Optimal Sample Complexity for Average Reward Markov Decision Processes

Shengbo Wang, Jose Blanchet, Peter Glynn

shengbo.wang@stanford.edu, jose.blanchet@stanford.edu, glynn@stanford.edu

Stanford MS&E

Our Main Contributions

We resolve the open question regarding the sample complexity of policy learning for maximizing the long-run average reward associated with a uniformly ergodic Markov decision process (MDP), assuming a generative model. In this context, the existing literature provides a sample complexity upper bound of $\widetilde{O}(|S||A|t_{\mathrm{mix}}^2\epsilon^{-2})$ and a lower bound of $\Omega(|S||A|t_{\mathrm{mix}}\epsilon^{-2})$. In these expressions, $|S|$ and $|A|$ denote the cardinalities of the state and action spaces respectively, t_{mix} serves as a uniform upper limit for the total variation mixing times, and ϵ signifies the error tolerance. Therefore, a notable gap of t_{mix} still remains to be bridged. Our primary contribution is the development of an estimator for the optimal policy of average reward MDPs with a sample complexity of $\widetilde{O}(|S||A|t_{\mathrm{mix}}\epsilon^{-2})$. This marks the first algorithm and analysis to reach the literature's lower bound.

Table 1: Sample complexities of AMDP algorithms. When t_{mix} appears in the sample complexity, an assumption of uniform ergodicity is being made, while the presence of H^3 is associated with an assumption that the MDP is weakly communicating.

Algorithm	Origin	Sample complexity upper bound ($\widetilde{O}$)
Primal-dual π learning	Wang (2017)	$\lvert S\rvert\lvert A\rvert\tau^2 t_{\mathrm{mix}}^2\epsilon^{-2}$ [2]
Primal-dual SMD[1]	Jin and Sidford (2020)	$\lvert S\rvert\lvert A\rvert t_{\mathrm{mix}}^2\epsilon^{-2}$
Reduction to DMDP[1]	Jin and Sidford (2021)	$\lvert S\rvert\lvert A\rvert t_{\mathrm{mix}}\epsilon^{-3}$
Reduction to DMDP	Wang et al. (2022)	$\lvert S\rvert\lvert A\rvert H\epsilon^{-3}$
Refined Q-learning	Zhang and Xie (2023)	$\lvert S\rvert\lvert A\rvert H^2\epsilon^{-2}$
Reduction to DMDP	This paper	$\lvert S\rvert\lvert A\rvert t_{\mathrm{mix}}\epsilon^{-2}$
Lower bound	Jin and Sidford (2021)	$\Omega(\lvert S\rvert\lvert A\rvert t_{\mathrm{mix}}\epsilon^{-2})$
	Wang et al. (2022)	$\Omega(\lvert S\rvert\lvert A\rvert H\epsilon^{-2})$

Markov Decision Processes

An MDP model $\mathcal{M}$ is denoted by $\mathcal{M} = (S, A, P, r)$. Here, S, A denote the finite state and action spaces, respectively. The transition kernel is $P = \{p_{s,a} \in \mathcal{P}(S), s \in S, a \in A\}$. The reward function is $r : S \times A \to [0, 1]$. To achieve optimal decision-making in the context of infinite horizon average reward MDPs (AMDPs) or discounted MDPs (DMDPs), it suffices to consider the policy class Π consisting of stationary, Markov, and deterministic policies. Under policy $\pi \in \Pi$, the state process $\{X_t, t \geq 0\}$ is a Markov chain with transition matrix P_π defined by $P_\pi(s, s') = p_{s,\pi(s)}(s')$.

Uniform Ergodicity and Mixing Time

The transition kernel P_π is uniformly ergodic if for some $m \geq 0$,

$$\max_{s\in S} \|P_\pi^m(s,\cdot) - \eta_\pi(\cdot)\|_1 \leq \frac{1}{2}.$$

Here $\eta_\pi(\cdot)$ is the unique stationary distribution of P_π and $\|\cdot\|_1$ is the ℓ_1 distance.

