Sample-Efficient Reinforcement Learning for Linearly-Parameterized MDPs with a Generative Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

The curse of dimensionality is a widely known issue in reinforcement learning (RL). In the tabular setting where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite, to obtain a nearly optimal policy with sampling access to a generative model, the minimax-optimal sample complexity scales linearly with $|\mathcal{S}| \times |\mathcal{A}|$, which can be prohibitively large when $\mathcal{S}$ or $\mathcal{A}$ is large. This paper considers a Markov decision process (MDP) that admits a set of state-action features, which can linearly express (or approximate) its probability transition kernel. We show that a model-based approach (resp. Q-learning) provably learns an ε-optimal policy (resp. Q-function) with high probability as soon as the sample size exceeds the order of $\frac{K}{(1-\gamma)^3\varepsilon^2}$ (resp. $\frac{K}{(1-\gamma)^4\varepsilon^2}$), up to some logarithmic factor. Here K is the feature dimension and $\gamma \in (0,1)$ is the discount factor of the MDP. The results is applicable to the tabular MDPs by taking the coordinate basis with $K = |\mathcal{S}| \times |\mathcal{A}|$. Both sample complexity bounds are provably tight, and our result for the model-based approach matches the minimax lower bound. Our results show that for arbitrarily large-scale MDP, both the model-based approach and Q-learning are sample-efficient when K is relatively small, and hence the title of this paper.

1 Introduction

Reinforcement learning (RL) studies the problem of learning and decision making in a Markov decision process (MDP). Recent years have seen exciting progress in applications of RL in real world decision-making problems such as AlphaGo [44, 45] and autonomous driving [33]. Specifically, the goal of RL is to search for an optimal policy that maximizes the cumulative reward, based on sequential noisy data. There are two popular approaches to RL: model-based and model-free ones.

- The model-based approaches start with formulating an empirical MDP by learning the probability transition model from the collected data samples, and then estimating the optimal policy / value function based on the empirical MDP.
- The model-free approaches (e.g. Q-learning) learn the optimal policy or the optimal (action-)value function from samples. As its name suggests, model-free approaches do not attempt to learn the model explicitly.

Generally speaking, model-based approaches enjoy great flexibility since after the transition model is learned in the first place, it can then be applied to any other problems without touching the raw data samples. In comparison, model-free methods, due to its online nature, are usually memory-efficient and can interact with the environment and update the estimate on the fly.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

This paper is devoted to investigating the sample efficiency of both model-based RL and Q-learning (arguably one of the most commonly adopted model-free RL algorithms). It is well known that MDPs suffer from the curse of dimensionality. For example, in the tabular setting where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite, to obtain a near optimal policy or value function given sampling access to a generative model, the minimax optimal sample complexity scales linearly with $|\mathcal{S}| \times |\mathcal{A}|$ [5, 2]. However contemporary applications of RL often encounters environments with exceedingly large state and action spaces, whilst the data collection might be expensive or even high-stake. This suggests a large gap between the theoretical findings and practical decision-making problems where $|\mathcal{S}|$ and $|\mathcal{A}|$ are large or even infinite.

To close the aforementioned theory-practice gap, one natural idea is to impose certain structural assumption on the MDP. In this paper we follow the feature-based linear transition model studied in [61], where each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ admits a K dimensional feature vector $\phi(s, a) \in \mathbb{R}^K$ that expresses the transition dynamics $\mathbb{P}(\cdot|s, a) = \Psi\phi(s, a)$ for some unknown matrix $\Psi \in \mathbb{R}^{|\mathcal{S}| \times K}$ which is common for all (s, a). This model encompasses both the tabular case and the homogeneous model in which the state space can be partitioned into K equivalent classes. Assuming access to a generative model [31, 32], under this structural assumption, this paper aims to answer the following two questions:

> *How many samples are needed for model-based RL and Q-learning to learn an optimal policy under the feature-based linear transition model?*

In what follows, we will show that the answer to this question scales linearly with the dimension of the feature space K and is independent of $|\mathcal{S}|$ and $|\mathcal{A}|$ under the feature-based linear transition model. With the aid of this structural assumption, model-based RL and Q-learning becomes significantly more sample-efficient than that in the tabular setting.

