Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Gen Li, Yuting Wei, Yuejie Chi, Yuxin Chen

To cite this article:
Gen Li, Yuting Wei, Yuejie Chi, Yuxin Chen (2024) Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. Operations Research 72(1):203-221. https://doi.org/10.1287/opre.2023.2451

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 72, No. 1, January–February 2024, pp. 203–221
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

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

Gen Li,[a] Yuting Wei,[a,*] Yuejie Chi,[b] Yuxin Chen[a,c]

[a] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [c] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104

*Corresponding author

Contact: ligen@wharton.upenn.edu (GL); ytwei@wharton.upenn.edu, https://orcid.org/0000-0003-1488-4647 (YW); yuejiechi@cmu.edu, https://orcid.org/0000-0002-6766-5459 (YueC); yuxinc@wharton.upenn.edu, https://orcid.org/0000-0001-9256-5815 (YuxC)

Received: December 25, 2021
Revised: September 5, 2022
Accepted: March 16, 2023
Published Online in Articles in Advance: April 21, 2023

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2023.2451

Abstract. This paper is concerned with the sample efficiency of reinforcement learning, assuming access to a generative model (or simulator). We first consider γ-discounted infinite-horizon Markov decision processes (MDPs) with state space $\mathcal{S}$ and action space $\mathcal{A}$. Despite a number of prior works tackling this problem, a complete picture of the trade-offs between sample complexity and statistical accuracy has yet to be determined. In particular, all prior results suffer from a severe sample size barrier in the sense that their claimed statistical guarantees hold only when the sample size exceeds at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$. The current paper overcomes this barrier by certifying the minimax optimality of two algorithms—a *perturbed* model-based algorithm and a *conservative* model-based algorithm—as soon as the sample size exceeds the order of $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ (modulo some log factor). Moving beyond infinite-horizon MDPs, we further study time-inhomogeneous finite-horizon MDPs and prove that a plain model-based planning algorithm suffices to achieve minimax-optimal sample complexity given any target accuracy level. To the best of our knowledge, this work delivers the first minimax-optimal guarantees that accommodate the entire range of sample sizes (beyond which finding a meaningful policy is information theoretically infeasible).

Funding: Y. Wei is supported in part by the Google Research Scholar Award and the National Science Foundation [Grants CCF-2106778, DMS-2147546, and DMS-2143215]. Y. Chi is supported in part by the Office of Naval Research [Grants N00014-18-1-2142 and N00014-19-1-2404] and the National Science Foundation [Grants CCF-1806154, CCF-2007911, and CCF-2106778]. Y. Chen is supported in part by the Alfred P. Sloan Foundation [research fellowship], Google [research scholar award], the Air Force Office of Scientific Research [Grants FA9550-19-1-0030 and FA9550-22-1-0198], the Office of Naval Research [Grant N00014-22-1-2354], and the National Science Foundation [Grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713, and IIS-2218773].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.2451.

Keywords: model-based reinforcement learning • minimaxity • policy evaluation • generative model

1. Introduction

Reinforcement learning (RL) (Szepesvári 2010, Sutton and Barto 2018), which is frequently modeled as learning and decision making in a Markov decision process (MDP), has garnered growing interest in recent years because of 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 cumulative 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 fundamental trade-off

between sample complexity and statistical accuracy and to design efficient algorithms that provably achieve the optimal trade-off.

Broadly speaking, there are at least two common algorithmic approaches: a model-based approach and a model-free one. The model-based approach decouples model estimation and policy learning tasks; more specifically, one first estimates 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). A notable 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 well to scenarios when a model is difficult to estimate or changes on the fly. Characterizing the sample efficiency of both approaches has been the focal point of a large body of recent works (e.g., Kearns and Singh 1999; Azar et al. 2013; Jin et al. 2018; Sidford et al. 2018a, b; Tu and Recht 2019; Wainwright 2019a, b; Agarwal et al. 2020; Li et al. 2023).

In this paper, we pursue a comprehensive understanding of model-based RL given access to a generative model—that is, a simulator that produces samples based on the transition kernel of the true MDP for each state-action pair (Kearns and Singh 1999, Kakade 2003). To allow for more precise discussions, we first look at an infinite-horizon discounted MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, and discount factor $0 < \gamma < 1$ and pay particular attention to the scenarios where the sizes of the state/action spaces and the effective horizon $\frac{1}{1-\gamma}$ are all quite large. 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 sublinear 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 is in general infeasible to estimate the model accurately).

1.1. Motivation: Sample Size Barriers

Several prior works were dedicated to investigating model-based RL for γ-discounted infinite-horizon MDPs with a generative model, which uncovered the minimax optimality of this approach for an already wide regime (Azar et al. 2013, Agarwal et al. 2020). 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 of Agarwal et al. (2020) 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 (Pananjady and Wainwright 2019, Agarwal et al. 2020). Furthermore, an even more severe barrier emerged in prior theory for model-free methods; for instance, Sidford et al. (2018a) and Wainwright (2019b) required the sample size to exceed $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$ 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, thus resulting in a gap between upper and lower bounds in this sample-starved regime. Noteworthily, such a sample size barrier is not only present for discounted infinite-horizon MDPs; the situation is similar for finite-horizon MDPs (Yin et al. 2021).

1.2. Our Contributions

The current paper seeks to achieve optimal sample complexity even below the aforementioned sample size barrier. For γ-discounted infinite-horizon MDPs, we propose two model-based algorithms: (i) *perturbed model-based planning*, which performs planning based on an empirical MDP learned from samples with mild random *reward perturbation*, and (ii) *conservative model-based planning*, which computes approximately optimal policies for the empirical MDP without reward perturbation. These two proposed algorithms provably find an ε-optimal policy with an order of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ samples (up to log factor), thereby matching the minimax lower bound (Azar et al. 2013). Our result accommodates the full range of accuracy level ε $\left(\text{namely, } \varepsilon \in \left(0, \frac{1}{1-\gamma}\right]\right)$, thus unveiling the minimaxity of our algorithms 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 literature. Along the way, we also derive minimax-optimal statistical guarantees for policy evaluation, which strengthen state-of-the-art results by broadening the applicable sample size range.

Moving beyond discounted infinite-horizon MDPs, we further characterize the sample efficiency of model-based planning for time-inhomogeneous finite-horizon MDPs, which provably achieves minimax-optimal sample complexity as well for the full range of target accuracy levels (Domingues et al. 2021). No reward perturbation or conservative action selection is needed for this finite-horizon scenario. See Table 2 for detailed comparisons with prior literature.

Table 1. Comparisons with Prior Results (up to Log Factors) Regarding Finding an ε-Optimal Policy in a γ-Discounted Infinite-Horizon MDP with a Generative Model

Algorithm	Sample size range	Sample complexity	ε range
Phased Q learning (Kearns and Singh 1999)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\varepsilon^2}$	$\left(0,\frac{1}{1-\gamma}\right]$
Empirical QVI (Azar et al. 2013)	$\left[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\left(0,\frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}\right]$
Sublinear randomized value iteration (Sidford et al. 2018a)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\left(0,\frac{1}{1-\gamma}\right]$
Variance-reduced QVI (Sidford et al. 2018b)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0,1]$
Randomized primal-dual method (Wang 2020)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\left(0,\frac{1}{1-\gamma}\right]$
Empirical MDP + planning (Agarwal et al. 2020)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\left(0,\frac{1}{\sqrt{1-\gamma}}\right]$
Perturbed empirical MDP + planning (this paper)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{1-\gamma},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\left(0,\frac{1}{1-\gamma}\right]$
Empirical MDP + *conservative* planning (this paper)	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{1-\gamma},\infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\left(0,\frac{1}{1-\gamma}\right]$

Notes. 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) and Wang (2020) only hold for a restricted family of MDPs satisfying certain ergodicity assumptions. In addition, Azar et al. (2013) (Wainwright 2019b) showed that empirical QVI (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 $\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3},\infty\right)$, which did not translate directly to an ε-optimal policy.

On the technical side, our theory for infinite-horizon MDPs 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. (2020)—that help decouple the complicated statistical dependency between the empirically optimal policy (as opposed to value functions) and data samples; and (3) a tiebreaking argument guaranteeing that the empirically optimal policy is sufficiently separated from all other policies under reward perturbation. The case with finite-horizon MDPs is also established based on certain high-order expansion of the value estimation errors, in addition to careful variance control for the terms in the expansion.

1.3. Key Contributions that Extend the Neural Information Processing Systems Version

Partial results of this paper have been presented in *Advances in Neural Information Processing Systems 33* (Li et al. 2020). Compared with Li et al. (2020), the current paper includes the following key extensions.

- The current version proposes a new variant of the model-based approach (i.e., the conservative model-based algorithm) for discounted inifinite-horizon MDPs. This new approach achieves minimax optimality without the need of reward perturbation, with a more streamlined analysis.
- We demonstrate how to overcome the sample size barrier in finite-horizon Markov decision processes, which was not studied in Li et al. (2020).
- Complete proof details of all theoretical findings are included and elucidated in the current paper, and they consist of highly sophisticated/delicate schemes. These proof ideas were only briefly mentioned in Li et al. (2020) (and most details were missing therein).

2. Problem Formulation

The current paper studies both discounted infinite-horizon MDPs and finite-horizon MDPs, which will be

Table 2. Comparisons with Prior Results (up to Log Factors) Regarding Finding an ε-Optimal Policy in a Time-Inhomogeneous Finite-Horizon MDP with a Generative Model

Algorithm	Sample size range	Sample complexity	ε range
Sublinear randomized value iteration (Sidford et al. 2018a)	$[\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^3,\infty)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{\varepsilon^2}$	$(0,H]$
Variance-reduced QVI (Sidford et al. 2018b)	$[\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4,\infty)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4}{\varepsilon^2}$	$(0,1]$
Empirical MDP + planning (Yin et al. 2021)	$[\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^3,\infty)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4}{\varepsilon^2}$	$(0,\sqrt{H}]$
Empirical MDP + planning (this paper)	$[\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^2,\infty)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4}{\varepsilon^2}$	$(0,H]$

Notes. 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. The results in Sidford et al. (2018a, b) were originally stated for the time-homogeneous case; we translate them into the time-inhomogeneous case with an additional factor of H. In addition, Li et al. (2021c) proved that Q learning finds an ε-optimal Q-function estimate with sample complexity $\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4}{\varepsilon^2}$ ($\varepsilon\in(0,1]$) in a sample size range $[\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4,\infty)$, which did not translate directly to an ε-optimal policy.

introduced separately in the sequel. Here and throughout, we adopt the standard notation $[H] := \{1, \ldots, H\}$.

