On the Sample Complexity of Average-reward MDPs

Yujia Jin [1] **Aaron Sidford** [1]

Abstract

In this work we study the sample complexity for solving average-reward Markov decision processes (AMDPs). Given access to a generative model for an AMDP with $\mathrm{A}_{\mathrm{tot}}$ state-action pairs, where every stationary policy has mixing time at most t_{mix}, we present two new methods for finding approximately-optimal stationary policies. Together, these methods yield a sample complexity upper bound of $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}/\epsilon^2 \cdot \min(t_{\mathrm{mix}}, 1/\epsilon))$. We also provide a sample complexity lower bound of $\Omega(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}/\epsilon^2)$ oblivious samples. This work makes progress toward designing new algorithms with better sample complexity for solving AMDPs and points to a key open problem of closing the gap between our upper and lower bounds.

1 Introduction

Average-reward Markov decision processes (AMDPs) are a fundamental mathematical tool for modeling decision making under uncertainty and are foundational to reinforcement learning. As opposed to discounted Markov decision processes (DMDPs) and finite-horizon Markov decision processes, AMDPs consider infinite undiscounted horizons. Such a setting has received interest in optimal control, learning automata, and various real-world reinforcement learning tasks (Mahadevan, 1996; Auer & Ortner, 2007; Ouyang et al., 2017).

There has been extensive research on the sample complexity for finding approximately optimal policies for MDPs, assuming access to a generative model (Kakade et al., 2003). In settings like DMDPs (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020) and finite-horizon MDPs (Sidford et al., 2018), researchers have proposed methods that achieve near-optimal sample complexity (up to logarithmic factors) to close the gap with established sample complexity lower bound (Azar et al., 2013; Sidford et al., 2018), answering a major problem for these two classes of models. In contrast, relatively few methods and techniques to are known to be immediately generalizable to give complexity bounds for provably solving AMDPs. A prominent counterexample, is the primal-dual method on the linear programming formulation of AMDP in (Wang, 2017), which requires (i) a bounded mixing time condition andd (ii) an ergodicity assumption that bounds the stationary distribution under any policy to be τ-multiplicatively close to a uniform distribution over all states.

We take steps in improving the sample complexity of solving AMDPs. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ denote an AMDP with state space $\mathcal{S}$, action space $\mathcal{A} = \cup_{s\in\mathcal{S}} \mathcal{A}_s$ with total state-action pair size $|\mathcal{A}| = \mathrm{A}_{\mathrm{tot}}$, probability transition matrix $\mathbf{P} \in \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ capturing the transition kernel under different choices of action for each state, reward vector $\mathbf{r} \in [0,1]^{\mathcal{A}}$ capturing the instantaneous rewards received for each state-action pair (s, a_s) chosen. A (stationary) policy is defined as a mapping $\pi : \mathcal{S} \to \Delta^{\mathcal{A}}$, under which there is induced probability transition matrix and reward vector defined as

$$\begin{aligned} \mathbf{P}^\pi \text{ where } [\mathbf{P}^\pi]_{s,\cdot} &= \sum_{a\in\mathcal{A}_s} [\pi(s)]_a [\mathbf{P}]_{(s,a),\cdot}, \\ \mathbf{r}^\pi \text{ where } [\mathbf{r}^\pi]_s &= \sum_{a\in\mathcal{A}_s} [\pi(s)]_a [\mathbf{r}]_{s,a}. \end{aligned} \tag{1}$$

We consider AMDPs under the *mixing* assumption, a standard regularity assumption for studying AMDPs. (Wang, 2017; Ortner, 2020).

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

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

Further, an AMDP is d-mixing*, if the above argument holds true for all deterministic stationary policies. Note this is a weaker assumption than mixing.*

Our goal is to find an approximation of the optimal policy, which is defined to maximize the following cumulative re-

[1]Management Science and Engineering, Stanford University, CA, United States. Correspondence to: Yujia Jin <yujiajin@stanford.edu>.

