Proceedings of Machine Learning Research vol 125:1–17, 2020

Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research

Sham Kakade SHAM@CS.WASHINGTON.EDU
University of Washington

Lin F. Yang LINYANG@EE.UCLA.EDU
University of California, Los Angeles

Abstract

This work considers the sample and computational complexity of obtaining an ε-optimal policy in a discounted Markov Decision Process (MDP), given only access to a generative model. In this model, the learner accesses the underlying transition model via a sampling oracle that provides a sample of the next state, when given any state-action pair as input. We are interested in a basic and unresolved question in model based planning: is this naïve "plug-in" approach — where we build the maximum likelihood estimate of the transition model in the MDP from observations and then find an optimal policy in this empirical MDP — non-asymptotically, minimax optimal? Our main result answers this question positively. With regards to computation, our result provides a simpler approach towards minimax optimal planning: in comparison to prior model-free results, we show that using *any* high accuracy, black-box planning oracle in the empirical model suffices to obtain the minimax error rate. The key proof technique uses a leave-one-out analysis, in a novel "absorbing MDP" construction, to decouple the statistical dependency issues that arise in the analysis of model-based planning; this construction may be helpful more generally.

1. Introduction

How best to plan across a long-horizon with access to an approximate model of a Markov Decision Process? This is a fundamental question at the heart of reinforcement learning, and understanding it is essential to tackling even more complex challenges such as sample-efficient exploration (see e.g. (Kakade et al., 2003; Strehl et al., 2006; Strehl, 2007; Jaksch et al., 2010; Osband and Van Roy, 2014; Azar et al., 2017; Sidford et al., 2018b,a)). When the approximate model is arbitrary, these questions are studied, for example, in the approximate dynamic programming literature (Bertsekas, 1976). Before moving to approximation questions, a more basic question is an information theoretic one: how many samples from the Markov Decision Process are required to yield a near optimal policy? Our work studies this question in the generative model framework introduced in the work of Kearns and Singh (1999).

In the generative model setting, the learning agent has sampling access to a generative model of the Markov Decision Process (henceforth MDP), and it can query the next state s' sampled from the transition process,

given as input any state-action pair. The information theoretic question is to quantify how many samples from the generative model are required in order to obtain a near optimal policy; this question is analogous to the classical question of *sample complexity* in the supervised learning setting.

Arguably, the simplest approach here is a *model-based* one: the approach is to first build the maximum likelihood estimate of the transition model in the MDP from observations and then find an optimal policy in this empirical MDP. This work seeks to address the following unresolved question: is the naïve "plug-in" approach, non-asymptotically, minimax optimal in the quality of the policy it finds, given a fixed sample size? Throughout, we refer to the non-asymptotic regime as one where the sample size is sublinear in the model size. This work answers this question affirmatively showing that a model based planning approach is non-asymptotically minimax optimal.

We note that the first provably, non-asymptotically, minimax optimal algorithm is the Variance Reduced Q-value iteration algorithm (Sidford et al., 2018a), a model free approach. The significance of the optimality of our model-based result is that it allows the use of *any* efficient planning algorithm in the empirical MDP, which simplifies algorithm design, as the algorithm utilized need not be tied to the sampling procedure. We now discuss our contributions and the related work more broadly.

1.1. Our Contributions

There exists a large body of literature on MDPs and RL (see e.g. (Kakade et al., 2003; Strehl et al., 2009; Kalathil et al., 2014; Dann and Brunskill, 2015) and reference therein). A summary of our result relative to the prior works using a generative model is presented in Table 1. Here, ϵ is a desired accuracy parameter; $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the (finite) state and actions spaces; γ is a discount factor. We refer to ε-optimal policy the one whose discounted cumulative value in the MDP is ε close to the optimal value.

Before discussing the sample complexity of finding an ε-optimal policy, let us review the results on computing an ε-optimal value function. This refers to the problem of finding a function $\widehat{Q}$ which approximates $Q^\star$ to an error of ϵ at all states. The work of (Azar et al., 2012) shows that for $\epsilon \in (0,1)$ it suffices to use at most $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\big)$ calls to the generative model in order to return an ε-optimal *value* function [1]. Furthermore, the work of (Azar et al., 2012) shows this sample complexity is minimax optimal.

Obtaining an ε-optimal policy (rather than just estimating the value itself) is more subtle; naïvely, a policy obtained in a greedy manner from an ε-optimal value will incur a further degradation in its quality by a factor of $1/(1-\gamma)$ (Singh and Yee, 1994). The work of (Azar et al., 2013) shows that this additional error amplification is avoidable provided that the number of samples is at least $O(|\mathcal{S}|^2|\mathcal{A}|)$ (see Table 1); note that such a sample size is actually linear in the model size.

Our work avoids this error amplification and shows that for a desired accuracy threshold of ϵ, we can find an ϵ-optimal policy for any $\epsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$ using at most $\widetilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\big)$ samples. Our result holds for *any planning algorithm* that finds a near optimal policy in the empirically constructed MDP. Due to existing lower bounds (Azar et al., 2012; Sidford et al., 2018a), this bound is known to be *minimax optimal* for $\epsilon \in (0,1]$. Notably, this sample complexity is $o(S^2A)$ whenever $\epsilon^2 \geq 1/((1-\gamma)^3|\mathcal{S}|)$, meaning that we can

1. We conjecture that our techniques can be used to broaden the range of ε to go beyond $\epsilon \in (0,1)$, as needed in (Azar et al., 2012). In particular, the proof of Lemma 12 (used to prove Theorem 1) uses a self-bounding approach which we conjecture can be used to broaden the range of ε to allow for $\epsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$. We do not focus on this improvement in this work, as our main focus is on the value of the policy itself.

Algorithm	Sample Complexity	ϵ-Range	References
Phased Q-Learning	$C\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^2}$	$(0,(1-\gamma)^{-1}]$	(Kearns and Singh, 1999)
Empirical QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\epsilon^2}$	$(0,1]$	(Azar et al., 2013)
Empirical QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}$	$\left(0,\frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}\right]$	(Azar et al., 2013)
Randomized Primal-Dual Method	$C\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}$	$(0,(1-\gamma)^{-1}]$	(Wang, 2017)
Sublinear Randomized Value Iteration	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\cdot\operatorname{poly}\log\epsilon^{-1}$	$(0,1]$	(Sidford et al., 2018b)
Variance Reduced QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\cdot\operatorname{poly}\log\epsilon^{-1}$	$(0,1]$	(Sidford et al., 2018a)
Empirical MDP + *any* accurate black-box planner	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}$	$(0,(1-\gamma)^{-1/2}]$	This work

Table 1: **Sample Complexity to Compute ϵ-Optimal Policies Using the Generative Sampling Model**: Here $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions per state, $\gamma \in (0,1)$ is the discount factor, and C is an upper bound on the ergodicity. We ignore $\operatorname{poly}\log(|\mathcal{S}||\mathcal{A}|/\delta/(1-\gamma))$ factors in the sample complexity. Rewards are bounded between 0 and 1.

use the model to find a near optimal policy even in sample regimes where an accurate (say in total variation) approximation to the actual transition probabilities cannot be constructed.

