2$$

for some quantities B_1 and B_2, where neither B_1 nor B_2 depends on the value of τ. Armed with this observation, we can easily show that: for any $\omega>0$, the interval $\mathcal{I}_{1,\omega}$ (cf. (69a)) obeys

$$\mathcal{I}_{1,\omega}\subseteq\{\tau\mid|\tau+B_1-B_2|<\omega\},$$

and hence has length (or Lebesgue measure) at most 2ω.

Step 2. We then move on to $\mathcal{I}_{2,\omega}$ associated with the range $\tau > \tau_{\rm th}$ in which case $\pi^\star_\tau(s) = a_1$. Towards this, we first make some useful observations.

- To begin with, given the relation $\boldsymbol{Q}^\star_\tau = \boldsymbol{r}_\tau + \gamma \boldsymbol{P}\boldsymbol{V}^\star_\tau$, it is easily seen that for any $\tau_2 > \tau_1 > \tau_{\rm th}$,

$$\boldsymbol{0} \le \boldsymbol{Q}^\star_{\tau_2} - \boldsymbol{Q}^\star_{\tau_1} \le \boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1} + \gamma \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty. \tag{70}$$

  In addition, for any state-action pair $(\tilde{s}, a) \neq (s, a_1)$, by construction we have $r_{\tau_2}(\tilde{s}, a) - r_{\tau_1}(\tilde{s}, a) = 0$, which together with (70) indicates that

$$\forall (\tilde{s}, a) \neq (s, a_1): \quad 0 \le Q^\star_{\tau_2}(\tilde{s}, a) - Q^\star_{\tau_1}(\tilde{s}, a) \le \gamma \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty. \tag{71}$$

- Next, observe that for any $\tau_2 > \tau_1 > \tau_{\rm th}$,

$$\forall s' \in \mathcal{S}: \quad 0 \le V^\star_{\tau_2}(s') - V^\star_{\tau_1}(s') = \max_a Q^\star_{\tau_2}(s', a) - \max_a Q^\star_{\tau_1}(s', a) \le \left\|\boldsymbol{Q}^\star_{\tau_2} - \boldsymbol{Q}^\star_{\tau_1}\right\|_\infty \tag{72}$$

  and hence $\|\boldsymbol{Q}^\star_{\tau_2} - \boldsymbol{Q}^\star_{\tau_1}\|_\infty \ge \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty$. This combined with (71) and the fact $\gamma < 1$ implies that

$$Q^\star_{\tau_2}(s, a_1) - Q^\star_{\tau_1}(s, a_1) \ge \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty, \tag{73}$$

  which together with the facts $V^\star_{\tau_1}(s) = Q^\star_{\tau_1}(s, a_1)$ and $V^\star_{\tau_2}(s) = Q^\star_{\tau_2}(s, a_1)$ (by virtue of (67b)) yields

$$\begin{aligned} V^\star_{\tau_2}(s) - V^\star_{\tau_1}(s) = Q^\star_{\tau_2}(s, a_1) - Q^\star_{\tau_1}(s, a_1) &\ge \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty \\ \Longrightarrow \quad Q^\star_{\tau_2}(s, a_1) - Q^\star_{\tau_1}(s, a_1) &= \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty. \end{aligned}$$

  Invoke the Bellman equation to further derive

$$\begin{aligned} Q^\star_{\tau_2}(s, a_1) - Q^\star_{\tau_1}(s, a_1) = \|\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{V}^\star_{\tau_1}\|_\infty &= \left\|\boldsymbol{r}_{\tau_2} + \gamma \boldsymbol{P}\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{r}_{\tau_1} - \gamma \boldsymbol{P}\boldsymbol{V}^\star_{\tau_1}\right\|_\infty \\ &\ge \left\|\boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1}\right\|_\infty = \tau_2 - \tau_1, \end{aligned} \tag{74}$$

  where the last inequality holds since $\boldsymbol{r}_{\tau_2} + \gamma \boldsymbol{P}\boldsymbol{V}^\star_{\tau_2} - \boldsymbol{r}_{\tau_1} - \gamma \boldsymbol{P}\boldsymbol{V}^\star_{\tau_1} \ge \boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1} \ge \boldsymbol{0}$ (due to the monotonicity properties $\boldsymbol{r}_{\tau_2} \ge \boldsymbol{r}_{\tau_1}$ and $\boldsymbol{V}^\star_{\tau_2} \ge \boldsymbol{V}^\star_{\tau_1}$), and the last identity follows from the definition of $\boldsymbol{r}_\tau$.

With the above two properties (71) and (74) in mind, we are ready to locate $\mathcal{I}_{2,\omega}$ by showing that

$$\mathcal{I}_{2,\omega} \subseteq \left[\tau_{\rm th}, \tau_{\rm th} + \frac{\omega}{1-\gamma}\right]. \tag{75}$$

Given that $Q^\star_{\tau_{\rm th}}(s, a_1) \ge Q^\star_{\tau_{\rm th}}(s, a_2)$ (in view of (67)), we have for any $\tau \ge \tau_{\rm th}$ and any $a_2 \neq a_1$ that

$$\begin{aligned} Q^\star_\tau(s, a_1) - Q^\star_\tau(s, a_2) &\ge \big(Q^\star_\tau(s, a_1) - Q^\star_{\tau_{\rm th}}(s, a_1)\big) - \big(Q^\star_\tau(s, a_2) - Q^\star_{\tau_{\rm th}}(s, a_2)\big) \\ &\ge (1-\gamma)\big(Q^\star_\tau(s, a_1) - Q^\star_{\tau_{\rm th}}(s, a_1)\big). \end{aligned} \tag{76}$$