2.1. Discounted Infinite-Horizon Markov Decision Processes

2.1.1. Models and Background. Consider a discounted infinite-horizon 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 $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the reward function, namely that $r(s, a)$ 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 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 (or action selection rule) is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ that maps a state to an action. The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π is defined by

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

which is the expected discounted total reward starting from the initial state $s^0 = s$; here, the sample trajectory $\{(s^t, a^t)\}_{t \geq 0}$ is generated based on the transition kernel (namely, $s^{t+1} \sim P(\cdot|s^t, a^t)$), with the actions taken according to policy π (namely, $a^t = \pi(s^t)$ for all $t \geq 0$). 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

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

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 $V^\pi(s)$ ($Q^\pi(s, a)$) for all states $s \in \mathcal{S}$ (state-action pairs $(s, a) \in (\mathcal{S} \times \mathcal{A})$) (Sutton and Barto 2018). The corresponding value function $V^\star := V^{\pi^\star}$ (action-value function $Q^\star := Q^{\pi^\star}$) is called the optimal value function (optimal action-value function).

2.1.2. A Generative Model and an Empirical MDP. The current paper focuses on a stylized generative model (also called a simulator) as studied in Kearns et al. (2002) and Kakade (2003). Assuming access to this generative model, we collect N independent samples

$$s^i_{s,a} \overset{\text{i.i.d.}}{\sim} P(\cdot|s, a), \qquad i = 1, \ldots, N$$

for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, which allows us to construct an empirical transition kernel $\widehat{P}$ as follows:

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

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, therefore, be understood as $N^{\text{total}} := N|\mathcal{S}||\mathcal{A}|$. 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.

2.1.3. Learning the Optimal Policy via Model-Based Planning. Given a few data samples in hand, the task of policy learning 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

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}:$$
$$V^{\pi_{\text{est}}}(s) \geq V^\star(s) - \varepsilon, \quad Q^{\pi_{\text{est}}}(s, a) \geq Q^\star(s, a) - \varepsilon. \tag{4}$$

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 $\left[0, \frac{1}{1-\gamma}\right]$; this means that the range of the target accuracy level ε should be set to $\varepsilon \in \left[0, \frac{1}{1-\gamma}\right]$. 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).

2.1.4. Aside: Policy Evaluation. A related task is policy evaluation, which 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

$$\forall s \in \mathcal{S}: |V^\pi_{\text{est}}(s) - V^\pi(s)| \leq \varepsilon. \tag{5}$$

2.2. Finite-Horizon Markov Decision Processes

2.2.1. Models and Background. Another type of model considered in this paper is a finite-horizon MDP, which

can be represented and denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote, respectively, the state space and the action space as before, and H represents the horizon length of the MDP. For any $1 \le h \le H$, we let $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denote the probability transition kernel at step h. That is, $P_h(s'|s,a)$ is the probability of transiting to s' from (s, a) at step h; $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ indicates the reward function at step h, namely that $r_h(s,a)$ is the immediate reward gained at step h in response to (s, a). As before, we assume normalized rewards throughout the paper, so that all the $r_h(s,a)$'s reside within the interval $[0,1]$.

Let $\pi = \{\pi_h\}_{1\le h\le H}$ represent a deterministic policy, such that for any $1 \le h \le H$ and any $s \in \mathcal{S}$, $\pi_h(s)$ specifies the action selected at step h in state s. Note that π could be nonstationary, meaning that the π_h's might be different across different time steps h. The value function and the Q function associated with policy π are defined, respectively, by

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{k=h}^{H} r_k(s_k,a_k) \,\middle|\, s_h = s\right]$$

for all $s \in \mathcal{S}$ and all $1 \le h \le H$ and

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{k=h}^{H} r_k(s_k,a_k) \,\middle|\, s_h = s, a_h = a\right]$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and all $1 \le h \le H$. As usual, the expectations are taken over the randomness of the MDP trajectory $\{(s_k,a_k)\}_{1\le k\le H}$ induced by the transition kernel $\{P_h\}_{h=1}^H$ when policy π is adopted. With slight abuse of notation, we let $Q_{H+1}^\pi(s,a) = 0$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $V_{H+1}^\pi(s) = 0$ for every $s \in \mathcal{S}$. In view of the assumed range of the immediate rewards, it is easily seen that

$$0 \le V_h^\pi(s) \le H \qquad \text{and} \qquad 0 \le Q_h^\pi(s,a) \le H$$

for any π, any state-action pair (s,a), and any step h. Akin to the infinite-horizon counterpart, the optimal value functions $\{V_h^\star\}_{1\le h\le H}$ and optimal Q functions $\{Q_h^\star\}_{1\le h\le H}$ are defined, respectively, by

$$V_h^\star(s) := \max_\pi V_h^\pi(s) \qquad \text{and} \qquad Q_h^\star(s,a) := \max_\pi Q_h^\pi(s,a)$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $1 \le h \le H$. It is well known that there exists at least one policy that allows one to simultaneously achieve the optimal value function and optimal Q functions for all state-action pairs and all time steps. Throughout this paper, we shall denote by $\pi^\star = \{\pi_h^\star\}_{1\le h\le H}$ an optimal policy.

2.2.2. A Generative Model and an Empirical MDP. Similar to the infinite-horizon setting, we assume access to a generative model, which is able to generate N independent samples for each triple $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ as follows:

$$s_h^i(s,a) \overset{\text{i.i.d.}}{\sim} P_h(\cdot|s,a), \qquad i = 1,\dots,N.$$

The empirical transition kernel $\{\widehat{P}_h\}_{h=1}^H$ is thus given by

$$\forall s' \in \mathcal{S}, \qquad \widehat{P}_h(s'|s,a) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s_h^i(s,a) = s'\}, \tag{6}$$

which records the empirical frequency of transitions from (s, a) to state s' at step h. This gives rise to a total sample size $N^{\text{total}} := NH|\mathcal{S}||\mathcal{A}|$. We shall let $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H)$ represent the empirical MDP constructed from the data samples. The value function and the Q function of a policy π for $\widehat{\mathcal{M}}$ can be defined analogously, which shall be denoted by $\{\widehat{V}_h^\pi\}_{h=1}^H$ and $\{\widehat{Q}_h^\pi\}_{h=1}^H$, respectively. We denote by $\widehat{\pi}^\star$ the optimal policy of $\widehat{\mathcal{M}}$, and the resulting optimal value function and Q function are denoted by $\widehat{V}_h^\star := \widehat{V}_h^{\widehat{\pi}^\star}$ and $\widehat{Q}_h^\star := \widehat{Q}_h^{\widehat{\pi}^\star}$, respectively.

2.2.3. Learning the Optimal Policy via Model-Based Planning. Given the data samples in hand, the task of policy learning in the finite-horizon case can be defined similarly as the infinite-horizon counterpart. Specifically, for any target level $\varepsilon > 0$, the aim of policy learning is to compute an ε-accurate policy π_{est} obeying

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]:$$
$$V_h^{\pi_{\text{est}}}(s) \ge V_h^\star(s) - \varepsilon, \quad Q_h^{\pi_{\text{est}}}(s,a) \ge Q_h^\star(s,a) - \varepsilon. \tag{7}$$

With the normalized range of the reward function, it is easily seen that the value function and the Q function reside within the interval $[0,H]$, thus implying that the range of the target accuracy level should be $\varepsilon \in [0,H]$.

2.3. Notation

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

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

3. Model-Based Planning in Discounted Infinite-Horizon MDPs

As summarized in Table 1, the theory of all prior works required the sample size per state-action pair to at least exceed $N \geq \Omega\left(\frac{1}{(1-\gamma)^2}\right)$. In order to break this sample size barrier, we develop two model-based algorithms that provably overcome such a sample size barrier.

3.1. Model-Based Reinforcement Learning: Two Algorithms

Algorithm 1 (Perturbed Model-Based Planning)

The first algorithm applies model-based planning to an empirical MDP with *randomly perturbed rewards*. Specifically, for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we randomly perturb the immediate reward by

$$r_{\mathrm{p}}(s,a) = r(s,a) + \zeta(s,a), \qquad \zeta(s,a) \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(0,\xi), \tag{8}$$

where $\mathsf{Unif}(0,\xi)$ denotes the uniform distribution between zero 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 $\widehat{P}$ (cf. (3)) and the perturbed reward function r_{p}. Let $\widehat{\pi}_{\mathrm{p}}^{\star}$ represent the optimal policy of $\widehat{\mathcal{M}}_{\mathrm{p}}$: that is,

$$\widehat{\pi}_{\mathrm{p}}^{\star} := \arg\max_{\pi} \widehat{V}_{\mathrm{p}}^{\pi}. \tag{9}$$

Algorithm 2 (Conservative Model-Based Planning)

An alternative approach that eliminates the need of reward perturbation is to select *approximately optimal* actions for the empirical MDP instead of the absolute optimal actions. To be precise, denote by $\widehat{Q}^{\star}$ ($\widehat{V}^{\star}$) the optimal action-value (value) function of the empirical MDP $\widehat{\mathcal{M}} = (\mathcal{S},\mathcal{A},\widehat{P},r,\gamma)$ with the probability transition kernel $\widehat{P}$ (cf. (3)) and the original reward function r. By producing a random draw from $\varsigma \sim \mathsf{Unif}(0,\xi)$ (with ξ specified shortly), we can generate the following policy $\widehat{\pi}_{\mathrm{c}}$:

$$\forall s \in \mathcal{S}: \qquad \widehat{\pi}_{\mathrm{c}}(s) := \min\{a \in \mathcal{A} : \widehat{Q}^{\star}(s,a) > \widehat{V}^{\star}(s) - \varsigma\}. \tag{10}$$

Note that there is an index assigned to each action as $\mathcal{A} = \{1,\ldots,|\mathcal{A}|\}$, which induces a natural order for all actions. In words, this approach is more conservative and does not stick to the optimal actions with respect to (w.r.t.) the empirical MDP; instead, the policy $\widehat{\pi}_{\mathrm{c}}$ picks out—for each state $s \in \mathcal{S}$—the smallest indexed action that is within a gap of ς from optimal.