Prior to this work, the only other non-asymptotically minimax optimal approach takes a different algorithmic path: (Sidford et al., 2018a) (also see Sidford et al. (2018b)) use a modification of the Q-value iteration method, with explicit control of variance in value estimates, to obtain an optimal sample complexity for $\epsilon \in (0,1]$. Our guarantees hold for a broader range of ϵ values (though we conjecture that our techniques could also improve the ε dependence in (Sidford et al., 2018a). See Footnote 1.).

Importantly, our work highlights that the sub-optimality of the prior model-based results was not due to any inherent limitation of the approach, but instead due to a matter of analysis. As a by-product, we retain a conceptually and algorithmically simpler solution strategy relative to Sidford et al. (2018a). On a technical note, our analysis is based on a novel absorbing MDP construction to deal with the dependence issues which arise in the analysis of Azar et al. (2012, 2013). This argument is a leave-one-out analysis in spirit; it leaves out one *state* out at a time (as opposed to one sample), cycling the through the contributions to the error at each state; this technique may prove more generally useful to decouple statistical dependency issues.

2. Setting

Markov Decision Process We denote a discounted Markov decision process (MDP) as a tuple $M = (\mathcal{S}, \mathcal{A}, P_M, r_M, \gamma)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $P_M : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathcal{S}}$ is the transition kernel (that is, $P_M(s' \mid s, a)$ is the probability of obtaining state s' when we take action a in state

s), $r_M : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function[2], and $\gamma \in (0,1)$ is a discount factor. For any (s,a), we denote $P_M(\cdot \mid s,a) \in \mathbb{R}^{|\mathcal{S}|}$ as the probability vector conditioning on state-action pair (s,a). A (deterministic) stationary policy is a map $\pi : \mathcal{S} \to \mathcal{A}$ that maps a state to an action. The value function of a policy π is a vector $V_M^\pi \in \mathbb{R}^{|\mathcal{S}|}$, defined as follows.

$$\forall s \in \mathcal{S}: \quad V_M^\pi(s) := \mathbb{E}\Big[\sum_{i=0}^{\infty} \gamma^t r_M(s^t, a^t) | s^0 = s\Big], \tag{1}$$

where $a^t = \pi(s^t)$ and $s^1, s^2, s^3, \ldots$ are generated from the distribution $s^{t+1} \sim P_M(\cdot|s^t, a^t)$. We also define an action value function $Q_M^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q_M^\pi(s,a) = r_M(s,a) + \gamma P_M(\cdot|s,a)^\top V^\pi.$$

We slightly abuse notation to use π to represent a stochastic stationary policy, i.e., $\pi : \mathcal{S} \to \mathbb{R}^{\mathcal{A}}$ maps a state to a distribution on $\mathcal{A}$ and $\pi(a|s)$ is the probability of playing action a at state s. When the MDP M is clear from the context, we drop the subscript to avoid clutter. The goal of a planning algorithm is to find a stationary policy in the MDP which maximizes the expected reward, denoted by $\pi^\star$. The famous theorem of Bellman and Dreyfus (1959) shows that there exists a policy $\pi^\star$ which simultaneously maximizes $V^\pi(s_0)$ for all $s_0 \in \mathcal{S}$. We also use $Q^\star$ and $V^\star$ to denote the value functions induced by $\pi^\star$. We call a policy, π, ϵ-optimal, if $V^\pi(s) \geq V^\star(s) - \epsilon$ for all $s \in \mathcal{S}$.

Generative Model Assume we have a access to a *generative model* or a *sampler*, which can provide us with samples $s' \sim P(\cdot \mid s,a)$. Suppose we call our sampler N times at each state action pair. Let $\widehat{P}$ be our empirical model, defined as follows:

$$\widehat{P}(s'|s,a) = \frac{\text{count}(s',s,a)}{N},$$

where $\text{count}(s',s,a)$ is the number of times the state-action pair (s,a) transitions to state s'. We define $\widehat{M}$ to be the empirical MDP that is identical to the original M, except that it uses $\widehat{P}$ instead of P for the transition kernel. We let $\widehat{V}^\pi$ and $\widehat{Q}^\pi$ to denote the value functions of a policy π in $\widehat{M}$, and $\widehat{\pi}^\star$, $\widehat{Q}^\star$ and $\widehat{V}^\star$ refer to the optimal policy and its value functions in $\widehat{M}$. The reward function r is assumed to be known and deterministic[3], and hence is identical in M and $\widehat{M}$.

Optimization Oracle Our goal in this paper is to determine the smallest sample size N, such that a planner run in $\widehat{M}$ returns a near-optimal policy in M. In order to decouple the statistical and computational aspects of planning with respect to an approximate model $\widehat{M}$, we will make use of an *optimization oracle* which takes as input an MDP M and returns a policy π satisfying: $\|Q_M^\pi - Q_M^\star\|_\infty \leq \epsilon_{\text{opt}}$.

We will use this optimization oracle for the empirical MDP $\widehat{M}$, and analyze the performance of the returned policy in the original MDP M. Classical algorithms such as value or policy iteration (Puterman, 2014) are the most common examples, though we discuss more sophisticated oracles as well in the next section.

2. We consider the setting where the rewards are in $[0,1]$. Our results can be generalized to other ranges of reward function via a standard reduction (see e.g. (Sidford et al., 2018a))
3. If r is unknown, we can use additional $|\mathcal{S}||\mathcal{A}|$ samples to obtain the exact value of r. If r is stochastic, we can query $|\mathcal{S}||\mathcal{A}|/\epsilon^2/(1-\gamma)^2$ samples to obtain a sufficiently accurate estimate of its mean. In both cases, the complexity contributed by r is only a lower order term to the present case. We can therefore assume, without loss of generality, r is known and deterministic.

