

---

# Towards Tight Bounds on the Sample Complexity of Average-reward MDPs

---

Yujia Jin<sup>1</sup> Aaron Sidford<sup>1</sup>

## Abstract

We prove new upper and lower bounds for sample complexity of finding an  $\epsilon$ -optimal policy of an infinite-horizon average-reward Markov decision process (MDP) given access to a generative model. When the mixing time of the probability transition matrix of all policies is at most  $t_{\text{mix}}$ , we provide an algorithm that solves the problem using  $\tilde{O}(t_{\text{mix}}\epsilon^{-3})$  (oblivious) samples per state-action pair. Further, we provide a lower bound showing that a linear dependence on  $t_{\text{mix}}$  is necessary in the worst case for any algorithm which computes oblivious samples. We obtain our results by establishing connections between infinite-horizon average-reward MDPs and discounted MDPs of possible further utility.

## 1. Introduction

In this paper we consider the fundamental problem of computing an approximately optimal policy in a Markov decision process (MDP) given by a generative model. We consider a standard MDP model with a known set of states and actions. If an agent chooses an action at a given state, a known reward is immediately given to the agent and the agent probabilistically transitions to a new state by an (unknown) fixed distribution (as a function of the state-action pair). Given access to a *generative model* (Kakade et al., 2003), i.e. an oracle which when queried by a state-action pair returns an independent sample from the distribution over next states, our goal is to find a policy, i.e. a choice of action per state, that approximately maximizes a measure of (cumulative) reward over time, e.g. discounted reward, average reward, etc.

Solving MDPs with a generative model is a fundamental problem in learning theory and a classic model for decision

making under uncertainty and reinforcement learning (Puterman, 2014; Sutton & Barto, 2018). It is a prominent theoretical test-bed for learning algorithms and has been studied extensively over past 20 years (Kakade et al., 2003). Multiple algorithms have been proposed for the problem (Puterman, 2014) and there has been extensive research on improving the sample complexity for finding an approximately optimal policy (Wang, 2017a; Sidford et al., 2018a; Wainwright, 2019; Agarwal et al., 2020).

In certain settings, the sample complexity of the problem is settled. For example, consider discounted MDPs (DMDPs) where the goal is to minimize the  $\gamma$ -discounted reward, i.e. the sum of the rewards where the reward at step  $t \geq 0$  is discounted by  $\gamma^t \in [0, 1)$ . For a DMDP with  $A_{\text{tot}}$  total state-action pairs, it is known that  $\tilde{\Omega}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  samples are necessary to find an  $\epsilon$ -optimal policy (Azar et al., 2013; Feng et al., 2019) and there are algorithms which solve the problem with near-optimal  $\tilde{O}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  samples (Sidford et al., 2018a; Agarwal et al., 2020; Li et al., 2020) for certain ranges of  $\epsilon$ . Similarly, in the case of finite-horizon MDPs, again optimal sample and query complexities are known (Sidford et al., 2018a) for sufficiently small  $\epsilon$ .

Another popular class of MDPs are infinite-horizon average-reward MDPs (AMDPs) (Mahadevan, 1996). Here the reward function (also known as the *gain* for AMDPs) is the infinite-horizon average reward and it is assumed that for any given policy the transition matrix it induces has a mixing time bounded by  $t_{\text{mix}} > 0$ . AMDPs arise naturally in controlling computer systems and communication networks, where a controller makes frequent decisions, and for inventory systems with frequent restocking decisions (Puterman, 2014). However, despite advancements in the theory of DMDPs and multiple proposed algorithms (Wang, 2017b; Jin & Sidford, 2020), the optimal sample complexity of AMDPs has resisted similar characterization. The best

---

<sup>1</sup>Many results in literature instead consider MDPs where at each state there are a fixed  $A$  actions and use  $|\mathcal{S}|A$  to denote the total number of state-action pair in their result. These results typically generalize to non-uniform action set among states, and thus we replace this total number of state-action pair with  $A_{\text{tot}}$  throughout.

---

<sup>1</sup>Management Science and Engineering, Stanford University, CA, United States. Correspondence to: Yujia Jin <yujia-jin@stanford.edu>.

known sample complexity is  $\tilde{O}(A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$  (Jin & Sidford, 2020) and there is no known lower-bound.

In comparison to existing near-optimal DMDP methods for generative models (Kakade et al., 2003), recent methods for provably solving AMDPs suffer from two additional limitations. First, they work with randomized policies, i.e. ones which choose a distribution over actions in every state. Correspondingly, these methods require stronger condition of mixing bound on all randomized policies. Second, they use dynamic samples from the model, as opposed to the case for DMDPs where the sampling process can be completely oblivious, i.e. a fixed number of samples can be generated per state-action pair.

In this paper we make progress on these problems. First we provide a new method that finds an  $\epsilon$ -optimal deterministic stationary policy using  $\tilde{O}(A_{\text{tot}} t_{\text{mix}} / \epsilon^3)$  oblivious samples, thereby overcoming these limitations of previous methods. Further, we provide a lower bound showing that  $\Omega(A_{\text{tot}} t_{\text{mix}} / \epsilon^2)$  oblivious samples are necessary. Consequently, we resolve the question of optimal dependence on  $t_{\text{mix}}$  for oblivious sampling methods. We achieve these results by establishing a connection between infinite-horizon average-reward MDPs and discounted MDPs which may be of utility for further research in this area.

### 1.1. Problem Setup

We define a Markov decision process (MDP) as a tuple  $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$  with the following interpretations:

- state space  $\mathcal{S}$  - a finite discrete set of states with size  $|\mathcal{S}|$  that the process transits on. In particular,  $s \in \mathcal{S}$  denotes a single state.
- total action space  $\mathcal{A}$  - the union of all actions that an agent can take at any state  $s \in \mathcal{S}$ , i.e.  $\mathcal{A} = \cup_{s \in \mathcal{S}} \mathcal{A}_s$ , where  $\mathcal{A}_s$  is the action space at state  $s$  (and the  $\mathcal{A}_s$  are disjoint). We denote the total number of state-action pairs as  $A_{\text{tot}} = \sum_{s \in \mathcal{S}} |\mathcal{A}_s|$ .
- transition probability matrix  $\mathbf{P} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$  - When choosing action  $a_s \in \mathcal{A}_s$  at state  $s \in \mathcal{S}$  the next state is chosen from the distribution  $\mathbf{p}_{s, a_s} = \mathbf{P}((s, a_s), \cdot) \in \Delta^{\mathcal{S}}$ . We assume that  $\mathbf{P}$  is unknown but can be queried by a generative model which when queried at any state action pair  $(s, a_s)$  outputs an independent sample from the distribution  $\mathbf{p}_{s, a_s}$ .
- reward vector  $\mathbf{r} \in [0, 1]^{\mathcal{A}}$  - the reward at any state  $s$  when playing action  $a_s$  is denoted as  $r_{s, a_s}$ . As is common practice we assume the reward doesn't depend on the state it transits to (Sidford et al., 2018a), and is bounded in  $[0, 1]$  (by uniform rescaling).