Our contributions. We focus our attention on an infinite horizon MDP with discount factor $\gamma \in (0, 1)$. We use ε-optimal policy to indicate the policy whose expected discounted cumulative rewards are ε close to the optimal value of the MDP. Our contributions are two-fold:

- We demonstrate that model-based RL provably learns an ε-optimal policy by performing planning based on an empirical MDP constructed from a total number of

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3 \varepsilon^2}\right)$$

  samples, for all $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Here $\widetilde{O}(\cdot)$ hides logarithmic factors compared to the usual $O(\cdot)$ notation. To the best of our knowledge, this is the first theoretical guarantee for model-based RL under the feature-based linear transition model. This sample complexity bound matches the minimax limit established in [61] up to logarithmic factor.

- We also show that Q-learning provably finds an entrywise ε-optimal Q-function using a total number of

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^4 \varepsilon^2}\right)$$

  samples, for all $\varepsilon \in (0, 1]$. This sample complexity upper bound improves the state-of-the-art result in [61] and the dependency on the effective horizon $(1-\gamma)^{-4}$ is sharp in view of [34].

These results taken collectively show the minimax optimality of model-based RL and the sub-optimality of Q-learning in sample complexity.

2 Problem formulation

This paper focuses on tabular MDPs in the discounted infinite-horizon setting [10]. Here and throughout, $\Delta_{d-1} := \{v \in \mathbb{R}^d : \sum_{i=1}^d v_i = 1, v_i \geq 0, \forall i \in [d]\}$ stands for the d-dimensional probability simplex and $[N] := \{1, 2, \cdots, N\}$ for any $N \in \mathbb{N}^+$.

Discounted infinite-horizon MDPs. Denote a discounted infinite-horizon MDP by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$ is a finite set of states, $\mathcal{A} = \{1, \cdots, |\mathcal{A}|\}$ is a finite set of actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{|\mathcal{S}|-1}$ represents the probability transition kernel where $P(s'|s, a)$ denotes the probability of transiting from state s to state s' when action a is taken, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the reward function where $r(s, a)$ is the instantaneous reward received when taking action $a \in \mathcal{A}$ while in state $s \in \mathcal{S}$, and $\gamma \in (0, 1)$ is the discount factor.

Value function and Q-function. Recall that the goal of RL is to learn a policy that maximizes the cumulative reward, which corresponds to value functions or Q-functions in the corresponding MDP. For a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ and a starting state $s \in \mathcal{S}$, we define the value function as

$$V^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \;\middle|\; s_0 = s\right]$$

for all $s \in \mathcal{S}$. Here, the trajectory is generated by $a_k = \pi(s_k)$ and $s_{k+1} \sim P(s_{k+1}|s_k, a_k)$ for every $k \geq 0$. This function measures the expected discounted cumulative reward received on the trajectory $\{(s_k, a_k)\}_{k\geq 0}$ and the expectation is taken with respect to the randomness of the transitions $s_{k+1} \sim P(\cdot|s_k, a_k)$ on the trajectory. Recall that the immediate rewards lie in $[0, 1]$, it is easy to derive that $0 \leq V^{\pi}(s) \leq \frac{1}{1-\gamma}$ for any policy π and state s. Accordingly, we define the Q-function for policy π as

$$Q^{\pi}(s, a) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \,|s_0 = s, a_0 = a\right]$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, the actions are chosen by the policy π except for the initial state (i.e. $a_k = \pi(s_k)$ for all $k \geq 1$). Similar to the value function, we can easily check that $0 \leq Q^{\pi}(s, a) \leq \frac{1}{1-\gamma}$ for any π and (s, a). To maximize the value function or Q function, previous literature [9, 49] establishes that there exists an optimal policy $\pi^{\star}$ which simultaneously maximizes $V^{\pi}(s)$ (resp. $Q^{\pi}(s, a)$) for all $s \in \mathcal{S}$ (resp. $(s, a) \in \mathcal{S} \times \mathcal{A}$). We define the optimal value function $V^{\star}$ and optimal Q-function $Q^{\star}$ respectively as

$$V^{\star}(s) \coloneqq \max_{\pi} V^{\pi}(s) = V^{\pi^{\star}}(s), \qquad Q^{\star}(s, a) \coloneqq \max_{\pi} Q^{\pi}(s, a) = Q^{\pi^{\star}}(s, a)$$

for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Linear transition model. Given a set of K feature functions $\phi_1, \phi_2, \cdots, \phi_K$: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define ϕ to be a feature mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^K$ such that

$$\phi(s, a) = [\phi_1(s, a), \cdots, \phi_K(s, a)] \in \mathbb{R}^K.$$

Then we are ready to define the linear transition model [61] as follows.