3. Main results

In this section we present our main results. Before presenting our main theorem, we review some of the key challenges and our approach. Our high-level approach is to invoke any reasonable optimization oracle for the sample-based MDP $\widehat{M}$, and understand the sub-optimality of the returned policy π in the original MDP M. The key challenge is that π depends on the randomness in $\widehat{M}$, and hence, its value estimate from $\widehat{M}$ is not an unbiased estimator of its value in M. A usual way to address such issues is via uniform convergence, that is, we first establish that the values of all policies are similar in $\widehat{M}$ and M. This then implies that the high value of π in $\widehat{M}$ translates to a high value in M. Unfortunately, a naïve application of this argument yields bounds scaling as $|\mathcal{S}|^2$. Azar et al. (2013) do establish uniform convergence, but use a more careful argument which yields a bound scaling linearly in $|\mathcal{S}|$, but only when the desired accuracy $\epsilon \leq \sqrt{1/((1-\gamma)|\mathcal{S}|)}$, where the $|\mathcal{S}|$ factor in the condition of ϵ is due to uniform convergence. Sidford et al. (2018a,b) instead use a more complex algorithmic modification using variance reduction to get a sharper uniform convergence over a smaller class of policies with small variance in their value functions. In our result, we develop a novel leave-one-out technique, adapted to MDPs via an absorbing construction, to establish uniform convergence of our value estimates; we use the most natural algorithmic primitive of running any high accuracy black-box optimization oracle on the sample-based MDP $\widehat{M}$. We show the following result for this scheme.

Theorem 1 *Suppose $\delta > 0$ and $\epsilon \in (0, (1-\gamma)^{-1/2}]$. Let $\widehat{\pi}$ be any ϵ_{opt}-optimal policy for $\widehat{M}$, i.e. $\|\widehat{Q}^{\widehat{\pi}} - \widehat{Q}^{\star}\|_\infty \leq \epsilon_{opt}$. If*

$$N \geq \frac{c\gamma \log\left(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1}\delta^{-1}\right)}{(1-\gamma)^3\epsilon^2}, \quad \textit{we have} \quad Q^{\widehat{\pi}} \geq Q^{\star} - \epsilon - \frac{5\epsilon_{opt}}{(1-\gamma)},$$

with probability at least $1-\delta$, where c is an absolute constant.

Thus, the theorem shows that if ϵ_{opt} is made suitably small (roughly $(1-\gamma)\epsilon$), then we will find an $O(\epsilon)$ sub-optimal policy with $O\left(\log \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta}/(1-\gamma)^3/\epsilon^2\right)$ samples in each s, a pair. The total number of samples from the generative model then is $|\mathcal{S}||\mathcal{A}|N$ which amounts to $O\left(|\mathcal{S}||\mathcal{A}| \log \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta}/(1-\gamma)^3/\epsilon^2\right)$ samples. As remarked before, this is known to be unimprovable (up to a logarithmic factor) in the regime $\epsilon \in (0, 1]$ due to the lower bounds of (Azar et al., 2012; Sidford et al., 2018a).

We have so far focused on the statistical aspects of our estimators, since the use of a black-box optimization method in $\widehat{M}$ allows us to leverage the best possible solutions available. We now discuss some specific implications on the computational complexity of sparse model-based planning, instantiating the bound for some of the natural methods that may be used. Throughout we focus on attaining $\epsilon_{\text{opt}} = O((1-\gamma)\epsilon)$, since that equates the statistical and optimization errors. A very natural idea is to use value iteration (see e.g. (Puterman, 2014)), which requires $O\big((1-\gamma)^{-1} \cdot \log \epsilon_{\text{opt}}^{-1}\big)$ iterations, with each iteration taking $O(|\mathcal{S}||\mathcal{A}|N)$ time. Thus the overall running time for this algorithm is

$$O\big(|\mathcal{S}||\mathcal{A}|N \cdot (1-\gamma)^{-1} \cdot \log \epsilon_{\text{opt}}^{-1}\big) = O\left(\frac{|\mathcal{S}||\mathcal{A}| \cdot \log \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta} \cdot \log \frac{1}{(1-\gamma)\epsilon}}{(1-\gamma)^4\epsilon^2}\right).$$

Policy iteration methods (see again (Puterman, 2014)) can obtain an ϵ_{opt}-optimal policy within the same iteration complexity bound as value iteration. However, each iteration of the policy iteration requires solving

a linear system of size $|\mathcal{S}|^2$, which can be expensive. This computation time can be additionally improved. For instance, after initial phase of reading $O(|\mathcal{S}||\mathcal{A}|N)$ data points, (Sidford et al., 2018b) give a randomized algorithm to obtain an ϵ_{opt}-optimal policy with probability at least $1-\delta$ in time

$$\widetilde{O}\left[\left(\text{nnz}(\widehat{P})+\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right)\cdot\log\left(\frac{1}{\epsilon_{\text{opt}}}\right)\cdot\log\frac{1}{\delta}\right]=\widetilde{O}\left[\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\cdot\left(\frac{\log\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta}}{\min(\epsilon^2,1)}\right)\cdot\log\left(\frac{1}{(1-\gamma)\epsilon}\right)\cdot\log\frac{1}{\delta}\right],$$

where $\widetilde{O}$ hides $\text{poly}\log\log$ factors and $\text{nnz}(P)$ means the number of non-zero entries in P. Thus, the computational complexity of this scheme is nearly-linear in the total sample size up to additional logarithmic factors. There are other results for obtaining an exactly optimal policy for the MDP $\widehat{M}$ as well, for instance the SIMPLEX policy iteration (Ye, 2011), which runs in time $O(\text{poly}(|\mathcal{S}||\mathcal{A}|N/(1-\gamma))$.

4. Analysis

We begin with some notation needed for our analysis, and then give a high-level outline of the proof, along with some basic lemmas. We then present our main technical novelty, which is a construction of an auxiliary MDP as a device to guarantee uniform convergence of value functions. We conclude by providing the proof of the theorem in terms of the key lemmas, deferring the proofs of the lemmas to the appendix.