Given MDP  $\mathcal{M}$  at any state  $s \in \mathcal{S}$  an agent can take action  $a_s \in \mathcal{A}_s$  after which it receives instant reward  $r_{s, a_s}$  and transits to some other state  $s'$  with probability  $\mathbf{p}_{s, a_s}(s')$ . A (deterministic stationary) policy of an MDP is a mapping  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ , i.e. it maps each state  $s \in \mathcal{S}$  to a fixed action  $\pi(s) \in \mathcal{A}_s$ . A randomized stationary policy of an MDP is a mapping  $\pi : \mathcal{S} \rightarrow \Delta^{\mathcal{A}}$ , i.e. it maps a state  $s \in \mathcal{S}$  to a fixed distribution over actions  $\pi(s) \in \Delta^{\mathcal{A}_s}$ . Under a given fixed policy  $\pi$  and initial distribution over states  $\mathbf{q} \in \Delta^{\mathcal{S}}$ , the MDP generates a sample path  $\{(s_1, a_1) \in (\mathcal{S} \times \mathcal{A}), (s_2, a_2) \in (\mathcal{S} \times \mathcal{A}), \dots\}$  where  $a_i$  is chosen from  $\pi(s_i)$  and  $s_{i+1}$  is chosen from each  $\mathbf{p}_{s_i, a_i}$ . Following its sample path, it receives the (cumulative) reward  $V_{\mathbf{q}}^{\pi}$  defined as

$$V_{\mathbf{q}, \gamma}^{\pi} = \mathbb{E}^{\pi} \left[ \sum_{t \geq 1} \gamma^{t-1} r_{s_t, a_t} \mid s_1 \sim \mathbf{q} \right], \quad \gamma\text{-DMDPs};$$

$$V_{\mathbf{q}}^{\pi} = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[ \sum_{t \geq 1} r_{s_t, a_t} \mid s_1 \sim \mathbf{q} \right], \quad \text{AMDPs}, \quad (1)$$

given  $\mathbf{q}$  as the initial distribution over states. We consider discount factor  $\gamma < 1$  for DMDPs, and an average reward for AMDPs. For brevity, under a policy  $\pi$  we use  $\mathbf{P}^{\pi} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$  to denote the transition matrix of the underlying Markov chain where  $\mathbf{P}^{\pi}(s, s') = \sum_{a_s \in \mathcal{A}_s} \pi_s(a_s) \mathbf{p}_{s, a_s}(s')$ , and  $\mathbf{r}^{\pi} \in [0, 1]^{\mathcal{S}}$  is the reward vector where  $\mathbf{r}^{\pi}(s) = \sum_{a_s \in \mathcal{A}_s} \pi_s(a_s) r_{s, a_s}$ . We define  $\nu^{\pi}$  to denote the stationary distribution of a transition matrix  $\mathbf{P}^{\pi}$  under policy  $\pi$ , i.e.  $(\nu^{\pi})^{\top} \mathbf{P}^{\pi} = \nu^{\pi}$ .

Our goal in solving an MDP is to find an optimal policy  $\pi^*$ , that maximizes the cumulative reward. We say (stationary) policy  $\pi$  is  $\epsilon$ -optimal, if  $V_{\mathbf{q}}^{\pi} \geq V_{\mathbf{q}}^{\pi^*} - \epsilon$ , for any initial distribution  $\mathbf{q} \in \Delta^{\mathcal{S}}$ . In this paper our main goal is to characterize the number of samples that need to be collected for each state-action pair (in the worst case) to find an  $\epsilon$ -optimal policy for AMDPs. We further restrict our attention to mixing AMDPs, which we define as those satisfying the following assumption.

**Assumption A.** An AMDP instance is mixing if for any policy  $\pi$ , there exists a stationary distribution  $\nu^{\pi}$  so that for any initial distribution  $\mathbf{q} \in \Delta^{\mathcal{S}}$ , the induced Markov chain has mixing time bounded by  $t_{\text{mix}} < \infty$ , where  $t_{\text{mix}}$  is defined as

$$t_{\text{mix}} := \max_{\pi} \left[ \arg \min_{t \geq 1} \left\{ \max_{\mathbf{q} \in \Delta^{\mathcal{S}}} \|(\mathbf{P}^{\pi})^t \mathbf{q} - \nu^{\pi}\|_1 \leq \frac{1}{2} \right\} \right].$$

This is a natural and widely used regularity assumption (Wang, 2017b; Jin & Sidford, 2020) for AMDPs to ensure the existence of cumulative reward of AMDPs in (1). Under Assumption A one can show that  $V^{\pi} = \langle \nu^{\pi}, \mathbf{r}^{\pi} \rangle$ ,

where  $\nu^\pi$  is the stationary distribution over states  $\mathcal{S}$  under the given policy  $\pi$ . Thus,  $V^\pi$  exists independent of  $\mathbf{q}$  for arbitrary policy  $\pi$ , allowing us to omit the subscript  $\mathbf{q}$  for brevity. We remark that the assumption we use is weaker than the ones in Wang (2017b); Jin & Sidford (2020), as we only assume mixing bounds for deterministic stationary policies, not the randomized ones.

## 1.2. Results

The main result of the paper is the following upper and lower bounds for finding an  $\epsilon$ -optimal policy for mixing AMDPs assuming a generative model access.

**Theorem 1.** *There exists an algorithm that, given a mixing AMDP with  $A_{\text{tot}}$  state-action pairs and mixing time bounded by  $t_{\text{mix}}$  and accuracy parameter  $\epsilon \in (0, 1)$ , finds an  $\epsilon$ -optimal deterministic policy with probability  $1 - \delta$  with  $O(A_{\text{tot}} \log(A_{\text{tot}}/\epsilon\delta)t_{\text{mix}}/\epsilon^3)$  oblivious samples.*

The prior state-of-the-art sample complexity for this problem is  $\tilde{O}(A_{\text{tot}}t_{\text{mix}}^2/\epsilon^2)$  due to the primal-dual method (Jin & Sidford, 2020). Our method compares favorably in the following settings:

- $t_{\text{mix}} \gg \omega(1/\epsilon)$ : In certain large-scale settings,  $t_{\text{mix}}$  may be large and increase with dimension and whereas  $\epsilon \in (0, 1)$  need not. In such settings, the improved dependence on  $t_{\text{mix}}$  achieved by our method, at the expense of larger dependence on  $1/\epsilon$ , may be desirable.
- one only has access to oblivious samples: Our method is the first one that uses oblivious instead of dynamic samples, which can be easier to access and cheaper to collect. Note this also implies our method has  $O(1)$  depth and is desirable for parallel computing, improving the prior parallel method for this model in Tiapkin et al. (2021).
- one only has mixing condition on deterministic stationary policies: Our mixing condition is weaker than the standard mixing condition considered in prior work (Wang, 2017b; Jin & Sidford, 2020), which requires a  $t_{\text{mix}} < \infty$  mixing time bound for all randomized stationary policies.

We show that our upper bound on sample complexity for mixing AMDPs is tight up to logarithmic and poly- $\epsilon$  factors, by proving the lower bound on oblivious samples formally as follows.

**Theorem 2.** *There are constants  $\epsilon_0, \delta_0 \in (0, 1/2)$  such that for all  $\epsilon \in (0, \epsilon_0)$  and any algorithm  $\mathcal{K}$ , which on input mixing AMDP  $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$  given by a generative model outputs a policy  $\pi$  satisfying  $V^\pi \geq V^{\pi^*} - \epsilon$  with probability at least  $1 - \delta_0$ ,  $\mathcal{K}$  makes at least  $\Omega(A_{\text{tot}}t_{\text{mix}}/\epsilon^2)$  deterministic oblivious queries to the generative model on some instance with  $A_{\text{tot}}$  total states and mixing time at most  $t_{\text{mix}}$ .*

## 1.3. Approach

Both our algorithm design and lower bound construction leverage ideas from research on the complexity of solving DMDPs (Azar et al., 2013; Sidford et al., 2018a; Li et al., 2020). AMDPs are the limiting case for DMDPs with  $\gamma \rightarrow 1$  and this connection has been leveraged previously to prove that Bellman equations are optimality conditions for AMDPs, obtain value iteration methods for AMDPs (Puterman, 2014), and efficiently compute stationary distributions (Cohen et al., 2016). However, to the best of our knowledge, quantitative application of this connection has not been performed prior for a finite  $\gamma$  in studying the sample complexity of AMDPs given by a generative model. We obtain our main theorems by bridging this gap and showing how to transfer between AMDPs and DMDPs with bounded losses in the value.

