LEARNING POLICY FROM A SINGLE TRAJECTORY IN AVERAGE-REWARD MARKOV DECISION PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

While there is an extensive body of work characterizing the sample complexity of discounted cumulative-reward MDPs, finite sample analyses for average-reward MDPs have been limited, and most existing works rely on restrictive assumptions such as ergodicity or access to a generative model. In this work, we establish the first finite sample complexity guarantees from a single trajectory for weakly communicating average-reward MDPs. To this end, we study the dynamics of a single trajectory in weakly communicating MDPs, and based on this analysis, we develop novel model-free methods. Notably, our value-based and policy-based methods provide finite sample complexity guarantees of $\widetilde{O}(1/\varepsilon^2)$ and $\widetilde{O}(1/\varepsilon^4)$ from a single trajectory in weakly communicating MDPs, respectively. Furthermore, we show that our value-based method requires no prior knowledge of problem-dependent quantities for communicating MDPs.

1 INTRODUCTION

Average-reward Markov decision processes (MDPs) provide a fundamental framework for modeling sequential decision-making problems whose objective is to maximize long-term steady-state performance. Under the average-reward criterion, characterizing the sample complexity for finding an ε-optimal policy has been an active topic in the reinforcement learning (RL) theory literature. However, most existing finite sample analyses rely either on restrictive sampling models, such as access to a generative model or i.i.d. offline data, or on restrictive structural assumptions, such as ergodicity. Overall, finite sample complexity analysis for average-reward RL from a single trajectory remains unexplored.

Contribution. In this work, we establish finite sample complexity analyses from a single trajectory in weakly communicating average-reward MDPs for the first time (see Table 1). To guarantee a PAC (probably approximately correct) bound for obtaining an ε-optimal policy in this setup, we analyze the structure of weakly communicating MDPs through recurrent-transient classification and develop theoretical machinery based on probabilistic arguments with the stopping time and strong Markov property. Incorporating this machinery, we propose a novel value-based method SAVIC and a policy-based method SCPMA which achieve sample complexities of $\widetilde{O}(1/\varepsilon^2)$ and $\widetilde{O}(1/\varepsilon^4)$, respectively. We also introduce a first model-free method, SAVIC+, that requires no prior knowledge of problem-dependent quantities and achieves $\widetilde{O}(1/\varepsilon^2)$ sample complexity from a single trajectory for communicating MDPs.

1.1 PRELIMINARIES AND NOTATIONS

Average reward. Let $\mathcal{M}(\mathcal{X})$ be the space of probability distributions over a set $\mathcal{X}$. We write $(\mathcal{S}, \mathcal{A}, P, r)$ to denote the infinite-horizon undiscounted MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, transition matrix $P\colon \mathcal{S}\times\mathcal{A}\to\mathcal{M}(\mathcal{S})$, and bounded reward $r\colon \mathcal{S}\times\mathcal{A}\to[-B,B]$. Let $\pi\colon \mathcal{S}\to\mathcal{M}(\mathcal{A})$ denote a policy,

$$J^\pi(s,a)=\lim_{H\to\infty}\frac{1}{H}\mathbb{E}_\pi\left[\sum_{h=0}^{H-1} r(s_h,a_h)\,\middle|\, s_0=s, a_0=a\right]$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior works	MDP class	Complexity	Method type	Additional assumptions
Li et al. (2024)	ergodic	ε^{-2}	model-free	-
Tuynman et al. (2024)	communicating	ε^{-2}	model-based	-
(Lee & Ryu, 2025b)	weakly communicating	ε^{-8}	model-free	finite coverage & ergodicity on behavior policy
Our work	weakly communicating	$\varepsilon^{-2}, \varepsilon^{-4}$	model-free	-

Table 1: Comparison of analyses for finding an ε-optimal policy from a single trajectory in average-reward MDPs. (Only the dependence on ε is shown for the sample complexity.) For ergodic MDPs, Li et al. (2024) establish finite sample complexity for stochastic first-order policy optimization methods that combine policy mirror descent with variance-reduced temporal-difference learning. For communicating MDPs, Tuynman et al. (2024) propose a model-based fixed-confidence algorithm without any prior knowledge of problem-dependent quantities and establish corresponding lower bounds. For weakly communicating MDPs, Lee & Ryu (2025b) study the sample complexity of average-reward fitted Q-iteration using function approximation and offline data generated from a single trajectory under an additional finite coverage assumption and an ergodicity assumption on the behavior policy.

for the average reward of a given policy (Puterman, 2014, Section 9.1), and

$$Q^\pi(s,a) = \lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} \mathbb{E}_\pi \left[\sum_{i=0}^{h-1} \big(r(s_i,a_i) - J^\pi(s_i,a_i) \big) \,\middle|\, s_0 = s, a_0 = a \right]$$

for the state-action value function (under an aperiodicity assumption, averaging with respect to H can be omitted in the definition (Puterman, 2014, Section 8.2)), where $\mathbb{E}_\pi$ denotes the expected value over all trajectories $(s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1})$ induced by P and π. Denote $r^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[r(s,a)\right]$ for the reward induced by policy π. For $x \in \mathbb{R}^d$ we denote by $\|x\|_\infty = \max_i |x_i|$ its infinity norm and by $\|x\|_{\mathrm{sp}} = \max_i x_i - \min_i x_i$ its span seminorm. We say $J^\star$ is an optimal average reward if $J^\star = \max_\pi J^\pi$ and an optimal average reward always exists in an average-reward MDP (Puterman, 2014, Section 9.1). We say π is an ε-optimal policy if $\|J^\star - J^\pi\|_\infty \le \varepsilon$. We denote $P_\star = \lim_{H\to\infty} \frac{1}{H} \sum_{i=0}^{H-1} P^i$ as the Cesàro limit of a stochastic matrix P. (Cesàro limit of a stochastic matrix always exists (Puterman, 2014, Theorem A.6)). We denote the transition matrix induced by policy π and P as

$$P^\pi((s,a),(s',a')) = \mathrm{Prob}((s,a) \to (s',a') \,|\, s' \sim P(\cdot\,|\,s,a), a' \sim \pi(\cdot\,|\,s')),$$

where $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$. Then, by definition, we can write (Puterman, 2014, Section 8.2)

$$J^\pi = P_\star^\pi r, \qquad Q^\pi = (I - P^\pi + P_\star^\pi)^{-1}(I - P_\star^\pi) r$$

For every policy π, it is known that Q^π and J^π satisfy the following *Bellman equations* (Puterman, 2014, Theorem 8.2.6):

$$P^\pi J^\pi = J^\pi, \qquad r + P^\pi Q^\pi = J^\pi + Q^\pi.$$

MDP classes. MDPs are classified according to the structure of their transition matrices. (For definitions of irreducible classes, accessibility, and aperiodicity of transition matrices, refer to (Puterman, 2014, Appendix A.2).) An MDP is *ergodic* if, for every policy π, the induced transition matrix has a single recurrent class and is aperiodic. An MDP is *communicating* if there exists a set of states such that every state in the set is accessible from every other state in the set under some policy. An MDP is *weakly communicating* if there is a set of states where each state in the set is accessible from every other state in the set under some policy, together with a possibly empty set of states that are transient under all policies. Every ergodic MDP is communicating, and every communicating MDP is weakly communicating.

Value Iteration. Given an undiscounted MDP $(\mathcal{S}, \mathcal{A}, P, r)$, for $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, define the Bellman optimality and consistency operators $\mathbf{T}$ and $\mathbf{T}^\pi$ as

$$\mathbf{T}Q(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(\cdot\,|\,s,a)}\Big[\max_{a'\in\mathcal{A}} Q(s',a')\Big]$$

$$\mathbf{T}^\pi Q(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(\cdot\,|\,s,a), a'\sim\pi(\cdot\,|\,s')}[Q(s',a')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, respectively. We define the standard Value Iteration (VI) for Bellman consistency and optimality operators as

$$Q^k = \mathbf{T}Q^{k-1}, \qquad Q^k = \mathbf{T}^\pi Q^{k-1} \qquad \text{for } k = 1, 2, \ldots, K,$$

where Q^0 is an initial point. For notational conciseness, we write $\mathbf{T}^\pi Q = r + P^\pi Q$ and $\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}[Q(s,a)] = Q(s,\pi(s))$.