Accepted at *ICML 2021 Workshop on Reinforcement Learning Theory*.

ward of AMDP given some initial distribution $\mathbf{q}$, policy π, and its induced stationary distribution $\boldsymbol{\nu}^\pi$,

$$\max_\pi V_{\mathbf{q}}^\pi := \lim_{T\to\infty} \frac{1}{T} \sum_{t\in[T]} \mathbf{q}^\top (\mathbf{P}^\pi)^t \mathbf{r}^\pi \overset{(\star)}{=} \langle \boldsymbol{\nu}^\pi, \mathbf{r}^\pi \rangle. \quad (2)$$

Note the mixing assumption $(\star)$ ensures that the cumulative reward is well-defined and doesn't depend on initial distribution; as a consequence, we may omit $\mathbf{q}$ when writing V^π for brevity.

We study the sample complexity for finding an ϵ-optimal policy π such that $V^\pi \geq \max_{\pi'} V^{\pi'} - \epsilon$, in terms of the problem dimension $\mathrm{A}_{\mathrm{tot}}$ (total state-action pairs), and mixing time bound t_{mix} (under all randomized / deterministic stationary policies). We first provide a primal-dual method based on stochastic mirror descent (SMD) that finds an ϵ-optimal randomized policy using $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2 \epsilon^{-2})$ dynamic samples. This improves upon prior art (Wang, 2017) that achieves sample complexity $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2 \tau^2 \epsilon^{-2})$ by removing an ergodicity condition (assumption (ii)) and an extra τ^2 dependence (see Table 1). Additionally, we provide an alternative approach by reducing AMDPs to solving DMDPs with $\gamma \approx 1$ to find an ϵ-optimal deterministic policy within $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}} \epsilon^{-3})$ oblivious samples. Our alternative approach crucially uses the latest state-of-the-art DMDP solver (Li et al., 2020) which achieves near-optimal sample complexity for DMDPs for all range of desired accuracy $\epsilon > 0$. Compared with the first method, the second algorithm we propose comes with some additional advantages including requiring weaker access model and weaker model assumptions; for example, it only uses oblivious samples and improves the depth for efficient parallel computing. In Table 1 we provide a detailed comparison in the features for the methods mentioned above.

This extended abstract is organized as follows: In Section 2, we provide the techniques and results for our first algorithm for solving AMDP using stochastic mirror descent. In Section 3, we provide the techniques and results for our second AMDP solver based on reduction to DMDPs. In Section 4, we present a lower bound of the problem together with the construction of hard instances. In Section 5, we point out some open problems for studying sample complexity of AMDP solvers.

We remark that Section 2 includes a subset of results in Jin & Sidford (2020), and Section 3 and 4 are based the recent work of Jin & Sidford (2021), all published by the same authors as this extended abstract.

2 A method based on stochastic mirror descent

Given a mixing AMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ with mixing time bound t_{mix}, we define $\mathbf{1}$, $\mathbf{0}$ as the all-1 and all-0 vectors respectively, and the extended identity matrix $\mathbf{E} \in \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ as $[\mathbf{E}]_{(s,a),\cdot} = \mathbf{1}_s$ where $\mathbf{1}_s$ denote the basic vector that is 1 on coordinate $s \in \mathcal{S}$ and all-0 on the rest coordinates. We consider the linear programming formulation in the following primal and dual form equivalently,

$$\begin{aligned} \text{(P)} \quad & \min_{\bar{v},\mathbf{v}} && \bar{v} \\ & \text{subject to} && \bar{v}\cdot\mathbf{1} + (\mathbf{E}-\mathbf{P})\mathbf{v} - \mathbf{r} \geq \mathbf{0}, \\ \text{(D)} \quad & \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} && \boldsymbol{\mu}^\top \mathbf{r} \\ & \text{subject to} && (\mathbf{E}-\mathbf{P})^\top \boldsymbol{\mu} = \mathbf{0}. \end{aligned} \quad (3)$$