To obtain our upper bound, we prove a key fact relating AMDPs and DMDPs given the same MDP tuple  $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ . In Lemma 2 we show that for any initial distribution  $\mathbf{q}$ , the cumulative rewards as defined in (1) satisfy  $|V^\pi - (1 - \gamma)V_{\mathbf{q}, \gamma}^\pi| \leq O((1 - \gamma)t_{\text{mix}})$ . This fact follows from carefully examining the matrix expressions of  $V^\pi$  for AMDP and  $(1 - \gamma)V_{\mathbf{q}, \gamma}^\pi$  for DMDP. It also utilizes the algebraic implications of the mixing property of transition matrices  $\mathbf{P}^\pi$ . In prior work, Cohen et al. (2016) used such algebraic techniques on mixing properties of directed graphs; and Jin & Sidford (2020) used it on MDP analysis. With that lemma, we show that a  $\gamma = 1 - \Theta(\epsilon/t_{\text{mix}})$ -discounted MDP well approximates an AMDP in terms of their (approximately) optimal policy, without incurring an error larger than order of  $\epsilon$ . Also, we show that it suffices to solve the corresponding DMDP to accuracy  $\epsilon = \epsilon/(1 - \gamma)$ . By plugging in the sample complexity of most recent DMDP solvers (Li et al., 2020), we thus obtain a sample complexity of

$$\tilde{O}\left(\frac{A_{\text{tot}}}{(1 - \gamma)^3 \epsilon^2}\right) \stackrel{\text{choice of } \epsilon}{\sim} \tilde{O}\left(\frac{A_{\text{tot}}}{(1 - \gamma)\epsilon^2}\right) \stackrel{\text{choice of } \gamma}{\sim} \tilde{O}\left(\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}\right).$$

For our lower bound, we modify the hard DMDP instances considered in (Azar et al., 2013). In these instances we add to most state-action pairs a small probability of  $O(1/t_{\text{mix}})$  to “restart” the Markov chain, leading to an AMDP with  $t_{\text{mix}}$  mixing bound under the modified transition probabilities. In the class of hard instances, each state  $i^1 \in \mathcal{X}^1$  transits to a different state  $i_{(i^1, a^1)}^2$  when the agent takes a different action  $a^1 \in \mathcal{A}_{i^1} := [K]$ , and there is one action among these that contributes to a larger cumulative reward than all the rest. We show to find an  $\epsilon$ -optimal policy for the constructed AMDP, one needs to identify the correct actions for at least a constant fraction of the states  $i^1 \in \mathcal{X}^1$ , which

each requires  $\Omega(Kt_{\text{mix}}/\epsilon^2)$  samples over all actions in  $\mathcal{A}_{i^1}$ , giving the desired  $\Omega(A_{\text{tot}}t_{\text{mix}}/\epsilon^2)$  sample lower bound.

#### 1.4. Previous Work

The study of sample complexities for finding approximately optimal policies for MDPs dates back to the proposal of generative models in 2000s (Kakade et al., 2003). Ever since, the area has seen vast progress in terms of understanding the hardness of solving different types of MDPs (Azar et al., 2013) and in designing efficient algorithms with improved sample complexities. Here we briefly survey advances in the complexity of computing approximately optimal policies for three typical types of MDPs given by a generative model (see Table 1 for a summary of these relevant prior results in each setup).<sup>2</sup>

**DMDPs.** Azar et al. (2013) and Feng et al. (2019) prove a lower bound of  $\tilde{\Omega}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  for  $\gamma$ -discounted MDPs. On the upper bound side, Azar et al. (2013) also obtain a  $Q$ -value-iteration algorithm with a (sub-)optimal sample complexity of either  $\tilde{O}(A_{\text{tot}}/(1-\gamma)^5\epsilon^2)$  or  $\tilde{O}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  for  $\epsilon \in (0, 1/\sqrt{(1-\gamma)|\mathcal{S}|})$ . Later, a sequence of work (Sidford et al., 2018b;a) provide a variance-reduced ( $Q$ -)value iteration that has near-optimal sample complexity and runtime of  $\tilde{O}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  for  $\epsilon \in (0, 1)$ . Their method incorporates variance reduction in estimating the value iteration step, and performs a fine-grained analysis of the error growth through the iterative process using a tight variance bound of Markov decision processes. Similarly, Wainwright (2019) shows that variance reduction can be applied to  $Q$ -learning and obtain a method with competing sample complexity guarantees for the same range of accuracy  $\epsilon \in (0, 1)$ . From a more statistical and less algorithmic perspective, another work of Agarwal et al. (2020) shows that  $\tilde{O}(A_{\text{tot}}/(1-\gamma)^3\epsilon^2)$  samples suffice to build an empirical MDP such that the optimal policy of it yields a near-optimal policy for the original MDP. Their work also utilizes the fine-grained variance bound using Bernstein inequality and extends the near-optimal sample complexity dependence to all  $\epsilon \in (0, 1/\sqrt{1-\gamma})$ . Recent work (Li et al., 2020) fully settled the sample complexity for DMDPs for all  $\epsilon$ , i.e.  $\epsilon \in (0, 1/(1-\gamma))$ , with a perturbed empirical MDP construction. We leverage this result crucially to obtain our results in Section 2.

**Finite-horizon MDPs.** For finite-horizon MDPs, cumulative reward is measured as the sum of the rewards obtained within first  $H$  steps for a given finite horizon  $H > 0$ . To the best of our knowledge, the only near-optimal al-

gorithm given for finite-horizon MDP is in Sidford et al. (2018b). There, the authors show how to apply the near-optimal variance-reduced value iteration method to MDPs with finite horizon of length  $H$ . They formally prove an upper bound of  $\tilde{O}(A_{\text{tot}}H^3/\epsilon^2)$  in sample complexity, for  $\epsilon \in (0, 1)$ . Through reduction to DMDP lower bounds, they also obtain a lower bound of  $\tilde{\Omega}(A_{\text{tot}}H^3/\epsilon^2)$ .

**AMDPs.** Average-reward MDPs with bounded mixing time are another fundamental class of MDPs (Kearns & Singh, 2002; Ortner & Auer, 2007), though less studied in terms of sample complexity. The first sample complexity bounds in the setting of a generative model is Wang (2017b), which applies a primal-dual method for the minimax problem related to the linear programming formulation, and proves an upper bound of  $O(\tau^2 A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$  where  $\tau$  denotes an upper bound on the ergodicity of all stationary distribution under arbitrary policies, i.e. there exists some distribution  $\mathbf{q} \in \Delta^{\mathcal{S}}$  satisfying  $\sqrt{1/\tau} \cdot \mathbf{q} \leq \boldsymbol{\nu}^\pi \leq \sqrt{\tau} \cdot \mathbf{q}$  for all policies  $\pi$  and its induced stationary distribution  $\boldsymbol{\nu}^\pi$ . Recently, Jin & Sidford (2020) design a similar primal-dual stochastic mirror descent and improve the bound to  $O(A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$ , removing the ergodicity assumption through an improved analysis of the optimality conditions of the minimax problem. In contrast to value iteration,  $Q$ -learning, and sample-based methods for DMDPs, both known efficient methods for AMDPs use a linear programming formulation, dynamic sampling, a stronger mixing condition, and only compute randomized stationary policies. On the hardness side, there is no known lower bound for AMDPs with bounded mixing time.