Policy mirror ascent. Given a weakly communicating average-reward MDP $(\mathcal{S}, \mathcal{A}, P, r)$, for a positive policy $\pi \in \Pi_+$, the policy gradient ∇J^π is

$$\nabla_\theta J^{\pi_\theta} = \sum_{s\in\mathcal{R}} \sum_{a\in\mathcal{A}} d^{\pi_\theta}(s) \nabla_\theta \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a),$$

where $\mathcal{R}$ is a single recurrent class and d^π is the unique stationary distribution Lee & Ryu (2025a). Next, consider the direct parameterization $\pi_\theta(a\,|\,s) = \theta_{s,a}$, where $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ satisfies $\sum_{a\in\mathcal{A}} \theta_{s,a} = 1$ and $\theta_{s,a} \geq 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Let $h : \mathcal{M}(\mathcal{A}) \to \mathbb{R}$ be a strictly convex function and continuously differentiable on the relative interior of $\mathcal{M}(\mathcal{A})$, denoted as $\operatorname{rint} \mathcal{M}(\mathcal{A})$. Define the Bregman divergence generated by h as

$$D(p, p') \;=\; h(p) - h(p') - \langle \nabla h(p'),\, p - p' \rangle, \qquad \forall\, p \in \mathcal{M}(\mathcal{A}),\, p' \in \operatorname{rint}\, \mathcal{M}(\mathcal{A}).$$

Then, we define policy mirror ascent as

$$\pi_{k+1}(\cdot\,|\,s) = \operatorname*{arg\,max}_{p\in\mathcal{M}(\mathcal{A})} \left\{ b_k \sum_{a\in\mathcal{A}} Q^{\pi_k}(s,a) p(a) - D\big(p(\cdot), \pi_k(\cdot\,|\,s)\big) \right\}$$

for all $s \in \mathcal{R}$ where π_0 is an initial policy, and b_k is a step size.

2 A Single Trajectory in Weakly Communicating MDPs

In this section, we provide stochastic analysis for a single trajectory in weakly communicating MDPs. We first study the structure of weakly communicating MDPs through recurrent-transient classification. For the analysis, we define the class of policies

$$\Pi = \text{set of all policies} = \mathcal{M}(\mathcal{A})^{\mathcal{S}}, \qquad \Pi_+ = \{\pi \in \Pi \,|\, \pi(a\,|\,s) > 0 \text{ for all } (s,a) \in \mathcal{S}\times\mathcal{A}\},$$

so that Π_+ is the (relative) interior of Π.

2.1 Recurrent-transient classifications

Definition 2.1. Given a policy $\pi \in \Pi$, a state $s \in \mathcal{S}$ is recurrent if its return time starting from s is finite with probability 1. Otherwise, s is transient.

With the policy class Π_+ instead of Π, (Lee & Ryu, 2025c) shows that the recurrent–transient classification of states is invariant across all $\pi \in \Pi_+$ on finite state and action spaces. Together with the structure of weakly communicating MDPs, the following fact follows.

Fact 1 (Fact 1, Lemma 26, (Lee & Ryu, 2026)). In a weakly communicating MDP, $\mathcal{S} = \mathcal{R} \cup \mathcal{T}$, where $\mathcal{R}$ is a single recurrent class and $\mathcal{T}$ is a transient set, and this decomposition is invariant for all policies $\pi \in \Pi_+$. Moreover, $\mathcal{T}$ continues to be a transient set for $\pi \in \Pi$.

Fact 1 implies that, for $\pi \in \Pi_+$, P^π has only a single recurrent class for a weakly communicating MDP, and the recurrent-transient classification is determined by the transition kernel P, not by the particular choice of $\pi \in \Pi_+$. In contrast, a policy $\pi \in \Pi \setminus \Pi_+$, i.e., a policy that assigns zero probability to some actions, may induce a different classification and a multichain structure. But $\mathcal{T}$ continues to be a transient set for a policy $\pi \in \Pi \setminus \Pi_+$ for weakly communicating MDP.

2.2 Escaping from transient states and listing the recurrent states

Consider a behavior policy $\pi_b \in \Pi_+$ and a single trajectory $\{s_t, a_t, r_t\}_{t=0}^\infty$, where $a_t \sim \pi_b(\cdot\,|\,s_t)$, $r_t = r(s_t, a_t)$, and $s_{t+1} \sim P(\cdot\,|\,s_t, a_t)$.

Algorithm 1: RECUR(n_1, n_2)

Input: $n_1, n_2 \in \mathbb{N}$
Initialize: $R = \{\}$
for $t = 0, 1, \ldots, n_1 + n_2$ **do**
  $a_t \sim \pi_b(\cdot \mid s_t),\ s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  if $t > n_1$ **then**
    $R \leftarrow R \cup \{s_{t+1}\}$
  end if
end for
if $R = \{\}$ **then**
  $R \leftarrow \mathcal{S}$
end if
Output: R

Algorithm 2: SAMPLE(n, m, d, R)

Input: $d \in \mathbb{R}^R, n, m \in \mathbb{N}, R \subset \mathcal{S}$
Initialize: $m_{s,a}, D_{s,a} = 0 \quad \forall (s,a) \in R \times \mathcal{A}$
for $t = 0, 1, \ldots, n$ **do**
  $a_t \sim \pi_b(\cdot \mid s_t),\ s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  if $s_t, a_t \in R \times \mathcal{A},\ m_{s_t,a_t} < m$ **then**
    $D_{s_t,a_t} \leftarrow d(s_{t+1}) + D_{s_t,a_t}$
    $m_{s_t,a_t} \leftarrow m_{s_t,a_t} + 1$
  end if
end for
for $(s,a) \in R \times \mathcal{A}$ **do**
  $D_{s,a} \leftarrow \frac{1}{m} D_{s,a}$
end for
Output: D

If the starting state s_0 is transient, the trajectory will reach the recurrent states in finite time by the definition of transient states. To specify this time, we define the hitting time and expected hitting time as

$$\tau_{\text{hit}} = \inf_{t \geq 0} \{s_t \in \mathcal{R}\}, \qquad t_{\text{hit}} = \max_{s \in \mathcal{S}} \mathbb{E}_{\pi_b}[\tau_{\text{hit}} \mid s_0 = s].$$

Then, the following lemma holds by Markov's inequality and the Markov property.

Lemma 2.2. *Consider a single trajectory generated from a weakly communicating MDP under a behavior policy $\pi_b \in \Pi_+$. Let $\delta \in (0,1)$ and $L \geq \lceil e t_{\text{hit}} \rceil \lceil \ln(\frac{1}{\delta}) \rceil$. With probability at least $1 - \delta$, $s_L \in \mathcal{R}$.*

After hitting the recurrent class, the trajectory will remain in the single recurrent class. In a single recurrent class, every recurrent state can be reached in a finite time, and to specify the time for reaching all recurrent states, we define the covering time and expected covering time as

$$\tau_{\text{cov}} = \inf_{t \geq 0} \{\cup_{j=0}^{t-1} \{s_j\} = \mathcal{R}\}, \qquad t_{\text{cov}} = \max_{s \in \mathcal{R}} \mathbb{E}[\tau_{\text{cov}} \mid s_0 = s].$$

Again, by Markov's inequality and the Markov property, the following fact holds.

Fact 2 (Chen et al. (2022); Lee & Ryu (2025a)). Consider a single trajectory generated by a finite irreducible Markov process with a single recurrent class. Let $\delta \in (0,1)$ and $L \geq \lceil e t_{\text{cov}} \rceil \lceil \ln(\frac{1}{\delta}) \rceil$. With probability at least $1 - \delta$, $\bigcup_{j=0}^{L-1} \{s_j\} = \mathcal{R}$.

Based on Lemma 2.2 and Fact 2, Algorithm 1 generates a list of all recurrent states with high probability under an appropriate choice of parameters. As the analysis in the next section shows, the cardinality of the recurrent class gives refined sample complexity. We also note that the expected hitting time and covering time are finite when the state space is finite (Lee & Ryu, 2025c; Levin & Peres, 2017).

2.3 SAMPLING FROM A SINGLE TRAJECTORY

The main technical challenge with single trajectory samples is handling the dependency between samples and related random variables. For a single recurrent class, we can obtain i.i.d. samples from a single trajectory by using stopping times and the strong Markov property as the following fact shows.

Fact 3. Consider a single trajectory generated by a Markov process with a single recurrent class. For a given state $s \in \mathcal{S}$, define the stopping time $\tau_i = \inf\{k > \tau_{i-1} : s_k = s\}$ with $\tau_0 = -1$. Then, for $M \in \mathbb{N}$, $\{s_{\tau_i + 1}\}_{i=1}^{M}$ are i.i.d. and $\tau_M < \infty$ with probability 1.

Building on Fact 3, we introduce our key inequality used in our analysis. To this end, we define $d_{\min}$ as $\min_{s \in \mathcal{R}} d(s)$, where d is the unique stationary distribution of the given transition probability.

Lemma 2.3. *Suppose $\{s_i\}_{i=0}^{\infty}$ is a trajectory generated by a finite irreducible Markov process. For each state $s \in \mathcal{S}$ and $N, M \in \mathbb{N}$, define the stopping time $\tau_i = \inf\{k > \tau_{i-1} \mid s_k = s\}$ with $\tau_0 = -1$,*

and random variable $I_s = \min\{M, \max\{k \mid \tau_k \leq N\}\}$. *Then, for* $\lambda \in \mathbb{R}$ *and* $f : \mathcal{S} \to [a, b]$, *we have*

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{M}\sum_{i=1}^{I_s}(f(s_{\tau_i+1}) - \mathbb{E}[f(s_{\tau_i+1})]\right)\right] \leq \exp\left(\frac{\lambda^2(a-b)^2}{8M}\right).$$

Moreover, if $N \geq \lceil e^2(\ln|\mathcal{S}| + e)(t_{\text{cov}} + (M-1)/d_{\min})\rceil\lceil\ln\left(\frac{1}{\delta}\right)\rceil$, *with probability at least* $1-\delta$, $I_s = M\ \forall s \in \mathcal{S}$.

