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

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).

1 Introduction

Reinforcement learning (RL) [33, 35], 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] To enable reliable RL 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. Characterizing the sample complexities of RL algorithms (including both model-based and model-free approaches) has been the focal point of a large body of recent work, e.g. [1, 2, 15, 19, 29, 30, 38, 39, 40, 41].

This paper pursues a comprehensive understanding about the sample efficiency 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 [16, 19]. We focus 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 $\varepsilon > 0$, the goal is to compute 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).

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

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Algorithm	Sample size range	Sample complexity	ε-range
Phased Q-learning [19]	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^5}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^7\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical QVI [2]	$\big[\frac{\vert\mathcal{S}\vert^2\vert\mathcal{A}\vert}{(1-\gamma)^2}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\vert\mathcal{S}\vert}}]$
Sublinear randomized VI [30]	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^2}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^4\varepsilon^2}$	$\big(0, \frac{1}{1-\gamma}]$
Variance-reduced QVI [29]	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^3}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
Randomized primal-dual alg. [41]	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^2}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^4\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical MDP + planning [1]	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^2}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$
Perturbed empirical MDP + planning **This paper**	$\big[\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{1-\gamma}, \infty\big)$	$\frac{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$

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 [19, 41] only hold for a restricted family of MDPs satisfying certain ergodicity assumptions.

Motivation: sample size barriers. 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 regime [1, 2]. However, their results often suffered from a sample size barrier that prevents us from obtaining a complete trade-off curve between sample size and accuracy. For instance, the state-of-the-art result [1] required the total sample size to at least exceed $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$ (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 [1, 26]. Furthermore, an even higher barrier emerged in prior theory for model-free methods; for instance, [30, 40] required the sample size to exceed $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$ (up to some log factor). In stark contrast, however, no lower bounds developed thus far preclude us from attaining reasonable statistical accuracy when going below such sample size 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 finds an ε-optimal policy with an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ samples (up to log factor), which matches the minimax lower bound [2]. Our result accommodates the full range of accuracy level ε (i.e. $\varepsilon \in (0, \frac{1}{1-\gamma}]$), thus unveiling the minimaxity of our algorithm as soon as the sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ (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 work. Along the way, we also derive (instance-dependent) 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 value function estimation error, 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 [1] — 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 perturbations.

2 Problem formulation

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 $\mathcal{S} := \{1, 2, \dots, |\mathcal{S}|\}$ denotes a finite set of states, $\mathcal{A} := \{1, 2, \dots, |\mathcal{A}|\}$ is a finite set of actions, $\gamma \in (0, 1)$ is the discount factor, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the reward function, i.e. $r(s, a)$ 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, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the probability transition kernel, 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 $\pi : \mathcal{S} \to \mathcal{A}$, which maps a state to an action. The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π is defined by

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

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) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π is defined by

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

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 known that there exists an optimal policy, denoted by $\pi^\star$, that maximizes $V^\pi(s)$ (resp. $Q^\pi(s,a)$) for all $s \in \mathcal{S}$ (resp. $(s,a) \in (\mathcal{S} \times \mathcal{A})$) [33]. The corresponding value function $V^\star := V^{\pi^\star}$ (resp. Q-function $Q^\star := Q^{\pi^\star}$) is called the optimal value function (resp. Q-function).

A generative model and an empirical MDP. The current paper focuses on a stylized generative model (also called a simulator) as studied in [16, 18]. Assuming access to this generative model, we collect N independent samples $s_{s,a}^i \sim P(\cdot|s,a)$, $i = 1, \ldots, N$ for any state-action pair (s,a), which allows us to construct an empirical transition kernel $\widehat{P}$ given by

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

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 this policy. To be precise, for any target level $\varepsilon > 0$, the goal is to find an ε-accurate estimate V_{est}^π such that

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

In contrast, the task of planning seeks a policy that (approximately) maximizes the expected discounted reward. Specifically, for any target $\varepsilon > 0$, we aim to find an ε-optimal policy π_{est} obeying

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

Naturally, one would hope to accomplish these tasks with as few samples as possible. Recall that for the normalized reward setting ($0 \leq r \leq 1$), the value function and Q-function fall in the range $[0, \frac{1}{1-\gamma}]$; this means that the range of accuracy level ε should be set to $\varepsilon \in [0, \frac{1}{1-\gamma}]$. A model-based approach 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 [5] to perform policy evaluation or planning, with prominent examples including Q-value iteration (QVI) and policy iteration (PI).

