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 $\gamma$-discounted infinite-horizon Markov decision process (MDP) with state space $S$ and action space $A$, 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 $|S||A| (1 - \gamma)^2$ (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 $|S||A|$ (1 - $\gamma$) (modulo some log factor). More specifically, a perturbed model-based planning algorithm provably finds an $\varepsilon$-optimal policy with an order of $|S||A| (1 - \gamma)^3 \varepsilon^2 \log |S||A| (1 - \gamma)$ 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. 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 $S$, action space $A$, 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 $\varepsilon$-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 $|S|^2|A|$ of model parameters (so that it might be infeasible to estimate the model accurately).

1Here, the “model” refers to the probability transition kernel and the rewards of the MDP taken collectively.
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 \( \frac{|S||A|}{(1-\gamma)^2} \) (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 \( \frac{|S||A|}{(1-\gamma)^2} \) (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 \( \varepsilon \)-optimal policy with an order of \( \frac{|S||A|}{(1-\gamma)^2} \) samples (up to some log factor), which matches the minimax lower bound [2]. Our result accommodates 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 \( \mathcal{M} = (S, A, P, r, \gamma) \), where \( S := \{1, 2, \ldots, |S|\} \) denotes a finite set of states, \( A := \{1, 2, \ldots, |A|\} \) a finite set of actions, \( \gamma \in (0, 1) \) the discount factor, and \( r : S \times A \to [0, 1] \) 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, \( P : S \times A \to \Delta(S) \) represents the probability transition kernel, where...
\( P(s'|s, a) \) denotes the probability of transitioning from state \( s \) to state \( s' \) when action \( a \) is executed, and \( \Delta(S) \) denotes the probability simplex over \( S \).

A deterministic policy is a mapping \( \pi : S \rightarrow A \), which maps a state to an action. The value function

\[
V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s^t, a^t) \Big| s^0 = s \right],
\]

which is the expected discounted total reward starting from the initial state \( s^0 = s \), with the actions taken according to the policy \( \pi \) (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) \( Q^\pi : S \times A \rightarrow \mathbb{R} \) of a policy \( \pi \) is defined by

\[
Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s^t, a^t) \Big| s^0 = s, a^0 = a \right],
\]

where the actions are taken according to the policy \( \pi \) after the initial action (i.e. \( a^0 = \pi(s^0) \) for all \( t \geq 1 \)). It is known that there exists an optimal policy, denoted by \( \pi^* \), that maximizes \( V^\pi(s) \) (resp. \( Q^\pi(s, a) \)) for all \( s \in S \) (resp. \( (s, a) \in (S \times A) \)) [33]. The corresponding value function \( V^* := V^{\pi^*} \) (resp. Q-function \( Q^* := Q^{\pi^*} \)) 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 \( s^i, a^i \sim P(\cdot|s^i, a^i), i = 1, \ldots, N \) for any state-action pair \( (s, a) \), which allows us to construct an empirical transition kernel \( \hat{P} \) given by

\[
\forall s' \in S, \quad \hat{P}(s'|s, a) = \frac{1}{N} \sum_{i=1}^{N} I\{s_i = s'\},
\]

where \( I\{\cdot\} \) is the indicator function. In words, \( \hat{P}(s'|s, a) \) counts the empirical frequency of transitions from \( (s, a) \) to state \( s' \). The total sample size should be understood as \( N^{\text{total}} := N|S||A| \). This leads to an empirical MDP \( \hat{M} = (S, A, \hat{P}, r, \gamma) \) constructed from the data samples. We can define the value function and the action-value function of a policy \( \pi \) for \( \hat{M} \) analogously, which we shall denote by \( \hat{V}^\pi \) and \( \hat{Q}^\pi \), respectively. The optimal policy of \( \hat{M} \) is denoted by \( \hat{\pi}^* \), with the optimal value function and Q-function denoted by \( \hat{V}^*: = \hat{V}^{\hat{\pi}^*} \) and \( \hat{Q}^*: = \hat{Q}^{\hat{\pi}^*} \), 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 \( V^\pi \) under this policy. To be precise, for any target level \( \varepsilon > 0 \), the goal is to find an \( \varepsilon \)-accurate estimate \( V^\pi_{\text{est}} \) such that

\[
(\text{policy evaluation}) \quad \forall s \in S : \quad |V^\pi_{\text{est}}(s) - V^\pi(s)| \leq \varepsilon.
\]