Remark 1. The perturbed model-based approach seems more natural than the conservative model-based approach in terms of the algorithm design, and we recommend it in practice. Note, however, that the conservative model-based approach admits simpler and cleaner analyses, which might shed light on the algorithm design and analysis for other settings as well.

3.2. Theoretical Guarantees

Indeed, both the approaches result in a value function (Q function) that well approximates the true optimal value function $V^{\star}$ (optimal Q function $Q^{\star}$). We start by presenting our results for the perturbed model-based approach.

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 (9) obeys*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad V^{\widehat{\pi}_{\mathrm{p}}^{\star}}(s) \geq V^{\star}(s) - \varepsilon \quad \text{and} \quad Q^{\widehat{\pi}_{\mathrm{p}}^{\star}}(s,a) \geq Q^{\star}(s,a) - \gamma\varepsilon, \tag{11}$$

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

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

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

Remark 2. Theorem 1 holds unchanged if ξ is taken to be $\frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^{\alpha}|\mathcal{A}|^{\alpha}}$ for any fixed constant $\alpha \geq 1$. This paper picks the specific choice $\alpha = 5$ merely to convey that a very small degree of perturbation suffices for our purpose.

Remark 3. 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 optimization 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 $O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\log^2\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta}\right)\right)$.

Further, similar performance guarantees can be established for the conservative model-based approach without reward perturbation, as stated.

Theorem 2 (Conservative Model-Based Planning). *Under the same assumptions of Theorem 1 (including both the sample size and the choice of ξ), the policy $\widehat{\pi}_{\mathrm{c}}$ defined in*

(10) *achieves*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \widehat{V^{\pi_\epsilon}}(s) \geq V^\star(s) - \varepsilon \quad \text{and} \quad \widehat{Q^{\pi_\epsilon}}(s,a) \geq Q^\star(s,a) - \gamma\varepsilon, \tag{13}$$

with probability at least $1-\delta$.

In a nutshell, the theorems demonstrate that both model-based algorithms we introduce succeed in finding 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 with 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 because its validity is guaranteed as long as

$$N \geq \tilde{\Omega}\left(\frac{1}{1-\gamma}\right). \tag{14}$$

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

We remark that the work of Azar et al. (2013) established a minimax lower bound of the same order as (12) (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 $\varepsilon = O\left(\frac{1}{1-\gamma}\right)$. 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 $N \leq \tilde{O}\left(\frac{1}{1-\gamma}\right)$. Put another way, Condition (14) contains the full range of "meaningful" sample sizes.

Finally, we single out an intermediate result in the analysis of our theorems 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 3 (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*

$$\forall s \in \mathcal{S} : |\widehat{V}^\pi(s) - V^\pi(s)| \leq \varepsilon, \tag{15}$$

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

$$N \geq c_0 \frac{\log\left(\frac{|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^3\varepsilon^2}. \tag{16}$$

In words, this theorem reveals that $\widehat{V}^\pi$ begins to outperform a random guess as soon as $N \geq \tilde{\Omega}\left(\frac{1}{1-\gamma}\right)$. The sample complexity bound (16) enjoys *full* coverage of the ε range $\left(0, \frac{1}{1-\gamma}\right]$ 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 (Pananjady and Wainwright 2019, Khamaru et al. 2020). Although 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.

3.3. Comparisons with Prior Works 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.

3.3.1. Prior Bounds for Planning and Policy Learning. None of the prior results with a generative model (including both model-based and model-free approaches) were capable of efficiently finding the desired policy while accommodating the full sample size range (14). For instance, the state-of-the-art analysis for the model-based approach in Agarwal et al. (2020) required the sample size to at least exceed

$$N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^2}\right), \tag{17}$$

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

$$N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^3}\right). \tag{18}$$

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 ε range and the sample size range.

Remark 4. We briefly point out why the sample size barrier (17) appeared in the analysis of Agarwal et al. (2020).

Take Agarwal et al. (2020, 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 one, thereby requiring $N \geq \tilde{\Omega}((1-\gamma)^{-2})$.

3.3.2. Prior Bounds for Policy Evaluation. Regarding value function estimation for any fixed policy π, the prior results in Azar et al. (2013), Pananjady and Wainwright (2019), and Agarwal et al. (2020) for the plug-in approach all operated under the assumption that $N \geq \tilde{\Omega}\left(\frac{1}{(1-\gamma)^2}\right)$, 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 $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (Pananjady and Wainwright 2019, Agarwal et al. 2020) while extending them to the range $\varepsilon \in \left[\frac{1}{\sqrt{1-\gamma}}, \frac{1}{1-\gamma}\right]$ uncovered in these previous papers.

4. Model-Based Planning in Finite-Horizon MDPs

Moving beyond discounted infinite-horizon MDPs, our theoretical framework is also able to accommodate finite-horizon MDPs, which we detail in this section.

4.1. Algorithm: Model-Based Planning

The algorithm considered in this section is model-based planning (without reward perturbation). Specifically, this model-based approach returns a policy $\widehat{\pi}^\star = \{\widehat{\pi}_h^\star\}_{1\leq h\leq H}$ by means of the following two steps.

1. Construct the empirical MDP $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}_{1\leq h\leq H}, \{r_h\}_{1\leq h\leq H}, H)$ based on the data samples in hand (see (6) for the computation of the empirical transition kernel $\widehat{P}_h$).
2. Run a classical dynamic programming algorithm (Bertsekas 2017) to find an optimal policy $\widehat{\pi}^\star$ of the empirical MDP $\widehat{\mathcal{M}}$.

Note that $\widehat{\pi}_h^\star$ is an optimal policy of $\widehat{\mathcal{M}}$ at step h, computed by the dynamic programming algorithm calculated backward from $h=H$. Because $\widehat{\pi}_h^\star$ is calculated solely based on what happens after step h, $\widehat{\pi}_h^\star$ is independent of the empirical transitions $\{\widehat{P}_j\}_{1\leq j<h}$.

It is noteworthy that, in contrast to the infinite-horizon counterpart in Section 3, we do not need to enforce random reward perturbation for this finite-horizon case.

4.2. Theoretical Guarantees and Implications

The model-based algorithm described turns out to be nearly minimax optimal as asserted by the following theorem.

Theorem 4 (Model-Based Planning). *There exist some universal constants $c_0, c_1 > 0$ such that for any $\delta > 0$ and any $0 < \varepsilon \leq H$, the aforementioned policy $\widehat{\pi}^\star$ returned by model-based planning obeys*

$$\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]: \quad V_h^{\widehat{\pi}^\star}(s) \geq V_h^\star(s) - \varepsilon \quad \text{and} \quad Q_h^{\widehat{\pi}^\star}(s,a) \geq Q_h^\star(s,a) - \varepsilon, \tag{19}$$

with probability at least $1-\delta$, provided that the sample size for every triple (s,a,h) exceeds

$$N \geq \frac{c_0 H^3 \log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{\varepsilon^2}. \tag{20}$$

Akin to the discounted infinite-horizon scenario, the model-based approach manages to achieve ε accuracy as long as the sample size per (s,a,h) exceeds the order of

$$\frac{H^3}{\varepsilon^2} \quad \text{(up to some log factor)}.$$

This result, which is valid for the full ε range $(0, H]$, is reminiscent of the bound (12), except that the effective horizon $\frac{1}{1-\gamma}$ needs to be replaced by the horizon length H. Given that there are in total $|\mathcal{S}||\mathcal{A}|H$ different combinations of (s,a,h), the total sample complexity is on the order of $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^4}{\varepsilon^2}\right)$.

The quadruple scaling H^4 of this total sample complexity—as opposed to the cubic scaling in the discounted infinite-horizon case—is because of time inhomogeneity; that is, the P_h's might be different across h, resulting in an additional H factor. Again, our result kicks in as soon as the sample size satisfies

$$N \geq \tilde{\Omega}(H), \tag{21}$$

improving upon the sample size requirement

$$N \geq \tilde{\Omega}(H^2) \tag{22}$$

in the state-of-the-art analysis for the model-based approach of Yin et al. (2021).