The paper considers the uniformly ergodic MDPs: an MDP is uniformly ergodic if for all $\pi \in \Pi$, P_π is uniformly ergodic. Then, define the mixing time as

$$t_{\mathrm{mix}} := \max_{\pi\in\Pi} \inf\left\{m \geq 1 : \max_{s\in S} \|P_\pi^m(s,\cdot) - \eta_\pi(\cdot)\|_1 \leq \frac{1}{2}\right\} < \infty. \quad (1)$$

Discounted MDPs: Optimal Sample Complexity

The *discounted value function* $v^\pi(s)$ of a DMDP is defined via

$$v^\pi(s) := E^\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \middle| X_0 = s\right].$$

It can be seen as a vector $v^\pi \in \mathbb{R}^{|S|}$, and computed using the formula $v^\pi = (I - \gamma P_\pi)^{-1} r_\pi$. The optimal discounted value function is defined as $v^*(s) := \max_{\pi\in\Pi} v^\pi(s)$, for every $s \in S$.

It is well known that v^* is the unique solution of the following *Bellman equation*:

$$v^*(s) = \max_{a\in A} \left(r(s,a) + \gamma p_{s,a}[v^*]\right). \quad (2)$$

Moreover, the greedy policy $\pi^*(s) \in \arg\max_{a\in A} (r(s,a) + \gamma p_{s,a}[v^*])$ is optimal.

We modify the Perturbed Model-based Planning in (Li et al., 2020):

Algorithm Perturbed Model-based Planning: PMBP(γ, ζ, n)

Input: Discount $\gamma \in (0, 1)$. Perturbation size $\zeta > 0$. Sample size $n \geq 1$.

Sample small perturbation $Z(s, a) \sim U(0, \zeta)$ and compute $R = r + Z$.

Sample i.i.d. $S_{s,a}^{(1)}, S_{s,a}^{(2)}, , S_{s,a}^{(n)}$, for each $(s, a) \in S \times A$. Then, compute the *empirical kernel* $\widehat{P} := \{\hat{p}_{s,a}(s') : (s, a) \in S \times A, s' \in S\}$ where

$$\hat{p}_{s,a}(s') := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{S_{s,a}^{(i)} = s'\right\}, \quad (s,a) \in S \times A.$$

Compute the solution $\hat{v}_0$ to the empirical version of the Bellman equation (2); i.e. $\forall s \in S$, $\hat{v}_0(s) = \max_{a\in A} (R(s,a) + \gamma\hat{p}_{s,a}[\hat{v}_0])$. Then, extract the greedy policy

$$\hat{\pi}_0(s) \in \arg\max_{a\in A} \left(R(s,a) + \gamma\hat{p}_{s,a}[\hat{v}_0]\right), \quad s \in S.$$

return $\hat{\pi}_0$.

We choose a perturbation size $\zeta = (1 - \gamma)\epsilon/4$ and a total sample size

$$|S||A|n = \widetilde{O}\left(\frac{|S||A|t_{\mathrm{mix}}}{(1-\gamma)^2\epsilon^2}\right)$$

where $\widetilde{O}$ hides log factors (in particular $\log(1/\delta)$). Then, we show that w.p at least $1 - \delta$, the output $\hat{\pi}_0$ satisfies $0 \leq v^* - v^{\hat{\pi}_0} \leq \epsilon$. This is optimal as it achieves the lower bound in Wang et al. (2023).

Average Reward MDPs: Optimal Sample Complexity

Under uniform ergodicity, the *long-run average reward* of any policy $\pi \in \Pi$ is defined as

$$\alpha^\pi := \lim_{T\to\infty} \frac{1}{T} E^\pi\left[\sum_{t=0}^{T-1} r(X_t, A_t) \middle| X_0 = s\right]$$

where the limit always exists and doesn't depend on s. The long-run average reward α^π can be characterized via any solution pair (u, α), $u : S \to \mathbb{R}$ and $\alpha \in \mathbb{R}$ to the *Poisson's equation*,

$$r_\pi - \alpha = (I - P_\pi)u. \quad (3)$$

A solution pair (u, α) always exists and is unique up to a shift in u; i.e. $\{(u + ce, \alpha) : c \in \mathbb{R}\}$, where $e(s) = 1, \forall s \in S$, are all the solution pairs to (3).

Define the optimal long-run average reward $\bar{\alpha}$ as $\bar{\alpha} := \max_{\pi\in\Pi} \alpha^\pi$. Then, for any $\bar{\pi}$ that achieve the above maximum, $(u^{\bar{\pi}}, \bar{\alpha})$ solves $r_{\bar{\pi}} - \bar{\alpha} = (I - P_{\bar{\pi}})u^{\bar{\pi}}$.