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 \( \varepsilon \)-optimal policy \( \pi_{\text{est}} \) obeying

\[
(\text{planning}) \quad \forall (s, a) \in S \times A : \quad V^{\pi_{\text{est}}}(s) \geq V^*(s) - \varepsilon, \quad Q^{\pi_{\text{est}}}(s, a) \geq Q^*(s, a) - \varepsilon.
\]

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 \([0, 1]\); this means that the range of accuracy level \( \varepsilon \) should be set to \( \varepsilon \in [0, \frac{1}{1 - \gamma}] \). A model-based approach starts by constructing an empirical MDP \( \hat{M} \) 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 \( X := ([S], |A|, \frac{1}{1 - \gamma}, \frac{1}{1 - \gamma}) \). The notation \( f(X) = O(g(X)) \) means that there exists a constant \( C_f > 0 \) such that \( f \leq C_f g \), whereas \( f(X) = \Omega(g(X)) \) means \( g(X) = O(f(X)) \). The notation \( \tilde{O}(\cdot) \) (resp. \( \tilde{\Omega}(\cdot) \)) is defined in the same way as \( O(\cdot) \) (resp. \( \Omega(\cdot) \)) except that it ignores log factors. For any vector \( a = [a_i]_{1 \leq i \leq n} \in \mathbb{R}^n \), we overload the notation \( \sqrt{a} \) and \( |a| \) in an entry-wise manner s.t. \( \sqrt{a} := [\sqrt{a_i}]_{1 \leq i \leq n} \) and \( |a| := |a_i|_{1 \leq i \leq n} \). For any vectors \( a = [a_i]_{1 \leq i \leq n} \) and \( b = [b_i]_{1 \leq i \leq n} \), the notation \( a \geq b \) (resp. \( a \leq b \)) means \( a_i \geq b_i \) (resp. \( a_i \leq b_i \)) for all \( 1 \leq i \leq n \), and we let \( a \odot b := [a_i b_i]_{1 \leq i \leq n} \) represent the Hadamard product. We denote by \( 1 \) 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 \((s, a) \in S \times A\), let us perturb the immediate reward (with uniform distribution) to obtain

\[
r_p(s, a) = r(s, a) + \xi(s, a), \quad \xi(s, a) \sim \text{Unif}(0, \xi),
\]

where \(\xi > 0\) is a parameter to be specified shortly.\(^2\) For any policy \(\pi\), denote by \(\hat{V}_p^\pi\) the corresponding value function of the perturbed empirical MDP \(\hat{M}_p = (S, A, \hat{P}, r_p, \gamma)\) with the probability transition kernel \(\hat{P}\) and the perturbed reward function \(r_p\). Let \(\hat{\pi}_p^*\) represent the optimal policy of \(\hat{M}_p\), i.e.

\[
\hat{\pi}_p^* := \arg \max_{\pi} \hat{V}_p^\pi.
\]

Encouragingly, this policy results in a value function \(V^\pi\) (resp. Q-function \(Q^\pi\)) that is \(\varepsilon\)-close to the true optimal value function \(V^*\) (resp. optimal Q-function \(Q^*\)), 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 \(\hat{\pi}_p^*\) (cf. (7)) obeys

\[
\forall (s, a) \in S \times A, \quad V^\pi(s) \geq V^*(s) - \varepsilon \quad \text{and} \quad Q^\pi(s, a) \geq Q^*(s, a) - \gamma \varepsilon,
\]

provided that the perturbation obeys \(\xi = \frac{c_1(1-\gamma)\varepsilon}{|S||A|^2}\) and the sample size per state-action pair obeys

\[
N \geq c_0 \log \left( \frac{|S||A|}{(1-\gamma)^2}\right).
\]

In addition, the empirical QVI (cf. \([2]\)) recovers \(\hat{\pi}_p^*\) perfectly within \(O\left( \frac{1}{1-\gamma} \log \left( \frac{|S||A|}{(1-\gamma)^2}\right) \right)\) iterations.

**Remark 1.** Theorem 1 holds unchanged if \(\xi\) is taken to be \(\frac{c_1(1-\gamma)\varepsilon}{|S||A|^2}\) 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 \(\hat{\pi}_p^*\) of the perturbed empirical MDP \(\hat{M}_p\), 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 \(\hat{P}\) (which is a sparse matrix with at most \(N|S||A|\) nonzeros), hence the resulting computational complexity can be as low as \(O\left( \frac{|S||A|}{(1-\gamma)^2}\right)\).