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 that there exists a constant $C_1 > 0$ such that $f \leq C_1 g$, whereas $f(\mathcal{X}) = \Omega(g(\mathcal{X}))$ means $g(\mathcal{X}) = O(f(\mathcal{X}))$. 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 log 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 s.t. $\sqrt{\boldsymbol{a}} := [\sqrt{a_i}]_{1\leq i\leq n}$ 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. We denote by $\mathbf{1}$ the all-one vector.

3 Main results

To break the aforementioned 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 perturb the immediate reward (with uniform distribution) 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 $\xi > 0$ is a parameter to be specified shortly.[2] For any policy π, 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 is ε-close to the true optimal value function $V^{\star}$ (resp. optimal Q-function $Q^{\star}$), as asserted below.

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

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

provided that the perturbation obeys $\xi = \frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ and the sample size per state-action pair obeys

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

In addition, the empirical QVI (cf. [2]) recovers $\widehat{\pi}_{\mathrm{p}}^{\star}$ 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.

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 few iterations without incurring further optimization errors. To give a flavor of the overall computational cost, recall that each iteration of QVI takes time proportional to the time taken to read $\widehat{P}$ (which is a sparse matrix with at most $N|\mathcal{S}||\mathcal{A}|$ nonzeros), hence the resulting computational complexity can be as low as $O\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\log^2(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta})\big)$.

Theorem 1 demonstrates that: the proposed perturbed model-based 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). 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 (particularly the scenario when $N \in [\frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2}]$) requires us to develop new analysis ideas beyond prior theory, which we shall discuss in detail momentarily.

We remark that [2] established a minimax lower bound of the same order as (9) (up to log factor) in the regime $\varepsilon = O(1)$. A closer inspection of their analysis, however, reveals that their argument holds 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 $N \leq \widetilde{O}\big(\frac{1}{1-\gamma}\big)$. 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 [2]), focusing on the effect of the discount complexity $\frac{1}{1-\gamma}$

[2] Perturbation is only invoked when running the planning algorithms and does not require new samples.

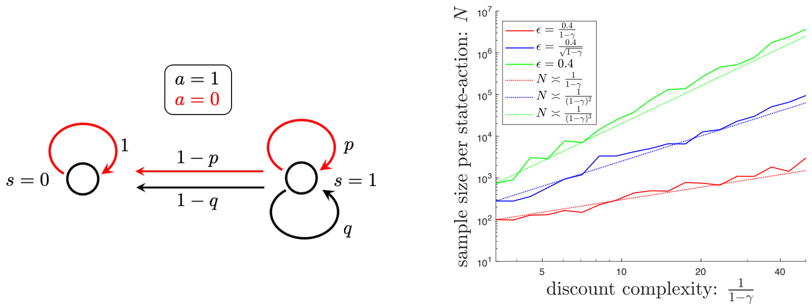

Figure 1: Numerical sample complexity per state-action pair N vs. discount complexity $\frac{1}{1-\gamma}$.

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 [2]. For each point $(N, \frac{1}{1-\gamma}, \varepsilon)$, we perform 100 Monte Carlo simulations, and we declare N to successfully attain ε accuracy if this accuracy is achieved in at least 95% of the trials. In Fig. 1, the solid lines depict the minimum sample size per state-action pair needed to achieve ε accuracy numerically as the discount complexity varies. As can be seen, the numerical sample complexity per state-action pair 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, 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 large enough. Our result is this:

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 $0 < \varepsilon \leq \frac{1}{1-\gamma}$, one has*

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

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 $N \geq \widetilde{\Omega}\left(\frac{1}{1-\gamma}\right)$. The sample complexity bound (12) enjoys *full* coverage of the ε-range $(0, \frac{1}{1-\gamma}]$, and matches the minimax lower bound [26, 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 ([20, 26]). While this is not our focus, our analysis does uncover an improved instance-dependent bound. See Lemma 1.

Prior bounds for planning and policy learning. None of the prior results with a generative model (including both model-based and model-free methods) 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 [1] required the sample size to exceed $N \geq \widetilde{\Omega}\big(\frac{1}{(1-\gamma)^2}\big)$, whereas the theory for the variance-reduced model-free approach [29, 40] imposed the sample size requirement $N \geq \widetilde{\Omega}\big(\frac{1}{(1-\gamma)^3}\big)$. 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.