5. Other Related Works

Classical analyses of reinforcement learning algorithms have largely focused on asymptotic performance (e.g., Jaakkola et al. 1994, Tsitsiklis 1994, Tsitsiklis and Van Roy 1997, Szepesvári 1998). Leveraging the tool kit of concentration inequalities, a number of recent papers have shifted attention toward understanding the performance in the nonasymptotic and finite-time settings. A highly incomplete list includes Bradtke and Barto (1996), Kearns and Singh (1999), Even-Dar and Mansour (2003), Strehl et al. (2006), Beck and Srikant (2012), Azar et al. (2017), Bhandari et al. (2018), Dalal et al. (2018), Jin et al. (2018), Lakshminarayanan and Szepesvari (2018), Shah and Xie (2018), Sidford et al. (2018b), Cai et al. (2019),

Fan et al. (2019), Gupta et al. (2019), Srikant and Ying (2019), Wainwright (2019b), Xu et al. (2019), Chen et al. (2020), Kaledin et al. (2020), Khamaru et al. (2020), Mou et al. (2020), Qu and Wierman (2020), Xu and Gu (2020), Li et al. (2021b, 2022c, 2023), Shi et al. (2022), and Yan et al. (2022b), 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) and has been invoked in Kearns and Singh (1999), Kearns et al. (2002), Kakade (2003), Azar et al. (2012, 2013), Lattimore and Hutter (2012), Sidford et al. (2018a, b), Pananjady and Wainwright (2019), Wainwright (2019b), Yang and Wang (2019), Agarwal et al. (2020), Khamaru et al. (2020), Wang (2020), Wang et al. (2021), and Li et al. (2022a), to name just a few. In particular, Azar et al. (2013) developed the minimax lower bound on the sample complexity $N = \Omega\left(\frac{|\mathcal{S}\|\mathcal{A}|\log(|\mathcal{S}\|\mathcal{A}|)}{(1-\gamma)^3\varepsilon^2}\right)$ 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 sublinear in the model complexity. A recent breakthrough in Agarwal et al. (2020) substantially improved the model-based guarantee with the aid of auxiliary state-absorbing MDPs, extending the range of sample complexity to $\left[\frac{|\mathcal{S}\|\mathcal{A}|\log(|\mathcal{S}\|\mathcal{A}|)}{(1-\gamma)^2}, \infty\right)$. Our analysis is motivated in part by Agarwal et al. (2020) 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; see El Karoui (2018), Chen et al. (2019a, b, 2021), Pananjady and Wainwright (2019), Agarwal et al. (2020), Ma et al. (2020), and Yan et al. (2021). The application of such an analysis framework to MDPs should be attributed to Agarwal et al. (2020). Other applications to Markov chains include Chen et al. (2019a) and Pananjady and Wainwright (2019). More recently, several follow-up works have further generalized the leave-one-out analysis idea to accommodate broader RL settings, including offline RL (Li et al. 2022b), RL with linear function approximation (Wang et al. 2021), Markov games (Cui and Yang 2021, Yan et al. 2022a), and so on.

6. Analysis: Infinite-Horizon MDPs

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

6.1. Matrix Notation and Bellman Equations

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

- $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|}$: 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

$$\boldsymbol{\Pi}^\pi = \begin{pmatrix} \boldsymbol{e}^\top_{\pi(1)} & \boldsymbol{0}^\top & \cdots & \boldsymbol{0}^\top \\ \boldsymbol{0}^\top & \boldsymbol{e}^\top_{\pi(2)} & \cdots & \boldsymbol{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}^\top & \boldsymbol{0}^\top & \cdots & \boldsymbol{e}^\top_{\pi(|\mathcal{S}|)} \end{pmatrix}. \tag{23}$$

- $\boldsymbol{P}^\pi \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}| \times |\mathcal{S}\|\mathcal{A}|}$ 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

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \quad \text{and} \quad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}. \tag{24}$$

- $\boldsymbol{r}_\pi \in \mathbb{R}^{|\mathcal{S}|}$: 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 matrix notation, we can write, for any policy π, the *Bellman consistency equation* as

$$\boldsymbol{Q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}^\pi \boldsymbol{Q}^\pi, \tag{25}$$

which implies that

$$\boldsymbol{Q}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}^\pi)^{-1} \boldsymbol{r}; \tag{26}$$

$$\boldsymbol{V}^{\pi} = \boldsymbol{r}_{\pi} + \gamma \boldsymbol{P}_{\pi}\boldsymbol{V}^{\pi} \qquad \text{and} \qquad \boldsymbol{V}^{\pi} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_{\pi}. \tag{27}$$

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

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

$$[\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V})]_{(s,a)} := \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{s'}^2 - \left(\sum_{s'\in\mathcal{S}} P(s'|s,a)V_{s'}\right)^2$$

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

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V})\circ(\boldsymbol{P}\boldsymbol{V}). \tag{28}$$

Similarly, for any given policy π, we define

$$\begin{aligned}\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V}) &= \boldsymbol{\Pi}^{\pi}\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \\ &= \boldsymbol{P}_{\pi}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_{\pi}\boldsymbol{V})\circ(\boldsymbol{P}_{\pi}\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}|}.\end{aligned} \tag{29}$$

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.

6.2. 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 3, 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

$$\boldsymbol{V}^{\pi} = (\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi})^{-1}\boldsymbol{r}_{\pi} \quad \text{and} \quad \widehat{\boldsymbol{V}}^{\pi} = (\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}_{\pi}. \tag{30}$$

Lemma 1. *Fix any policy π. Consider any $0<\delta<1$, and suppose $N \ge \frac{32e^2}{1-\gamma}\log\left(\frac{4|\mathcal{S}|\log\left(\frac{e}{1-\gamma}\right)}{\delta}\right)$. Then, with probability at least $1-\delta$, the vectors defined in (30) obey*

$$\begin{aligned}\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\|_{\infty} &\le 4\gamma\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\left(\frac{e}{1-\gamma}\right)}{\delta}\right)}{N}}\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}[\boldsymbol{V}^{\pi}]}\|_{\infty} + \frac{2\gamma\log\left(\frac{4|\mathcal{S}|\log\left(\frac{e}{1-\gamma}\right)}{\delta}\right)}{(1-\gamma)N}\|\boldsymbol{V}^{\pi}\|_{\infty} \\ &\le 6\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\left(\frac{e}{1-\gamma}\right)}{\delta}\right)}{N(1-\gamma)^3}}.\end{aligned} \tag{31}$$

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 Online Appendix EC.2.1. □

Clearly, Theorem 3 is a straightforward consequence of Lemma 1. Further, we strengthen the result by providing an additional instance-dependent bound (see the first line of (31) 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 (31). 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 $N \ge \tilde{\Omega}\left(\frac{1}{1-\gamma}\right)$. In contrast, Pananjady and Wainwright (2019, theorem 1) is only valid when $N \ge \tilde{\Omega}\left(\frac{1}{(1-\gamma)^2}\right)$.

6.2.1. Proof Ideas. We now briefly and informally describe the key proof ideas. As a starting point, the elementary identities (30) allow us to obtain

$$\begin{aligned}\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi} &= (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}\boldsymbol{r}_{\pi} - \boldsymbol{V}^{\pi} \\ &= (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi} - (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})\boldsymbol{V}^{\pi} \\ &= \gamma(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi}.\end{aligned} \tag{32}$$

Because of the complicated dependency between $(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}$ and $(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi}$, a natural strategy is to control these two terms separately and then, to combine bounds; see Agarwal et al. (2020, lemma 5) for an introduction. This simple approach, however, leads to suboptimal statistical guarantees.

In order to refine the statistical analysis, we propose to further expand (32) in a similar way to deduce

$$\begin{aligned}(32) &= \gamma(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi} + \gamma\{(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1} - (\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi} \\ &= \gamma(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi} + \gamma^2(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_{\pi})^{-1}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi},\end{aligned} \tag{33}$$

where the last line holds because of the same reason as (32) (basically, it can be seen by replacing $\boldsymbol{r}_{\pi}$ with $(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi}$ in (32)). This can be viewed as a "second-order" expansion, with (32) being a "first-order" counterpart. The advantage is that the first term in (33) becomes easier to cope with than its counterpart (32), owing to the independence between $(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}$ and $(\widehat{\boldsymbol{P}}_{\pi}-\boldsymbol{P}_{\pi})\boldsymbol{V}^{\pi}$. However, the second term in (33) 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 entry-wise bound on a quantity taking the form $(\boldsymbol{I}-\gamma \boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{P_\pi}(\boldsymbol{V})}$ (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 EC.2 in the online appendix. Details are deferred to Online Appendix EC.2.1.

6.3. Analysis: Perturbed 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:

$$\begin{aligned}\boldsymbol{V}^\star-\boldsymbol{V}^{\hat{\pi}^\star}&=(\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}-\boldsymbol{V}^{\hat{\pi}^\star})+(\widehat{\boldsymbol{V}}^{\pi^\star}-\widehat{\boldsymbol{V}}^{\hat{\pi}^\star})+(\boldsymbol{V}^\star-\widehat{\boldsymbol{V}}^{\pi^\star})\\&\le(\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}-\boldsymbol{V}^{\hat{\pi}^\star})+(\boldsymbol{V}^{\pi^\star}-\widehat{\boldsymbol{V}}^{\pi^\star}),\end{aligned}\tag{34}$$

where the inequality follows from the optimality of $\widehat{\pi}^\star$ w.r.t. $\widehat{\boldsymbol{V}}$ (so that $\widehat{\boldsymbol{V}}^{\pi^\star}\le\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}$) 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 $1-\delta$,

$$\|\widehat{\boldsymbol{V}}^{\pi^\star}-\boldsymbol{V}^{\pi^\star}\|_\infty\le 6\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{N(1-\gamma)^3}}.\tag{35}$$

Step 2 (Bounding $\|\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}-\boldsymbol{V}^{\hat{\pi}^\star}\|_\infty$). Extending the result in Step 1 to $\|\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}-\boldsymbol{V}^{\hat{\pi}^\star}\|_\infty$ is considerably more challenging, primarily because of the complicated statistical dependency between $(\boldsymbol{V}^{\hat{\pi}^\star},\widehat{\boldsymbol{V}}^{\hat{\pi}^\star})$ and the data matrix $\widehat{\boldsymbol{P}}$. The recent work by Agarwal et al. (2020) 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. (2020).

- Decoupling statistical dependency between $\widehat{\pi}^\star$ and $\widehat{\boldsymbol{P}}$. Instead of attempting to decouple the statistical dependency between $\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}$ and $\widehat{\boldsymbol{P}}$ as in Agarwal et al. (2020), 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 6.3.1). A key ingredient of this step lies in the construction of a collection of auxiliary state-action–absorbing MDPs (motivated by Agarwal et al. 2020), which allows us to get hold of $\|\boldsymbol{V}^{\widehat{\pi}^\star}-\widehat{\boldsymbol{V}}^{\hat{\pi}^\star}\|_\infty$. See Section 6.3.2 for details, with a formal bound delivered in Lemma 5.
- Tiebreaking via reward perturbation. A shortcoming of the 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 nonunique 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 6.3.3 for details.