Additional Notation For a vector v, we let $(v)^2$, $\sqrt{v}$, and $|v|$ be the component-wise square, square root, and absolute value operations. We let $\mathbb{1}$ denotes the vector of all ones (adapting to dimensions based on the context). It is helpful to overload notation and let P be a matrix of size $(\mathcal{S}\times\mathcal{A})\times\mathcal{S}$ where the entry $P_{(s,a),s'}$ is equal to $P(s'\mid s,a)$. Also, let $P_{s,a}$ denote the vector $P(\cdot\mid s,a)$. We also define P^π to be the transition matrix on state-action pairs induced by a stochastic stationary policy π. In particular,

$$P^\pi_{(s,a),(s',a')}=P(s'\mid s,a)\pi(a'|s') \quad \text{for all } s,a,s',a'.$$

With this notation, we have

$$Q^\pi=r+\gamma PV^\pi=r+\gamma P^\pi Q^\pi, \quad \text{and} \quad Q^\pi=(I-\gamma P^\pi)^{-1}r.$$

Slightly abusing the notation, for $V\in\mathbb{R}^{\mathcal{S}}$, we define the vector $\text{Var}_P(V)\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as:

$$\text{Var}_P(V)(s,a):=\text{Var}_{P(\cdot|s,a)}(V), \quad \text{so that} \quad \text{Var}_P(V)=P(V)^2-(PV)^2,$$

where the squares are applied componentwise. We define Σ^π_M as the variance of the discounted reward, i.e.

$$\Sigma^\pi_M(s,a):=\mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)-Q^\pi_M(s,a)\right)^2\,\middle|\,s_0=s,a_0=a\right]$$

where the expectation is induced under the trajectories induced by π in M. It can be verified that, for all π, Σ^π satisfies the following Bellman style, self-consistency conditions (see Lemma 6 in (Azar et al., 2013)):

$$\Sigma^\pi_M=\gamma^2\text{Var}_P(V^\pi_M)+\gamma^2P^\pi\Sigma^\pi_M \tag{2}$$

Lemma 2 (Lemma 7 in Azar et al. (2013)) *The matrix Σ^π_M satisfies $\|\Sigma^\pi_M\|_\infty\le\gamma^2/(1-\gamma)^2$.*

4.1. Errors in empirical estimates

We begin the analysis by stating some basic results about empirical estimates of values derived from $\widehat{M}$ relative to their true values in M. We start with stating a lemma on componentwise error bounds. Its proof has been postponed to the appendix.

Lemma 3 (Componentwise bounds) *For any policy π, we have*

$$Q^\pi - \widehat{Q}^\pi = \gamma(I - \gamma P^\pi)^{-1}(P - \widehat{P})\widehat{V}^\pi.$$

In addition, we have:

$$Q^\pi \geq Q^\star - \|Q^\pi - \widehat{Q}^\pi\|_\infty - \|\widehat{Q}^\pi - \widehat{Q}^\star\|_\infty - \|\widehat{Q}^{\pi^\star} - Q^\star\|_\infty.$$

We hope to invoke the second part of the lemma to establish Theorem 1, where the middle term is the optimization error, and we will focus on bounding the other two terms. We next state another basic lemma.

Lemma 4 *For any policy π, MDP M and vector $v \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have $\|(I-\gamma P^\pi)^{-1}v\|_\infty \leq \|v\|_\infty/(1-\gamma)$.*

Proof Note that $v = (I - \gamma P^\pi)(I - \gamma P^\pi)^{-1}v = (I - \gamma P^\pi)w$, where $w = (I - \gamma P^\pi)^{-1}v$. By triangle inequality, we have

$$\|v\| = \|(I - \gamma P^\pi)w\| \geq \|w\|_\infty - \gamma\|P^\pi w\|_\infty \geq \|w\|_\infty - \gamma\|w\|_\infty,$$

where the final inequality follows since $P^\pi w$ is an average of the elements of w by the definition of P^π so that $\|P^\pi w\|_\infty \leq \|w\|_\infty$. Rearranging terms completes the proof. ■

Our next lemma is a key observation in Lemma 7 of Azar et al. (2012), namely the Bellman property of a policy's variance and its accumulation under the transition operator of the corresponding policy. We provide a short proof in the appendix for completeness.

Lemma 5 (Lemma 7 in Azar et al. (2013)) *For any policy π and MDP M,*

$$\left\|(I - \gamma P^\pi)^{-1}\sqrt{\mathrm{Var}_P(V_M^\pi)}\right\|_\infty \leq \sqrt{\frac{2}{(1-\gamma)^3}},$$

where P is the transition model of M.

Finally, it will be useful to also have more direct bounds on the errors in our value estimates which follow directly from Hoeffding's inequality, even though we are eventually after more careful bounds that account for variance. This result can be also be found as Lemma 4 in Azar et al. (2013), and is a standard concentration argument. For completeness, we provide its proof in Section A.4.

Lemma 6 (Crude Value Bounds, Lemma 4 in Azar et al. (2013)) *Let $\delta \geq 0$. With probability greater than $1-\delta$,*

$$\|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty \leq \Delta_{\delta,N} \quad \textit{and} \quad \|Q^\star - \widehat{Q}^\star\|_\infty \leq \Delta_{\delta,N}\,, \quad \textit{where} \quad \Delta_{\delta,N} := \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}.$$

We observe that these simple bounds are worse than what Theorem 1 posits by a factor of $\sqrt{1/(1-\gamma)}$, and removing this additional factor requires a significantly more careful analysis as we will see in the remainder of this section.

4.2. An s-absorbing MDP

In order to improve upon the crude bounds in Lemma 6, we would like to directly bound the errors in our value estimates using the componentwise bounds of Lemma 3. Doing so requires an understanding of quantities such as $|(P-\widehat{P})\widehat{V}^\star|$ and $|(P-\widehat{P})\widehat{V}^{\pi^\star}|$, which we will do next. However $\widehat{V}^\star$ and $\widehat{V}^{\pi^\star}$ depend on $\widehat{P}$, so that we are not able to directly apply a standard concentration argument. We now address this challenge by providing a method to decouple these dependencies.

For a state s and a scalar u, define the MDP $M_{s,u}$ as follows: $M_{s,u}$ is identical to M except that state s is absorbing in $M_{s,u}$, i.e. $P_{M_{s,u}}(s|s,a) = 1$ for all a, and the instantaneous reward at state s in $M_{s,u}$ is $(1-\gamma)u$; the remainder of the transition model and reward function are identical to those in M. In order to avoid notational clutter, we use $V^\pi_{s,u}$ to denote the value function $V^\pi_{M_{s,u}}$ and correspondingly for Q and reward and transition functions. This implies that for all policies π:

$$V^\pi_{s,u}(s) = u,$$

since s is absorbing with instantaneous reward $(1-\gamma)u$. For some state s, we will only consider $M_{s,u}$ for u in a finite set U_s, where

$$U_s \subset \left[V^\star(s) - \Delta_{\delta,N}\, V^\star(s) + \Delta_{\delta,N}\right].$$

In particular, we will set U_s to consist of evenly spaced elements in this interval, where we set the size of $|U_s|$ appropriately later on. As before, we let $\widehat{M}_{s,u}$ denote the MDP that uses the empirical model $\widehat{P}$ instead of P, at all non-absorbing states and abbreviate the value functions in $\widehat{M}_{s,u}$ as $\widehat{V}^\pi_{s,u}$. We now show concentration of value estimates in $M_{s,u}$, before doing a union bound over the set U_s to achieve a uniform convergence result.