Here, the last inequality holds since

$$Q^\star_\tau(s, a_2) - Q^\star_{\tau_{\rm th}}(s, a_2) \overset{\text{(i)}}{\le} \gamma \left\|\boldsymbol{V}^\star_\tau - \boldsymbol{V}^\star_{\tau_{\rm th}}\right\|_\infty \overset{\text{(ii)}}{\le} \gamma\big(Q^\star_\tau(s, a_1) - Q^\star_{\tau_{\rm th}}(s, a_1)\big),$$

where (i) follows from (71) and (ii) is due to (73). As a result, for any $\tau > \tau_{\rm th} + \frac{\omega}{1-\gamma}$, one can invoke (76) and (74) to see that

$$\forall a_2 \neq a_1: \quad Q^\star_\tau(s, a_1) - Q^\star_\tau(s, a_2) \ge (1-\gamma)\big(Q^\star_\tau(s, a_1) - Q^\star_{\tau_{\rm th}}(s, a_1)\big) \ge (1-\gamma)(\tau - \tau_{\rm th}) > \omega,$$

which necessarily implies that such a τ does not lie within the interval $\mathcal{I}_{2,\omega}$ as defined in (69b). This establishes the claimed relation (75).

Step 3. Putting the results in the above two steps together, we see the set $\mathcal{I}_{0,\omega}$ (cf. (68)) has total length (or Lebesgue measure) at most $\frac{3\omega}{1-\gamma}$. Given that $r_{\mathrm{p}}(s,a) = r(s,a) + \zeta(s,a)$ with $\zeta(s,a) \sim \mathsf{Unif}(0,\xi)$, one has

$$\mathbb{P}\left\{\left|Q_{\mathrm{p}}^{\star}(s,a_1) - Q_{\mathrm{p}}^{\star}(s,a_2)\right| < \omega\right\} \leq \mathbb{P}\left\{r(s,a) + \zeta(s,a) \in \mathcal{I}_{0,\omega}\right\} \leq \frac{3\omega}{\xi(1-\gamma)}.$$

By setting $\omega = \frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}$, we arrive at

$$\mathbb{P}\left\{\left|Q_{\mathrm{p}}^{\star}(s,a_1) - Q_{\mathrm{p}}^{\star}(s,a_2)\right| < \frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}\right\} \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|^2}.$$

Finally, taking the union bound over all s, a_1, a_2, we conclude that

$$\mathbb{P}\left\{\exists s, a_1 \neq a_2 : \ \left|Q_{\mathrm{p}}^{\star}(s,a_1) - Q_{\mathrm{p}}^{\star}(s,a_2)\right| < \frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}\right\} \leq \delta,$$

thus establishing Lemma 6 as long as the claim (67) is valid.

Proof of the claim (67). To establish the claim, it suffices to take

$$\tau_{\mathrm{th}} = \sup\left\{u \mid \pi_{\tau}^{\star}(s) \neq a_1 \text{ for all } \tau < u\right\}. \tag{77}$$

It thus suffices to verify (67b) for our choice (77). Towards this, suppose instead that there exist some $\tau_3 < \tau_{\mathrm{th}} \leq \tau_2 < \tau_1$ such that

$$\pi_{\tau_3}^{\star}(s) \neq a_1, \qquad \pi_{\tau_2}^{\star}(s) = a_1, \qquad \text{and} \qquad \pi_{\tau_1}^{\star}(s) \neq a_1.$$

It is straightforward to see that $V_{\tau_1}^{\star} = V_{\tau_3}^{\star}$, since in both cases, the reward $r_{\tau}(s,a_1)$ does not enter the calculation of the optimal value function (while the rewards in other state-action pairs are identical in both cases). In view of the monotonicity of the value function w.r.t. the reward vector, we have

$$V_{\tau_1}^{\star} = V_{\tau_2}^{\star} = V_{\tau_3}^{\star}.$$

However, this contradicts our assumption that a_1 is the optimal action for state s at τ_2 but not at τ_3, since enlarging τ_2 to τ_3 otherwise will enlarge the optimal value function $V_{\tau_3}^{\star}$. We have thus established (67). □

References

Agarwal, A., Kakade, S., and Yang, L. F. (2019). Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804*.

Azar, M. G., Munos, R., and Kappen, B. (2012). On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*.

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

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

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

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

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

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

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57.

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

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204.

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

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

Duan, Y. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*.

El Karoui, N. (2015). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81.

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

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

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

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

Jiang, N. and Agarwal, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

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

Kaledin, M., Moulines, E., Naumov, A., Tadic, V., and Wai, H.-T. (2020). Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. *arXiv preprint arXiv:2002.01268*.

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

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

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

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355.

Lattimore, T. and Hutter, M. (2012). PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:arXiv:2006.03041*.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632.

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

Pananjady, A. and Wainwright, M. J. (2019). Value function estimation in Markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. *arXiv preprint arXiv:1909.08749*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*.

Shah, D. and Xie, Q. (2018). Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121.

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

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM.

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

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. (2006). PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103.

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

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

Tu, S. and Recht, B. (2018). The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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

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

Wang, M. (2019). Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. (2020). Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*.

Xu, P. and Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. *accepted to International Conference on Machine Learning*.

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

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.