Prior bounds for policy evaluation. Prior results [1, 2, 26] for the plug-in approach 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

$\frac{1}{1-\gamma}$. Our sample complexity matches the state-of-the-art guarantees in the regime where $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ [1, 26], while extending them to the range $\varepsilon \in \big[\frac{1}{\sqrt{1-\gamma}}, \frac{1}{1-\gamma}\big]$ uncovered previously.

Other related work. Classical analyses of reinforcement learning algorithms have largely focused on asymptotic performance (e.g. [14, 34, 36, 37]). 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 [3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 15, 17, 19, 20, 21, 24, 25, 27, 28, 29, 31, 32, 40, 42, 43, 44], a large fraction of which is concerned with model-free algorithms.

The generative model adopted herein was first proposed in [19], which has been invoked in [1, 2, 16, 18, 19, 20, 22, 26, 29, 30, 40, 41, 45], to name a few. In particular, [2] 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 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 extra factor of $\frac{1}{1-\gamma}$ in estimation accuracy and of $\frac{1}{(1-\gamma)^2}$ in sample complexity. In light of this, [2] 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 [1] substantially improved the model-based guarantee with the aid of auxiliary state-absorbing MDPs, extending the sample complexity range to $[\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2}, \infty)$. Our analysis is motivated in part by [1], but also relies on several other novel techniques to complete the picture.

4 Analysis

This section sketches the key proof ideas, with full details postponed to the full version [23].

Matrix notation. We first introduce a few matrix notation. Denoting by $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$ the standard basis vectors, we can define $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{V}^{\pi} \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^{\pi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as the vector representations of the reward function r, the value function V^{π} and the Q-function Q^{π}. Let $\boldsymbol{V}^{\star}$ and $\boldsymbol{Q}^{\star} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the optimal value function $V^{\star}$ and optimal Q-function $Q^{\star}$. Let $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ be a matrix representing the probability transition kernel P, where the (s,a)-th row of $\boldsymbol{P}$ is a probability vector representing $P(\cdot|s,a)$.

We introduce $\boldsymbol{\Pi}^{\pi} = \begin{pmatrix} \boldsymbol{e}^{\top}_{\pi(1)} & & & \\ & \boldsymbol{e}^{\top}_{\pi(2)} & & \\ & & \ddots & \\ & & & \boldsymbol{e}^{\top}_{\pi(|\mathcal{S}|)} \end{pmatrix} \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ as the projection matrix associated with policy π. The matrices $\boldsymbol{P}^{\pi} := \boldsymbol{P}\boldsymbol{\Pi}^{\pi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{P}_{\pi} := \boldsymbol{\Pi}^{\pi}\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ are two *square* probability transition matrices induced by policy π over the state-action pairs and the states, respectively. Let $\boldsymbol{r}_{\pi} \in \mathbb{R}^{|\mathcal{S}|}$ be a reward vector restricted to the actions chosen by policy π, i.e. $\boldsymbol{r}_{\pi} = \boldsymbol{\Pi}^{\pi}\boldsymbol{r}$. For a vector $\boldsymbol{V} = [V_i]_{1\leq i\leq|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$, we define the vectors

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

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. $\widehat{\mathcal{M}}$ analogously.

4.1 Key ideas of the proofs

Analysis: model-based policy evaluation. We start with the simpler task of policy evaluation, which is crucial in subsequent analysis. To establish Theorem 2, we aim to prove the following result. In view of the Bellman equation [5] and with our matrix notation in place, the true value function under policy π and the model-based empirical estimate can be written respectively as

$$\boldsymbol{V}^{\pi} = (\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_{\pi} \qquad \text{and} \qquad \widehat{\boldsymbol{V}}^{\pi} = (\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}_{\pi}. \tag{13}$$

Lemma 1. *Fix any policy π. Consider any $0 < \delta < 1$, and suppose $N \geq \frac{32e^2}{1-\gamma} \log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)$. Then with probability at least $1-\delta$, the vectors defined in (13) obey*

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} &\leq \gamma\sqrt{\frac{32\log\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}}{N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\left[\boldsymbol{V}^{\pi}\right]}\Big\|_{\infty} + \frac{2\gamma\log\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}}{(1-\gamma)N}\big\|\boldsymbol{V}^{\pi}\big\|_{\infty} \\ &\leq 6\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log(\frac{e}{1-\gamma})}{\delta}\big)}{N(1-\gamma)^3}}. \end{aligned} \tag{14}$$

Here, we strengthen Theorem 2 by providing an additional instance-dependent bound (see the first line of (14) that depends on the true instance $\boldsymbol{P}_{\pi}, \boldsymbol{V}^{\pi}$), which is often tighter and more adaptive than the worst-case bound stated in the second line of (14). Our contribution can be better understood when compared with [26]. Assuming that there is no noise in the rewards, our instance-dependent guarantee matches [26, 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, [26, Theorem 1] is only valid when $N \geq \widetilde{\Omega}(\frac{1}{(1-\gamma)^2})$.