Based on Lemma 2.3 and the dynamics of a single trajectory in weakly communicating MDPs, we will show that Algorithm 2 successfully approximates the expectation of given quantities with respect to transition probability for a given $R \subset \mathcal{S}$. We note that while prior analyses for average-reward MDPs Wei et al. (2020); Bai et al. (2024) rely on mixing assumptions to address the dependency of random variables related to algorithms, we utilize insights into the structure of weakly communicating MDPs based on recurrent-transient classifications, the strong Markov property, and stopping time arguments to resolve this technical issue.

In the next section, we will combine this machinery with model-free algorithms and establish finite sample complexity from a single trajectory in weakly communicating average-reward MDPs.

3 Stochastic Anchored Value Iteration

In this section, we present a novel value-based method that provides finite sample complexity guarantees from a single trajectory for weakly communicating MDPs, and a prior-knowledge-free variant for communicating MDPs.

Algorithm 3 SAVIC $(n_{\text{esc}}, n_{\text{rec}}, n, \{t_k\}_{k=0}^n, \varepsilon, \delta)$
W.C. = WEAKLY COMMUNICATING, C. = COMMUNICATING

Input: $n_{\text{esc}}, n_{\text{rec}}, n \in \mathbb{N}, \varepsilon > 0, \delta \in (0,1)$
Initialize: $\beta_k = k/(k+2)$, $c_k = 5(k+2)\ln^2(k+2)\ \forall k \in \mathbb{N}$
$R = \text{RECUR}(n_{\text{esc}}, n_{\text{rec}})$, $\eta = \ln(8|R||\mathcal{A}|(n+1)/\delta)$, $Q^0 = \mathbf{0}_{R\times\mathcal{A}}$, $T^{-1} = r_{R\times\mathcal{A}}$, $h^{-1} = \mathbf{0}_R$
for $k = 0, \ldots, n$ **do**
  $Q^k = (1-\beta_k)\,Q^0 + \beta_k\,T^{k-1}$
  $h^k(\cdot) = \max_{a\in\mathcal{A}} Q^k(\cdot, a)$
  $d^k = h^k - h^{k-1}$
  $m_k = \max\{\lceil \eta\, c_k \|d^k\|_{\text{sp}}^2/\varepsilon^2\rceil, 1\}$
  $D^k = \begin{cases}\text{SAMPLE}(t_k, m_k, d^k, R) & \text{w.c.}\\ \text{SAMPLE+}(m_k, d^k) & \text{c.}\end{cases}$
  $T^k = T^{k-1} + D^k$
end for
$\pi^n(s) \in \arg\max_{a\in\mathcal{A}} Q^n(s,a) \quad (\forall s \in R)$
$C_{\text{pur}} = \begin{cases}1/((|\mathcal{A}|+1)(\|Q^n\|_\infty + 1)) & \text{w.c.}\\ 0 & \text{c.}\end{cases}$
$\pi_\varepsilon^n(a\,|\,s) = \begin{cases}\dfrac{\varepsilon C_{\text{pur}} + \pi^n(a\,|\,s)}{1 + |\mathcal{A}|\varepsilon C_{\text{pur}}} & (\forall s \in R),\\ 1/|\mathcal{A}| & (\forall s \in R^{\complement})\end{cases}$
Output: $(Q^n, T^n, \pi_\varepsilon^n)$

3.1 Framework overview

Our novel value-based method, SAVIC, is divided into three stages. In the first stage, as discussed in the previous section, the agent escapes from transient states and collects the full list of recurrent states through Algorithm 1. Based on R obtained from RECUR, we focus on R instead of $\mathcal{S}$, and we initialize parameters as described in Algorithm 3, where $r_{R\times\mathcal{A}} \in \mathbb{R}^{R\times\mathcal{A}}$ is defined by $r_R(s,a) = r(s,a)\ \forall (s,a) \in R \times \mathcal{A}$ and $\mathbf{0}_R$ and $\mathbf{0}_{R\times\mathcal{A}}$ denote $|R|$- and $|R\times\mathcal{A}|$- dimensional zero vectors, respectively. Note that in the previous section, we showed that with high probability, $R = \mathcal{R}$.

In the next stage, during the first for-loop of Algorithm 3, we utilize an anchoring mechanism and a variance-reduction technique to obtain a near-optimal policy for the recurrent class with efficient sample complexity. The anchoring mechanism, also classically known as the Halpern iteration Halpern (1967), has been widely studied in minimax optimization and fixed-point problems Sabach & Shtern (2017); Lieder (2021); Yoon & Ryu (2021); Park & Ryu (2022); Bravo et al. (2026). Recently, Anchored Value Iteration has been studied to obtain finite-time bounds on policy error for average-reward MDPs Bravo & Contreras (2024); Lee & Ryu (2025a); Lee et al. (2025); Lee & Ryu (2025b); Zurek & Chen (2025), and in particular, *Anchored Q-Value Iteration* is

$$Q^k = (1-\beta_k)Q^0 + \beta_k \mathbf{T}Q^{k-1} \qquad \text{for } k = 1, 2, \dots. \tag{Anc-QI}$$

where the parameter β_k is chosen appropriately. Compared to standard Value Iteration, Anc-QI obtains the next iterate as a convex combination of the output of $\mathbf{T}$ and the starting point Q^0. We call the $(1-\beta_k)Q^0$ term the *anchor term*, since it pulls the iterates back toward the starting point Q^0. With this Anc-QI, Sabach & Shtern (2017); Lee & Ryu (2025a) establish non-asymptotic convergence in the average-reward setup. In weakly communicating MDPs, Anc-QI with $\beta_k = k/(k+2)$ exhibits the following convergence rate:

$$\|J^\star - J^{\pi_k}\|_\infty \le \|Q^k - \mathbf{T}(Q^k)\|_{\text{sp}} \le \tfrac{4}{k+1}\|Q^0 - Q^\star\|_{\text{sp}},$$

where $Q^\star$ is the optimal state-action value function, and π_k is the greedy policy induced by $\mathbf{T}(Q^k)$.

Adding to this anchoring mechanism, we utilize a *recursive sampling* technique borrowed from Jin et al. (2024b); Lee et al. (2025) to obtain the efficient sample complexity. The basic idea is to exploit the previous approximation $T^{k-1} \approx \mathbf{T}(Q^{k-1})$ to approximate $\mathbf{T}(Q^k)$ by estimating the difference $\mathbf{T}(Q^k) - \mathbf{T}(Q^{k-1}) = Pd^k$, where $d^k = h^k - h^{k-1}$, and adding it to T^{k-1}. Then, denoting by $D^k \approx Pd^k$ the matrix sampled at the k-th step, we have $T^k = r + \sum_{i=0}^k D^i$ so that all previous estimates $\{D^i\}_{i=0}^k$ are used to approximate $\mathbf{T}(Q^k)$. In the last stage, during the second for-loop of Algorithm 3, we perturb the policy to obtain a complete ε-optimal policy for the whole state space. Note that in the second stage, we only utilized samples from the single recurrent class. In other words, there is no guarantee that a near-optimal policy on the recurrent class is also near-optimal on the whole state space. To address this issue, we introduce the following lemma proved by utilizing the structure of weakly communicating MDPs and the fact that our Anc-QI simultaneously reduces Bellman error and policy error.

Lemma 3.1. *Consider a weakly communicating MDP. Let $Q \in \mathbb{R}^{\mathcal{R}\times\mathcal{A}}$ and π be a policy satisfying $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a) \,\forall s \in \mathcal{R}$. For $\epsilon > 0$, if policy $\pi_\epsilon \in \Pi_+$ satisfies $\max_{s\in\mathcal{R}} \|\pi(\cdot \mid s) - \pi_\epsilon(\cdot \mid s)\|_1 \le \epsilon$, we have $\|J^\star - J^{\pi_\epsilon}\|_\infty \le \|\mathbf{T}Q - Q\|_{\text{sp}} + \epsilon \|Q\|_\infty$.*

Intuitively speaking, this lemma ensures that a smaller Bellman error on the recurrent class induces a small policy error over all states if we properly choose the perturbation in weakly communicating MDP. Note that $J^\star, J^{\pi_\epsilon} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

Through these three stages, as shown in the next subsection, SAVIC provides an ε-optimal policy with efficient sample complexity from a single trajectory in weakly communicating average-reward MDPs. We lastly note that the framework of the second stage basically follows Lee et al. (2025), which analyzes sample complexity under the generative model.