Theorem 1 demonstrates that: the proposed perturbed model-based approach finds an \(\varepsilon\)-optimal policy as soon as the total sample complexity exceeds the order of \(\frac{|S||A|}{(1-\gamma)^2}\) (modulo some log factor). Compared to prior literature, our result imposes no restriction on the range of \(\varepsilon\) and, in particular, we allow the accuracy level \(\varepsilon\) 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

\[
N \geq \Omega\left( \frac{1}{1-\gamma} \right).
\]

Tackling the sample-limited regime (particularly the scenario when \(N \in \left[ \frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2} \right]\)) 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 \(\varepsilon = O\left( \frac{1}{1-\gamma}\right)\). This in turn corroborates the minimax optimality of our perturbed model-based approach for the full \(\varepsilon\)-range (which is previously unavailable), and demonstrates the information-theoretic infeasibility to learn a policy strictly better than a random guess if \(N \leq \tilde{O}\left( \frac{1}{1-\gamma} \right)\). 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.
As can be seen, the numerical sample complexity per state-action pair scales on the order of $1$ (including both model-based and model-free methods) was capable of efficiently finding the desired accuracy 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 $\varepsilon$ 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 $\varepsilon$, which is consistent with our theory.

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

$$\max_{s \in S} |\hat{V}_\pi(s) - V_\pi(s)| \leq \varepsilon$$

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

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

In words, this theorem reveals that $\hat{V}_\pi$ begins to outperform a random guess as soon as $N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^3}\right)$. The sample complexity bound (12) enjoys full coverage of the $\varepsilon$-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 $N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^2}\right)$, whereas the theory for the variance-reduced model-free approach [29, 40] imposed the sample size requirement $N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^3}\right)$. In fact, it was previously unknown what is achievable in the sample size range $N \in \left[\frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2}\right]$. In contrast, our results confirm the minimax-optimal statistical performance of the model-based approach with full coverage of the $\varepsilon$-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 $N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^3}\right)$, which is more stringent than our result by a factor of at least.

![Figure 1: Numerical sample complexity per state-action pair $N$ vs. discount complexity $\frac{1}{1-\gamma}$](image)
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 \( \varepsilon \in \left[ \frac{1}{\sqrt{1-\gamma}}, \frac{1}{1-\gamma} \right] \) 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 \( N = \Omega(\frac{|S||A| \log(|S||A|)}{(1-\gamma)^2}) \) necessary for finding an \( \varepsilon \)-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 \( \varepsilon \)-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 \( \frac{1}{1-\gamma} \) 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 \( \frac{|S^2||A|}{(1-\gamma)^2} \) 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 \( \Omega(\frac{|S||A| \log(|S||A|)}{(1-\gamma)^2}) \), \( \infty \). 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 \( e_1, \ldots, e_{|S|} \in \mathbb{R}^{|S|} \) the standard basis vectors, we can define \( r \in \mathbb{R}^{|S||A|} \), \( V^\pi \in \mathbb{R}^{|S|} \), \( Q^\pi \in \mathbb{R}^{|S||A|} \) as the vector representations of the reward function \( r \), the value function \( V^\pi \) and the Q-function \( Q^\pi \). Let \( V^\pi \) and \( Q^\pi \in \mathbb{R}^{|S||A|} \) be the optimal value function \( V^* \) and optimal Q-function \( Q^* \). Let \( P \in \mathbb{R}^{|S||A| \times |S|} \) be a matrix representing the probability transition kernel \( P \), where the \((s, a)\)-th row of \( P \) is a probability vector representing \( P(\cdot | s, a) \).

We introduce \( \Pi^\pi = \left( \begin{array}{c} e^\pi_{r(1)} \vspace{1em} \vdots e^\pi_{r(|S|)} \end{array} \right) \in \{0, 1\}^{|S| \times |S||A|} \) as the projection matrix associated with policy \( \pi \). The matrices \( P^\pi := \Pi^\pi P \in \mathbb{R}^{|S||A| \times |S|} \) and \( P_\pi := \Pi^\pi P \in \mathbb{R}^{|S| \times |S|} \) are two square probability transition matrices induced by policy \( \pi \) over the state-action pairs and the states, respectively. Let \( r_\pi \in \mathbb{R}^{|S|} \) be a reward vector restricted to the actions chosen by policy \( \pi \), i.e. \( r_\pi = \Pi^\pi r \). For a vector \( V = [V_i]_{1 \leq i \leq |S|} \in \mathbb{R}^{|S|} \), we define the vectors

\[
\text{Var}_P(V) := P(V \circ P) - (PV) \circ (PV) \in \mathbb{R}^{|S||A|}; \quad \text{Var}_{P_\pi}(V) := \Pi^\pi \text{Var}_P(V) \in \mathbb{R}^{|S|}.
\]