Lemma 7 *Fix a state s, an action a, a finite set U_s, and $\delta \geq 0$. With probability greater than $1-\delta$, it holds that for all $u \in U_s$,*

$$\begin{aligned}
|(P_{s,a} - \widehat{P}_{s,a}) \cdot \widehat{V}^\star_{s,u}| &\leq \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^\star_{s,u})} + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N} \\
|(P_{s,a} - \widehat{P}_{s,a}) \cdot \widehat{V}^{\pi^\star}_{s,u}| &\leq \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^{\pi^\star}_{s,u})} + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N}
\end{aligned}$$

Proof The random variables $\widehat{P}_{s,a}$[4] and $\widehat{V}^\star_{s,u}$ are independent. The result now follows from Bernstein's inequality along with a union bound over all U_s. ■

4. Note that $P_{s,a}$ and $\widehat{P}_{s,a}$ are from the original MDPs M and $\widehat{M}$ and not the absorbing versions, as the latter induce degenerate transitions in s for all actions a.

This independence of $\widehat{P}_{s,a}$ from the value function $\widehat{V}^\star_{s,u}$ is the biggest upshot of our construction. Note that a similar statement does not hold for $\widehat{V}^\star$. We next need to understand how to construct U_s so that $\widehat{V}^\star_{s,u}$ provides a good approximation for $\widehat{V}^\star$, for some $u \in U_s$. The following two lemmas provide helpful properties of these absorbing state MDPs to build towards this goal.

Lemma 8 *Let $u^\star = V^\star_M(s)$ and $u^\pi = V^\pi_M(s)$. We have*

$$V^\star_M = V^\star_{s,u^\star}\,, \quad \text{and for all policies } \pi, \quad V^\pi_M = V^\pi_{M_{s,u^\pi}}\,.$$

Proof To prove the first claim, it suffices to verify that $V^\star_M$ satisfies the Bellman optimality conditions in $M_{s,u^\star}$. To see this, observe that at state s, the Bellman equations are trivially satisfied as s is absorbing with value $u^\star = V^\star_M(s)$ at state s by construction. For state $s' \neq s$, the outgoing transition model at s' in $M_{s,u^\star}$ is identical to that in M. Since $V^\star_M$ satisfies the Bellman optimality conditions at state s' in M, it must also satisfy Bellman optimality conditions at state s' in $M_{s,u^\star}$. The proof of the second claim is analogous. ■

This lemma gives a good setting for u, but we also need robustness to misspecification of u as we seek to construct a cover. The next lemma provides this result.

Lemma 9 *For all states s, $u, u' \in \mathbb{R}$, and policies π,*

$$\|Q^\star_{s,u} - Q^\star_{s,u'}\|_\infty \leq |u - u'| \quad \text{and} \quad \|Q^\pi_{s,u} - Q^\pi_{s,u'}\|_\infty \leq |u - u'|\,.$$

Proof First observe

$$\|r_{s,u} - r_{s,u'}\|_\infty = (1-\gamma)|u - u'|,$$

since these two reward functions differ only in state s, in which case $r_{s,u}(s,a) = (1-\gamma)u$ and $r_{s,u'}(s,a) = (1-\gamma)u'$. Let $\pi_{s,u}$ be the optimal policy in $M_{s,u}$. Note

$$\begin{aligned} Q^\star_{s,u} - Q^\star_{s,u'} &= Q^\star_{s,u} - \max_\pi (I - \gamma P^\pi_{s,u'})^{-1} r_{s,u'} \leq Q^\star_{s,u} - (I - \gamma P^{\pi_{s,u}}_{s,u'})^{-1} r_{s,u'} \\ &\overset{(a)}{=} (I - \gamma P^{\pi_{s,u}}_{s,u'})^{-1}(r_{s,u} - r_{s,u'}) \leq \frac{1}{1-\gamma}\|r_{s,u'} - r_{s,u}\|_\infty = |u - u'|, \end{aligned}$$

where the equality (a) follows since $P_{s,u}$ only depends on the state s and not the value u. The proof of the lower bound is analogous, which completes the proof of the first claim. The proof of the second claim can be obtained with a similar argument. ■

With these two lemmas, we now show the main result of this section.

Proposition 10 *Fix a state s, an action a, a finite set U_s, and $\delta \geq 0$. With probability greater than $1 - 2\delta$, it holds that for all $u \in U_s$,*

$$\begin{aligned} |(P_{s,a} - \widehat{P}_{s,a}) \cdot \widehat{V}^\star| \;\leq\; & \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^\star)} \\ & + \min_{u \in U_s} |\widehat{V}^\star(s) - u| \left(1 + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right) + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N} \end{aligned}$$

$$
\begin{aligned}
|(P_{s,a} - \widehat{P}_{s,a}) \cdot \widehat{V}^{\pi^\star}| \quad &\leq \quad \sqrt{\frac{2\log(4|U_s|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^{\pi^\star})} \\
&\quad + \min_{u \in U_s} |\widehat{V}^{\pi^\star}(s) - u| \left(1 + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right) + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N}
\end{aligned}
$$

Proof By Lemma 7, with probability greater than $1-\delta$, we have that for all $u \in U_s$.

$$
\begin{aligned}
& |(P_{s,a} - \widehat{P}_{s,a}) \cdot \widehat{V}^\star| = |(P_{s,a} - \widehat{P}_{s,a}) \cdot (\widehat{V}^\star - \widehat{V}^\star_{s,u} + \widehat{V}^\star_{s,u})| \\
\leq \quad & |(P_{s,a} - \widehat{P}_{s,a}) \cdot (\widehat{V}^\star - \widehat{V}^\star_{s,u})| + |(P_{s,a} - \widehat{P}_{s,a}) \cdot (\widehat{V}^\star_{s,u})| \\
\leq \quad & \|\widehat{V}^\star - \widehat{V}^\star_{s,u}\|_\infty + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^\star_{s,u})} + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N} \\
\leq \quad & \|\widehat{V}^\star - \widehat{V}^\star_{s,u}\|_\infty + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^\star - \widehat{V}^\star_{s,u} - \widehat{V}^\star)} + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N} \\
\leq \quad & \|\widehat{V}^\star - V^\star_{\widehat{M}_{s,u}}\|_\infty \left(1 + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right) + \sqrt{\frac{2\log(4|U_s|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}^\star)} + \frac{2\log(4|U_s|/\delta)}{(1-\gamma)3N}
\end{aligned}
$$

using the triangle inequality, $\sqrt{\mathrm{Var}_{P_{s,a}}(V_1 + V_2)} \leq \sqrt{\mathrm{Var}_{P_{s,a}}(V_1)} + \sqrt{\mathrm{Var}_{P_{s,a}}(V_2)}$.