3.2 Finite sample complexity for weakly communicating MDPs

To establish the sample complexity of SAVIC, with high probability, we first ensure that our sampling process provides a good approximation $T^k \approx \mathbf{T}(Q^k)$ for all k. To this end, we define $d_{\min}^{\pi_b}$ as $\min_{s,a\in\mathcal{R}\times\mathcal{A}} d^{\pi_b}(s,a)$, where d^{π_b} is the unique stationary distribution of P^{π_b}, and define t_{cov} and t_{hit} with respect to P^{π_b} and its recurrent set, $\mathcal{R}\times\mathcal{A}$. We note that lemmas and facts in Section 2 continue to hold with P^{π_b} as addressed in Appendix C.2.

Proposition 3.2. *Consider a weakly communicating MDP. Let $n_{\text{esc}} \ge \lceil et_{\text{hit}}\rceil\lceil\ln(\frac{4}{\delta})\rceil$, $n_{\text{rec}} \ge \lceil et_{\text{cov}}\rceil\lceil\ln(\frac{4}{\delta})\rceil$, $t_k \ge \left\lceil e^2(\ln(|R||\mathcal{A}|)+e)\left(t_{\text{cov}} + \frac{m_k - 1}{d_{\min}^{\pi_b}}\right)\right\rceil \lceil\ln(\frac{4(n+1)}{\delta})\rceil$, and let T^k, Q^k be the iterates generated by* $\text{SAVIC}(n_{\text{esc}}, n_{\text{rec}}, n, \{t_k\}_{k=0}^n, \varepsilon, \delta)$. *Then, with probability at least $1-\delta$, we have $\left\|T^k - \mathbf{T}(Q^k)\right\|_\infty \le \varepsilon$ for all $k = 0, \dots, n$.*

We defer the proof to the Appendix C, but briefly mention that the proof is an adaptation of the Azuma-Hoeffding inequality combined with the sampling arguments in Section 2 and t_k is chosen so that trajectories are rolled out until every recurrent state-action is visited m_k times with high probability.

We now present the sample complexity bound for SAVIC.

Theorem 3.3. *Consider a weakly communicating MDP. Let* $n_{\mathrm{esc}} = \lceil et_{\mathrm{hit}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n_{\mathrm{rec}} = \lceil et_{\mathrm{cov}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n = \lceil 32 \|Q^\star\|_{\mathrm{sp}} / \varepsilon \rceil$, *and* $t_k = \left\lceil e^2 (\ln(|R||\mathcal{A}|) + e) \left(t_{\mathrm{cov}} + \frac{m_k - 1}{d_{\min}^{\pi_b}} \right) \right\rceil \lceil \ln(\frac{4(n+1)}{\delta}) \rceil$. *Let* $(Q^n, T^n, \pi_\varepsilon^n)$ *be the output computed by SAVIC*$(n_{\mathrm{esc}}, n_{\mathrm{rec}}, n, \{t_k\}_{k=0}^n, \varepsilon/16, \delta)$ *with* $\varepsilon \leq 1 \leq \|Q^\star\|_{\mathrm{sp}}$. *Then, with probability at least* $1 - \delta$, *we have* $\left\| J^\star - J^{\pi_\varepsilon^n} \right\|_\infty \leq \varepsilon$ *with sample complexity*

$$\widetilde{O}\left(t_{\mathrm{hit}} + t_{\mathrm{cov}} \|Q^\star\|_{\mathrm{sp}} / \varepsilon + 1 / d_{\min}^{\pi_b} \|Q^\star\|_{\mathrm{sp}}^2 / \varepsilon^2 \right),$$

where $\widetilde{O}(\cdot)$ *hides logarithmic factors including* $\log(1/\delta)$.

The proof exploits Proposition 3.2 and anchoring mechanism to establish an upper bound of the Bellman residual error on the recurrent class with efficient sample complexity. We also utilize Lemma 3.1 to obtain a complete ϵ-optimal policy for all states. To the best of our knowledge, this is the first finite sample complexity result from a single trajectory in weakly communicating MDPs.

3.3 Without prior knowledge for communicating MDPs

One issue with Theorem 3.3 is that knowledge of constants such as t_{cov}, t_{hit}, $d_{\min}^{\pi_b}$, and $\|Q^\star\|_{\mathrm{sp}}$ is required to set parameters for obtaining an ε-optimal policy. We can alleviate this issue by restricting the MDP to be a communicating MDP.

Fact 4. In a communicating MDP, $\mathcal{S} = \mathcal{R}$, where $\mathcal{R}$ is a single recurrent class, and this decomposition is invariant for all policies $\pi \in \Pi_+$.

In a communicating MDP, as Fact 4 shows, every state is recurrent for any $\pi \in \Pi_+$. Thus, the first and third stages from our previous framework can be omitted. Also, we replace Algorithm 2 with Algorithm 6, which automatically stops when the adapted number of samples has been collected. Then, following the framework of (Lee et al., 2025), we apply a doubling trick and an explicit stopping rule based on the empirical Bellman residual error $\|Q^{n_i} - T^{n_i}\|_{\mathrm{sp}}$ as described in Algorithm 7. (We note that the omitted algorithms are presented in Appendix E.)

The following theorem shows that SAVIC+ does not require any prior knowledge to obtain an ε-optimal policy.

Theorem 3.4. *Consider a communicating MDP. Let* (Q^N, T^N, π^N) *be the output of* SAVIC+$(\varepsilon/16, \delta/2)$ *with* $\varepsilon \leq 1 \leq \|Q^\star\|_{\mathrm{sp}}$. *Then, with probability at least* $1 - \delta$, *we have*

$$\|J^\star - J^{\pi_N}\|_\infty \leq \|Q^N - \mathbf{T}(Q^N)\|_{\mathrm{sp}} \leq \varepsilon,$$

with sample complexity

$$\widetilde{O}\left(t_{\mathrm{cov}} \|Q^\star\|_{\mathrm{sp}} / \varepsilon + 1 / d_{\min}^{\pi_b} \|Q^\star\|_{\mathrm{sp}}^2 / \varepsilon^2 \right),$$

where $\widetilde{O}(\cdot)$ *hides logarithmic factors including* $\log(1/\delta)$.

To the best of our knowledge, SAVIC+ is the first model-free method that guarantees finite sample complexity from a single trajectory in communicating MDPs.

4 Stochastic α-clipped Policy Mirror Ascent

In this section, we introduce policy gradient methods that guarantee finite sample complexity from a single trajectory in weakly communicating MDPs.

4.1 Framework overview

Our novel policy-based method, SCPMA, consists of two stages. During the first stage, the agent escapes from transient states and collects the full list of recurrent states through Algorithm 1. Then,

Algorithm 4 SCPMA$(\alpha, n_{\mathrm{esc}}, n_{\mathrm{rec}}, K, \pi_0, \{n_k\}_{k=0}^K, \varepsilon, \delta)$
c.s. = constant step size, a.s. = adaptive step size

Input: $\alpha \in (0, 1/|\mathcal{A}|), n_{\mathrm{esc}}, n_{\mathrm{rec}}, K \in \mathbb{N}, \pi_0 \in \Pi_\alpha, \{n_k\}_{k=0}^K \subset \mathbb{N}^{K+1}, \varepsilon > 0, \delta \in (0,1)$
$R = \mathrm{RECUR}(n_{\mathrm{esc}}, n_{\mathrm{rec}})$
for $k = 0, 1, \ldots, K-1$ **do**
  $\eta_k = \ln(8K|R||\mathcal{A}|(n_k+1)/\delta)$
  $\varepsilon' = \begin{cases} \varepsilon/(16(2C_\alpha + K + 2)) & \text{c.s.} \\ \varepsilon/(64C_\alpha) & \text{a.s.} \end{cases}$
  $\hat{Q}^{\pi_k} = \mathrm{SAVIPE}(\pi_k, n_k, \varepsilon'/2, \delta/(2K+2), \eta_k, R)$
  for $s \in R$ **do**
    $\pi_{k+1}(\cdot \,|\, s) = \underset{p \in \mathcal{M}_\alpha(\mathcal{A})}{\arg\max} \left\{ b_k \sum_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(s,a) p(a) - D\big(p(\cdot), \pi_k(\cdot \,|\, s)\big) \right\}$
  end for
end for
$\eta_K = \ln(8K|R||\mathcal{A}|(n_K+1)/\delta)$
$\hat{Q}^{\pi_K} = \mathrm{SAVIPE}(\pi_K, n_K, \varepsilon'/2, \delta/(2K+2), \eta_K, R)$
$\pi(s) \in \arg\max_{a \in \mathcal{A}} \hat{Q}^{\pi_K}(s,a) \quad (\forall s \in R)$
$\pi(a \,|\, s) = 1/|\mathcal{A}| \quad (\forall s \in R^{\complement})$
Output: π

during the second stage, in the for-loop of Algorithm 4, we alternately perform policy evaluation and policy improvement at every iteration.