Definition 1 (Linear transition model)**.** . *Given a discounted infinite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$, M admits the linear transition model if there exists some (unknown) functions $\psi_1, \cdots, \psi_K : \mathcal{S} \to \mathbb{R}$, such that*

$$P(s'|s, a) = \sum_{k=1}^{K} \phi_k(s, a)\, \psi_k(s') \tag{1}$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$.

Readers familiar with linear MDP literatures might immediately recognize that the above definition is the same as the structure imposed on the probability transition kernel P in the linear MDP model [61, 30, 63, 26, 50, 54, 58]. However unlike linear MDP which also requires the reward function $r(s, a)$ to be linear in the feature mapping $\phi(s, a)$, here we do not impose any structural assumption on the reward.

Example 1 (Tabular MDP)**.** *Each tabular MDP can be viewed as a linear transition model with feature mapping $\phi(s, a) = \boldsymbol{e}_{(s,a)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ (i.e. the vector with all entries equal to 0 but the one corresponding to (s, a) equals to 1) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. To see this, we can check that Definition 1 is satisfied with $K = |\mathcal{S}| \times |\mathcal{A}|$ and $\psi_{(s,a)}(s') = \mathbb{P}(s'|s, a)$ for each $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. This example is a sanity check of Definition 1, which also shows that our results (Theorem 1 and 2) can recover previous results on tabular MDP [2, 34] by taking $K = |\mathcal{S}| \times |\mathcal{A}|$.*

Example 2 (Simplex Feature Space)**.** *If all feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ fall in the probability simplex Δ_{K-1}, a linear transition model can be constructed by taking $\psi_k(\cdot)$ to be any probability measure over $\mathcal{S}$ for all $k \in [K]$.*

A key observation is that the model size of linear transition model with known feature mapping ϕ is $|\mathcal{S}|K$ (the number of coefficients $\psi_k(s')$ in (1)), which is still large when the state space $\mathcal{S}$ is large. In contrast, it will be established later that to learn a near-optimal policy or Q-function, we only need a much smaller number of samples, which depends linearly on K and is independent of $|\mathcal{S}|$.

Next, we introduce a critical assumption employed in prior literature [61].

Assumption 1 (Anchor state-action pairs)**.** *Assume there exists a set of anchor state-action pairs $\mathcal{K} \subset \mathcal{S}\times\mathcal{A}$ with $|\mathcal{K}| = K$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, its corresponding feature vector can be expressed as a convex combination of the feature vectors of anchor state-action pairs $\{(s,a) : (s,a) \in \mathcal{K}\}$:*

$$\phi(s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i(s,a)\,\phi(s_i,a_i) \quad \textit{for} \quad \sum_{i=1}^{K} \lambda_i(s,a) = 1 \quad \textit{and} \quad \lambda_i(s,a) \geq 0. \tag{2}$$

Further, we assume that the vectors in $\{\phi(s,a) : (s,a) \in \mathcal{K}\}$ are linearly independent.

We pause to develop some intuition of this assumption using Examples 1 and 2. In Example 1, it is straightforward to check that tabular MDPs satisfies Assumption 1 with $\mathcal{K} = \mathcal{S}\times\mathcal{A}$. In terms of Example 2, without loss of generality we can assume that the subspace spanned by the features has full rank, i.e. $\mathsf{span}\{\phi(s,a) : (s,a) \in \mathcal{S}\times\mathcal{A}\} = \mathbb{R}^K$ (otherwise we can reduce the dimension of feature space). Then we can also check that Example 2 satisfies Assumption 1 with arbitrary $\mathcal{K} \subseteq \mathcal{S}\times\mathcal{A}$ such that the vectors in $\{\phi(s,a) : (s,a) \in \mathcal{K}\}$ are linearly independent. In fact, this sort of "anchor" notion appears widely in the literature. Besides [61], [3] considers "anchor word" in topic modeling; [19] defines "separability" in their study of non-negative matrix factorization; [46] introduces "aggregate" in reinforcement learning; [21] studies "anchor state" in soft state aggregation models. These concepts all bear some kind of resemblance to our definition of anchor state-action pairs here.

A useful implication of Assumption 1 is that we can represent the transition kernel as

$$P(\cdot|s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i(s,a)\,P(\cdot|s_i,a_i), \tag{3}$$

This follows simply from substituting (2) into (1) (see (14) in Appendix A for a formal proof).