We shall also define \( \hat{V}^\pi, \hat{Q}^\pi, \hat{V}^*, \hat{Q}^*, \hat{P}, \hat{P}_\pi, \hat{V}_\pi, \text{Var}_P(V), \text{Var}_{P_\pi}(V) \) w.r.t. \( \hat{M} \) 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 \( \pi \) and the model-based empirical estimate can be written respectively as

\[
V^\pi = (I - \gamma P_\pi)^{-1} r_\pi \quad \text{and} \quad \hat{V}^\pi = (I - \gamma \hat{P}_\pi)^{-1} r_\pi . \tag{13}
\]
Lemma 1. Fix any policy $\pi$. Consider any $0 < \delta < 1$, and suppose $N \geq \frac{2e^2}{3\gamma} \log \left( \frac{4|S| \log(e/\delta)}{\delta} \right)$. Then with probability at least $1 - \delta$, the vectors defined in (13) obey

$$
\|\hat{V}^* - V^*\|_\infty \leq 32 \log \frac{4|S| \log(e/\delta)}{N \delta} \left\| (I - \gamma P_{\pi})^{-1} \right\|_\infty \sqrt{\text{Var}_{P_{\pi}}[V^\pi]} \|_{\infty} + 2 \gamma \log \frac{4|S| \log(e/\delta)}{1 - \gamma} N \|V^*\|_{\infty} 
\leq 6 \sqrt{2 \log \frac{4|S| \log(e/\delta)}{N (1 - \gamma)^3}}.
$$

(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 $P_{\pi}, V^\pi$), 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 $\log \frac{1}{1 - \gamma}$ factor, while being capable of covering the full sample size range $N \geq \Omega(\frac{1}{(1 - \gamma)^3})$. In contrast, [26, Theorem 1] is only valid when $N \geq \Omega(\frac{1}{(1 - \gamma)^{2\gamma}})$.

The key proof idea is to resort to a high-order successive expansion of $\hat{V}^* - V^*$, 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, perturbation relation

$$
\hat{V}^* - V^* = (I - \gamma P_{\pi})^{-1} r_{\pi} - V^* = \gamma (I - \gamma \hat{P}_{\pi})^{-1} (\hat{P}_{\pi} - P_{\pi}) V^*. 
$$

(15)

Due to the complicated statistical dependency between $(I - \gamma \hat{P}_{\pi})^{-1}$ and $(\hat{P}_{\pi} - P_{\pi}) V^*$, 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

$$
(15) = \gamma (I - \gamma P_{\pi})^{-1} (\hat{P}_{\pi} - P_{\pi}) V^* + \gamma \left\{ (I - \gamma \hat{P}_{\pi})^{-1} - (I - \gamma P_{\pi})^{-1} \right\} (\hat{P}_{\pi} - P_{\pi}) V^*, 
$$

(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 $(I - \gamma P_{\pi})^{-1}$ and $(\hat{P}_{\pi} - P_{\pi}) V^*$. 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:

$$
V^* - \hat{V}^\pi = (\hat{V}^\pi - V^\pi) + (\hat{V}^\pi - \hat{V}^{\hat{\pi}}) + (V^* - \hat{V}^{\hat{\pi}}) 
\leq (\hat{V}^\pi - V^\pi) + (V^* - \hat{V}^{\hat{\pi}}),
$$

(17)

which uses $\hat{V}^{\hat{\pi}} \leq \hat{V}^\pi$ due to the optimality of $\hat{\pi}^\star$. This leaves us with two terms to control.

Step 1: bounding $\|V^\pi - \hat{V}^\pi\|_\infty$. 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 $1 - \delta$,

$$
\|\hat{V}^{\hat{\pi}} - V^\pi\|_\infty \leq 6 \sqrt{2 \log \frac{4|S| \log(e/\delta)}{N (1 - \gamma)^3}}.
$$

(18)

Step 2: bounding $\|\hat{V}^\pi - \hat{V}^{\hat{\pi}}\|_\infty$. Extending the result in Step 1 to $\|\hat{V}^\pi - \hat{V}^{\hat{\pi}}\|_\infty$ is considerably more challenging, primarily due to the complicated statistical dependency between $(V^\pi, \hat{V}^{\hat{\pi}})$ and $\hat{P}$. 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 $\varepsilon$-range. To address this challenge, our analysis consists of the following steps, all of which require new techniques beyond [1].
We shall now flesh out the key ideas mentioned above for Step 2, consisting of three sub-steps.