For the policy evaluation step, we utilize our novel policy evaluation algorithm, SAVIPE, which approximates the state-action value function given a policy π with an anchoring mechanism and a variance-reduction technique. Compared with SAVIC, SAVIPE uses a different coefficient $\beta_k = \frac{k}{k+1}$ and the estimator $\frac{1}{n+1}\sum_{k=0}^{n-1} T^k$ to efficiently approximate the state-action value function $\hat{Q}^{\pi_k} \approx Q_{\mathcal{R}}^{\pi_k}$ (see Appendix D) as described in Algorithm 5. With the approximated state-action value function $\hat{Q}^{\pi_k}$, we perform policy improvement following the policy mirror ascent framework Shani et al. (2020); Xiao (2022). Policy mirror ascent utilizes a Bregman divergence and the state-action value function to update the policy effectively, as formulated in the preliminaries section. Specifically, we apply α-clipping to policy mirror ascent motivated by (Lee & Ryu, 2025a), and to this end, we define

$$\Pi_\alpha = \{\pi \,|\, \pi(a \,|\, s) \geq \alpha \text{ for all } s \in \mathcal{S},\, a \in \mathcal{A}\}, \qquad \mathcal{M}_\alpha(\mathcal{A}) = \{p \,|\, p(a) \geq \alpha \text{ for all } a \in \mathcal{A}\}$$

with $\alpha \in (0, 1/|\mathcal{A}|), p \in \mathcal{M}(\mathcal{A})$, and

$$\pi_\alpha = \underset{\pi \in \Pi_\alpha}{\arg\max}\, J^\pi.$$

By a slight abuse of notation, let d^π denote the unique stationary distribution of the state-transition matrix P_s^π (see Appendix C.2). On Π_α, we can ensure the finiteness of the coefficients appearing in the sample-complexity results of Theorem 4.2 as the following fact shows.

Fact 5 (Lemma 3, Lee & Ryu (2025a)). For $\alpha \in (0, \frac{1}{|\mathcal{A}|})$,

$$\max_{\pi \in \Pi_\alpha} \|Q^\pi\|_\infty := Q_\alpha < \infty,$$

$$\max_{s \in \mathcal{R}, \pi, \pi' \in \Pi_\alpha} \left\| \frac{d^\pi(s)}{d^{\pi'}(s)} \right\|_\infty := C_\alpha < \infty.$$

Algorithm 5: SAVIPE$(\pi, n, \varepsilon, \delta, \eta, R)$

Input: $\pi \in \Pi_+, n \in \mathbb{N}, \varepsilon > 0, \delta \in (0,1), \eta \in \mathbb{R}, R \subset \mathcal{S}$
Initialize: $Q^0 = \mathbf{0}_{R \times \mathcal{A}}, T^{-1} = r_{R \times \mathcal{A}}, h^{-1} = \mathbf{0}_R, \beta_k = k/(k+1), c_k = 5(k+2)\ln^2(k+2) \forall\, k \in \mathbb{N}$
for $k = 0, \ldots, n-1$ **do**
  $h^k(s) = Q^k(s, \pi(\cdot \,|\, s)) \quad (\forall s \in R)$
  $d^k = h^k - h^{k-1}$
  $m_k = \max\{\lceil \eta\, n^2 c_k \|d^k\|_{\mathrm{sp}}^2 / \varepsilon^2 \rceil, 1\}$
  $t_k = \left\lceil e^2 (\ln|R||\mathcal{A}| + e) \left(t_{\mathrm{cov}} + \frac{m_k - 1}{d_{\min}^{\pi_b}} \right) \right\rceil \lceil \ln(\frac{4n}{\delta}) \rceil$
  $D^k = \mathrm{SAMPLE}(t_k, m_k, d^k, R)$
  $T^k = T^{k-1} + D^k$
  $Q^{k+1} = (1 - \beta_{k+1})\, Q^0 + \beta_{k+1}\, T^k$
end for
Output: Q^n

We note that our SCPMA can be viewed as an extension of α-clipped policy mirror ascent (Lee & Ryu, 2025a) to the single-trajectory setting, incorporating a variance-reduced policy evaluation algorithm. In contrast, (Lee & Ryu, 2025a) considers the generative model setting and employs a naive policy evaluation algorithm.

4.2 FINITE SAMPLE COMPLEXITY FOR WEAKLY COMMUNICATING MDPS

To establish the sample complexity of SCPMA, we first ensure that with high probability SAVIPE provides a good approximation of the state-action value function. To this end, we define coefficients related to the target time in a weakly communicating MDP for $\pi \in \Pi_+$ and a recurrent class $\mathcal{R}$.

$$t_{s,a} := \inf\{t \geq 0 \,|\, s_t, a_t = s, a\}, \;\; t^\pi_{\mathrm{tar}} := \sum_{(s,a),(s',a') \in \mathcal{R}\times\mathcal{A}} d^\pi(s',a')\, \mathbb{E}_\pi[t_{s',a'} \,|\, s_0, a_0 = s, a], \;\; t_\alpha = \max_{\pi\in\Pi_\alpha} t^\pi_{\mathrm{tar}}.$$

We note that these coefficients are always finite when the state and action spaces are finite (Levin & Peres, 2017; Lee & Ryu, 2025a).

Proposition 4.1. *Consider a weakly communicating MDP. Let* $n_{\mathrm{esc}} \geq \lceil et_{\mathrm{hit}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n_{\mathrm{rec}} \geq \lceil et_{\mathrm{cov}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n_k \geq \left\lceil \frac{8t^{\pi_k}_{\mathrm{tar}} \|Q^{\pi_k}\|_{\mathrm{sp}}}{\varepsilon'} \right\rceil$, *and* $\hat{Q}^{\pi_k}$ *be the iterates generated by SCPMA*$(\alpha, n_{\mathrm{esc}}, n_{\mathrm{rec}}, K, \pi_0, \{n_k\}_{k=0}^K, \varepsilon, \delta)$. *Then, with probability at least* $1-\delta$, $\|\hat{Q}^{\pi_k} - Q^{\pi_k}_{\mathcal{R}}\|_{\mathrm{sp}} \leq \varepsilon'$ *for all* $k = 0, \ldots, K$.

As clarified in the preliminaries section, policy gradient in weakly communicating MDPs requires the state-action value function only on recurrent states. Next, to choose the specific α, we define

$$J^\pi_{\mathcal{R}_i} = \sum_{(s,a)\in\mathcal{R}_i\times\mathcal{A}} P^\pi_\star((s,a),(s,a)) r(s,a), \qquad G = \min_{\{\pi\in D, \mathcal{R}_i\in\mathcal{R}_\pi \,|\, J^\star > J^\pi_{\mathcal{R}_i}\}} J^\star - J^\pi_{\mathcal{R}_i},$$

where $\mathcal{R}_\pi$ is the set of recurrent states induced by P^π, $\mathcal{R}_i$ is a single recurrent class included in $\mathcal{R}_\pi$, and D is the set of all deterministic policies. We consider the case where the optimizing set in the definition of G is nonempty, so that $G > 0$ because if the set is empty, every deterministic policy is optimal. Also, assume $C_\alpha > 1$ to avoid trivial MDP. Then, based on Proposition 4.1 and the convergence analysis of inexact α-clipped policy mirror ascent in Lee & Ryu (2025a), we establish the complexity bound for SCPMA with constant and adaptive step sizes.

Theorem 4.2. *Consider a weakly communicating MDP. Let* $\alpha = \min\left\{\frac{1}{2|\mathcal{A}|}, \frac{G}{16|\mathcal{A}|(\|Q^{\pi^\star}\|_\infty + 1)}\right\}$, $n_{\mathrm{esc}} = \lceil et_{\mathrm{hit}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n_{\mathrm{rec}} = \lceil et_{\mathrm{cov}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n_k = \lceil 8t^{\pi_k}_{\mathrm{tar}} \|Q^{\pi_k}\|_{\mathrm{sp}}/\varepsilon' \rceil$, $K = \left\lceil 16\left(\frac{D_{d^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{b} + C_\alpha \|J^{\pi_\alpha} - J^{\pi_0}\|_\infty\right)/\varepsilon \right\rceil$. *Then, for* $\varepsilon \leq 1 \leq \|Q^\star\|_{\mathrm{sp}}$ *and any* $\pi_0 \in \Pi_\alpha$, *with probability at least* $1-\delta$, *the iterates generated by SCPMA*$(\alpha, n_{\mathrm{esc}}, n_{\mathrm{rec}}, K, \pi_0, \{n_k\}_{k=0}^K, \varepsilon, \delta)$ *with constant step size* $b > 0$ *generate an* ε*-optimal policy with sample complexity*

$$\widetilde{O}\left(t_{\mathrm{hit}} + g t_{\mathrm{cov}} t_\alpha C^2_\alpha Q^2_\alpha/\varepsilon^3 + (t_{\mathrm{cov}} t_\alpha g^3 Q^2_\alpha + 1/d^{\pi_b}_{\min} g C^4_\alpha t^2_\alpha Q^4_\alpha)/\varepsilon^5 + 1/d^{\pi_b}_{\min} t^2_\alpha Q^4_\alpha g^5/\varepsilon^9\right)$$

where $g = C_\alpha \|J^{\pi_\alpha} - J^{\pi_0}\|_\infty + \frac{D_{d^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{b}$. *With adaptive step sizes satisfying* $b_{k+1}(C_\alpha - 1) \geq b_k C_\alpha > 0$ *and* $K = \left\lceil \log\left(16(\|J^{\pi_\alpha} - J^{\pi_0}\|_\infty + \frac{D_{d^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{b_0(C_\alpha-1)})/\varepsilon\right) / \log(C_\alpha/(C_\alpha-1)) \right\rceil$ *with probability at least* $1-\delta$, *the iterates generate an* ε*-optimal policy with sample complexity*

$$\widetilde{O}\left(t_{\mathrm{hit}} + K(t_\alpha t_{\mathrm{cov}} Q^2_\alpha C^2_\alpha/\varepsilon^2 + 1/d^{\pi_b}_{\min} t^2_\alpha Q^4_\alpha C^4_\alpha/\varepsilon^4)\right),$$

where $\widetilde{O}(\cdot)$ *hides logarithmic factors including* $\log(1/\delta)$.