By standard Lagrangian multiplier and duality theory, we can write problem (3) equivalently in its bilinear saddle-point form,

$$\min_{\mathbf{v}\in 4t_{\mathrm{mix}}\cdot[-1,1]^{\mathcal{S}}} \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} \quad f(\mathbf{v},\boldsymbol{\mu}), \quad (4)$$

$$\begin{aligned} \text{where } f(\mathbf{v},\boldsymbol{\mu}) &:= \bar{v} + \boldsymbol{\mu}^\top(-\bar{v}\cdot\mathbf{1} + (\mathbf{P}-\mathbf{E})\mathbf{v} + \mathbf{r}) \\ &= \boldsymbol{\mu}^\top((\mathbf{P}-\mathbf{E})\mathbf{v} + \mathbf{r}), \end{aligned} \quad (5)$$

for which we impose the constraint $\mathbf{v} \in 4t_{\mathrm{mix}}\cdot[-1,1]^{\mathcal{S}}$ for primal variables by observing the optimal primal solution satisfies $\mathbf{v}^\star \in 2t_{\mathrm{mix}}\cdot[-1,1]^{\mathcal{S}}$.

The method builds on a primal-dual stochastic mirror descent method for the bilinear saddle point objective constraining the a box-simplex space. Formally, we design stochastic estimators for the gradient operator as

$$\begin{aligned} &\text{Sample } (s,a) \sim [\boldsymbol{\mu}]_{s,a}, s' \sim \mathbf{P}((s,a),s'), \\ &\quad \text{set } \tilde{g}^{\mathrm{v}}(\mathbf{v},\boldsymbol{\mu}) = \mathbf{1}_{s'} - \mathbf{1}_s; \\ &\text{Sample } (s,a) \sim \frac{1}{\mathrm{A}_{\mathrm{tot}}}, s' \sim \mathbf{P}((s,a),s'), \\ &\quad \text{set } \tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu}) = \mathrm{A}_{\mathrm{tot}}(v_s - v_{s'} - r_{s,a})\mathbf{1}_{s,a}. \end{aligned} \quad (6)$$

Note these are unbiased estimators of $g^{\mathrm{v}} = (\mathbf{P}-\mathbf{E})^\top\boldsymbol{\mu}$ and $g^{\mu} = (\mathbf{E}-\mathbf{P})\mathbf{v} - \mathbf{r}$. They also have bounded moments (in the local-norm sense, cf. Lemma 13 in Carmon et al. (2019)) such that $\|\tilde{g}^{\mathrm{v}}\|_2^2 \leq 2$ and that $\|\tilde{g}^{\mu}\|_{\boldsymbol{\mu}'}^2 \leq O(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2)$, for any $\boldsymbol{\mu}'$. Applying standard mirror descent technique this implies one can find an ϵ-optimal saddle point $(\mathbf{v}^\epsilon, \boldsymbol{\nu}^\epsilon)$ satisfying $\max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} f(\mathbf{v}^\epsilon,\boldsymbol{\mu}) - \min_{\mathbf{v}\in 4t_{\mathrm{mix}}\cdot[-1,1]^{\mathcal{S}}} f(\mathbf{v},\boldsymbol{\mu}^\epsilon) \leq \epsilon$ with sample complexity $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2/\epsilon^2)$.

Further, the primal-dual optimality of $\max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} f(\mathbf{v}^\epsilon,\boldsymbol{\mu}) - \min_{\mathbf{v}\in 4t_{\mathrm{mix}}\cdot[-1,1]^{\mathcal{S}}} f(\mathbf{v},\boldsymbol{\mu}^\epsilon) \leq \epsilon$, together with the choice of primal space $\|\mathbf{v}\|_\infty \leq 4t_{\mathrm{mix}}$ in comparison with $\|\mathbf{v}^\star\|_\infty \leq 2t_{\mathrm{mix}}$, implies that the approximate dual variables $\boldsymbol{\lambda}^\epsilon$ of linear programming (D) in (3) satisfy the dual constraints approximately, i.e. $\mathbb{E}\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\|_1 \leq \frac{1}{M}\epsilon$. Now consider policy π^ϵ induced by $\boldsymbol{\nu}^\epsilon$ satisfying $\pi^\epsilon_{s,a} = \frac{\nu^\epsilon_{s,a}}{\sum_{a\in\mathcal{A}_s}\nu^\epsilon_{s,a}}$, we can formally show that the cumulative (average) reward under policy π^ϵ satisfies $V^{\pi^\epsilon} \geq \max_\pi V^\pi - 3\epsilon$.