3 Model-based RL with a generative model

We start with studying model-based RL with a generative model in this section. We propose a model-based planning algorithm and show that it returns an ε-optimal policy with minimax optimal sample size.

3.1 Main results

A generative model and an empirical MDP. We assume access to a generative model that provides us with independent samples from M. For each anchor state-action pair $(s_i,a_i) \in \mathcal{K}$, we collect N independent samples $s_i^{(j)} \sim P(\cdot|s_i,a_i)$, $j \in [N]$. This allows us to construct an empirical transition kernel $\widehat{P}$ where

$$\widehat{P}(s' \mid s,a) = \sum_{i=1}^{K} \lambda_i(s,a) \cdot \left(\frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\left\{ s_i^{(j)} = s' \right\} \right), \tag{4}$$

for each $(s,a) \in \mathcal{S}\times\mathcal{A}$. Here, $\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{s_i^{(j)} = s'\}$ is an empirical estimate of $P(s'|s_i,a_i)$ and then (3) is employed. With $\widehat{P}$ in hand, we can construct an empirical MDP $\widehat{M} = (\mathcal{S},\mathcal{A},\widehat{P},r,\gamma)$. Our goal here is to derive the sample complexity which guarantees that the optimal policy of $\widehat{M}$ is an ε-optimal policy for the true MDP M. The algorithm is summarized below.

Algorithm 1 Model-based RL with a generative model

Inputs: a set of anchor state-action pairs $\mathcal{K} = \{(s_i, a_i) : i \in [K]\}$, feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$, any planning algorithm $\mathcal{P}$, target algorithmic errror level $\varepsilon_{\mathsf{opt}}$.
For $i = 1, \cdots, K$ **do**
  Draw N independent samples $s_i^{(j)} \sim P(\cdot|s_i, a_i)$, $j = 1, \cdots, N$.
End for
Construct an empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$ where $\widehat{P}$ can be computed by (4).
Output $\widehat{\pi}$ as an $\varepsilon_{\mathsf{opt}}$-optimal policy of $\widehat{M}$ computed by the planning algorithm $\mathcal{P}$.

Then the following theorem provides theoretical guarantees for the output policy $\widehat{\pi}$ of the chosen optimization algorithm on the empirical MDP $\widehat{M}$.

Theorem 1. *Suppose that $\delta > 0$ and $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Let $\widehat{\pi}$ be the policy returned by Algorithm 1. Assume that*

$$N \geq \frac{C \log\left(K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)^3\,\varepsilon^2} \tag{5}$$

for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - \delta$,

$$Q^\star(s,a) - Q^{\widehat{\pi}}(s,a) \leq \varepsilon + \frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma}, \tag{6}$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$.

We first remark that the two terms on the right hand side of (6) can be viewed as statistical error and algorithmic error, respectively. The first term ε denotes the statistical error coming from the deviation of the empirical MDP $\widehat{M}$ from the true MDP M. As the sample size N grows, ε could decrease towards 0. The other term $4\varepsilon_{\mathsf{opt}}/(1-\gamma)$ represents the algorithmic error where $\varepsilon_{\mathsf{opt}}$ is the target accuracy level of the planning algorithm applied to $\widehat{M}$. Note that $\varepsilon_{\mathsf{opt}}$ can be arbitrarily small if we run the planning algorithm (e.g. value iteration) for enough iterations. A few implications of this theorem are in order.

- *Minimax-optimal sample complexity*. Assume that $\varepsilon_{\mathsf{opt}}$ is made negligibly small, e.g. $\varepsilon_{\mathsf{opt}} = O((1-\gamma)\varepsilon)$ to be discussed in the next point. Note that we draw N independent samples for each state-action pair $(s, a) \in \mathcal{K}$, therefore the requirement (5) for finding an $O(\varepsilon)$-optimal policy translates into the following sample complexity requirement

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3\,\varepsilon^2}\right).$$

  This matches the minimax optimal lower bound (up to a logarithm factor) established in [61, Theorem 1] for feature-based MDP. In comparison, for tabular MDP the minimax optimal sample complexity is $\widetilde{\Omega}((1-\gamma)^{-3}\varepsilon^{-2}|\mathcal{S}||\mathcal{A}|)$ [5, 2]. Our sample complexity scales linearly with K instead of $|\mathcal{S}||\mathcal{A}|$ for tabular MDP as desired.