Note that the α-related terms $C_\alpha, Q_\alpha, t_\alpha$ in the sample complexity are independent of ε. Although the adaptive step size yields a more efficient sample complexity, it requires knowledge of C_α to set the step sizes. We lastly clarify that to the best of our knowledge, Theorem 4.2 is the first finite sample complexity guarantee for policy gradient methods from a single trajectory in weakly communicating MDPs.

5 CONCLUSION

In this work, we propose novel model-free methods that achieve the first finite sample complexity from a single trajectory for weakly communicating average-reward MDPs. For communicating MDPs, we further introduce the first model-free method that requires no prior knowledge of problem-dependent quantities. One future research direction is to close the sample-complexity gap between the value-based and policy-based methods by using a refined variance-reduction technique. Another future direction is to extend our results to parameterized function-approximation settings, where handling representation error and stability under single trajectory sampling poses additional challenges.

REFERENCES

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes. *International Conference on Artificial Intelligence*, 2024.

Jonathan Baxter and Peter L Bartlett. Direct gradient-based reinforcement learning. *IEEE International Symposium on Circuits and Systems*, 3:271–274, 2000.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5): 679–684, 1957.

Abraham Berman and Robert J Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM, 1994.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33: 719–726, 1962.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward MDPs. *Neural Information Processing Systems*, 2024.

Hippolyte Bourel, Anders Jonsson, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Exploration in reward machines with low regret. *International Conference on Artificial Intelligence and Statistics*, 2023.

Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

Mario Bravo and Juan Pablo Contreras. Stochastic Halpern iteration in normed spaces and applications to reinforcement learning. *arXiv preprint arXiv:2403.12338*, 2024.

Mario Bravo, Roberto Cominetti, and Jongmin Lee. Minimax-optimal halpern iterations for lipschitz maps. *arXiv preprint arXiv:2601.15996*, 2026.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Siu On Chan, Qinghua Ding, and Sing Hei Li. Learning and testing irreducible markov chains via the k-cover time. *Algorithmic Learning Theory*, 2021.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pp. 3246–3270. PMLR, 2022.

Zaiwei Chen. Non-asymptotic guarantees for average-reward Q-learning with adaptive stepsizes. *Neural Information Processing Systems*, 2025.

Eugenio Della Vecchia, Silvia Di Marco, and Alain Jean-Marie. Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs. *Annals of Operations Research*, 199:193–214, 2012.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

Awi Federgruen, Paul J Schweitzer, and Hendrik Cornelis Tijms. Contraction mappings underlying undiscounted Markov decision problems. *Journal of Mathematical Analysis and Applications*, 65 (3):711–730, 1978.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in MDPs with options without prior knowledge. *Neural Information Processing Systems*, 2017.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. Offline primal-dual reinforcement learning for linear MDPs. *International Conference on Artificial Intelligence and Statistics*, 2024.

Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain MDPs via an actor-critic approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal regret with novel policy gradient approaches in infinite-horizon average reward MDPs. *International Conference on Artificial Intelligence and Statistics*, 2025a.

Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. In *Forty-second International Conference on Machine Learning*, 2025b.

Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

Ronald A Howard. *Dynamic Programming and Markov Processes.* John Wiley and Sons, 1960.

Gerhard Hübner. Improved procedures for eliminating suboptimal actions in markov programming by the use of contraction properties. *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 1977.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Yuchen Jiao, Jiin Woo, Gen Li, Gauri Joshi, and Yuejie Chi. Sample complexity of average-reward q-learning: From single-agent to federated reinforcement learning. *arXiv preprint arXiv:2601.13642*, 2026.

Ying Jin, Ramki Gummadi, Zhengyuan Zhou, and Jose Blanchet. Feasible Q-learning for average reward reinforcement learning. *International Conference on Artificial Intelligence and Statistics*, 2024a.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward MDPs. *International Conference on Machine Learning*, 2021.

Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. *Neural Information Processing Systems*, 2024b.

Sham M Kakade. A natural policy gradient. *Neural Information Processing Systems*, 2001.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Neural Information Processing Systems*, 1999.

Navdeep Kumar, Yashaswini Murthy, Itai Shufaro, Kfir Y Levy, R Srikant, and Shie Mannor. On the global convergence of policy gradient in average reward Markov decision processes. *International Conference on Learning Representations*, 2024a.

Navdeep Kumar, Kaixin Wang, Kfir Yehuda Levy, and Shie Mannor. Efficient value iteration for s-rectangular robust Markov decision processes. *International Conference on Machine Learning*, 2024b.

Jongmin Lee and Ernest Ryu. Optimal non-asymptotic rates of value iteration for average-reward MDPs. *International Conference on Learning Representations*, 2025a.

Jongmin Lee and Ernest K Ryu. Finite-time bounds for average-reward fitted Q-iteration. *Neural Information Processing Systems*, 2025b.

Jongmin Lee and Ernest K Ryu. Why policy gradient algorithms work for undiscounted total-reward MDPs. *arXiv preprint arXiv:2510.18340*, 2025c.

Jongmin Lee and Ernest K Ryu. Policy gradient algorithms in average-reward multichain MDPs. *arXiv preprint arXiv:2602.18003*, 2026.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. *International Conference on Machine Learning*, 2025.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research*, 50(4):3125–3160, 2024.

Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1):159–195, 1996.

Peter Marbach and John N Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *International Conference on Machine Learning*, 2020.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, and et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Washim U Mondal and Vaneet Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 3097–3105. PMLR, 2024.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Yashaswini Murthy and R Srikant. On the convergence of natural policy gradient and mirror descent-like policy methods for average-reward MDPs. *IEEE Conference on Decision and Control*, 2023.

Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Offline reinforcement learning via linear-programming with error-bound induced constraints. *arXiv preprint arXiv:2212.13861*, 2024.

Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. *International Conference on Machine Learning*, 2022.

Bhrij Patel, Wesley A Suttle, Alec Koppel, Vaneet Aggarwal, Brian M. Sadler, Dinesh Manocha, and Amrit Bedi. Towards global optimality for practical average reward reinforcement learning without mixing time oracles. In *Forty-first International Conference on Machine Learning*, 2024.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 2nd edition, 2014.

Gareth O Roberts and Jeffrey S Rosenthal. Shift-coupling and convergence rates of ergodic averages. *Stochastic Models*, 13(1):147–165, 1997.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored MDPs with unknown structure. *Neural Information Processing Systems*, 2021.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. *International Conference on Machine Learning*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Paul J Schweitzer and Awi Federgruen. The asymptotic behavior of undiscounted value iteration in Markov decision problems. *Mathematics of Operations Research*, 2(4):360–381, 1977.

Paul J Schweitzer and Awi Federgruen. Geometric convergence of value-iteration in multichain Markov decision problems. *Advances in Applied Probability*, 11(1):188–217, 1979.

E Seneta. *Non-Negative Matrices and Markov Chains*. Springer Science & Business Media, 3th edition, 2006.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *AAAI Conference on Artificial Intelligence*, 2020.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Naval Research Logistics*, 70(5):423–442, 2023.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3: 9–44, 1988.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems*, 1999.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward Markov decision processes without prior knowledge. *Neural Information Processing Systems*, 2024.

Johannes Van Der Wal. *Stochastic Dynamic Programming: Successive Approximations and Nearly Optimal Strategies for Markov Decision Processes and Markov games*. The Mathematical Centre, Amsterdam, 2nd edition, 1981.

Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

Yi Wan, Huizhen Yu, and Richard S Sutton. On convergence of average-reward Q-learning in weakly communicating Markov decision processes. *arXiv preprint arXiv:2408.16262*, 2024.

Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv:1710.06100*, 2017.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. *International Conference on Machine Learning*, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *International Conference on Machine Learning*, 2021.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Neural Information Processing Systems*, 2019.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. *Conference on Learning Theory*, 2023.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. *Neural Information Processing Systems*, 2024.