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

6.3.1. 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 (37)—is satisfied. To make it precise, we need to introduce a set of auxiliary vectors as follows:

$$\begin{aligned}&\boldsymbol{r}^{(0)}:=\boldsymbol{r}_\pi, && \boldsymbol{V}^{(0)}:=(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(0)},\\&\boldsymbol{r}^{(l)}:=\sqrt{\mathsf{Var}_{P_\pi}[\boldsymbol{V}^{(l-1)}]}, && \boldsymbol{V}^{(l)}:=(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(l)}, \quad l\ge 1.\end{aligned}\tag{36}$$

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 $|\boldsymbol{z}|:=[|z_1|,\ldots,|z_n|]^\top$ and $\sqrt{\boldsymbol{z}}:=[\sqrt{z_1},\ldots,\sqrt{z_n}]^\top$ 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. (36)) obeys*

$$\left|(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi)\boldsymbol{V}^{(l)}\right|\le\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{P_\pi}[\boldsymbol{V}^{(l)}]}+\frac{\beta_1\|\boldsymbol{V}^{(l)}\|_\infty}{N}\boldsymbol{1},\quad \text{for all } 0\le l\le\log\left(\frac{e}{1-\gamma}\right).\tag{37}$$

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

$$\|\widehat{\boldsymbol{V}}^\pi-\boldsymbol{V}^\pi\|_\infty\le\frac{6}{1-\gamma}\sqrt{\frac{\beta_1}{N(1-\gamma)}}.\tag{38}$$

Although the Bernstein-type condition (37) 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 $\widehat{\pi}^\star$.

6.3.2. Decoupling Statistical Dependency via (*s*, *a*)-Absorbing MDPs. We are now positioned to demonstrate how to control $\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\|_\infty$ 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). Toward 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. (2020, section 4.2). In comparison with Agarwal et al. (2020), which 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.

6.3.2.1. 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

$$\begin{aligned} &P_{\mathcal{M}_{s,a,u}}(s|s,a) = 1, \\ &P_{\mathcal{M}_{s,a,u}}(s'|s,a) = 0, && \text{for all } s' \neq s, \\ &P_{\mathcal{M}_{s,a,u}}(\cdot|s',a') = P_{\mathcal{M}}(\cdot|s',a'), && \text{for all } (s',a') \neq (s,a), \end{aligned} \tag{39}$$

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, whereas 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 notations $\boldsymbol{V}^\pi_{s,a,u}$ 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 5. The careful reader will remark that the instant reward u is constrained to reside within $\left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]$ rather than the usual range $[0, 1]$. Fortunately, none of the subsequent steps that involve u require u to lie within $[0, 1]$.

6.3.2.2. 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. (2020, lemma 7), and its proof is deferred to Online Appendix EC.2.2.

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

$$\boldsymbol{Q}^\star_{s,a,u^\star} = \boldsymbol{Q}^\star \quad \text{and} \quad \boldsymbol{V}^\star_{s,a,u^\star} = \boldsymbol{V}^\star. \tag{40}$$

Remark 6. 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 $u = u^\star$ $\left(\text{which can be easily shown to reside within } \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]\right)$, 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.

6.3.2.3. 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 ε 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:

$$\mathcal{B}_\omega := \{\widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a: a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s,a) \ge \omega \ \text{ for all } s \in \mathcal{S}\}. \tag{41}$$

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

$$\mathcal{N}_\epsilon := \{-n_\epsilon\epsilon, \ldots, -\epsilon, 0, \epsilon, \ldots, n_\epsilon\epsilon\}, \quad \text{for the largest integer } n_\epsilon \text{ obeying } n_\epsilon\epsilon < \frac{1}{1-\gamma}, \tag{42}$$

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

Lemma 4. *Consider any $\omega > 0$, and suppose the event $\mathcal{B}_\omega$ (cf. (41)) 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*

$$\widehat{\pi}^\star = \widehat{\pi}^\star_{s,a,u_0}. \tag{43}$$

Proof. See Online Appendix EC.2.3. □

6.3.2.4. Deriving an Optimal Error Bound Under the Separation Condition.

Armed with these bounds, we are ready to derive the desired error bound by combining Lemmas 2 and 4.

Lemma 5. *Given $0 < \omega < 1$ and $\delta > 0$, suppose that $\mathcal{B}_\omega$ (defined in (41)) occurs with probability at least $1 - \delta$. Then, with probability at least $1 - 3\delta$,*

$$\|\widehat{\boldsymbol{V}}^{\hat{\pi}^\star} - \boldsymbol{V}^{\hat{\pi}^\star}\|_\infty \le 6\sqrt{\frac{2\log\left(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\omega\delta}\right)}{N(1-\gamma)^3}} \quad \text{and}$$

$$\boldsymbol{V}^\star - \boldsymbol{V}^{\hat{\pi}^\star} \le 12\sqrt{\frac{2\log\left(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\omega\delta}\right)}{N(1-\gamma)^3}}\mathbf{1}, \tag{44}$$

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

Proof. See Online Appendix EC.2.4. □

6.3.3. A Tiebreaking Argument.

Unfortunately, the separation condition specified in $\mathcal{B}_\omega$ (cf. (41)) 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 (8) in a vector form as follows:

$$\boldsymbol{r}_{\mathrm{p}} := \boldsymbol{r} + \boldsymbol{\zeta}, \tag{45}$$

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 by $Q^\star_{\mathrm{p}}$ 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 (45). With probability at least $1 - \delta$,*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A} \text{ with } a \neq \pi^\star_{\mathrm{p}}(s): \quad Q^\star_{\mathrm{p}}(s, \pi^\star_{\mathrm{p}}(s)) - Q^\star_{\mathrm{p}}(s,a) > \frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2}. \tag{46}$$

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

Proof. See Online Appendix EC.2.5. □

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

6.3.4. Proof of Theorem 1. Proof of Theorem 1.

Let us consider the randomly perturbed reward function as in (45). For any policy π, we denote by $\boldsymbol{V}^\pi_{\mathrm{p}}$ ($\widehat{\boldsymbol{V}}^\pi_{\mathrm{p}}$) the corresponding value function vector in the MDP with probability transition matrix $\boldsymbol{P}$ ($\widehat{\boldsymbol{P}}$) and reward vector $\boldsymbol{r}_{\mathrm{p}}$. Note that $\pi^\star_{\mathrm{p}}$ ($\widehat{\pi}^\star_{\mathrm{p}}$) denotes the optimal policy that maximizes $\boldsymbol{V}^\pi_{\mathrm{p}}$ ($\widehat{\boldsymbol{V}}^\pi_{\mathrm{p}}$).

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

$$\left|\widehat{Q}^\star_{\mathrm{p}}(s, \widehat{\pi}^\star_{\mathrm{p}}(s)) - \widehat{Q}^\star_{\mathrm{p}}(s,a)\right| > \frac{\xi\delta(1-\gamma)}{3|\mathcal{S}||\mathcal{A}|^2} \tag{47}$$

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

$$\|\boldsymbol{V}^{\pi^\star_{\mathrm{p}}}_{\mathrm{p}} - \boldsymbol{V}^{\hat{\pi}^\star_{\mathrm{p}}}_{\mathrm{p}}\|_\infty \le 12\sqrt{\frac{2\log\left(\frac{96|\mathcal{S}|^2|\mathcal{A}|^3}{(1-\gamma)^4\xi\delta^2}\right)}{N(1-\gamma)^3}}. \tag{48}$$

In addition, the value functions under any policy π obey

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

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

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

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

$$\|\boldsymbol{V}^{\pi^\star} - \boldsymbol{V}_{\mathrm{p}}^{\pi^\star}\|_\infty \le \frac{1}{1-\gamma}\xi \quad \text{and}$$
$$\|\boldsymbol{V}^{\widehat{\pi}_{\mathrm{p}}^\star} - \boldsymbol{V}_{\mathrm{p}}^{\widehat{\pi}_{\mathrm{p}}^\star}\|_\infty \le \frac{1}{1-\gamma}\xi. \tag{49}$$

Now, let us consider the following decomposition:

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

where the last step follows from the optimality of $\pi_{\mathrm{p}}^\star$ w.r.t. $\boldsymbol{V}_{\mathrm{p}}$. Taking this collectively with Inequalities (48) and (49), one shows that with probability greater than $1-3\delta$,

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

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

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

Consequently, one has

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

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

$$\|\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k} - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\|_\infty < \frac{\xi\delta(1-\gamma)}{8|\mathcal{S}||\mathcal{A}|^2}. \tag{50}$$

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

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

where the last line results from (47) and (50). 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 (50) is satisfied. In addition, it has been shown that (Azar et al. 2013, lemma 2) the greedy policy induced by the kth iteration of both algorithms—denoted by π_k—satisfies $\|\widehat{\boldsymbol{Q}}_{\mathrm{p}}^{\pi_k} - \widehat{\boldsymbol{Q}}_{\mathrm{p}}^\star\|_\infty \le \frac{2\gamma^{k+1}}{(1-\gamma)^2}$. Taking $\xi = \frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ and $k = \frac{c_2}{1-\gamma}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta}\right)$ for some constant $c_2 > 0$ large enough, one guarantees that π_k satisfies (50), which in turn, ensures perfect recovery of $\widehat{\pi}_{\mathrm{p}}^\star$. □

6.4. Analysis: Conservative Model-Based Planning

In view of the conservative model-based planning (10), we begin with the following decomposition:

$$\boldsymbol{V}^\star - \boldsymbol{V}^{\widehat{\pi}_{\mathrm{c}}} = (\boldsymbol{V}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\pi^\star}) + (\widehat{\boldsymbol{V}}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}}) + (\widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} - \boldsymbol{V}^{\widehat{\pi}_{\mathrm{c}}}). \tag{51}$$

In order to control the second term on the right-hand side of the identity, we resort to the following lemma, whose proof is postponed to Online Appendix EC.2.6.