#### 1.5. Notation

We use unbold letters, e.g.  $V$ , to denote scalars, and bold letters, e.g.  $\mathbf{v}$  and  $\mathbf{P}$ , to denote vectors and matrices. We use  $\mathbf{e}_i$  to denote the basis vector that is 1 on coordinate  $i$ , and 0 elsewhere. We use  $\mathbf{1}_{\mathcal{S}}$  to denote the all-ones vector in  $\mathbb{R}^{\mathcal{S}}$ , and omit the subscript when it is clear from context. We use  $\|\cdot\|_{\infty}$  to denote the  $\ell_{\infty}$ -norm of vectors and  $\ell_{\infty}$ -operator norm of matrices, e.g.  $\|\mathbf{v}\|_{\infty} := \max_{s \in \mathcal{S}} |\mathbf{v}(s)|$  and  $\|\mathbf{P}\|_{\infty} := \max_{\|\mathbf{x}\|_{\infty}=1} \|\mathbf{P}\mathbf{x}\|_{\infty} = \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} |\mathbf{P}(s, s')|$ .

## 2. Upper Bound

In this section, we prove the sample complexity upper bound for obtaining an  $\epsilon$ -optimal deterministic policy for mixing AMDPs. We first provide Lemma 2 that relates the value of AMDPs and DMDPs under the same policy. Then we reduce solving AMDPs to DMDPs with the proper discount factor (Lemma 3) and use the state-of-the-art DMDP solver (restated in Lemma 4) to obtain our result.

Throughout the section, we consider some mixing AMDP, and its corresponding DMDP with the same tuple

<sup>2</sup>Another typical setting (outside the scope of this work) for studying these MDPs is to design efficient algorithms to minimize the *regret* compared with the optimal policy (Kearns & Singh, 2002; Ortner & Auer, 2007; Ortner, 2020).



Table 1. **Upper and lower bounds on sample complexity to get  $\epsilon$ -optimal policy for different type of MDPs.** Here  $A_{\text{tot}}$  denotes the total size of all state-action pairs,  $\gamma$  is discount factor for DMDP,  $H$ -MDP corresponds to finite-horizon MDP with length  $H$ ,  $t_{\text{mix}}$  is mixing time for mixing AMDP, and  $\tau$  is an ergodicity parameter showing up whenever the designed algorithm requires additional ergodic condition for MDP, i.e. there exists some distribution  $\mathbf{q}$  and  $\tau > 0$  satisfying  $\sqrt{1/\tau}\mathbf{q} \leq \nu^\pi \leq \sqrt{\tau}\mathbf{q}$ , for any policy  $\pi$  and its induced stationary distribution  $\nu^\pi$ .

Type	Method	Sample Complexity	Accuracy
	<i>lower bound</i> (Azar et al., 2013; Feng et al., 2019)	$\tilde{\Omega}(A_{\text{tot}}(1-\gamma)^{-3}\epsilon^{-2})$	N/A
	Empirical QVI (Azar et al., 2013)	$\tilde{O}(A_{\text{tot}}(1-\gamma)^{-3}\epsilon^{-2})$	$\epsilon \in (0, 1/\sqrt{(1-\gamma) S })$
	Primal-Dual Method (Wang, 2017a)	$\tilde{O}(A_{\text{tot}}\tau^4(1-\gamma)^{-4}\epsilon^{-2})$	$\epsilon \in (0, 1/(1-\gamma))$
	Variance-reduced QVI (Sidford et al., 2018a)	$\tilde{O}(A_{\text{tot}}(1-\gamma)^{-3}\epsilon^{-2})$	$\epsilon \in (0, 1)$
DMDP	Empirical MDP Sampler (Agarwal et al., 2020)	$\tilde{O}(A_{\text{tot}}(1-\gamma)^{-3}\epsilon^{-2})$	$\epsilon \in (0, \sqrt{1/(1-\gamma)})$
	Primal-Dual SMD (Jin & Sidford, 2020)	$\tilde{O}(A_{\text{tot}}(1-\gamma)^{-4}\epsilon^{-2})$	$\epsilon \in (0, 1/(1-\gamma))$
	Perturbed Empirical MDP Sampler (Li et al., 2020)	$\tilde{O}(A_{\text{tot}}(1-\gamma)^{-3}\epsilon^{-2})$	$\epsilon \in (0, 1/(1-\gamma))$
H-MDP	<i>lower bound</i> (Sidford et al., 2018a)	$\tilde{\Omega}(A_{\text{tot}}H^3\epsilon^{-2})$	N/A
	Variance-reduced QVI (Sidford et al., 2018a)	$\tilde{O}(A_{\text{tot}}H^3\epsilon^{-2})$	$\epsilon \in (0, 1)$
AMDP	<i>lower bound</i> (Theorem 2)	$\Omega(A_{\text{tot}}t_{\text{mix}}\epsilon^{-2})$	N/A
	Primal-Dual Method (Wang, 2017b)	$\tilde{O}(\tau^2A_{\text{tot}}t_{\text{mix}}^2\epsilon^{-2})$	$\epsilon \in (0, 1)$
	Primal-Dual SMD (Jin & Sidford, 2020)	$\tilde{O}(A_{\text{tot}}t_{\text{mix}}^2\epsilon^{-2})$	$\epsilon \in (0, 1)$
	<b>Our method</b> (Theorem 1)	$\tilde{O}(A_{\text{tot}}t_{\text{mix}}\epsilon^{-3})$	$\epsilon \in (0, 1)$

$(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$  and some discount factor  $\gamma \in (0, 1)$  to be specified.

**Characterization of value vectors.** We introduce the value vector  $\mathbf{v}^\pi \in \mathbb{R}^S$  under a given policy  $\pi$  for all states  $s \in \mathcal{S}$ . To distinguish between the value vectors of a DMDP and AMDP, we use  $\mathbf{v}_\gamma^\pi$  for DMDPs and  $\mathbf{v}^\pi$  for AMDPs respectively. For the discounted case, we let  $\mathbf{v}_\gamma^\pi(s) = V_{\mathbf{q}=\mathbf{e}_s, \gamma}^\pi$ , i.e. the cumulative reward of the MDP with initial distribution only on state  $s$ . We first give the following equations for computing value vectors  $\mathbf{v}^\pi$ . These are known results widely used in literature (see also Puterman (2014); Wang (2017b); Jin & Sidford (2020)).