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. To catch a glimpse of the motivation, we begin with the following elementary, and commonly used, perterbation relation

$$\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi} = \big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\boldsymbol{r}_{\pi} - \boldsymbol{V}^{\pi} = \gamma\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}. \tag{15}$$

Due to the complicated statistical 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 [1, Lemma 5]. This simple approach, however, leads to sub-optimal bounds. To refine analysis, we further expand (15) as

$$(15) = \gamma\big(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi}\big)^{-1}\big(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi} + \gamma\left\{\big(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi}\big)^{-1} - \big(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi}\big)^{-1}\right\}\big(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{\pi}, \tag{16}$$

which can be viewed as a "second-order" expansion. The advantage is that the first term in (16) becomes easier to cope with than its counterpart (15), 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 (16) remains difficult to control optimally, and it turns out that this issue can be remedied by resorting to further expansion. We shall continue to expand it to some log order, which eventually allows for optimal control of the estimation error.

Analysis: model-based planning. We then move on to outline the key ideas for model-based planning (Theorem 1). We shall start with the unperturbed version, which will elucidate the rationale of reward perturbation. First, we make note of the following elementary decomposition:

$$\begin{aligned}\boldsymbol{V}^{\star} - \boldsymbol{V}^{\widehat{\pi}^{\star}} &= \big(\widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}} - \boldsymbol{V}^{\widehat{\pi}^{\star}}\big) + \big(\widehat{\boldsymbol{V}}^{\pi^{\star}} - \widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}\big) + \big(\boldsymbol{V}^{\star} - \widehat{\boldsymbol{V}}^{\pi^{\star}}\big) \\ &\leq \big(\widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}} - \boldsymbol{V}^{\widehat{\pi}^{\star}}\big) + \big(\boldsymbol{V}^{\pi^{\star}} - \widehat{\boldsymbol{V}}^{\pi^{\star}}\big), \end{aligned} \tag{17}$$

which uses $\widehat{\boldsymbol{V}}^{\pi^{\star}} \leq \widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}$ due to the optimality of $\widehat{\pi}^{\star}$. 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 a fixed policy independent of the data, we can carry out this step using our result for model-based policy evaluation (cf. Lemma 1). Specifically, taking $\pi = \pi^{\star}$ in Lemma 1 yields that, with probability at least $1-\delta$,

$$\big\|\widehat{\boldsymbol{V}}^{\pi^{\star}} - \boldsymbol{V}^{\pi^{\star}}\big\|_{\infty} \leq 6\sqrt{\frac{2\log\big(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\big)}{N(1-\gamma)^3}}. \tag{18}$$

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 $\widehat{\boldsymbol{P}}$. The recent work [1] developed a clever "leave-one-out" type argument by constructing some auxiliary state-absorbing MDPs to decouple the statistical dependency. However, their argument requires $\varepsilon < 1/\sqrt{1-\gamma}$ and falls short of accommodating the full ε-range. To address this challenge, our analysis consists of the following steps, all of which require new techniques beyond [1].

- *Decoupling statistical dependency between $\widehat{\pi}^{\star}$ and $\widehat{\boldsymbol{P}}$.* Instead of decoupling the statistical dependency between $\widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}$ and $\widehat{\boldsymbol{P}}$ as in [1], we attempt to decouple the dependency between the policy $\widehat{\pi}^{\star}$ and $\widehat{\boldsymbol{P}}$. If this could be achieved, then the proof strategy proposed for fixed policies would be applicable. The key ingredient of this step lies in the construction of auxiliary state-action-absorbing MDPs (motivated by [1]), as well as the development of optimal value function estimation guarantees under certain Berstein-type conditions (rather than independence assumptions). These ingredients together allow us to get hold of $\|\boldsymbol{V}^{\widehat{\pi}^{\star}} - \widehat{\boldsymbol{V}}^{\widehat{\pi}^{\star}}\|_{\infty}$, provided a certain separation condition holds.