Table 1. **Upper and lower bounds on sample complexity to get ϵ-optimal policy for AMDPs.** Here $\mathrm{A}_{\mathrm{tot}}$ denotes the total size of all state-action pairs. Parameter τ shows up when the designed algorithm requires additional ergodic condition for MDP, i.e. there exists some distribution $\mathbf{q}$ and $\tau > 0$ satisfying $\sqrt{1/\tau}\mathbf{q} \le \boldsymbol{\nu}^\pi \le \sqrt{\tau}\mathbf{q}$, $\forall$ policy π and its induced stationary distribution $\boldsymbol{\nu}^\pi$.

Method	Sample Complexity	Mixing Condition	Generative Model
lower bound (our result)	$\Omega\left(\frac{\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}}{\epsilon^2}\right)$	N/A	N/A
Primal-Dual Method (Wang, 2017)	$\widetilde{O}\left(\frac{\tau^2 \mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2}{\epsilon^2}\right)$	mixing	dynamic
Primal-Dual SMD (our result)	$\widetilde{O}\left(\frac{\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2}{\epsilon^2}\right)$	mixing	dynamic
Reduction to DMDPs (our result)	$\widetilde{O}\left(\frac{\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}}{\epsilon^3}\right)$	d-mixing	oblivious

As a result, we show one can round the dual variable $\boldsymbol{\nu}^\epsilon$ in an ϵ-optimal saddle point to obtain an 3ϵ-optimal policy π^ϵ. Formally, this gives the following sample complexity guarantee on finding an ϵ-optimal policy for AMDPs.

Theorem 1. *Given a mixing AMDP with mixing time bounded by t_{mix} and accuracy parameter $\epsilon \in (0,1)$, Algorithm 1 with parameter choice $\eta^{\mathsf{v}} = O(\epsilon)$, $\eta^{\mu} = O(\epsilon t_{\mathrm{mix}}^{-2} {\mathrm{A}_{\mathrm{tot}}}^{-1})$ finds an ϵ-optimal randomized policy in expectation with sample complexity $O(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2} \log(\mathrm{A}_{\mathrm{tot}}))$.*

3 A method based on reduction to discounted MDP

Given a d-mixing AMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$, we study the relationship between its cumulative reward and the one of a corresponding discounted MDP (DMDP), with the same state, action space, transition probabilities, instantaneous rewards, and a discount factor $\gamma < 1$. We define the value vector of such DMDP under a policy π as

$$\mathbf{v}_{\mathrm{dis}}^\pi = \sum_{t\ge 0} \gamma^t \left(\mathbf{P}^\pi\right)^t \mathbf{r}^\pi.$$

Note here $[\mathbf{v}_{\mathrm{dis}}^\pi]_s$ represents the cumulative reward starting from an initial distribution $\mathbf{q} = \mathbf{1}_s$ for the given γ-discounted MDP.

One can define a similar value vector for AMDP as

$$\mathbf{v}_{\mathrm{avg}}^\pi = \langle \boldsymbol{\nu}^\pi, \mathbf{r}^\pi \rangle \cdot \mathbf{1},$$

where ν^π is the stationary distribution under policy π. Similarly each coordinate of $\mathbf{v}_{\mathrm{avg}}^\pi$ also represents the cumulative reward starting from an initial distribution, and are the same since it doesn't depend on the initial distribution under the d-mixing assumption.

Given any policy π, we again utilize the helper lemma on structure of transition kernel $\mathbf{P}^\pi$ to show the following relationship between the value vectors of AMDP and its corresponding DMDP.

$$\begin{aligned}
&\|\mathbf{v}_{\mathrm{avg}}^\pi - (1-\gamma)\mathbf{v}_{\mathrm{dis}}^\pi\|_\infty \\
=&\left\|(1-\gamma)\sum_{t\ge 0}\gamma^t \langle \mathbf{r}^\pi, \boldsymbol{\nu}^\pi\rangle \mathbf{1} - (1-\gamma)\sum_{t\ge 0}\gamma^t (\mathbf{P}^\pi)^t \mathbf{r}^\pi\right\|_\infty \\
\le&(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\|_\infty \cdot \|\mathbf{r}^\pi\|_\infty \\
\le&(1-\gamma)\sum_{t=0}^{\lceil t_{\mathrm{mix}}\rceil - 1} 2\gamma^t + (1-\gamma)\sum_{t\ge \lceil t_{\mathrm{mix}}\rceil}\frac{1}{2^{\lfloor k/t_{\mathrm{mix}}\rfloor}} \\
\le&3(1-\gamma)t_{\mathrm{mix}}\,.
\end{aligned}$$