Given a tuple  $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ , for DMDP with discount factor  $\gamma$  and a policy  $\pi$ , we have

$$\mathbf{v}_\gamma^\pi = \sum_{t \geq 0} \gamma^t (\mathbf{P}^\pi)^t \mathbf{r}^\pi \quad \text{and} \quad \|\mathbf{v}_\gamma^\pi\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]. \quad (2)$$

Similarly, for AMDP and a policy  $\pi$  that induces stationary distribution  $\nu^\pi$ , since the reward doesn't depend on initial distribution, we have

$$\mathbf{v}^\pi = \langle \mathbf{r}^\pi, \nu^\pi \rangle \mathbf{1} \quad \text{and} \quad \|\mathbf{v}^\pi\|_\infty \in [0, 1]. \quad (3)$$

To put the two value vectors on the same  $[0, 1]$  scale, we introduce the following rescaled value vectors denoted by

$\bar{\mathbf{v}}^\pi$ , one has

$$\bar{\mathbf{v}}^\pi = \mathbf{v}^\pi = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T (\mathbf{P}^\pi)^t \mathbf{r}^\pi = V^\pi \cdot \mathbf{1}$$

$$\text{and} \quad \bar{\mathbf{v}}_\gamma^\pi = (1-\gamma)\mathbf{v}_\gamma^\pi = (1-\gamma) \left[ \sum_{t=0}^{\infty} \gamma^t (\mathbf{P}^\pi)^t \mathbf{r}^\pi \right]. \quad (4)$$

We first state a helper lemma for  $\mathbf{P}^\pi$  following from the fact that it has mixing time bound  $t_{\text{mix}}$ , quoted from Lemma 23 of Cohen et al. (2016).

**Lemma 1** (see Lemma 14 of Jin & Sidford (2020), Lemma 23 of Cohen et al. (2016)). *For any policy  $\pi$  with induced probabilistic transition matrix  $\mathbf{P}^\pi$  of mixing time  $t_{\text{mix}}$  and stationary distribution  $\nu^\pi$  and any non-negative integer  $k \geq t_{\text{mix}}$ ,*

$$\|(\mathbf{P}^\pi)^k - \mathbf{1}(\nu^\pi)^\top\|_\infty \leq \left(\frac{1}{2}\right)^{\lfloor \frac{k}{t_{\text{mix}}} \rfloor}.$$

Now we provide a lemma that bounds the  $\ell_\infty$ -difference of the two rescaled value vectors  $\bar{\mathbf{v}}^\pi$  and  $\bar{\mathbf{v}}_\gamma^\pi$  under the given policy  $\pi$  and discount factor  $\gamma$ .

**Lemma 2.** *Given a same MDP tuple  $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ , a policy  $\pi$  and some discount factor  $\gamma$ , rescaled value vectors  $\bar{\mathbf{v}}^\pi$  for*

AMDP and  $\bar{\mathbf{v}}_\gamma^\pi$  for DMDP as defined in (4) satisfy

$$\|\bar{\mathbf{v}}^\pi - \bar{\mathbf{v}}_\gamma^\pi\|_\infty \leq 3(1-\gamma)t_{\text{mix}}.$$

*Proof.* Note that

$$\begin{aligned} & \|\bar{\mathbf{v}}^\pi - \bar{\mathbf{v}}_\gamma^\pi\|_\infty \\ &= \left\| (1-\gamma) \sum_{t \geq 0} \gamma^t \langle \mathbf{r}^\pi, \boldsymbol{\nu}^\pi \rangle \mathbf{1} - (1-\gamma) \sum_{t \geq 0} \gamma^t (\mathbf{P}^\pi)^t \mathbf{r}^\pi \right\|_\infty \\ &= (1-\gamma) \left\| \sum_{t=0}^{\infty} \gamma^t [(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top] \mathbf{r}^\pi \right\|_\infty \\ &\leq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \cdot \|\mathbf{r}^\pi\|_\infty. \end{aligned}$$

Now,  $\|\mathbf{r}^\pi\|_\infty \leq 1$  by assumption and for all  $t \geq 0$  we have

$$\|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \leq \|(\mathbf{P}^\pi)^t\|_\infty + \|\mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty = 2.$$

Further for all  $t \geq t_{\text{mix}}$  we have  $\|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \leq 2^{-\lfloor t/t_{\text{mix}} \rfloor}$  by Lemma 1. Combining yields the desired bound of

$$\begin{aligned} & \|\bar{\mathbf{v}}^\pi - \bar{\mathbf{v}}_\gamma^\pi\|_\infty \\ &\leq (1-\gamma) \sum_{t=0}^{t_{\text{mix}}-1} \gamma^t \|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \\ &\quad + (1-\gamma) \sum_{t \geq t_{\text{mix}}} \gamma^t \|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \\ &\leq (1-\gamma) \sum_{t=0}^{t_{\text{mix}}-1} 2\gamma^t + (1-\gamma) \sum_{t \geq t_{\text{mix}}} \frac{1}{2^{\lfloor t/t_{\text{mix}} \rfloor}} \\ &\leq 3(1-\gamma)t_{\text{mix}}. \end{aligned}$$

□

This lemma shows under the same policy, the values of AMDP and its corresponding DMDP are close up to  $\epsilon$  when choosing the discount factor  $\gamma = 1 - \Theta(\epsilon/t_{\text{mix}})$ . This allows us to formally reduce solving AMDPs to solving DMDPs with large enough discount factors in Lemma 3.

**Lemma 3.** *Given an AMDP with mixing time bounded by  $t_{\text{mix}}$ , accuracy parameter  $\epsilon \in (0, 1)$ , and an  $\frac{\epsilon}{3(1-\gamma)}$ -optimal policy  $\pi$  for the corresponding DMDP with  $\gamma = 1 - \frac{\epsilon}{9t_{\text{mix}}}$ ,  $\pi$  is also a  $\epsilon$ -optimal policy for the original AMDP.*

*Proof.* Consider a DMDP with the same transition matrix and discount factor  $\gamma = 1 - \frac{\epsilon}{9t_{\text{mix}}}$ , we have  $\|\bar{\mathbf{v}}^\pi - \bar{\mathbf{v}}_\gamma^\pi\|_\infty \leq \epsilon/3$  by Lemma 2.

Now let  $\pi_d$  and  $\pi_a$  denote optimal policies for the DMDP and AMDP respectively. By definition of  $\pi$  one has

$$\|\mathbf{v}_\gamma^\pi - \mathbf{v}_\gamma^{\pi_d}\|_\infty \leq \frac{\epsilon}{3(1-\gamma)} \iff \|\bar{\mathbf{v}}_\gamma^\pi - \bar{\mathbf{v}}_\gamma^{\pi_d}\|_\infty \leq \epsilon/3. \quad (5)$$

Consequently, one has that entrywise,

$$\begin{aligned} \bar{\mathbf{v}}^\pi + \frac{\epsilon}{3}\mathbf{1} &\stackrel{(i)}{\geq} \bar{\mathbf{v}}_\gamma^\pi \stackrel{(ii)}{\geq} \bar{\mathbf{v}}_\gamma^{\pi_d} - \frac{\epsilon}{3}\mathbf{1} \stackrel{(iii)}{\geq} \bar{\mathbf{v}}_\gamma^{\pi_a} - \frac{\epsilon}{3}\mathbf{1} \\ &\stackrel{(i)}{\geq} (\bar{\mathbf{v}}^{\pi_a} - \frac{\epsilon}{3}\mathbf{1}) - \frac{\epsilon}{3}\mathbf{1}, \end{aligned}$$

where we use (i) Lemma 2 together with the choice of  $\gamma = 1 - \frac{\epsilon}{9t_{\text{mix}}}$ , (ii) equation (5), and (iii) the optimality of  $\pi_d$  for DMDP by definition.