- *Computational complexity*. An advantage of Theorem 1 is that it incorporates the use of any efficient planning algorithm applied to the empirical MDP $\widehat{M}$. Classical algorithms include Q-value iteration (QVI) or policy iteration (PI) [42]. For example, QVI achieves the target level $\varepsilon_{\mathsf{opt}}$ in $O((1-\gamma)^{-1}\log\varepsilon_{\mathsf{opt}}^{-1})$ iterations, and each iteration takes time proportional to $O(NK + |\mathcal{S}||\mathcal{A}|K)$. To learn an $O(\varepsilon)$-optimal policy, which requires sample complexity (5) and the target level $\varepsilon_{\mathsf{opt}} = O((1-\gamma)\varepsilon)$, the overall running time is

$$\widetilde{O}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|\,K}{1-\gamma} + \frac{K}{(1-\gamma)^4\,\varepsilon^2}\right).$$

  In comparison, for the tabular MDP the corresponding running time is $\widetilde{O}((1-\gamma)^{-4}\varepsilon^{-2}|\mathcal{S}||\mathcal{A}|)$ [2]. This suggests that under the feature-based linear transition model, the computational complexity is $\min\{|\mathcal{S}||\mathcal{A}|/K, (1-\gamma)^{-3}\varepsilon^{-2}/K\}$ times lower than that for the tabular MDP (up to logarithm factors), which is significantly more efficient when K is not too large.

- *Stability vis-a-vis model misspecification*. A more general version of Theorem 1 (Theorem 3 in Appendix B) shows that when P approximately (instead of exactly) admits the linear transition model, we can still achieve some meaningful result. Specifically, if there exists a linear transition kernel $\widetilde{P}$ obeying $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\widetilde{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \le \xi$ for some $\xi \ge 0$, we can show that $\widehat{\pi}$ returned by Algorithm 1 (with slight modification) satisfies

$$Q^\star(s,a) - Q^{\widehat{\pi}}(s,a) \le \varepsilon + \frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma} + \frac{22\xi}{(1-\gamma)^2},$$

  for every $(s,a) \in \mathcal{S}\times\mathcal{A}$. This shows that the model-based method is stable vis-a-vis model misspecification. Interested readers are referred to Appendix B for more details.

4 Model-free RL—vanilla Q Learning

In this section, we turn to study one of the most popular model-free RL algorithms—Q-learning. We provide tight sample complexity bound for vanilla Q-learning under the feature-based linear transition model, which shows its sample-efficiency (depends on $|K|$ instead of $|\mathcal{S}|$ or $|\mathcal{A}|$) and sub-optimality in the dependency on the effective horizon.

4.1 Q-learning algorithm

The vanilla Q-learning algorithm maintains a Q-function estimate $Q_t : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ for all $t \ge 0$, with initialization Q_0 obeying $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$. Assume we have access to a generative model. In each iteration $t \ge 1$, we collect an independent sample $s_t(s,a) \sim P(\cdot|s,a)$ for every anchor state-action pair $(s,a) \in \mathcal{K}$ and define $Q_{\mathcal{K}}^{(t)} : \mathcal{K} \to \mathbb{R}$ to be

$$Q_{\mathcal{K}}^{(t)}(s,a) := \max_{a'\in\mathcal{A}} Q_t(s_t, a'), \qquad s_t \equiv s_t(s,a) \sim P(\cdot|s,a).$$

Then given the learning rate $\eta_t \in (0,1]$, the algorithm adopts the following update rule to update all entries of the Q-function estimate

$$Q_t = (1-\eta_t) Q_{t-1} + \eta_t \mathcal{T}_{\mathcal{K}}^{(t)}(Q_{t-1}).$$

Here, $\mathcal{T}_{\mathcal{K}}^{(t)}$ is an empirical Bellman operator associated with the linear transition model M and the set $\mathcal{K}$ and is given by

$$\mathcal{T}_{\mathcal{K}}^{(t)}(Q)(s,a) := r(s,a) + \gamma\lambda(s,a) Q_{\mathcal{K}}^{(t)},$$

where (3) is used in the construction. Clearly, this newly defined operator $\mathcal{T}_{\mathcal{K}}^{(t)}$ is an unbiased estimate of the famous Bellman operator $\mathcal{T}$ [8] defined as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad \mathcal{T}(Q)(s,a) := r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right].$$

A critical property is that the Bellman operator $\mathcal{T}$ is contractive with a unique fixed point which is the optimal Q-function $Q^\star$[8]. To solve the fixed-point equation $\mathcal{T}(Q^\star) = Q^\star$, Q-learning was then introduced by [56] based on the idea of stochastic approximation [43]. This procedure is precisely described in Algorithm 2.