By Lemmas 8 and 9,

$$
\|\widehat{V}^\star - V^\star_{s,u}\|_\infty = \|\widehat{V}^\star_{s,\widehat{V}^\star(s)} - V^\star_{s,u}\|_\infty \leq |\widehat{V}^\star(s) - u| \,.
$$

Since the above holds for all $u \in U_s$, we may take the best possible choice, which completes the proof of the first claim. The proof of the second claim is analogous. ■

For brevity in the remaining analysis, let us define the shorthand:

$$
L = \log(8|\mathcal{S}||\mathcal{A}|/((1-\gamma)\delta)). \tag{3}
$$

The proposition above, combined with an accounting of the discretization level yields the following result.

Lemma 11 *With probability greater than* $1-\delta$,

$$
\begin{aligned}
|(P - \widehat{P})\widehat{V}^\star| \quad &\leq \quad \sqrt{\frac{8L}{N}} \sqrt{\mathrm{Var}_P(\widehat{V}^\star)} + \Delta'_{\delta,N} \mathbb{1} \\
|(P - \widehat{P})\widehat{V}^{\pi^\star}| \quad &\leq \quad \sqrt{\frac{8L}{N}} \sqrt{\mathrm{Var}_P(\widehat{V}^{\pi^\star})} + \Delta'_{\delta,N} \mathbb{1}
\end{aligned}
$$

where

$$
\Delta'_{\delta,N} = \sqrt{\frac{cL}{N}} + \frac{cL}{(1-\gamma)N}
$$

with c being an absolute constant.

Proof We take U_s to be the evenly spaced elements in the interval $[V^\star(s) - \Delta_{\delta/2,N}\, V^\star(s) + \Delta_{\delta/2,N}]$, and we take the size of U_s to be $|U_s| = \frac{1}{(1-\gamma)^2}$. By Lemma 6, with probability greater than $1-\delta/2$, we have $\widehat{V}^\star(s) \in [V^\star(s) - \Delta_{\delta/2,N}\, V^\star(s) + \Delta_{\delta/2,N}]$ for all s. This implies:

$$
\min_{u \in U_s} |\widehat{V}^\star(s) - u| \leq \frac{2\Delta_{\delta/2,N}}{|U_s| - 1} = \frac{2}{|U_s| - 1} \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{4\log(4|\mathcal{S}||\mathcal{A}|/\delta)}{N}} \leq 4\gamma \sqrt{\frac{4\log(4|\mathcal{S}||\mathcal{A}|/\delta)}{N}}
$$

where we have used that that $\widehat{V}^\star(s)$ will land in one of $|U_s| - 1$ evenly sized sub-intervals of length $2\Delta_{\delta/2,N}/(|U_s| - 1)$. Now we use $\delta/(2|\mathcal{S}||\mathcal{A}|)$, so that the claims in Proposition 10 hold with probability greater than $1-\delta/2$ for all state action pairs. The first claim follows by substitution and noting that probability of either event failing is less than $\delta/2$. The proof of the second claim is analogous; note that Lemma 6 and Proposition 10 hold simultaneously with regards to the both claims regarding $\pi^\star$ and $\widehat{\pi}^\star$ so no further modifications to the failure probability are required. ■

4.3. The proof of Theorem 1

Theorem 1 immediately follows from the following lemma combined with Lemma 3.

Lemma 12 *Let $\widehat{\pi}$ be any policy satisfying the condition of Theorem 1. Then we have*

$$\begin{aligned}
\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty &\leq \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c}{(1-\gamma)^3}\frac{L}{N}} + \frac{cL}{(1-\gamma)^2 N}\right) + \frac{1}{1-\alpha_{\delta,N}} \cdot \frac{\gamma\epsilon_{opt}}{1-\gamma}\left(1 + \sqrt{\frac{L}{N}}\right) \\
\|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty &\leq \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c}{(1-\gamma)^3}\frac{L}{N}} + \frac{cL}{(1-\gamma)^2 N}\right)
\end{aligned}$$

where c is an absolute constant and where $\alpha_{\delta,N} = \frac{\gamma}{1-\gamma}\sqrt{8L/N}$.

Let us now show that Theorem 1 follows from this Lemma. From the condition on $\widehat{\pi}$ in the theorem statement, along with Lemma 3, we have

$$Q^{\widehat{\pi}} \geq Q^\star - \|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty - \epsilon_{\text{opt}} - \|\widehat{Q}^{\pi^\star} - Q^\star\|_\infty.$$

The condition on N in Theorem 1 (for an appropriately chosen absolute contant) implies that $\alpha_{\delta,N} = \frac{\gamma}{1-\gamma}\sqrt{\frac{8L}{N}} < 1/2$. This and Lemma 12 implies:

$$Q^{\widehat{\pi}} \geq Q^\star - 4\gamma\left(\sqrt{\frac{c}{(1-\gamma)^3}\cdot\frac{L}{N}} + \frac{c\cdot L}{(1-\gamma)^2 N}\right) - \frac{4\gamma\epsilon_{\text{opt}}}{1-\gamma} - \epsilon_{\text{opt}}.$$

Plugging in the choice of N in Theorem 1 (where the absolute constant in Theorem 1 need not be the same as that in Lemma 3) completes the proof of the theorem.