Lemma 7. *It holds that*

$$\widehat{\boldsymbol{V}}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} \le \frac{\xi}{1-\gamma}\mathbf{1}. \tag{52}$$

Combining Lemma 7 with (51), we arrive at

$$\begin{aligned}\boldsymbol{V}^\star - \boldsymbol{V}^{\widehat{\pi}_{\mathrm{c}}} &= (\boldsymbol{V}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\pi^\star}) + (\widehat{\boldsymbol{V}}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}}) + (\widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} - \boldsymbol{V}^{\widehat{\pi}_{\mathrm{c}}}) \\ &\le (\boldsymbol{V}^{\pi^\star} - \widehat{\boldsymbol{V}}^{\pi^\star}) + \frac{\xi}{1-\gamma}\mathbf{1} + (\widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} - \boldsymbol{V}^{\widehat{\pi}_{\mathrm{c}}}).\end{aligned} \tag{53}$$

Clearly, the first term of (53) has already been controlled in Section 6.2, whereas the second term of (53) is extremely small when we take $\xi = O\left(\frac{(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}\right)$. It thus suffices to bound the third term of (53), which again requires decoupling the statistical dependence between $\widehat{\pi}_{\mathrm{c}}$ and $\widehat{\boldsymbol{P}}$.

6.4.1. Representing $\widehat{\pi}_{\mathrm{c}}$ via a Small Set of Policies Independent of $\widehat{\boldsymbol{P}}_{s,a}$.

Akin to our analysis for the perturbed model-based planning algorithm in Section 6.3, a key step lies in demonstrating the connection between $\widehat{\pi}_{\mathrm{c}}$ and a reasonably small collection of leave-one-out auxiliary MDPs. Toward this, we are in need of the following lemma, which characterizes certain "stability" of our conservative model-based strategy and lies at the core of our analysis. The proof is deferred to Online Appendix EC.2.7.

Lemma 8. *Consider any given Q function $\widehat{Q}: \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and its associated value function $\widehat{V}: \mathcal{S} \to \mathbb{R}$ (i.e., $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$ for all s). Generate an independent random variable $\varsigma \sim \mathsf{Unif}(0,\xi)$. Then, with probability at least*

$1-\delta$,

$$\forall s \in \mathcal{S}, \qquad \{a \in \mathcal{A} : \widehat{Q}(s,a) > \widehat{V}(s) - \varsigma\} = \{a \in \mathcal{A} : Q(s,a) > V(s) - \varsigma\} \tag{54}$$

holds simultaneously for all Q function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (*and its associated value function* $V : \mathcal{S} \to \mathbb{R}$) *obeying*

$$\max_{s,a} |Q(s,a) - \widehat{Q}(s,a)| \le \frac{\xi\delta}{8|\mathcal{S}\|\mathcal{A}|}. \tag{55}$$

As an important implication of this lemma, the policy $\widehat{\pi}_{\mathrm{c}}$ computed in (10) remains unchanged upon slight perturbation of the Q-function estimates. Armed with Lemma 8 and the leave-one-out auxiliary MDPs $\{\widehat{\mathcal{M}}_{s,a,u}\}$ constructed in Section 6.3, our analysis proceeds as follows.

- For each $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega}$ (cf. (42)) such that the optimal Q function of $\widehat{Q}^{\star}_{s,a,u_0}$ of $\widehat{\mathcal{M}}_{s,a,u_0}$ obeys

$$\|\widehat{Q}^{\star}_{s,a,u_0} - \widehat{Q}^{\star}\|_{\infty} \le \omega. \tag{56}$$

This is a fact that has already been established in the proof of Lemma 4; see (EC.21) in the online appendix.

- Define a conservative policy for $\widehat{\mathcal{M}}_{s,a,u_0}$ as follows:

$$\forall s' \in \mathcal{S}: \qquad \widehat{\pi}_{s,a,u_0,\mathrm{c}}(s') = \min\{a' \in \mathcal{A} : \widehat{Q}^{\star}_{s,a,u_0}(s',a') > \widehat{V}^{\star}_{s,a,u_0}(s') - \varsigma\},$$

where $\widehat{Q}^{\star}_{s,a,u_0}$ and $\widehat{V}^{\star}_{s,a,u_0}$ denote the optimal Q function and optimal value function of $\widehat{\mathcal{M}}_{s,a,u_0}$, respectively. Taking $\omega = \frac{\xi\delta}{8|\mathcal{S}\|\mathcal{A}|}$ and invoking Lemma 8 and (56), we arrive at

$$\widehat{\pi}_{\mathrm{c}} = \widehat{\pi}_{s,a,u_0,\mathrm{c}}. \tag{57}$$

This result (57) parallels Lemma 4 for the perturbed model-based planning algorithm, revealing that $\widehat{\pi}_{\mathrm{c}}$ is representable using a policy independent of the randomness associated with (s,a). The remaining proof of Theorem 2 then follows from an identical argument as in the proof of Theorem 1 and is hence omitted here.

7. Analysis: Finite-Horizon MDPs

In this section, we outline the proof of Theorem 4. We shall start by introducing a set of convenient matrix notations before embarking on the main proof.

7.1. Matrix Notation and Bellman Equations

Akin to the infinite-horizon case, we introduce some matrix notation for finite-horizon MDPs. Analogous to Section 6.1, we introduce the following set of notation.

- $\boldsymbol{r}_h \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|}$: a vector representing the reward function r_h at step h.
- $\boldsymbol{V}^{\pi}_h \in \mathbb{R}^{|\mathcal{S}|}$: a vector representing the value function V^{π}_h of π at step h.
- $\boldsymbol{V}^{\star}_h \in \mathbb{R}^{|\mathcal{S}|}$: a vector representing the optimal value function $V^{\star}_h$ at step h.
- $\boldsymbol{Q}^{\pi}_h \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|}$: a vector representing the Q function Q^{π}_h of π at step h.
- $\boldsymbol{Q}^{\star}_h \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|}$: a vector representing the optimal Q function $Q^{\star}_h$ at step h.
- $\boldsymbol{P}_h \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|\times|\mathcal{S}|}$: a matrix representing the probability transition kernel P_h at step h.
- $\boldsymbol{P}_{h,\pi} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$: a submatrix of $\boldsymbol{P}_h$, which consists of the rows with indices coming from $\{(s,\pi_h(s))|s \in \mathcal{S}\}$.
- $\boldsymbol{r}^{\pi}_h \in \mathbb{R}^{|\mathcal{S}|}$: a subvector of $\boldsymbol{r}_h$, which consists of the rows with indices coming from $\{(s,\pi_h(s))|s \in \mathcal{S}\}$.

Armed with this notation, the Bellman equation here is given by

$$\boldsymbol{Q}^{\pi}_h = \boldsymbol{r}_h + \boldsymbol{P}_h \boldsymbol{V}^{\pi}_{h+1}, \qquad 1 \le h \le H, \tag{58}$$

where we recall that for all $s \in \mathcal{S}$,

$$V^{\pi}_h(s) = Q^{\pi}_h(s,\pi_h(s)) \quad \text{and} \quad V^{\pi}_{H+1}(s) = 0. \tag{59}$$

This also allows one to derive

$$\boldsymbol{V}^{\pi}_h = \boldsymbol{r}^{\pi}_h + \boldsymbol{P}_{h,\pi} \boldsymbol{V}^{\pi}_{h+1}. \tag{60}$$

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

7.2. An Auxiliary Value Function Sequence Obeying Bernstein-Type Conditions

Similar to the infinite-horizon case (in particular, Section 6.3.1), we find it convenient to introduce a collection of auxiliary vectors as follows. For any $l \ge 0$, define

$$\boldsymbol{V}^{(l)}_{H+1} := \boldsymbol{0} \quad \text{and} \quad \widehat{\boldsymbol{V}}^{(l)}_{H+1} := \boldsymbol{0}; \tag{61}$$

for any $1 \le h \le H$ and any policy π, define the following sequences recursively:

$$\boldsymbol{r}^{(0)}_h := \boldsymbol{r}^{\pi}_h, \qquad \boldsymbol{V}^{(0)}_h := \boldsymbol{r}^{(0)}_h + \boldsymbol{P}_{h,\pi}\boldsymbol{V}^{(0)}_{h+1},$$

$$\widehat{\boldsymbol{V}}^{(0)}_h := \boldsymbol{r}^{(0)}_h + \widehat{\boldsymbol{P}}_{h,\pi}\widehat{\boldsymbol{V}}^{(0)}_{h+1},$$

$$\boldsymbol{r}^{(l)}_h := \sqrt{\mathsf{Var}_{P_{h,\pi}}[\boldsymbol{V}^{(l-1)}_{h+1}]}, \qquad \boldsymbol{V}^{(l)}_h := \boldsymbol{r}^{(l)}_h + \boldsymbol{P}_{h,\pi}\boldsymbol{V}^{(l)}_{h+1},$$

$$\widehat{\boldsymbol{V}}^{(l)}_h := \boldsymbol{r}^{(l)}_h + \widehat{\boldsymbol{P}}_{h,\pi}\widehat{\boldsymbol{V}}^{(l)}_{h+1}, \qquad l \ge 1. \tag{62}$$

As can be easily verified, $\{\boldsymbol{V}^{(0)}_h\}$ coincides with the value function of policy π in the true MDP $\mathcal{M}$, whereas $\{\widehat{\boldsymbol{V}}^{(0)}_h\}$ corresponds to the value function of policy π in the empirical MDP $\widehat{\mathcal{M}}$.

As it turns out, if the auxiliary sequence satisfies certain Bernstein-type conditions, then we can establish a useful

upper bound on the entry-wise difference between $\boldsymbol{V}_h^{(0)}$ and $\widehat{\boldsymbol{V}}_h^{(0)}$, as stated. The proof of this lemma is deferred to Online Appendix EC.3.1.

Lemma 9. *Suppose that there exists some quantity $\beta_1 > 0$ such that the sequence $\{\boldsymbol{V}_{h+1}^{(l)}\}$ constructed in (62) obeys*

$$\left|(\widehat{\boldsymbol{P}}_{h,\pi} - \boldsymbol{P}_{h,\pi})\boldsymbol{V}_{h+1}^{(l)}\right| \le \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{h,\pi}}[\boldsymbol{V}_{h+1}^{(l)}]} + \frac{\beta_1\|\boldsymbol{V}_{h+1}^{(l)}\|_\infty}{N}\mathbf{1}, \quad \forall 0 \le l \le \log_2 H,\ 1 \le h \le H-1. \tag{63}$$

In addition, assume that $N > 12H\beta_1$. Then, we have

$$\|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\|_\infty \le 6H\sqrt{\frac{3\beta_1 H}{N}} \tag{64}$$

for all $1 \le h \le H$.