Algorithm 2 Vanilla Q-learning for infinite-horizon discounted MDPs

inputs: learning rates $\{\eta_t\}$, number of iterations T, discount factor γ, initial estimate Q_0.
for $t = 1, \ldots, T$ **do**
  Draw $s_t(s,a) \sim P(\cdot|s,a)$ for each $(s,a) \in \mathcal{K}$.
  Compute $\boldsymbol{Q}_t$ according to the update rule

$$Q_t = (1-\eta_t) Q_{t-1} + \eta_t \mathcal{T}_{\mathcal{K}}^{(t)}(Q_{t-1}).$$

end for

4.2 Main results

We are now ready to provide our main result for vanilla Q-learning, assuming sampling access to a generative model.

Theorem 2. *Consider any $\delta \in (0,1)$ and $\varepsilon \in (0,1]$. Assume that for any $0 \le t \le T$, the learning rates satisfy*

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}} \tag{7}$$

for some sufficiently small universal constants $c_1 \ge c_2 > 0$. Suppose that the total number of iterations T exceeds

$$T \ge \frac{C_3 \log\left(KT/\delta\right) \log^4 T}{\left(1-\gamma\right)^4 \varepsilon^2} \tag{8}$$

for some sufficiently large universal constant $C_3 > 0$. If the initialization obeys $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, then with probability exceeding $1-\delta$, the output Q_T of Algorithm 2 satisfies

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q_T\left(s,a\right) - Q^\star\left(s,a\right)\right| \le \varepsilon. \tag{9}$$

This theorem provides theoretical guarantees on the performance of Algorithm 2. A few implications of this theorem are in order.

- *Learning rate*. The condition (7) accommodates two commonly adopted choice of learning rates: (i) linearly rescaled learning rates $\eta_t = [1 + c_2(1-\gamma)t/\log^2 T]^{-1}$, and (ii) iteration-invariant learning rates $\eta_t \equiv [1 + c_1(1-\gamma)T/\log^2 T]$. Interested readers are referred to the discussions in [34, Section 3.1] for more details on these two learning rate schemes.
- *Tight sample complexity bound*. Note that we draw K independent samples in each iteration, therefore the iteration complexity (8) can be translated into the sample complexity bound TK in order for Q-learning to achieve ε-accuracy:

$$\widetilde{O}\left(\frac{K}{\left(1-\gamma\right)^4 \varepsilon^2}\right). \tag{10}$$

  As we will see shortly, this result improves the state-of-the-art sample complexity bound presented in [61, Theorem 2] . In addition, the dependency on the effective horizon $(1-\gamma)^{-4}$ matches the lower bound established in [34, Theorem 2] for vanilla Q-learning using either learning rate scheme covered in the previous remark, suggesting that our sample complexity bound (10) is sharp.
- *Stability vis-a-vis model misspecification*. Just like the model-based approach, we can also show that Q-learning is also stable vis-a-vis model misspecification when P approximately admits the linear transition model. We refer interested readers to Theorem 4 in Appendix B for more details.

Comparison with [61]. We compare our result with the sample complexity bounds for Q-learning under the feature-based linear transition model in [61].

- We first compare our result with [61, Theorem 2], which is, to the best of our knowledge, the state-of-the-art theory for this problem. When there is no model misspecification, [61, Theorem 2] showed that in order for their Phased Parametric Q-learning (Algorithm 1 therein) to learn an ε-optimal policy, the sample size needs to be

$$\widetilde{O}\left(\frac{K}{\left(1-\gamma\right)^7 \varepsilon^2}\right).$$

  Note that (10) is the sample complexity required for entrywise ε-accurate estimate of the optimal Q-function, thus a fair comparison requires to convert our bound (10) to the one required for learning an ε-optimal policy, which is

$$\widetilde{O}\left(\frac{K}{\left(1-\gamma\right)^6 \varepsilon^2}\right),$$

  according to [47]. Hence, our sample complexity improves upon previous work by a factor at least on the order of $(1-\gamma)^{-1}$.