Matthew Zurek and Yudong Chen. Faster fixed-point methods for multichain MDPs. *Neural Information Processing Systems*, 2025.

Matthew Zurek, Guy Zamir, and Yudong Chen. Optimal single-policy sample complexity and transient coverage for average-reward offline rl. *Neural Information Processing Systems*, 2025.

A PRIOR WORKS

Average-reward MDPs. The average-reward MDP setup was first introduced in the dynamic programming literature by Howard (1960), and Blackwell (1962) established a theoretical framework for its analysis. In reinforcement learning (RL), average-reward MDPs have mainly been considered in the sample-based setting, where the transition matrix is unknown (Mahadevan, 1996). For this sampling setup, various methods have been proposed, including model-based methods (Jin & Sidford, 2021; Tuynman et al., 2024; Zurek & Chen, 2024), value-based methods (Wan et al., 2024; Bravo & Cominetti, 2024; Chen, 2025; Jiao et al., 2026), and policy gradient methods (Bai et al., 2024; Murthy & Srikant, 2023; Kumar et al., 2024a). On the theoretical side, sample complexity analyses under generative model access (Wang, 2017; Zhang & Xie, 2023; Li et al., 2024; Jin et al., 2024a; Lee et al., 2025) and offline datasets (Ozdaglar et al., 2024; Gabbianelli et al., 2024; Zurek et al., 2025; Lee & Ryu, 2025b), as well as regret minimization frameworks (Burnetas & Katehakis, 1997; Jaksch et al., 2010; Zhang & Ji, 2019; Boone & Zhang, 2024), have been actively studied.

Value Iteration Value iteration (VI) was first introduced in the dynamic programming literature Bellman (1957) and serves as a fundamental algorithm based on computing the value functions. The sample-based variants, such as TD-Learning Sutton (1988), Fitted Value Iteration Ernst et al. (2005); Munos & Szepesvári (2008), and Deep Q-Network Mnih et al. (2015) are the workhorses of modern reinforcement learning algorithms Bertsekas & Tsitsiklis (1996); Sutton & Barto (2018); Szepesvári (2010), and are routinely applied in diverse settings, including factored MDPs Rosenberg & Mansour (2021), robust MDPs Kumar et al. (2024b), MDPs with reward machines Bourel et al. (2023), MDPs with options Fruit et al. (2017), and a generative model Wainwright (2019); Sidford et al. (2023); Lee et al. (2025). For the average-reward MDP, the convergence of VI in average-reward MDPs also has been extensively studied by (Seneta, 2006; Hübner, 1977; Federgruen et al., 1978; Van Der Wal, 1981; Della Vecchia et al., 2012; Schweitzer & Federgruen, 1977; 1979).

Policy gradient methods. Policy gradient methods (Williams, 1992; Sutton et al., 1999; Konda & Tsitsiklis, 1999; Kakade, 2001) are foundational reinforcement learning algorithms, commonly implemented with deep neural networks for policy parameterization (Schulman et al., 2015; 2017). In line with their practical success, convergence and sample complexity of policy gradient variants have been extensively studied across settings (Shani et al., 2020; Mei et al., 2020; Agarwal et al., 2021; Cen et al., 2022; Xiao, 2022; Bhandari & Russo, 2024; Mondal & Aggarwal, 2024). For the average-reward MDP, Sutton et al. (1999); Marbach & Tsitsiklis (2001); Baxter & Bartlett (2000) establish the policy gradient theorem in unichain MDPs, and recently, Lee & Ryu (2025a) establishes the policy gradient theorem for weakly communicating and multichain MDPs. The convergence of variants of policy gradient methods including natural policy gradient, projected policy gradient, and policy mirror ascent has been studied by Murthy & Srikant (2023); Kumar et al. (2024a); Lee & Ryu (2025a), as well as in parameterized settings Bai et al. (2024); Patel et al. (2024); Ganesh et al. (2025a;b); Ganesh & Aggarwal (2025).

B OMITTED PROOFS IN SECTION 2

B.1 PROOF OF LEMMA 2.2

Proof. First, by Markov inequality, $P(\tau_{\text{hit}} \geq \lceil et_{\text{hit}} \rceil) \leq e^{-1}$. By induction, we will show $P(\tau_{\text{hit}} \geq m\lceil et_{\text{hit}} \rceil) \leq e^{-m}$ for $m \in \mathbb{N}$.

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\tau_{\text{hit}} \geq m\lceil et_{\text{hit}} \rceil}] &= \mathbb{E}[\mathbf{1}_{\tau_{\text{hit}} \geq (m-1)\lceil et_{\text{hit}} \rceil} \mathbb{E}[\mathbf{1}_{\tau_{\text{hit}} \geq m\lceil et_{\text{hit}} \rceil} \mid \mathcal{F}_{(m-1)\lceil et_{\text{hit}} \rceil}]] \\
&= \mathbb{E}[\mathbf{1}_{\tau_{\text{hit}} \geq (m-1)\lceil et_{\text{hit}} \rceil} P(\tau_{\text{hit}} \geq \lceil et_{\text{hit}} \rceil \mid s_0)] \\
&\leq \mathbb{E}[\mathbf{1}_{\tau_{\text{hit}} \geq (m-1)\lceil et_{\text{hit}} \rceil}] e^{-1} \\
&\leq e^{-m}
\end{aligned}$$

where the second equality is from the Markov property and the definition of the hitting time, and the second inequality is from induction. □

B.2 Proof of Lemma 2.3

For a Markov chain, define $N_s(t)$ as the counting measure of state s appearing in the trajectory until t. Also, define the k cover time and expected k cover time as

$$\tau_{\text{cov}}^k = \inf\{t \,|\, N_s(t) \geq k \,\forall s \in \mathcal{S}\}, \qquad t_{\text{cov}}^k = \max_{s \in \mathcal{S}} \mathbb{E}[\tau_{\text{cov}}^k \,|\, s_0 = s].$$

Then, the following fact holds.

Fact 6 (Chan et al. (2021)). Consider a single trajectory generated by a Markov process with a single recurrent class. For $k \in \mathbb{N}$ and any initial state, $\mathbb{P}(\tau_{\text{cov}}^k \geq m\lceil et_{\text{cov}}^k \rceil) \leq e^{-m}$ and $t_{\text{cov}}^k \leq (e \ln |\mathcal{S}| + e^2)(t_{\text{cov}} + (k-1)/d_{\min})$.

Using Fact 6, we first prove our key lemma.

Proof. Define $\mathcal{F}_i = \sigma(s_0, s_1, \ldots, s_{\tau_i+1})$, and let $f_i = f(s_{\tau_i+1}) - \mathbb{E}\left[f(s_{\tau_i+1})\right]$. Then f_i is $\mathcal{F}_i$-measurable. Define $\mathcal{F}_{i-1}$ measurable function $H_i = \mathbf{1}_{\{\tau_i \leq N\}}$. Then, $\sum_{i=1}^I f_i = \sum_{i=1}^M f_i H_i$. By Hoeffding's lemma and strong Markov property, we have

$$\mathbb{E}[\exp(\lambda f_i H_i) \,|\, \mathcal{F}_{i-1}] \leq \mathbb{E}[\exp(H_i^2 \lambda^2 (a-b)^2/8) \,|\, \mathcal{F}_{i-1}] \leq \exp(\lambda^2 (a-b)^2/8).$$

Thus, by induction,

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{M}\sum_{i=1}^{I}(f(s_{\tau_i+1}) - \mathbb{E}[f(s_{\tau_i+1})]\right)\right] = \mathbb{E}\left[\exp\left(\frac{\lambda}{M}\sum_{i=1}^{M} f_i H_i\right)\right] \leq \exp\left(\frac{\lambda^2(a-b)^2}{8M}\right).$$

By Fact 6 and direct calculation, we obtain the second statement. □

C Omitted proofs in Section 3

C.1 Proof of Lemma 3.1

Proof. Since $\pi_\epsilon \in \Pi_+$, we have

$$\begin{aligned}
J^{\pi_\epsilon} &= P_\star^{\pi_\epsilon} r \\
&= P_\star^{\pi_\epsilon}(r + P^{\pi_\epsilon} Q - Q) \\
&= P_\star^{\pi_\epsilon}(\mathbf{T}^{\pi_\epsilon} Q - Q) \\
&\geq \min_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T}^{\pi_\epsilon} Q - Q)(s,a) \\
&\geq \min_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T}^{\pi} Q - Q)(s,a) - \epsilon \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} |Q(s,a)|,
\end{aligned}$$

where second equality is from $P_\star^{\pi_\epsilon} P^{\pi_\epsilon} = P_\star^{\pi_\epsilon}$, first inequality follows from $P_\star^{\pi_\epsilon}((s',a'),(s,a)) = 0$ for $(s,a) \in \mathcal{T} \times \mathcal{A}$, and second inequality is from $\max_{(s,a) \in \mathcal{R} \times \mathcal{A}} |\mathbf{T}^{\pi} Q(s,a) - \mathbf{T}^{\pi_\epsilon} Q(s,a)| \leq \epsilon \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} |Q(s,a)|$.