7.3. Proof of Theorem 4

Proof of Theorem 4. Let us begin with the following elementary decomposition:

$$\begin{aligned}\boldsymbol{V}_h^\star - \boldsymbol{V}_h^{\widehat{\pi}^\star} &= (\widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star} - \boldsymbol{V}_h^{\widehat{\pi}^\star}) + (\widehat{\boldsymbol{V}}_h^{\pi^\star} - \widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star}) + (\boldsymbol{V}_h^\star - \widehat{\boldsymbol{V}}_h^{\pi^\star}) \\ &\le (\widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star} - \boldsymbol{V}_h^{\widehat{\pi}^\star}) + (\boldsymbol{V}_h^{\pi^\star} - \widehat{\boldsymbol{V}}_h^{\pi^\star}).\end{aligned} \tag{65}$$

Here, the inequality follows from the definition $\boldsymbol{V}_h^\star = \boldsymbol{V}_h^{\pi^\star}$, as well as the fact that $\widehat{\boldsymbol{V}}_h^{\pi^\star} \le \widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star}$ (because $\widehat{\pi}^\star$ is the optimal policy of the empirical MDP). In light of (65), there are two terms that need to be controlled.

We intend to bound both $\widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star} - \boldsymbol{V}_h^{\widehat{\pi}^\star}$ and $\boldsymbol{V}_h^{\pi^\star} - \widehat{\boldsymbol{V}}_h^{\pi^\star}$ by means of Lemma 9. Toward this, we first note that for any policy π, the associated value functions of the MDP and the empirical MDP obey the following Bellman equations:

$$\boldsymbol{V}_h^\pi := \boldsymbol{r}_h^\pi + \boldsymbol{P}_{h,\pi}\boldsymbol{V}_{h+1}^\pi, \qquad \widehat{\boldsymbol{V}}_h^\pi := \boldsymbol{r}_h^\pi + \widehat{\boldsymbol{P}}_{h,\pi}\widehat{\boldsymbol{V}}_{h+1}^\pi,$$

along with the boundary conditions $\boldsymbol{V}_{H+1}^\pi = \widehat{\boldsymbol{V}}_{H+1}^\pi = \mathbf{0}$. This indicates that the vector $\boldsymbol{V}_h^{(0)}$ ($\widehat{\boldsymbol{V}}_h^{(0)}$) constructed in (62) is precisely the value function of policy π at step h in the true MDP (empirical MDP). As a result, in order to invoke Lemma 9, it is sufficient to verify the Bernstein-type Condition (63) w.r.t. policies $\pi^\star$ and $\widehat{\pi}^\star$ for some sufficiently small quantity β_1.

- Let us begin with the optimal policy $\pi^\star$, which is fixed and statistically independent of the data samples. As a result, if we take $\pi = \pi^\star$ during the construction of (62), then it is clearly seen that $\widehat{\boldsymbol{P}}_h$ is statistically independent of $\boldsymbol{V}_{h+1}^{(l)}$. Applying the Bernstein inequality together with the union bound then guarantees that with probability exceeding $1-\delta$,

$$\left|(\widehat{\boldsymbol{P}}_h - \boldsymbol{P}_h)\boldsymbol{V}_{h+1}^{(l)}\right| \le \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_h}[\boldsymbol{V}_{h+1}^{(l)}]} + \frac{\beta_1\|\boldsymbol{V}_{h+1}^{(l)}\|_\infty}{N}\mathbf{1} \tag{66}$$

holds uniformly over all $0 \le l \le \log_2 H$, $1 \le h \le H-1$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, where β_1 is given by

$$\beta_1 := 4\log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right). \tag{67}$$

Armed with Condition (66), we can readily invoke Lemma 9 to reach

$$\|\widehat{\boldsymbol{V}}_h^{\pi^\star} - \boldsymbol{V}_h^{\pi^\star}\|_\infty = \|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\|_\infty \le 6H\sqrt{\frac{3\beta_1 H}{N}} \le 6H\sqrt{\frac{12H\log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}}$$

with probability at least $1-\delta$.

- Next, we move on to the policy $\widehat{\pi}^\star$ by taking $\pi = \widehat{\pi}^\star$ during the construction (62). Note that $\boldsymbol{V}_{h+1}^{(l)}$ depends only on $\widehat{\pi}_j^\star (j \ge h+1)$. In view of our assumption on $\widehat{\pi}^\star$ (i.e., it is computed backward via dynamic programming), $\widehat{\pi}_i^\star$ is independent of any $\widehat{\boldsymbol{P}}_j$ with $j<i$, and hence, $\boldsymbol{V}_{h+1}^{(l)}$ is statistically independent from $\widehat{\boldsymbol{P}}_h$. Consequently, the preceding bounds (66) and (67) continue to hold. All of this immediately results in

$$\|\widehat{\boldsymbol{V}}_h^{\widehat{\pi}^\star} - \boldsymbol{V}_h^{\widehat{\pi}^\star}\|_\infty \le 6H\sqrt{\frac{12H\log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}}$$

with probability exceeding $1-\delta$.

Substituting the bounds into (65), we arrive at

$$\mathbf{0} \le \boldsymbol{V}_h^\star - \boldsymbol{V}_h^{\widehat{\pi}^\star} \le 12H\sqrt{\frac{12H\log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}}\mathbf{1}, \tag{68}$$

with probability greater than $1-2\delta$, provided that $N \ge 48H\log\left(\frac{H|\mathcal{S}||\mathcal{A}|}{\delta}\right)$. By taking the right-hand side of (68) to be smaller than $\varepsilon\mathbf{1}$, we immediately conclude the proof. □

8. Discussion

This paper has demonstrated that (some variants of) model-based planning algorithms achieve the minimax sample complexity in the presence of a generative model as soon as the sample size exceeds the order of $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ for γ-discounted infinite-horizon MDPs and $|\mathcal{S}||\mathcal{A}|H^2$ for time-inhomogeneous finite-horizon MDPs (modulo some log factor). Compared with prior literature, our result has considerably broadened the sample size range, allowing us to pin down a complete trade-off curve between sample complexity and statistical accuracy.

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

- Is perturbation or conservative action selection necessary for infinite-horizon MDPs? The planning

algorithm analyzed here for infinite-horizon MDPs is either applied to a perturbed variant of the empirical MDP (as in perturbed model-based planning) or run in a conservative manner (as in conservative model-based planning). This, however, gives rise to a natural question regarding the necessity of perturbation or conservative action selection. Can we achieve optimal performance directly using *plain* model-based planning on the empirical MDP? Although we conjecture that the answer is affirmative, settling this conjecture requires new techniques beyond the analysis framework of this paper.

- Improved analysis for model-free algorithms. As mentioned previously, an even more severe sample complexity barrier is present in all prior theory regarding model-free approaches (e.g., Sidford et al. 2018b, Wainwright 2019b, Li et al. 2022c). Our analysis might shed light on how to overcome such barriers for model-free approaches.
- Time-homogeneous finite-horizon MDPs. When it comes to finite-horizon MDPs, the present work concentrates on time-inhomogeneous MDPs, where the probability transition kernels may vary across time steps. Another important scenario is concerned with time-homogeneous MDPs, where $P_1 = P_2 = \cdots = P_H$. It remains unclear how to develop tight sample analysis for time-homogeneous MDPs because of the lack of statistical independence across time steps (namely, we shall use all samples to estimate the kernels across time steps as they are identical).
- Markovian sample trajectories. Going beyond the generative model, another common form of data samples takes the form of a Markovian sample trajectory, which is generated by taking actions according to a stationary behavior policy in the MDP. This is also referred to as the asynchronous setting in the context of Q learning (Tsitsiklis 1994). Although the sample complexity of several RL algorithms under this data-generating mechanism has been studied in prior literature (e.g., Qu and Wierman 2020; Li et al. 2022c, 2023), it remains unclear how to achieve minimax optimality for the full ε range because of the complicated statistical dependency across time. The recent work by Li et al. (2022b) demonstrated the plausibility of converting a finite-horizon Markovian trajectory into independent samples via twofold sample splitting in the context of offline RL. It would be interesting to investigate whether one could employ a similar idea—in conjunction with a proper leave-one-out analysis framework—to settle the sample complexity in the presence of Markovian samples.
- Online exploratory RL. In practice, there is no shortage of applications where the learner acquires data samples by executing the MDP in real time. This corresponds to an important setting, called online RL, that requires careful managing of the exploration-exploitation tradeoff (Jin et al. 2018, Bai et al. 2019, Li et al. 2021b). Interestingly, the model-based approach—with proper modification to implement optimism in the face of uncertainty—achieves minimax-optimal regret asymptotically (Azar et al. 2017), although its performance in the sample-starved regime remains largely unknown. It would be of great interest to see whether the analysis ideas developed herein could help characterize the sample efficiency of model-based online RL for the entire ε range.
- Beyond the tabular setting. The current paper focuses on the tabular setting with finite state and action spaces. Although we improve the sample size range, the sample complexities might still be prohibitively large when $|\mathcal{S}|$ and $|\mathcal{A}|$ are enormous. Therefore, it is desirable to further investigate settings where low-complexity function approximation is employed to improve the efficiency (e.g., Yang and Wang 2019, Jin et al. 2020, Li et al. 2021a).

Acknowledgments

The authors thank Qiwen Cui for pointing out an issue in an early version of this paper and Shicong Cen, Chen Cheng, and Cong Ma for numerous discussions. Partial results of this paper have been presented in *Advances in Neural Information Processing Systems 33* (Li et al. 2020).

Endnotes

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

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

References

Agarwal A, Kakade S, Yang LF (2020) Model-based reinforcement learning with a generative model is minimax optimal. *Conf. Learn. Theory (COLT)* (PMLR, New York), 67–83.