- Besides vanilla Q-learning, [61] also proposed a new variant of Q-learning called Optimal Phased Parametric Q-Learning (Algorithm 2 therein), which is essentially Q-learning with variance reduction. [61, Theorem 3] showed that the sample complexity for this algorithm is

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3\varepsilon^2}\right),$$

  which matches minimax optimal lower bound (up to a logarithm factor) established in [61, Theorem 1]. Careful reader might notice that this sample complexity bound is better than ours for vanilla Q-learning. We emphasize that as elucidated in the second implication under Theorem 2, our result is already tight for vanilla Q-learning. This observation reveals that while the sample complexity for vanilla Q-learning is provably sub-optimal, the variants of Q-learning can have better performance and achieve minimax optimal sample complexity.

We conclude this section by comparing model-based and model-free approaches. Theorem 1 shows that the sample complexity of the model-based approach is minimax optimal, whilst vanilla Q-learning, perhaps the most commonly adopted model-free method, is sub-optimal according to Theorem 2. However this does not mean that model-based method is better than model-free ones since (i) some variants of Q-learning (see [61, Algorithm 2] for example) also has minimax optimal sample complexity; and (ii) in many applications it might be unrealistic to estimate the model in advance.

5 A glimpse of our technical approaches

The establishment of Theorems 1 and 2 calls for a series of technical novelties in the proof. In what follows, we briefly highlight our key technical ideas and novelties.

- For the model-based approach, we employ "leave-one-out" analysis to decouple the complicated statistical dependency between the empirical probability transition model $\widehat{P}$ and the corresponding optimal policy. Specifically, [2] proposed to construct a collection of auxiliary MDPs where each one of them leaves out a single state s by setting s to be an absorbing state and keeping everything else untouched. We tailor this high level idea to the needs of linear transition model, then the independence between the newly constructed MDP with absorbing state s and data samples collected at state s will facilitate our analysis, as detailed in Lemma 1. This "leave-one-out" type of analysis has been utilized in studying numerous problems by a long line of work, such as [22, 37, 51, 13, 12, 14, 15], just to name a few.
- To obtain tighter sample complexity bound than the previous one $\widetilde{O}(\frac{K}{(1-\gamma)^7\varepsilon^2})$ in [61] for vanilla Q-learning, we invoke Freedman's inequality [24] for the concentration of an error term with martingale structure as illustrated in Appendix C, while the classical ones used in analyzing Q-learning are Hoeffding's inequality and Bernstein's inequality [61]. The use of Freedman's inequality helps us establish a recursive relation on $\{\|Q_t - Q^\star\|_\infty\}_{t=0}^T$, which consequently leads to the performance guarantee (9).

6 Additional related literature

To remedy the issue of prohibitively high sample complexity, there exists a substantial body of literature proposing and studying many structural assumptions and complexity notions under different settings. This current paper focuses on linear transition model which is studied in MDP by numerous previous works [61, 30, 62, 65, 39, 25, 54, 50, 26, 58]. Other popular structure assumptions include: [59] studied fully deterministic transition dynamics; [28] introduced Bellman rank and proposed an algorithm which needs sample size polynomially dependent on Bellman rank to obtain a near-optimal policy in contextual decision processes; [20] assumed the value function has low variance compared to the mean for all deterministic policy; [38, 41, 7, 64] used linear model to approximate the value function.

Apart from the linear transition model, another notion adopted in this work is the generative model, whose role in discounted MDP has been studied by extensive literature. The concept of generative model was originally introduced by [32], and then widely adopted in numerous works, including

[31, 5, 61, 52, 2, 40], to name a few. Specifically, it is assumed that a generative model of the studied MDP is available and can be queried for every state-action pair and output the next state. Among previous works, [5] proved that the minimax lower bound on the sample complexity to obtain an ε-optimal policy was $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$. [5] also showed that model-based approach can output an ε-optimal value function with near-optimal sample complexity for $\varepsilon \in (0, 1)$. Then [2] made significant progress on the challenging problem of establishing minimax optimal sample complexity in estimating an ε-optimal policy with the help of "leave-one-out" analysis.

In addition, after being proposed in [57], Q-learning has become the focus of a rich line of research [56, 11, 32, 23, 4, 29, 51, 16, 36, 60]. Among them, [16, 36, 60] studied Q-learning in the presence of Markovian data, i.e. a single sample trajectory. In contrast, under the generative setting of Q-learning where a fresh sample can be drawn from the simulator at each iteration, [52] analyzed a variant of Q-learning with the help of variance reduction, which was proved to enjoy minimax optimal sample complexity $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$. Then more recently, [34] improved the lower bound of the vanilla Q-learning algorithm in terms of its scaling with $\frac{1}{1-\gamma}$ and proved a matching upper bound $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2})$.