This shows under the same policy, the values of AMDP and its corresponding DMDP are close up to ϵ when choosing the discount factor $\gamma = 1 - \Theta(\epsilon/t_{\mathrm{mix}})$. Also, one can show that it suffices to solve the corresponding DMDP to an accuracy of $\varepsilon = \epsilon/(1-\gamma)$. By plugging in the sample complexity of most recent DMDP solvers (Li et al., 2020), we thus obtain an upper bound of sample complexity as

$$\widetilde{O}\left(\frac{\mathrm{A}_{\mathrm{tot}}}{(1-\gamma)^3\varepsilon^2}\right) \overset{\text{choice of } \varepsilon}{=\!=\!=} \widetilde{O}\left(\frac{\mathrm{A}_{\mathrm{tot}}}{(1-\gamma)\epsilon^2}\right) \overset{\text{choice of } \gamma}{=\!=\!=} \widetilde{O}\left(\frac{\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}}{\epsilon^3}\right).$$

Theorem 2. *Given a d-mixing AMDP with mixing time bounded by t_{mix} and accuracy parameter $\epsilon \in (0,1)$, there exists algorithm that finds an ϵ-optimal deterministic policy with probability $1-\delta$ with sample complexity $O(\mathrm{A}_{\mathrm{tot}} \log(\mathrm{A}_{\mathrm{tot}}/\epsilon\delta) t_{\mathrm{mix}}/\epsilon^3)$.*

We remark that for this reduction-based method to be an improvement over the SMD-based method in Section 2, we crucially rely on the fact that Li et al. (2020) obtains near-optimal sample complexity $\widetilde{O}(\mathrm{A}_{\mathrm{tot}}(1-\gamma)^{-2}\epsilon^{-2})$

for solving γ-discounted MDPs in the regime of accuracy parameter $\epsilon \in (0, 1/(1-\gamma))$, improving over the regime of accuracy $\epsilon \in (0, 1/\sqrt{1-\gamma})$ in prior work (Sidford et al., 2018; Agarwal et al., 2020). To see this, we note $\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}/\epsilon^3 \leq \mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}^2/\epsilon^2$ if and only if $\epsilon \gg \Omega(1/t_{\mathrm{mix}})$,i.e. this method improves over the SMD-based one if and only if $\varepsilon \gg 1/\sqrt{1-\gamma}$ for our choice of $\varepsilon = \epsilon/(1-\gamma), \gamma = 1 - \Theta(\epsilon/t_{\mathrm{mix}})$.

Finally, we mention a few advantages of this method over the prior one based on stochastic mirror descent (also see Table 1). First, it replaces a t_{mix} dependence with an extra $1/\epsilon$ dependence in the sample complexity bound. This is in favor in practice as t_{mix} might scale as problem dimension $\Omega(S)$ but the desired accuracy ϵ usually doesn't. Second, it requires a weaker assumption on the AMDP, i.e. a t_{mix} upper bound on mixing time for all deterministic stationary policies. On the contrary, the first method we propose requires the same mixing time bound for all randomized stationary policies. Third, it only uses oblivious samples instead of dynamic ones, which might be easier to access and collect through preprocessing. Further, it enables parallel algorithm for solving AMDP with $O(1)$ depth, improving the depth dependence of prior methods (Tiapkin et al., 2021) for this model.

4 A lower bound

Finally, we show a hard instance construction which implies an $\Omega(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}/\epsilon^2)$ lower bound in terms of sample complexity for solving mixing AMDPs. We utilize a variant of the hard instance for DMDP solvers constructed in Azar et al. (2013).