Altogether we conclude that  $\bar{\mathbf{v}}^\pi \geq \bar{\mathbf{v}}^{\pi_a} - \epsilon\mathbf{1}$  and therefore  $V^\pi \geq V^{\pi^*} - \epsilon$ , i.e.  $\pi$  is a  $\epsilon$ -optimal policy for the given AMDP. □

With the reduction, we can apply recent DMDP solvers to obtain an AMDP solver with the desired sample complexity. In order to solve the corresponding  $\gamma$ -discounted MDP to a desired accuracy, we use the following recent efficient DMDP solver (Li et al., 2020).

**Lemma 4** (Corollary of Theorem 1 of Li et al. (2020)). *There is an algorithm that, given a  $\gamma$ -discounted MDP, desired accuracy  $\epsilon \leq 1/(1-\gamma)$ , failure probability  $0 < \delta \ll 1$ , outputs an  $\epsilon$ -optimal policy with probability  $1 - \delta$  with a number of oblivious samples bounded by*

$$\tilde{O}\left(\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2} \log \frac{A_{\text{tot}}}{(1-\gamma)\epsilon\delta}\right).$$

We remark that what we state is an immediate corollary of Theorem 1 in Li et al. (2020) which works for non-uniform action space per state as well by expanding the space. Now we can apply this solver to find an  $\epsilon/(1-\gamma)$ -optimal policy for  $\gamma$ -discounted MDP to obtain our main result.

**Theorem 1.** *There exists an algorithm that, given a mixing AMDP with  $A_{\text{tot}}$  state-action pairs and mixing time bounded by  $t_{\text{mix}}$  and accuracy parameter  $\epsilon \in (0, 1)$ , finds an  $\epsilon$ -optimal deterministic policy with probability  $1 - \delta$  with  $O(A_{\text{tot}} \log(A_{\text{tot}}/\epsilon\delta)t_{\text{mix}}/\epsilon^3)$  oblivious samples.*

*Proof of Theorem 1.* By Lemma 3, it suffices to solve the corresponding DMDP to  $\epsilon/3(1-\gamma)$  accuracy with  $\gamma = 1 - \epsilon/9t_{\text{mix}}$ . Using the solver in Lemma 4 with  $\epsilon = \epsilon/3(1-\gamma)$ , it has sample complexity bounded by the following as stated.

$$\begin{aligned} & O\left(\frac{A_{\text{tot}}}{(1-\gamma)^3(\epsilon/3(1-\gamma))^2} \log \frac{A_{\text{tot}}}{(1-\gamma)\frac{\epsilon}{3(1-\gamma)}\delta}\right) \\ &= O\left(\frac{A_{\text{tot}}}{(1-\gamma)\epsilon^2} \log \frac{A_{\text{tot}}}{\epsilon\delta}\right) = O\left(\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3} \log \frac{A_{\text{tot}}}{\epsilon\delta}\right), \end{aligned}$$

where we use the choice of  $\gamma$  for the last equality. This proves the correctness of the method and yields the sample complexity bound as claimed. □

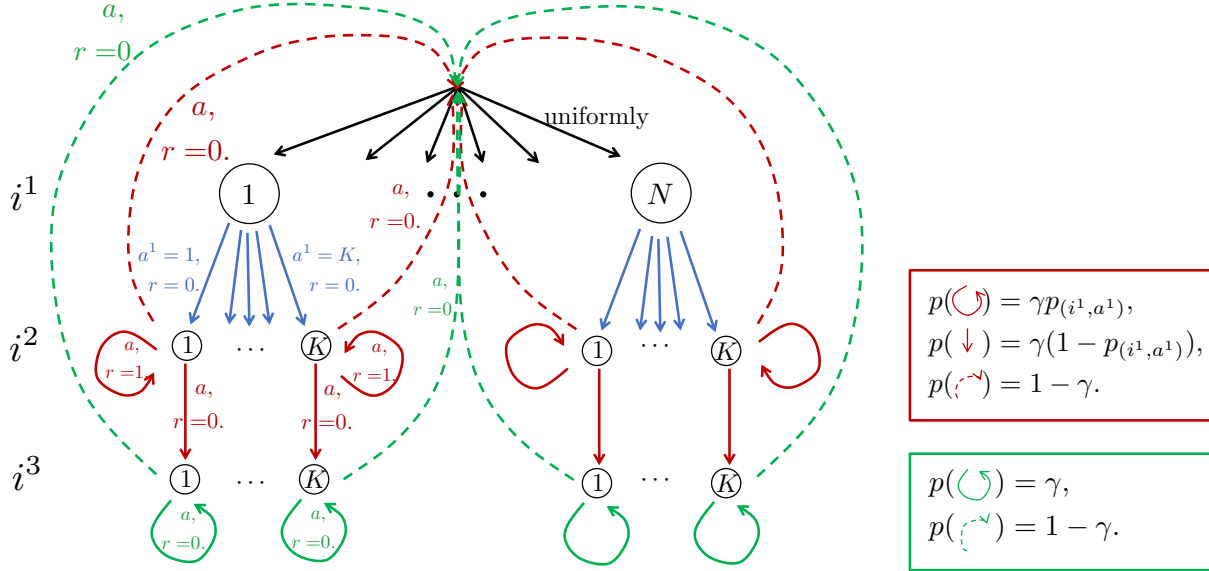


Figure 1. AMDP lower bound hard instance illustration.  $N$  states in  $\mathcal{X}^1$  (corresponding to first level),  $K$  action per state  $i^1 \in \mathcal{X}^1$ ,  $A_{\text{tot}} = O(NK)$  total state-action pairs.  $\gamma \in (0, 1)$ ,  $p_{i^1, a^1} \in [0, 1]$  for all  $i^1 \in \mathcal{X}^1$ ,  $a^1 \in [K]$  are tunable parameters.

**Remark 1** (Nearly-tight  $\epsilon$  regime.). *That Li et al. (2020) achieves near-optimal sample complexity for solving DMDPs for all  $\epsilon \in (0, 1/(1 - \gamma))$ , as opposed to the more restricted range of  $\epsilon \in (0, 1)$  in Sidford et al. (2018a) and  $\epsilon \in (0, 1/\sqrt{1 - \gamma})$  in Agarwal et al. (2020), is key for Theorem 1. This is due to the fact that we need to solve the corresponding DMDP to an accuracy of  $\epsilon = \epsilon/(1 - \gamma)$ , which is on the order of  $1/(1 - \gamma)$  for constant accuracy parameter  $\epsilon \in (0, 1)$ . The  $\epsilon$  regime where we can apply prior results, i.e.  $\epsilon \leq 1/\sqrt{1 - \gamma}$ , corresponds in AMDPs in the regime where our method’s sample complexity is no better than that of  $\tilde{O}(A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$  achieved by prior work (Jin & Sidford, 2020).*

### 3. Lower Bound

In this section, we show a lower bound of  $\Omega(A_{\text{tot}} t_{\text{mix}} / \epsilon^2)$  oblivious samples for finding an  $\epsilon$ -optimal policy for an AMDP with mixing time  $t_{\text{mix}}$ . This closes the gap (up to logarithmic and poly- $\epsilon$  terms) for finding an approximately-optimal policy for an AMDP given oblivious samples and shows that the method we propose in Theorem 1 with sample complexity  $\tilde{O}(A_{\text{tot}} t_{\text{mix}} / \epsilon^3)$  is near optimal for constant  $\epsilon$ . We defer some proofs in this section to Appendix A.