7 Discussion

This paper studies sample complexity of both model-based and model-free RL under a discounted infinite-horizon MDP with feature-based linear transition model. We establish tight sample complexity bounds for both model-based approaches and Q-learning, which scale linearly with the feature dimension K instead of $|\mathcal{S}| \times |\mathcal{A}|$, thus considerably reduce the required sample size for large-scale MDPs when K is relatively small. Our results are sharp, and the sample complexity bound for the model-based approach matches the minimax lower bound. The current work suggests a couple of directions for future investigation, as discussed in detail below.

- *Extension to episodic MDPs.* An interesting direction for future research is to study linear transition model in episodic MDP. This focus of this work is infinite-horizon discounted MDPs, and hopefully the analysis here can be extended to study the episodic MDP as well ([17, 18, 6, 27, 53, 26]).
- *Continuous state and action space.* The state and action spaces in this current paper are still assumed to be finite, since the proof relies heavily on the matrix operations. However, we expect that the results can be extended to accommodate continuous state and action space by employing more complicated analysis tool.
- *Accommodating entire range of ε.* Since both value functions and Q-functions can take value in $[0, (1-\gamma)^{-1}]$, ideally our theory should cover all choices of $\varepsilon \in (0, (1-\gamma)^{-1}]$. However we require that $\varepsilon \in (0, (1-\gamma)^{-1/2}]$ in Theorem 1 and $\varepsilon \in (0, 1]$ in Theorem 2. While most of the prior works like [2, 61] also impose these restrictions, a recent work [35] proposed a perturbed model-based planning algorithm and proved minimax optimal guarantees for any $\varepsilon \in (0, (1-\gamma)^{-1}]$. While their work only focused on model-based RL under tabular MDP, an interesting future direction is to improve our theory to accommodate any $\varepsilon \in (0, (1-\gamma)^{-1}]$.
- *General function approximation.* Another future direction is to extend the study to more general function approximation starting from linear structure covered in this paper. There exists a rich body of work proposing and studying different structures, such as linear value function approximation [38, 41, 7, 64], linear MDPs with infinite dimensional features [1], Eluder dimension [55], Bellman rank [28] and Witness rank [48], etc. Therefore, it is hopeful to investigate these settings and improve the sample efficiency.

References

[1] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

[4] Mohammad Gheshlaghi Azar, Remi Munos, M Ghavamzadaeh, and Hilbert J Kappen. Speedy q-learning. 2011.

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

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[7] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

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

[9] Richard Bellman and Stuart Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, pages 247–251, 1959.

[10] Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

[11] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[12] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

[13] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

[14] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020.

[15] Yuxin Chen, Jianqing Fan, Bingyan Wang, and Yuling Yan. Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution. *arXiv preprint arXiv:2008.01724*, 2020.

[16] Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.

[17] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *arXiv preprint arXiv:1510.08906*, 2015.

[18] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

[19] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *17th Annual Conference on Neural Information Processing Systems, NIPS 2003*. Neural information processing systems foundation, 2004.

[20] Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*, 2019.

[21] Yaqi Duan, Zheng Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

[23] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for q-learning. *Journal of machine learning Research*, 5(1), 2003.

[24] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[25] Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. *arXiv preprint arXiv:2011.04019*, 2020.

[26] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*, 2020.

[27] Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

[28] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[29] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

[30] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[31] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

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

[33] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[34] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[35] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

[36] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020.

[37] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[38] Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

[39] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[40] Ashwin Pananjady and Martin J Wainwright. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

[41] Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759, 2008.

[42] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[43] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[44] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[45] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[46] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems 7*, 7:361, 1995.

[47] Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

[48] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

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

[50] Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

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

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

[53] Ruosong Wang, Simon S Du, Lin Yang, and Sham Kakade. Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33, 2020.

[54] Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

[55] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

[56] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[57] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

[58] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

[59] Zheng Wen and Benjamin Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

[60] Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020.

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

[62] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[63] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[64] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[65] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The abstract concisely summarizes the motivations and contributions of this paper. The introduction provides more detailed statements and explanations in these aspects.
   (b) Did you describe the limitations of your work? [Yes] The limitations and potential directions for future work are discussed and stated in detail in Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The paper focuses solely on the theory of reinforcement learning. It is expected that this work will not have any potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our model is described in Definition 1 and the assumption is formalized in Assumption 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] All the proofs are provided in the Supplementary Materials. A sketch of the main proof ideas is provided in Section 5.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]