- *Tie-breaking via reward perturbation.* An inevitable shortcoming of the above-mentioned approach 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. We show that this can be guaranteed by randomly perturbing the reward function (so as to break the ties).

4.2 More detailed descriptions of Step 2

We shall now flesh out the key ideas mentioned above for Step 2, consisting of three sub-steps.

Step 2.1: value function estimation for a policy obeying Bernstein-type conditions. We start by pointing out that Lemma 1 can be generalized far beyond the family of fixed policies (namely, those independent of $\widehat{\boldsymbol{P}}$), as long as a certain Bernstein-type condition — which captures certain weak statistical dependency and will be formalized in (20) — is satisfied. To make it precise, we introduce:

$$\begin{aligned} \boldsymbol{r}^{(0)} &:= \boldsymbol{r}_{\pi}, & \boldsymbol{V}^{(0)} &:= (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}^{(0)}, & & \\ \boldsymbol{r}^{(l)} &:= \sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\big[\boldsymbol{V}^{(l-1)}\big]}, & \boldsymbol{V}^{(l)} &:= (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}^{(l)}, & l &\geq 1. \end{aligned} \tag{19}$$

Our generalization of Lemma 1 is as follows, which does *not* require independence between π and $\widehat{\boldsymbol{P}}$.

Lemma 2. *Suppose that there exists some quantity $\beta_1 > 0$ such that $\{\boldsymbol{V}^{(l)}\}$ (cf. (19)) obeys*

$$\Big|\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{(l)}\Big| \leq \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\big[\boldsymbol{V}^{(l)}\big]} + \frac{\beta_1\,\|\boldsymbol{V}^{(l)}\|_{\infty}}{N}\boldsymbol{1}, \qquad \forall\, 0 \leq l \leq \log\Big(\frac{e}{1-\gamma}\Big). \tag{20}$$

Suppose $N > \frac{16e^2}{1-\gamma}\beta_1$. Then one has $\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} \leq \frac{6}{1-\gamma}\sqrt{\frac{\beta_1}{N(1-\gamma)}}$.

Step 2.2: decoupling statistical dependency via (s,a)-absorbing MDPs. We are now positioned to show how to decouple the complicated statistical dependency between the optimal policy $\widehat{\pi}^{\star}$ and $\widehat{\boldsymbol{V}}^{\star}$. Towards this, we resort to a leave-one-row-out argument, largely motivated by the novel construction in [1]. In comparison to [1] that introduces state-absorbing MDPs (so that a state s is absorbing regardless of the subsequent actions), 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. Specifically, for each state-action pair (s,a) and each scalar u with $|u| \leq 1/(1-\gamma)$, construct an auxiliary MDP $\widehat{\mathcal{M}}_{s,a,u}$ — it is identical to the original $\widehat{\mathcal{M}}$ except that it is absorbing in state s if we always choose action a in state s; namely, its probability transition kernel obeys

$$\begin{aligned} P_{\widehat{\mathcal{M}}_{s,a,u}}(s \mid s,a) &= 1, & & \\ P_{\widehat{\mathcal{M}}_{s,a,u}}(s' \mid s,a) &= 0, & &\text{for all } s' \neq s, \\ P_{\widehat{\mathcal{M}}_{s,a,u}}(\cdot \mid s',a') &= P_{\widehat{\mathcal{M}}}(\cdot \mid s',a'), & &\text{for all } (s',a') \neq (s,a), \end{aligned} \tag{21}$$

where $P_{\widehat{\mathcal{M}}}$ is the transition kernel of $\widehat{\mathcal{M}}$. The instant reward received at (s,a) in $\widehat{\mathcal{M}}_{s,a,u}$ is set to be u, while all other rewards stay unchanged. Let $\widehat{\pi}^{\star}_{s,a,u}$ be the optimal policy in $\widehat{\mathcal{M}}_{s,a,u}$. The main advantage is: for any fixed u, the MDP $\widehat{\mathcal{M}}_{s,a,u}$ is statistically independent of the (s,a)-th row of $\widehat{\boldsymbol{P}}$.