The construction of the hard instance is: 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 by 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 one 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 γ it stays at the original state $i^3_{(i^1,a^1)} \in \mathcal{X}^3$. A reward 1 is only generated while the agent transfers from a state in $\mathcal{X}^2$ to itself, and all other transmissions generate 0 reward. See Figure 1 in Appendix for the detailed illustration of the hard instance.

We construct the instance 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. In a single chain, only the transition probability $p_{(i^1,a^1)}$ of transiting from $i^2_{(i^1,a^1)}$ to itself matters in terms of the average-reward when staying inside the chain. To create our family of hard AMDP intances, we hide a best action $a^1_\star$ for each $i^1 \in \mathcal{X}^1$ with $p_{(i^1,a^1_\star)} = p^\star = \gamma + \epsilon(1-\gamma)$, and let all other actions lead to transition probability $p_{(i^1,a^1)} = p = \gamma < p^\star$. We note one needs to find the best action for at least a constant portion of the states $i^1 \in \mathcal{X}^1$ to get an ϵ-optimal policy, and to find the best action for each $i^1 \in \mathcal{X}^1$ with constant probability requires $\Omega(K/(1-\gamma)\epsilon^2)$ samples on some consecutive states of i^1. Putting these together, we know for any algorithm finding an ϵ-optimal policy with constant probability would require at least $\Omega(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}} \epsilon^{-2})$ oblivious samples.

Theorem 3. *There are constants $\epsilon_0, \delta_0 \in (0, 1/2)$ such that for all $\epsilon \in (0, \epsilon_0)$ and any algorithm $\mathcal{K}$, on input mixing AMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ given by a generative model outputs a policy π satisfying $V^\pi \geq \max_{\pi'} V^{\pi'} - \epsilon$ with probability at least δ_0, $\mathcal{K}$ makes at least $\Omega(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}}/\epsilon^2)$ oblivious queries to the generative model on some mixing AMDP instance $\mathcal{M}_0$.*

5 Open problems

In this paper, we have shown an $\Omega(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}} \epsilon^{-2})$ sample complexity lower bound for AMDPs with mixing time bound t_{mix}, and two different algorithms that achieve upper bound $O(\mathrm{A}_{\mathrm{tot}} t^2_{\mathrm{mix}} \epsilon^{-2})$ (through stochastic mirror descent) or $O(\mathrm{A}_{\mathrm{tot}} t^2_{\mathrm{mix}} \epsilon^{-2})$ (through reduction to DMDP), under slightly different regularity conditions. Here we point out the main open problems for understanding the sample complexity of AMDPs under a generative model access:

Obtaining tight upper bound of sample complexity and runtime. The authors conjecture that the lower bound is tight, and an $\widetilde{O}(\mathrm{A}_{\mathrm{tot}} t_{\mathrm{mix}} \epsilon^{-2})$ upper bound on the required sample complexity may be attainable. However, it seems to possibly require new ideas in leveraging the mixing structure of AMDP more directly, instead of reducing it to DMDPs.

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 second method), 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_{mix} for certain types of AMDPs.

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.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 49–56, 2007.

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.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. *arXiv preprint arXiv:1907.02056*, 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.

Jin, Y. and Sidford, A. Towards tight bounds on the sample complexity of average-reward mdps. *arXiv preprint arXiv:2106.07046*, 2021.

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

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, R. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128, 2020.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pp. 1333–1342, 2017.

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, 2018.

Tiapkin, D., Stonyakin, F., and Gasnikov, A. Parallel stochastic mirror descent for mdps. *arXiv preprint arXiv:2103.00299*, 2021.

Wang, M. Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

A Supplementary Materials

Here we provide an algorithm pseudocode for SMD-based AMDP solver that we propose in Section 2.