**Step 2.2: decoupling statistical dependency via optimal policy.** More precisely, given any \( \hat{\pi}^* \) to show how to decouple the complicated statistical dependency between the optimal policy \( \hat{\pi}^* \) and \( \hat{V}^* \). To make it precise, we introduce:

\[
\begin{align*}
\hat{r}_0 & := r_\pi, \\
\hat{r}_l & := \sqrt{\text{Var}_{P_\pi}[V^{l-1}]}, \\
\hat{V}_0 & := (I - \gamma P_\pi)^{-1} r_0, \\
\hat{V}_l & := (I - \gamma P_\pi)^{-1} r_l, \quad l \geq 1.
\end{align*}
\]

Our generalization of Lemma 1 is as follows, which does not require independence between \( \pi \) and \( \hat{P} \).

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

\[
\left|\hat{P}_\pi - P_\pi\right| V^{(l)} \leq \sqrt{\frac{\beta_1}{N}} \sqrt{\text{Var}_{P_\pi}[V^{(l)}]} + \frac{\beta_1}{N} \frac{\|V^{(l)}\|_\infty}{1}, \quad 0 \leq l \leq \log\left(\frac{e}{1-\gamma}\right).
\]

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

**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 \( \hat{\pi}^* \) and \( \hat{V}^* \). 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 \( \hat{M}_{s,a,u} \) — it is identical to the original \( \hat{M} \) except that it is absorbing in state \( s \) if we always choose action \( a \) in state \( s \); namely, its probability transition kernel obeys

\[
\begin{align*}
P_{\hat{M}_{s,a,u}}(s'|s,a) & = 1, \\
P_{\hat{M}_{s,a,u}}(s'|s,a) & = 0, \quad \text{for all } s' \neq s, \\
P_{\hat{M}_{s,a,u}}(s'|s',a') & = P_{\hat{M}}(s'|s,a), \quad \text{for all } (s',a') \neq (s,a),
\end{align*}
\]

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

It turns out that we can represent \( \hat{\pi}^* \) via a few policies independent of \( \hat{P}_{s,a} \), as long as the original \( \tilde{Q}^* \) 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

\[
\mathcal{B}_\omega := \left\{ \tilde{Q}^*(s, \hat{\pi}^*(s)) - \max_{a: a \neq \hat{\pi}^*(s)} \tilde{Q}^*(s, a) \geq \omega \text{ for all } s \in S \right\}.
\]

Our result is stated below. Here, \( \mathcal{N} \) is an \( \epsilon \)-net of the interval \( \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right] \).
Lemma 3. Consider any $\omega > 0$, and suppose the event $B_\omega$ (cf. (22)) holds. Then for any pair $(s, a) \in S \times A$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$, such that
\[ \hat{\pi}^* = \hat{\pi}_{s,a,u_0}^*. \] (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 $\pi_p^*$ the optimal policy of the MDP $\mathcal{M}_p = (S, A, P, r_p, \gamma)$, and $Q_p^*$ its optimal state-action value function. We can define $\hat{Q}_p^*$ and $\hat{\pi}_p^*$ analogously for the MDP $\hat{\mathcal{M}}_p = (S, A, \hat{P}, r_p, \gamma)$. Our result is phrased as follows.

Lemma 4. Consider the perturbed reward defined in (6). With probability at least $1 - \delta$, \[ \forall (s, a) \in S \times A \text{ with } a \neq \pi_p^*(s) : \quad Q_p^*(s, \pi_p^*(s)) - Q_p^*(s, a) > \frac{\xi \delta (1 - \gamma)}{4 |S| |A|^2}. \] (24) This result holds unchanged if $(Q_p^*, \pi_p^*)$ is replaced by $(\hat{Q}_p^*, \hat{\pi}_p^*)$.

Lemma 4 reveals that at least a polynomially small degree of separation ($\omega = \frac{\xi \delta (1 - \gamma)}{4 |S| |A|^2}$) arises upon random perturbation (with size $\xi$) 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 $|S| |A|^{1/(1-\gamma)}$ (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 ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, and NSF CCF-1806154 and CCF-207911. 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-207911. 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