To obtain this result, we provide a family of AMDP instances that we prove are difficult to solve. Our construc-

tion is similar to that given in Azar et al. (2013); Feng et al. (2019) for lower bounding the sample complexity of DMDPs. Formally, we consider the state space to be  $\mathcal{S} = \mathcal{X}^1 \cup \mathcal{X}^2 \cup \mathcal{X}^3$ , denoting three disjoint subsets of states on different levels (see Figure 1). We denote the action space as  $\mathcal{A}_s = [K]$ , for all  $s = i^1 \in \mathcal{X}^1$ , and  $\mathcal{A}_s = \{\text{single fixed action}\}$ , for all  $s \in \mathcal{X}^2 \cup \mathcal{X}^3$ .

Let  $\mathcal{X}^1$  have  $N$  independent states, each with  $K$  independent actions. We assume for state  $i^1 \in \mathcal{X}^1$ , when taking action  $a^1$ , an agent gets to some state at second level, denoted as  $i^2_{(i^1, a^1)} \in \mathcal{X}^2$ . At state  $i^2_{(i^1, a^1)} \in \mathcal{X}^2$  the agent can only take one single action after which with probability  $1 - \gamma$  it goes uniformly random to a state at the first level in  $\mathcal{X}^1$ , with probability  $p_{(i^1, a^1)} \gamma$  it goes back to its own state, and with probability  $(1 - p_{(i^1, a^1)}) \gamma$  it gets to some state on the third level denoted as  $i^3_{(i^1, a^1)} \in \mathcal{X}^3$ . At  $i^3_{(i^1, a^1)} \in \mathcal{X}^3$ , the agent can take a single action after which with probability  $1 - \gamma$  it goes uniformly randomly to a state at first level in  $\mathcal{X}^1$  while with probability  $\gamma$  it stays at the original state  $i^3_{(i^1, a^1)} \in \mathcal{X}^3$ . A reward 1 is generated when the agent transfers from a state in  $\mathcal{X}^2$  to itself, and all other transmissions generate 0 reward. See Figure 1 for an illustration.

We construct the instances such that for each state-action pair  $(i^1, a^1)$ , a chain of length-2 composed of states  $i^2_{(i^1, a^1)}$ ,  $i^3_{(i^1, a^1)}$  follows. The probability  $(1 - \gamma)$  to go back uniformly

to a state  $i^1 \in \mathcal{X}^1$  from each chain allows the entire Markov chain to “restart” from  $i^1$  uniformly, and ensures a  $O(1/(1-\gamma))$  mixing time bound, as we show in Section 3.1. When in a single chain, only the transition probability  $p_{(i^1, a^1)}$  of transiting from  $i^1_{(i^1, a^1)}$  to itself effects the average-reward.

To create our family of hard AMDP instances, we consider all instances such that for each  $i^1 \in \hat{\mathcal{X}}^1$ , one of the following two cases occurs:

- Case (i): there is one action  $k \in [K]$  that leads to transition probability  $\gamma p'$ , and all other actions with probability  $\gamma p$ ; in this case the optimal action is  $k$ .
- Case (ii): there is one action  $a^1_\star$  that leads to transition probability  $\gamma p^\star$ , one action  $k \in [K]$  leading to probability  $\gamma p'$ , and all others leading to probability  $p$ ; in this case the optimal action is  $a^1_\star$ .

In Section 3.2 we argue one needs to find the best action for at least a constant fraction of the states  $i^1 \in \mathcal{X}^1$  to obtain an  $O(\epsilon)$ -optimal policy and we show this requires at least  $\Omega(NKt_{\text{mix}}\epsilon^{-2})$  oblivious samples for properly chosen  $p$ ,  $p'$ , and  $p^\star$ .

### 3.1. Stationary Distribution and Mixing Time

First, we characterize the stationary distribution under a policy  $\pi$ ; this is useful for bounding suboptimality for policies in Section 3.2.

**Lemma 5** (Stationary Distribution Characterization). *Consider a policy that chooses always a fixed  $a^1$  for each state in  $\mathcal{X}^1$ . The stationary distribution  $\nu$  is in the following form:*

$$\begin{aligned} \nu(i^1) &= \frac{1}{N} \cdot \frac{1-\gamma}{2-\gamma}, \forall i^1 \in \mathcal{X}^1; \\ \nu(i^2_{(i^1, a^1)}) &= \frac{1}{N} \cdot \frac{1-\gamma}{(1-\gamma p_{(i^1, a^1)})(2-\gamma)}, \text{ if } \pi_{i^1}(a^1) = 1, \\ \nu(i^2_{(i^1, a^1)}) &= 0, \quad \forall a^1 \in \mathcal{A}, \pi_{i^1}(a^1) = 0; \\ \nu(i^3_{(i^1, a^1)}) &= \frac{1}{N} \cdot \frac{\gamma(1-p_{(i^1, a^1)})}{(1-\gamma p_{(i^1, a^1)})(2-\gamma)}, \text{ if } \pi_{i^1}(a^1) = 1, \\ \nu(i^3_{(i^1, a^1)}) &= 0, \quad \forall a^1 \in \mathcal{A}, \pi_{i^1}(a^1) = 0. \end{aligned}$$

The proof of Lemma 5 follows by checking the definition of stationary distribution given the transition probabilities of the model; we defer it to Appendix A.

Next, we show that the mixing time of such a Markov chain under any policy  $\pi$  is  $O(1/(1-\gamma))$  formally; we defer the complete proof to Appendix A.

**Lemma 6.** *The AMDP constructed in this section has mixing time  $t_{\text{mix}} \leq O(1/(1-\gamma))$ .*

**Sketch of Proof** We first consider a regularized probability transition matrix in form

$$\hat{\mathbf{P}} = \gamma \mathbf{P} + (1-\gamma) \mathbf{1} \mathbf{p}^\top, \quad (6)$$

for some probability transition matrix  $\mathbf{P}$  and some density vector  $\mathbf{p}$ .

Such a probability transition matrix induces a Markov chain where each step moves according to  $\mathbf{P}$  with probability  $\gamma$ , and restart from a random state following a fixed distribution  $\mathbf{p}$  with probability  $1-\gamma$ . After  $\tilde{O}(1/(1-\gamma))$  steps the initial distribution doesn't affect where one is at because with high probability it has restarted following a fixed distribution for at least once. That the distribution isn't affected by the initial distribution ensures that the Markov chain is mixing.

Unfortunately, we cannot immediately apply this result as in our Markov chain one will only restart with certain probability when at a state in  $\mathcal{X}^2 \cup \mathcal{X}^3$  (as opposed to all states). Instead we show the 2-step probability transition matrix admits the structure of a regularized probability transition, i.e. (6). Thus we apply the result to the 2-step transition matrix to argue that it mixes within  $O(1/(1-\gamma))$  steps, which we show implies that the original Markov chain mixes within time steps  $O(1/(1-\gamma))$ , proving the statement as claimed.  $\square$

### 3.2. Lower Bound Proofs

Here we show a lower bound on the sample complexity for obtaining an  $\epsilon$ -optimal policy. Without loss of generality in this section we assume  $N, K$  are at least some sufficiently large constants,  $\epsilon \leq 1/32$ , and  $\gamma \geq 1/2$ .