Algorithm Reduction and Perturbed Model-based Planning

Input: Error tolerance $\epsilon \in (0, 1]$.

Assign

$$\gamma = 1 - \frac{\epsilon}{c_1 t_{\mathrm{mix}}}, \quad \zeta = \frac{1}{4}(1-\gamma)t_{\mathrm{mix}}, \quad \text{and } n = \frac{c_2\ell}{(1-\gamma)^2 t_{\mathrm{minorize}}}$$

where $c_1, c_2 > 1$ are a numerical constant, and ℓ is a log order term.

Run Algorithm 1 with parameter specification PMBP(γ, ζ, n) and obtain output $\hat{\pi}_0$.

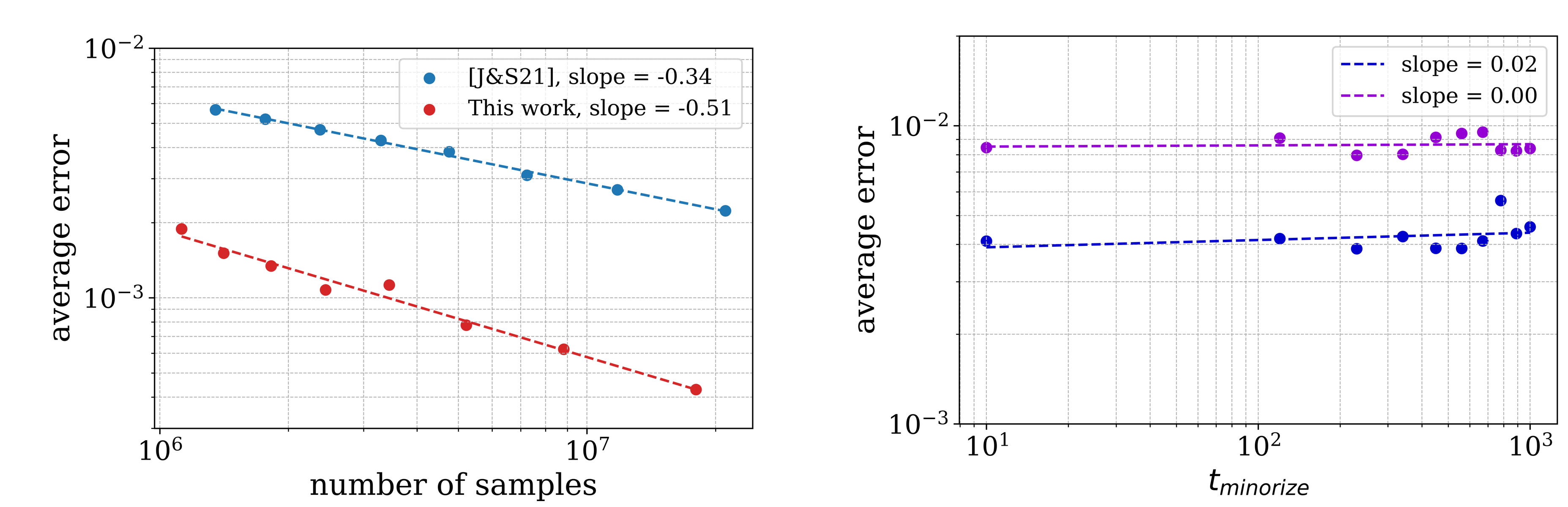
return $\hat{\pi}_0$.

By this algorithm, the total sample size is

$$|S||A|n = \widetilde{O}\left(\frac{|S||A|t_{\mathrm{mix}}}{\epsilon^2}\right)$$

where again $\widetilde{O}$ hides log factors. We show that w.p at least $1 - \delta$, the output $\hat{\pi}_0$ satisfies $0 \leq \bar{\alpha} - \alpha^{\hat{\pi}_0} \leq \epsilon$. This achieves the lower bound in Jin and Sidford (2021), hence optimal.

Numerical Validation



(a) Convergence rate comparison with Jin and Sidford (2021). A -0.5 slope verifies the $\widetilde{O}(\epsilon^{-2})$ dependence.

(b) Verification of t_{minorize} dependence. A 0 slope indicates the $\widetilde{O}(t_{\mathrm{minorize}})$ dependence.

Figure 1: Numerical experiments using the hard MDP instance in Wang et al. (2023).