Proof [**Proof of Lemma 12**] We have:

$$\begin{aligned}
&\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty \\
&\overset{(a)}{=} \gamma\|(I - \gamma P^{\widehat{\pi}})^{-1}(P - \widehat{P})\widehat{V}^{\widehat{\pi}}\|_\infty \\
&\overset{(b)}{\leq} \gamma\|(I - \gamma P^{\widehat{\pi}})^{-1}(P - \widehat{P})\widehat{V}^\star\|_\infty + \gamma\|(I - \gamma P^{\pi})^{-1}(P - \widehat{P})(\widehat{V}^{\widehat{\pi}} - \widehat{V}^\star)\|_\infty
\end{aligned}$$

$$\overset{(c)}{\leq} \gamma \|(I - \gamma P^{\widehat{\pi}})^{-1}(P - \widehat{P})\widehat{V}^\star\|_\infty + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma}$$
$$\overset{(d)}{\leq} \gamma \|(I - \gamma P^{\widehat{\pi}})^{-1}\left|(P - \widehat{P})\widehat{V}^\star\right|\|_\infty + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma},$$

where (a) uses Lemma 3; (b) is the triangle inequality; (c) uses Lemma 4; (d) uses that $(I - \gamma P^{\widehat{\pi}^\star})^{-1}$ has all positive entries.

Focusing on the first term, we see that

$$\begin{aligned}
&\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty \\
&\overset{(e)}{\leq} \gamma\sqrt{\frac{8L}{N}} \left\|(I - \gamma P^{\widehat{\pi}})^{-1}\sqrt{\mathrm{Var}_P(\widehat{V}^\star)}\right\|_\infty + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \\
&\overset{(f)}{\leq} \gamma\sqrt{\frac{8L}{N}} \left\|(I - \gamma P^{\widehat{\pi}})^{-1}\left(\sqrt{\mathrm{Var}_P(V^{\widehat{\pi}})} + \sqrt{\mathrm{Var}_P(V^{\widehat{\pi}} - \widehat{V}^{\widehat{\pi}})} + \sqrt{\mathrm{Var}_P(\widehat{V}^{\widehat{\pi}} - \widehat{V}^\star)}\right)\right\|_\infty \\
&\qquad + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \\
&\overset{(g)}{\leq} \gamma\sqrt{\frac{8L}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\sqrt{\|V^{\widehat{\pi}} - \widehat{V}^{\widehat{\pi}}\|_\infty^2}}{1-\gamma} + \frac{\epsilon_{\text{opt}}}{1-\gamma}\right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \\
&\leq \gamma\sqrt{\frac{8L}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty}{1-\gamma} + \frac{\epsilon_{\text{opt}}}{1-\gamma}\right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \\
&= \gamma\sqrt{\frac{8L}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty}{1-\gamma}\right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma}\left(1 + \sqrt{\frac{8L}{N}}\right),
\end{aligned}$$

where the inequality (e) uses Lemma 11; (f) uses $\sqrt{\mathrm{Var}_P(X+Y)} = \sqrt{\mathbb{E}_P[(X + Y - \mathbb{E}_P[X+Y])^2]} \leq \sqrt{\mathrm{Var}_P(X)} + \sqrt{\mathrm{Var}_P(Y)}$, by triangle inequality of norms, using $\sqrt{\mathbb{E}_P[Z^2]}$ as the norm; (g) uses Lemma 5. Solving for $\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\|_\infty$ proves the first claim. The proof of the second claim is analogous. ■

5. Conclusion

This paper sheds new light on a long-studied basic question in reinforcement learning, which is that of a good approach to planning, given an approximate model of the world. While this is a fundamental question in itself, previous advances have also resulted in improved algorithms for harder questions such as sample-efficient exploration. For instance, the Bellman structure of variances in an MDP, observed in Azar et al. (2013) has subsequently formed a crucial component of minimax optimal exploration algorithms (Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019; Wainwright, 2019). We hope that the new technical components in our work can be similarly reused in broader contexts in future work, beyond their utility in analyzing sparse, model-based planning. At the same time, we do recognize that the leave-one-out analysis here does use the uniform sampling of N samples in each s, a pair rather crucially, which might present challenges in exploration settings.

Acknowledgments

Sham Kakade thanks Rong Ge for numerous helpful discussions. We thank Csaba Szepesvari, Kaiqing Zhang, and Mohammad Gheshlaghi Azar for helpful discussions and pointing out typos in the initial version of the paper. We are grateful for the constructive comments from anonymous reviewers. S. K. gratefully acknowledges funding from the Washington Research Foundation for Innovation in Data-intensive Discover, the ONR award N00014-18-1-2247, and NSF Award CCF-1703574.

References

Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

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.

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

Richard Bellman and Stuart Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13(68):247–251, 1959.

Dimitri P Bertsekas. Dynamic programming and stochastic control. 1976.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

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

Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Dileep Kalathil, Vivek S Borkar, and Rahul Jain. Empirical q-value iteration. *arXiv preprint arXiv:1412.0180*, 2014.

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

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31*, pages 5186–5196. Curran Associates, Inc., 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018b.

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

Alexander L Strehl. *Probably approximately correct (PAC) exploration in reinforcement learning*. PhD thesis, Rutgers University-Graduate School-New Brunswick, 2007.

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

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ-infty -bounds for q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

Mengdi Wang. Randomized linear programming solves the discounted Markov decision problem in nearly-linear running time. *arXiv preprint arXiv:1704.01869*, 2017.

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, 2011.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Appendix A. Proofs of supporting Lemmas

A.1. Proof of Lemma 2

Proof [Proof of Lemma 2] By definition, we have

$$\begin{aligned}
\Sigma_M^\pi(s,a) &:= \mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) - Q_M^\pi(s,a)\right)^2 \,\middle|\, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) - r(s,a) - \gamma Q_M^\pi(s_1,a_1)\right)^2 \,\middle|\, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t) - \gamma Q_M^\pi(s_1,a_1)\right)^2 \,\middle|\, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2 \,\middle|\, s_0=s, a_0=a\right] - \mathbb{E}\left[\left(\gamma Q_M^\pi(s_1,a_1)\right)^2 \,\middle|\, s_0=s, a_0=a\right] \\
&\le \mathbb{E}\left[\left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2 \,\middle|\, s_0=s, a_0=a\right] \\
&\le \frac{\gamma^2}{(1-\gamma)^2}.
\end{aligned}$$

■

A.2. Proof of Lemma 3

Proof [Proof of Lemma 3] For any policy π,

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

For the second claim,

$$Q^\pi - Q^\star = Q^\pi - \widehat{Q}^\star + \widehat{Q}^\star - Q^\star \ge Q^\pi - \widehat{Q}^\star + \widehat{Q}^{\pi^\star} - Q^\star \ge -\|Q^\pi - \widehat{Q}^\star\|_\infty - \|\widehat{Q}^{\pi^\star} - Q^\star\|_\infty.$$