We consider the family of AMDPs  $\mathbf{M}$  where for each MDP instance and any fixed  $i^1 \in \hat{\mathcal{X}}^1$  either case (i) or case (ii) will happen, i.e.

$$\begin{aligned} \text{let } p &= \gamma, p' = \gamma + \epsilon(1-\gamma), p^\star = \gamma + 2\epsilon(1-\gamma), \\ \text{case (i)} : p_{(i^1, a^1)} &= \begin{cases} p' & \text{for some } k \in [K], \\ p & \text{for any } a^1 \neq k; \end{cases} \\ \text{case (ii)} : p_{(i^1, a^1)} &= \begin{cases} p^\star & \text{for some } a^1_\star \in [K], \\ p' & \text{for some } k \neq a^1_\star, \\ p & \text{for any } a^1 \notin \{a^1_\star, k\}. \end{cases} \end{aligned} \quad (7)$$

Following the characterization of stationary policy in Lemma 5, for one state  $i^1 \in \mathcal{X}^1$  the difference of rewards when choosing a suboptimal action that leads to the transition  $\gamma p_2$  of staying at its own state in  $i^2_{(i^1, a^1)}$  instead of an optimal action that leads to the transition  $\gamma p_1$ , with the



policy for all other states keeping the same will be

$$\begin{aligned}
 & \frac{1}{N} \cdot \frac{(1-\gamma)p_1\gamma}{(1-\gamma p_1)(2-\gamma)} - \frac{1}{N} \cdot \frac{(1-\gamma)p_2\gamma}{(1-\gamma p_2)(2-\gamma)} \\
 &= \frac{1}{N} \cdot \frac{1-\gamma}{2-\gamma} \frac{p_1(1-\gamma p_2) - p_2(1-\gamma p_1)}{(1-\gamma p_1)(1-\gamma p_2)} \quad (8) \\
 &= \frac{1}{N} \cdot \frac{1-\gamma}{2-\gamma} \frac{p_1 - p_2}{(1-\gamma p_1)(1-\gamma p_2)} \geq \frac{2}{N} \epsilon,
 \end{aligned}$$

where in the last inequality we use the choice of  $p_1, p_2 \in \{p, p', p^*\}$ ,  $p_1 > p_2$ , and the fact that  $\epsilon \leq 1/32 \leq \frac{1}{4(1+\gamma)^2(2-\gamma)}$ . Consequently, in order to obtain an  $\frac{1}{12}\epsilon$ -approximate deterministic policy, one must choose the optimal action, i.e.  $\pi(i) = k$  for case (i) and  $\pi(i) = a^1$  for case (ii) for a subset of  $i \in \mathcal{I} \subseteq [N]$  satisfying  $|\mathcal{I}| \geq \frac{23}{24}N$ .

Now the key argument follows from an information-theoretical lower bound for distinguishing between two binary variables with mean  $\gamma p$  and  $\gamma p^*$ , formally given as follows.

**Lemma 7.** *Given a random variable  $X$  drawn uniformly randomly from the family  $\{\text{bin}(\gamma p), \text{bin}(\gamma p^*)\}$  where  $p, p^*$  are as defined in (7). When taking fewer than  $T = \Omega(1/(1-\gamma)\epsilon^2)$  samples of  $X$ , any procedure with probability  $1/4$  will make a wrong prediction on which binary random variable  $X$  is.*

As an immediate corollary, one can show that for any algorithm  $\mathcal{K}$  taking fewer than  $T$  samples on  $a^1$  on any randomly permuted set of actions of a state  $i^1$ , it must fail to distinguish between case (i) and case (ii) in (7) with probability  $1/4$ .

Now given any algorithm that takes fewer than  $\Theta(NKT)$  deterministic<sup>3</sup> oblivious samples, we show the algorithm will not collect enough samples for a constant fraction of actions  $a^1 \in \mathcal{A}_{i^1}$  of a constant fraction of states  $i^1 \in \mathcal{X}^1$ . Thus, one can consider an adversarial instance in the family  $\mathbf{M}$  that hides the best action uniformly randomly among the actions with insufficient samples. With constant probability the algorithm will output a suboptimal action for such  $\Theta(N)$  states and output a suboptimal policy  $\pi$  satisfying  $V^\pi \leq V^* - \Theta(\epsilon)$  for some instance as we pick adversarially. By adjusting constants and combing with Lemma 6, this implies any algorithm will need at least  $\Omega(NKT) = \Omega(A_{\text{tot}}t_{\text{mix}}\epsilon^{-2})$  deterministic oblivious samples to find an  $\epsilon$ -optimal policy for all instances in the family.

Finally, we note that our lower bound statement is applicable to any algorithms yielding deterministic policy and using deterministic oblivious sampling, which already matches our upper bound results in Section 2 nearly tightly. We present our lower bound under this setting in our main paper

<sup>3</sup>Here deterministic means the algorithm takes fixed number of samples per state-action pair, regardless of the instance.

for clarity and simplicity. However, we believe one can extend the result to algorithms with randomized policies and using dynamic samples; we think these are interesting future directions to further strengthen the lower bound and discuss them in more detail in Appendix B.

## 4. Discussion

In this paper, we have shown an  $\Omega(A_{\text{tot}}t_{\text{mix}}\epsilon^{-2})$  sample complexity lower bound for AMDPs with mixing time bound  $t_{\text{mix}}$ , and a matching upper bound (up to logarithmic and poly( $1/\epsilon$ ) factors) obtained through reduction to DMDPs. Our work suggest a few open directions which we believe would help clarify the structure of AMDPs and its connection with DMDPs:

**Obtaining tight upper bound of sample complexity and runtime.** While the authors suspect that an  $\tilde{O}(A_{\text{tot}}t_{\text{mix}}\epsilon^{-2})$  upper bound on the required sample complexity may be attainable, it seems to require new ideas in leveraging the mixing structure of AMDP more directly, instead of reducing it to DMDPs. Further, it would be interesting to obtain algorithms with efficient running times as has been shown for DMDPs (Sidford et al., 2018a) in certain  $\epsilon$ -regimes.

**Relaxing the mixing bound assumption.** In certain cases, assuming global mixing time bound for all policies, even for all deterministic stationary policies (as we do in the paper), can be restrictive. We ask if it is possible to obtain sample complexity dependence in terms of the mixing time of the optimal policy, or in terms of some alternative parameters like diameter (Jaksch et al., 2010), or bias span (Bartlett & Tewari, 2012; Fruit et al., 2018) that can be smaller than  $t_{\text{mix}}$  for certain types of AMDPs.

## Acknowledgements

The authors thank Ron Amit and Lin F. Yang for helpful conversations. YJ was supported by Stanford Graduate Fellowship. AS was supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

## References

- Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83, 2020.
- Azar, M. G., Munos, R., and Kappen, H. J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.
- Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.
- Cohen, M. B., Kelner, J., Peebles, J., Peng, R., Sidford, A., and Vladu, A. Faster algorithms for computing the stationary distribution, simulating random walks, and more. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 583–592. IEEE, 2016.
- Feng, F., Yin, W., and Yang, L. F. How does an approximate model help in reinforcement learning? *arXiv preprint arXiv:1912.02986*, 2019.
- Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.
- Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.
- Jin, Y. and Sidford, A. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pp. 4890–4900. PMLR, 2020.
- Kakade, S. M. et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.
- Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.
- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.
- Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1-3):159–195, 1996.
- Ortner, P. and Auer, R. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19:49, 2007.
- Ortner, R. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128, 2020.
- Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.
- Sidford, A., Wang, M., Wu, X., and Ye, Y. 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*, pp. 770–787. SIAM, 2018b.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Tiapkin, D., Stonyakin, F., and Gasnikov, A. Parallel stochastic mirror descent for mdps. *arXiv preprint arXiv:2103.00299*, 2021.
- Wainwright, M. J. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.
- Wang, M. Randomized linear programming solves the discounted markov decision problem in nearly-linear running time. *arXiv preprint arXiv:1704.01869*, 2017a.
- Wang, M. Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017b.
- Yu, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.