It turns out that we can **represent $\widehat{\pi}^{\star}$ via a few policies independent of $\widehat{\boldsymbol{P}}_{s,a}$**, as long as the original $\widehat{\boldsymbol{Q}}^{\star}$ satisfies a sort of separation condition (which indicates that there is no tie when it comes to the optimal policy). More precisely, given any $0 < \omega < 1$ we define the separation condition as

$$\mathcal{B}_{\omega} := \Big\{ \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s,a) \geq \omega \;\text{ for all } s \in \mathcal{S} \Big\}. \tag{22}$$

Our result is stated below. Here, $\mathcal{N}_{\epsilon}$ is an ϵ-net of the interval $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$.

Lemma 3. *Consider any $\omega > 0$, and suppose the event $\mathcal{B}_\omega$ (cf. (22)) 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*

$$\widehat{\pi}^\star = \widehat{\pi}^\star_{s,a,u_0}. \tag{23}$$

Step 2.3: a tie-breaking argument. Unfortunately, the separation condition specified in (22) 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, 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 it, we need to introduce additional notation. Denote by $\pi^\star_{\mathrm{p}}$ the optimal policy of the MDP $\mathcal{M}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, r_{\mathrm{p}}, \gamma)$, 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 $\widehat{\mathcal{M}}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \widehat{\boldsymbol{P}}, r_{\mathrm{p}}, \gamma)$. Our result is phrased as follows.

Lemma 4. *Consider the perturbed reward defined in (6). With probability at least $1-\delta$,*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} \text{ with } a \neq \pi^\star_{\mathrm{p}}(s): \qquad Q^\star_{\mathrm{p}}(s, \pi^\star_{\mathrm{p}}(s)) - Q^\star_{\mathrm{p}}(s,a) > \frac{\xi\delta(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2}. \tag{24}$$

This result holds unchanged if $(Q^\star_{\mathrm{p}}, \pi^\star_{\mathrm{p}})$ is replaced by $(\widehat{Q}^\star_{\mathrm{p}}, \widehat{\pi}^\star_{\mathrm{p}})$.

Lemma 4 reveals that at least a polynomially small degree of separation ($\omega = \frac{\xi\delta(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2}$) arises upon random perturbation (with size ξ) of the reward function. As we shall see, this level of separation suffices for our purpose, which together with Steps 2.1-2.2 allows us to justify our main theorems.

5 Concluding remarks

This paper uncovers that a *perturbed* model-based planning algorithm enables minimax sample complexity in a generative model, as soon as the sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ (up to 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 accuracy. The present work opens up several directions for future investigation. For instance, a natural question arises regarding the necessity of perturbation — is it possible to achieve optimality by directly performing planning on the original empirical MDP without perturbation? Our analysis framework might also shed light on how to improve sample complexity guarantees for model-free algorithms and finite-horizon MDPs.

Broader Impact

This work is a theoretical contribution to characterize the minimax optimality of model-based reinforcement learning. The insights from the proposed algorithm can potentially be leveraged in various reinforcement learning tasks in the future.

Acknowledgments and Disclosure of Funding

G. Li and Y. Gu are supported in part by the grant NSFC-61971266. 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. We thank Qiwen Cui for pointing out an issue in the proof of Lemma 4 in an early version of this paper, and thank Shicong Cen, Chen Cheng and Cong Ma for numerous discussions.

References

[1] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804*, 2019.

[2] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[3] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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

[5] Dimitri P Bertsekas. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 2017.

[6] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

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

[8] Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, pages 11312–11322, 2019.

[9] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

[10] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

[11] Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

[12] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

[13] Harsh Gupta, R Srikant, and Lei Ying. 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, 2019.

[14] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

[15] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[16] SM Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, 2003.

[17] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. *arXiv preprint arXiv:2002.01268*, 2020.

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

[19] Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002, 1999.

[20] Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

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

[22] Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

[23] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

[24] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Neural Information Processing Systems*, 2020.

[25] Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *arXiv preprint arXiv:2004.04719*, 2020.

[26] Ashwin Pananjady and Martin J Wainwright. Value function estimation in Markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. *arXiv preprint arXiv:1909.08749*, 2019.

[27] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*, 2020.

[28] Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121, 2018.

[29] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. 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, 2018.

[30] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. 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, 2018.

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

[32] Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

[33] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[34] Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070, 1998.

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

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

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

[38] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*, 2018.

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

[40] Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

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

[42] Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

[43] Pan Xu and Quanquan Gu. A finite-time analysis of Q-learning with neural network function approximation. *accepted to International Conference on Machine Learning*, 2020.

[44] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643, 2019.

[45] Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.