Azar MG, Munos R, Kappen B (2012) On the sample complexity of reinforcement learning with a generative model. Preprint, submitted June 27, https://arxiv.org/abs/1206.6461.

Azar MG, Munos R, Kappen HJ (2013) Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learn.* 91(3):325–349.

Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. *Proc. 34th Internat. Conf. Machine Learn.*, vol. 70 (PMLR, New York), 263–272.

Bai Y, Xie T, Jiang N, Wang YX (2019) Provably efficient Q-learning with low switching cost. *Adv. Neural Inform. Processing Systems 32* (NeurIPS, San Diego).

Beck CL, Srikant R (2012) Error bounds for constant step-size Q-learning. *Systems Control Lett.* 61(12):1203–1208.

Bellman R (1952) On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA* 38(8):716–719.

Bertsekas DP (2017) *Dynamic Programming and Optimal Control*, 4th ed. (Athena Scientific, Nashua, NH).

Bhandari J, Russo D, Singal R (2018) A finite time analysis of temporal difference learning with linear function approximation. *Conf. Learn. Theory (COLT)* (PMLR, New York), 1691–1692.

Bradtke SJ, Barto AG (1996) Linear least-squares algorithms for temporal difference learning. *Machine Learn.* 22(1–3):33–57.

Cai Q, Yang Z, Lee JD, Wang Z (2019) Neural temporal-difference learning converges to global optima. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 11312–11322.

Chen Y, Chi Y, Fan J, Ma C (2021) Spectral methods for data science: A statistical perspective. *Foundations Trends Machine Learn.* 14(5):566–806.

Chen Y, Fan J, Ma C, Wang K (2019a) Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* 47(4):2204–2235.

Chen Y, Fan J, Ma C, Yan Y (2019b) Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* 116(46):22931–22937.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2020) Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. Preprint, submitted October 21, https://arxiv.org/abs/2002.00874v4.

Cui Q, Yang LF (2021) Minimax sample complexity for turn-based stochastic game. *Proc. Thirty-Seventh Conf. Uncertainty Artificial Intelligence*, vol. 161 (PMLR, New York), 1496–1504.

Dalal G, Szörényi B, Thoppe G, Mannor S (2018) Finite sample analyses for TD(0) with function approximation. *Thirty-Second AAAI Conf. Artificial Intelligence* (AAAI Press, Palo Alto, CA).

Domingues OD, Ménard P, Kaufmann E, Valko M (2021) Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. *Algorithmic Learn. Theory* (PMLR, New York), 578–598.

El Karoui N (2018) On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* 170:95–175.

Even-Dar E, Mansour Y (2003) Learning rates for Q-learning. *J. Mach. Learn. Res.* 5(December):1–25.

Fan J, Wang Z, Xie Y, Yang Z (2019) A theoretical analysis of deep Q-learning. Preprint, submitted May 29, https://arxiv.org/abs/1901.00137v2.

Gupta H, Srikant R, Ying L (2019) Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 4706–4715.

Jaakkola T, Jordan MI, Singh SP (1994) Convergence of stochastic iterative dynamic programming algorithms. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 703–710.

Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 4863–4873.

Jin C, Yang Z, Wang Z, Jordan MI (2020) Provably efficient reinforcement learning with linear function approximation. *Conf. Learn. Theory* (PMLR, New York), 2137–2143.

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

Kaledin M, Moulines E, Naumov A, Tadic V, Wai HT (2020) Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. Preprint, submitted February 4, https://arxiv.org/abs/2002.01268.

Kearns MJ, Singh SP (1999) Finite-sample convergence rates for Q-learning and indirect algorithms. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 996–1002.

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

Khamaru K, Pananjady A, Ruan F, Wainwright MJ, Jordan MI (2020) Is temporal difference learning optimal? An instance-dependent analysis. Preprint, submitted March 16, https://arxiv.org/abs/2003.07337.

Lakshminarayanan C, Szepesvari C (2018) Linear stochastic approximation: How far does constant step-size and iterate averaging go? *Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1347–1355.

Lattimore T, Hutter M (2012) PAC bounds for discounted MDPs. *Internat. Conf. Algorithmic Learn. Theory* (Springer, Berlin), 320–334.

Li G, Chi Y, Wei Y, Chen Y (2022a) Minimax-optimal multi-agent RL in zero-sum Markov games with a generative model. Preprint, submitted October 12, https://arxiv.org/abs/2208.10458.

Li G, Cai C, Chen Y, Wei Y, Chi Y (2023) Is Q-learning minimax optimal? A tight sample complexity analysis. *Oper. Res.* Forthcoming.

Li G, Chen Y, Chi Y, Gu Y, Wei Y (2021a) Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *Adv. Neural Inform. Processing Systems 34* (NeurIPS, San Diego), 16671–16685.

Li G, Shi L, Chen Y, Chi Y, Wei Y (2022b) Settling the sample complexity of model-based offline reinforcement learning. Preprint, submitted April 11, https://arxiv.org/abs/2204.05275v1.

Li G, Shi L, Chen Y, Gu Y, Chi Y (2021b) Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Adv. Neural Inform. Processing Systems 34* (NeurIPS, San Diego), 17762–17776.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2020) Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Adv. Neural Inform. Processing Systems 33* (NeurIPS, San Diego), 12861–12872.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2022c) Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Trans. Inform. Theory* 68(1):448–473.

Li G, Cai C, Chen Y, Gu Y, Wei Y, Chi Y (2021c) Tightening the dependence on horizon in the sample complexity of Q-learning. *Internat. Conf. Machine Learn.* (PMLR, New York), 6296–6306.

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

Mou W, Li CJ, Wainwright MJ, Bartlett PL, Jordan MI (2020) On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. Preprint, submitted April 9, https://arxiv.org/abs/2004.04719.

Pananjady A, Wainwright MJ (2019) Value function estimation in Markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. Preprint, submitted September 19, https://arxiv.org/abs/1909.08749v1.

Qu G, Wierman A (2020) Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conf. Learn. Theory*, 3185–3205.

Shah D, Xie Q (2018) Q-learning with nearest neighbors. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 3111–3121.

Shi L, Li G, Wei Y, Chen Y, Chi Y (2022) Pessimistic Q-learning for offline reinforcement learning: Toward optimal sample complexity. *Internat. Conf. Machine Learn.* (PMLR, New York).

Sidford A, Wang M, Wu X, Ye Y (2018a) Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Proc. Twenty-Ninth Annual ACM-SIAM Sympos. Discrete Algorithms*, 770–787.

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

Srikant R, Ying L (2019) Finite-time error bounds for linear stochastic approximation and TD learning. *Conf. Learn. Theory* (PMLR, New York), 2803–2830.

Strehl AL, Li L, Wiewiora E, Langford J, Littman ML (2006) PAC model-free reinforcement learning. *Internat. Conf. Machine Learn.* (PMLR, New York), 881–888.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Szepesvári C (1998) The asymptotic convergence-rate of Q-learning. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 1064–1070.

Szepesvári C (2010) *Algorithms for Reinforcement Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning (Springer, Cham, Switzerland).

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

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

Tu S, Recht B (2019) The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *Conf. Learn. Theory* (PMLR, New York), 3036–3083.

Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47 (Cambridge University Press, Cambridge, UK).

Wainwright MJ (2019a) Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. Preprint, submitted June 24, https://arxiv.org/abs/1905.06265v2.

Wainwright MJ (2019b) Variance-reduced Q-learning is minimax optimal. Preprint, submitted August 8, https://arxiv.org/abs/1906.04697.

Wang M (2020) Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Math. Oper. Res.* 45(2):517–546.

Wang B, Yan Y, Fan J (2021) Sample-efficient reinforcement learning for linearly-parameterized MDPs with a generative model. *Neural Inform. Processing Systems* (NeurIPS, San Diego), 23009–23022.

Xu P, Gu Q (2020) A finite-time analysis of Q-learning with neural network function approximation. *Internat. Conf. Machine Learn.* (PMLR, New York), 10555–10565.

Xu T, Zou S, Liang Y (2019) Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Adv. Neural Inform. Processing Systems* (NeurIPS, San Diego), 10633–10643.

Yan Y, Chen Y, Fan J (2021) Inference for heteroskedastic PCA with missing data. Preprint, submitted July 26, https://arxiv.org/abs/2107.12365.

Yan Y, Li G, Chen Y, Fan J (2022a) Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. Preprint, submitted June 8, https://arxiv.org/abs/2206.04044.

Yan Y, Li G, Chen Y, Fan J (2022b) The efficacy of pessimism in asynchronous Q-learning. Preprint, submitted March 14, https://arxiv.org/abs/2203.07368.

Yang L, Wang M (2019) Sample-optimal parametric Q-learning using linearly additive features. *Internat. Conf. Machine Learn.* (PMLR, New York), 6995–7004.

Yin M, Bai Y, Wang YX (2021) Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. *Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1567–1575.

Gen Li is a postdoctoral researcher in the Department of Statistics and Data Science at the Wharton School, University of Pennsylvania. His research interests include reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization. He has received the Excellent Graduate Award and the Excellent Thesis Award from Tsinghua University.

Yuting Wei is an assistant professor of Statistics and Data Science at the Wharton School, University of Pennsylvania. She was the recipient of the 2022 National Science Foundation CAREER Award, an honorable mention for the 2023 Bernoulli Society's New Researcher Award, and the 2018 Erich L. Lehmann Citation from the Berkeley Statistics Department. Her research interests include high-dimensional statistics, nonparametric statistics, statistical machine learning, and reinforcement learning.

Yuejie Chi is a professor in the Department of Electrical and Computer Engineering and a faculty affiliate with the Machine Learning Department and CyLab at Carnegie Mellon University. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning, and inverse problems, with applications in sensing, imaging, decision making, and societal systems, broadly defined.

Yuxin Chen is an associate professor of statistics and data science and of electrical and systems engineering at the University of Pennsylvania. His research interests include statistics, optimization, and machine learning. He has received the Alfred P. Sloan Research Fellowship, the International Consortium of Chinese Mathematicians Best Paper Award, the Princeton Graduate Mentoring Award, and was selected as a finalist for the Best Paper Prize for Young Researchers in Continuous Optimization.