Another application of triangle inequality completes the proof. ■

A.3. Proof of Lemma 5

Proof [Proof of Lemma 5] Note that $(1-\gamma)(I-\gamma P^\pi)^{-1}$ is matrix whose rows are a probability distribution. For a positive vector v and a distribution ν (where ν is vector of the same dimension of v), Jensen's inequality

implies that $\nu \cdot \sqrt{v} \leq \sqrt{\nu \cdot v}$. This implies:

$$\begin{aligned}
\|(I-\gamma P^\pi)^{-1}\sqrt{v}\|_\infty &= \frac{1}{1-\gamma}\|(1-\gamma)(I-\gamma P^\pi)^{-1}\sqrt{v}\|_\infty \\
&\leq \sqrt{\left\|\frac{1}{1-\gamma}(I-\gamma P^\pi)^{-1}v\right\|_\infty} \\
&\leq \sqrt{\left\|\frac{2}{1-\gamma}(I-\gamma^2 P^\pi)^{-1}v\right\|_\infty}.
\end{aligned}$$

where we have used that $\|(I-\gamma P^\pi)^{-1}v\|_\infty \leq 2\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty$ (which we will prove shortly). The proof is completed as follows: by Equation 2, $\Sigma_M^\pi = \gamma^2(I-\gamma^2 P^\pi)^{-1}\mathrm{Var}_P(V_M^\pi)$, so taking $v = \mathrm{Var}_P(V_M^\pi)$ and using that $\|\Sigma_M^\pi\|_\infty \leq \gamma^2/(1-\gamma)^2$ completes the proof.

Finally, to see that $\|(I-\gamma P^\pi)^{-1}v\|_\infty \leq 2\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty$, observe:

$$\begin{aligned}
\|(I-\gamma P^\pi)^{-1}v\|_\infty &= \|(I-\gamma P^\pi)^{-1}(I-\gamma^2 P^\pi)(I-\gamma^2 P^\pi)^{-1}v\|_\infty \\
&= \|(I-\gamma P^\pi)^{-1}\Big((1-\gamma)I+\gamma(I-\gamma P^\pi)\Big)(I-\gamma^2 P^\pi)^{-1}v\|_\infty \\
&= \|\Big((1-\gamma)(I-\gamma P^\pi)^{-1}+\gamma I\Big)(I-\gamma^2 P^\pi)^{-1}v\|_\infty \\
&\leq (1-\gamma)\|(I-\gamma P^\pi)^{-1}(I-\gamma^2 P^\pi)^{-1}v\|_\infty + \gamma\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty \\
&\leq \frac{1-\gamma}{1-\gamma}\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty + \gamma\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty \\
&\leq 2\|(I-\gamma^2 P^\pi)^{-1}v\|_\infty
\end{aligned}$$

which proves the claim. ■

A.4. Proof of Lemma 6

Proof Note that $V^\star$ is a fixed vector independent with the randomness in $\widehat{P}$. Moreover, $\|V^\star\|_\infty \leq (1-\gamma)^{-2}$. Thus, by Hoeffding bound and a union bound over all $\mathcal{S}\times\mathcal{A}$, we have, with probability at least $1-\delta$,

$$\|(\widehat{P}-P)V^\star\|_\infty \leq \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N\cdot(1-\gamma)^2}}.$$

For the rest of the proof, we condition on the event that the above inequality holds.

Next we show the first inequality. Note that for any π, we have,

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

Consider $\pi^\star$. Since $(I-\gamma\widehat{P}^\pi)^{-1}=\sum_{i=0}\gamma^i(\widehat{P}^\pi)^i$ and $(\widehat{P}^\pi)^i$ is a probability matrix, we have

$$\begin{aligned}\|\gamma(I-\gamma\widehat{P}^\pi)^{-1}(\widehat{P}-P)V^\star\|_\infty &\le \gamma\sum_{i=0}^{\infty}\|\gamma^i(\widehat{P}^\pi)^i(\widehat{P}-P)V^\star\|_\infty \le \gamma\sum_{i=0}^{\infty}\|\gamma^i(\widehat{P}-P)V^\star\|_\infty \\ &\le \frac{\gamma}{(1-\gamma)}\cdot\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N\cdot(1-\gamma)^2}}\end{aligned}$$

as desired.

Now we consider the second inequality. Let $\mathcal{T}$ be the Bellman optimality operator on M, i.e., for any $V\in\mathbb{R}^{\mathcal{S}}$

$$\begin{aligned}\forall s\in\mathcal{S}: \quad \mathcal{T}(V)(s) &= \max_a\big[r(s,a)+P(\cdot\mid s,a)^\top V\big], \quad \text{and} \\ \forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad \mathcal{T}(Q)(s,a) &= r(s,a)+\sum_{s'}P(s'\mid s,a)\max_{a'}Q(s',a').\end{aligned}$$

Let $\widehat{\mathcal{T}}$ be the Bellman optimality operator on $\widehat{M}$. Further recalling our notations P^π and $\widehat{P}^\pi$, we have

$$\begin{aligned}\|Q^\star-\widehat{Q}^\star\|_\infty &= \|\mathcal{T}Q^\star-\widehat{\mathcal{T}}\widehat{Q}^\star\|_\infty \\ &\le \|\mathcal{T}Q^\star-r-\widehat{P}^{\pi^\star}Q^\star\|_\infty+\|\widehat{P}^{\pi^\star}Q^\star+r-\widehat{\mathcal{T}}\widehat{Q}^\star\|_\infty \\ &= \gamma\|P^{\pi^\star}Q^\star-\widehat{P}^{\pi^\star}Q^\star\|_\infty+\gamma\|\widehat{P}^{\pi^\star}Q^\star-\widehat{P}^{\widehat{\pi}^\star}\widehat{Q}^\star\|_\infty \\ &= \gamma\|(P-\widehat{P})V^\star\|_\infty+\gamma\|\widehat{P}V^\star-\widehat{P}\widehat{V}^\star\|_\infty \\ &\le \gamma\|(P-\widehat{P})V^\star\|_\infty+\gamma\|V^\star-\widehat{V}^\star\|_\infty \\ &\le \gamma\|(P-\widehat{P})V^\star\|_\infty+\gamma\|Q^\star-\widehat{Q}^\star\|_\infty.\end{aligned}$$

Solving for $\|Q^\star-\widehat{Q}^\star\|$, we complete the proof. ■