Algorithm 1 SMD for mixing AMDP

Input: MDP tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{R})$, initial $(\mathbf{v}_0, \boldsymbol{\mu}_0) \in \mathcal{V} \times \mathcal{U}$, with $\mathcal{V} \coloneqq 4t_{\text{mix}} \cdot [-1, 1]^{\mathcal{S}}$.
Output: An expected ϵ-approximate solution $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ for problem (4).
Parameter: Step-size η^{v}, η^{μ}, number of iterations T, accuracy level ϵ.
for $t = 1, \ldots, T$ **do**
  get gradient estimators $\tilde{g}^{\mathsf{v}}_{t-1}$, $\tilde{g}^{\mu}_{t-1}$ as in (6)
  {*Stochastic mirror descent steps* (Π *as projection*)}
  $\mathbf{v}_t \leftarrow \Pi_{4t_{\text{mix}} \cdot [-1,1]^{\mathcal{S}}}(\mathbf{v}_{t-1} - \eta^{\mathsf{v}} \tilde{g}^{\mathsf{v}}_{t-1})$
  $\boldsymbol{\mu}_t \leftarrow \Pi_{\Delta^{\mathcal{A}}}(\boldsymbol{\mu}_{t-1} \circ \exp(-\eta^{\mu} \tilde{g}^{\mu}_{t-1}))$
end for
Return $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon) \leftarrow \frac{1}{T} \sum_{t \in [T]} (\mathbf{v}_t, \boldsymbol{\mu}_t)$

In Figure 1, we provide a detailed illustration of the hard instance construction for proving the lower bound in Section 4.

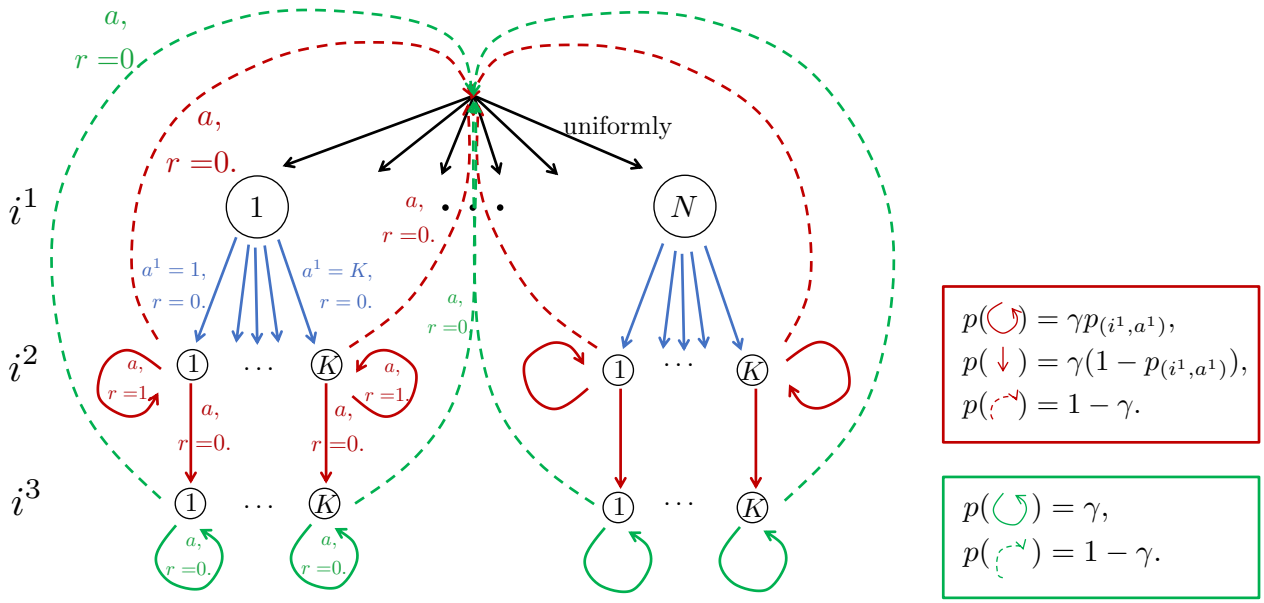

Figure 1. **AMDP lower bound hard instance illustration.** N states in $\mathcal{X}^1$ (corresponding to first level), K action per state $i^1 \in \mathcal{X}^1$, $\mathrm{A}_{\text{tot}} = O(NK)$ total state-action pairs. $\gamma \in (0, 1)$, $p_{i^1,a^1} \in [0, 1]$ for all $i^1 \in \mathcal{X}^1$,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