By Fact 7, for any $\epsilon' > 0$, there exist $\pi' \in \Pi_+$ such that $J^\star - J^{\pi'} \leq \epsilon'$. Then, with similar argument,

$$J^{\pi'} = P_\star^{\pi'} r = P_\star^{\pi'}(r + P^{\pi'} Q - Q) \leq \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T}^{\pi'} Q - Q)(s,a) \leq \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T} Q - Q)(s,a)$$

where the last inequality is from the definition of $\mathbf{T}$. Therefore, we have $J^\star \leq \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T} Q - Q)(s,a)$ and this implies

$$\|J^\star - J^{\pi_\epsilon}\|_\infty \leq \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T}^{\pi} Q - Q)(s,a) - \min_{(s,a) \in \mathcal{R} \times \mathcal{A}} (\mathbf{T}^{\pi} Q - Q)(s,a) + \epsilon \max_{(s,a) \in \mathcal{R} \times \mathcal{A}} |Q(s,a)|.$$

□

C.2 State action transition probability in weakly communicating MDP

Given an MDP, define $P^{\pi}_{\mathcal{SA}}(s,a)$ and $P^{\pi}_{\mathcal{S}}(s)$. The following lemma clarifies the correspondence between the recurrent-transient classifcation of $P^{\pi}_{\mathcal{S}}$ and $P^{\pi}_{\mathcal{SA}}$.

Lemma C.1. *For $\pi \in \Pi_+$, if $\{\mathcal{R}_i\}_{i=1}^k, \mathcal{T}$ are irreducible recurrent classes and the transient set of $P^{\pi}_{\mathcal{S}}$ respectively, then, $\{\mathcal{R}_i \times \mathcal{A}\}_{i=1}^k, \mathcal{T} \times \mathcal{A}$ are irreducible recurrent classes and the transient set of $P^{\pi}_{\mathcal{SA}}$ respectively.*

Proof. Suppose $s \in \mathcal{R}_i$. Then, for any $a \in \mathcal{A}$, if $s' \notin \mathcal{R}_i$, $P(s' \,|\, s,a) = 0$ since $\pi \in \Pi_+$ and a finite irreducible recurrent class is closed. Hence, there exist s' such that $s' \in \mathcal{R}_i$ and $P(s' \,|\, s,a) \neq 0$. By the definition of a communication class, there exists n satisfying $(P^{\pi}_{\mathcal{S}})^n(s', s'') \neq 0$ for any $s'' \in \mathcal{R}_i$. Then, $(P^{\pi}_{\mathcal{SA}})^{n+1} = P(P^{\pi}_{\mathcal{S}})^n \Theta_\pi$ implies that $(P^{\pi}_{\mathcal{SA}}((s,a) \to (s'', a')))^{n+1} \neq 0$ for all $a' \in \mathcal{A}$. This shows that $\mathcal{R}_i \times \mathcal{A}$ is a closed communicating class of $P^{\pi}_{\mathcal{SA}}$. Lastly, suppose $s \in \mathcal{T}$. Then, for any $s \in \mathcal{T}, a \in \mathcal{A}$, $\sum_{n=0}^{\infty}(P^{\pi}_{\mathcal{SA}})^n(s,a) = I + P(\sum_{n=0}^{\infty}(P^{\pi}_{\mathcal{S}})^n)\Theta_\pi(s,a) < \infty$ since $(\sum_{n=0}^{\infty}(P^{\pi}_{\mathcal{S}})^n)(s', s'') < \infty$ for all $s', s'' \in \mathcal{T}$ ((Berman & Plemmons, 1994, Lemma 8.3.20)). This shows that $\mathcal{T} \times \mathcal{A}$ is a transient set. □

In particular, Lemma C.1 implies that if $P^{\pi}_{\mathcal{S}}$ has a single recurrent class, then $P^{\pi}_{\mathcal{SA}}$ has a single recurrent class as well. For a single trajectory, (s_{i+1}, a_{i+1}) can be considered as a state of transition probability $P^{\pi}_{\mathcal{SA}}$, i.e., $(s_{i+1}, a_{i+1}) \sim P^{\pi}_{\mathcal{SA}}(\cdot, \cdot \,|\, s_i, a_i)$. Hence, facts and lemmas in section 2 continue to hold with respect to $P^{\pi_b}_{\mathcal{SA}}$ for $\pi_b \in \Pi_+$.

C.3 Proof of Proposition 3.2

Let $L_k = \sum_{i=0}^{k-1}(t_i + 1)$ and $\mathcal{F}_k = \sigma((s_i, a_i) \,|\, 0 \le i \le L_{k+1})$ denote the natural filtration generated by the sampling process in SAVIC, and $\mathbb{P}(\cdot)$ the probability distribution over the trajectories. Then, T^k is $\mathcal{F}_k$-measurable whereas Q^k, h^k, d^k and m_k, being functions of T^{k-1}, are $\mathcal{F}_{k-1}$-measurable.

Proof. First, by Lemma 2.2, Fact 2, and union bounds, with probability $1 - \delta/2$, $s_{n_{\text{esc}}+n_{\text{rec}}} \in \mathcal{R}$ and $R = \mathcal{R}$.

Suppose that $s_{n_{\text{esc}}+n_{\text{rec}}} \in \mathcal{R}$ and $R = \mathcal{R}$ so that we focus on $\mathcal{R} \times \mathcal{A}$, and for simplicity, consider $s_{n_{\text{esc}}+n_{\text{rec}}}$ as s_0. Let $s_{\tau_j^k(s,a)+1} \sim P(\cdot \,|\, s,a)$ for $j = 1, \dots, I^k_{s,a}$ be the samples at the k-th iteration where $\tau_i^k(s,a) = \inf\{l > \tau_{i-1}^k : (s^l, a^l) = (s,a)\}$ for $(s,a) \in \mathcal{R} \times \mathcal{A}$ and $\tau_0^k = L_k - 1$. Define $I^k_{s,a} = \min\{m_k, \max\{j \,|\, \tau_j^k(s,a) \le L_k + t_k\}\}$,

$$Y^k(s,a) = \tfrac{1}{m_k}\sum_{j=1}^{I^k_{s,a}} \big(d^k(s_{\tau_j^k(s,a)+1}) - Pd^k(s,a)\big) \qquad \forall (s,a) \in \mathcal{R} \times \mathcal{A},$$

and $X^k = \sum_{i=0}^{k} Y^i$. We proceed to estimate $\mathbb{P}(\|X^k\|_\infty \ge \varepsilon)$ by adapting the arguments of Azuma-Hoeffding's inequality with a stopping time argument. Note that d^k and m_k are $\mathcal{F}_{k-1}$-measurable. From Markov's inequality and the tower property of conditional expectations we get that for each $(s,a) \in \mathcal{R} \times \mathcal{A}$ and $\lambda > 0$

$$\begin{aligned}\mathbb{P}(X^k(s,a) \ge \varepsilon) &\le e^{-\lambda\varepsilon}\,\mathbb{E}\big[\exp(\lambda\, X^k(s,a))\big] \\ &= e^{-\lambda\varepsilon}\,\mathbb{E}\big[\exp(\lambda\, X^{k-1}(s,a))\,\mathbb{E}[\exp(\lambda\, Y^k(s,a)) \,|\, \mathcal{F}_{k-1}]\big]. \end{aligned} \tag{1}$$

Conditionally on $\mathcal{F}_{k-1}$, for fixed j, $d^k(s_{\tau_j^k(s,a)+1}) - Pd^k(s,a)$ is a random variable with zero mean and its absolute value is upper bounded by $\|d^k\|_{\text{sp}}$. Then, by Lemma 2.3, we have

$$\mathbb{E}\big[\exp(\lambda\, Y^k(s,a)) \,|\, \mathcal{F}_{k-1}\big] \le \exp\big(\tfrac{1}{2}\lambda^2(\|d^k\|_{\text{sp}})^2/m_k\big). \tag{2}$$

Using equation 1 and equation 2, together with the definition of m_k, a simple induction yields

$$\mathbb{E}[\exp(\lambda\, X^k(s,a))] \le \exp\big(\tfrac{1}{2}\lambda^2\varepsilon^2\textstyle\sum_{i=0}^{k} c_i^{-1}/\eta\big).$$

Then, since $\sum_{i=0}^{\infty} c_i^{-1} \le \frac{1}{2}$ we get $\mathbb{P}(X^k(s,a) \ge \varepsilon) \le \exp\big(-\lambda\,\varepsilon + \frac{1}{4}\lambda^2\varepsilon^2/\eta\big)$ and taking $\lambda = 2\eta/\varepsilon$ we deduce

$$\mathbb{P}(X^k(s,a) \ge \varepsilon) \le \exp(-\eta) = \delta/(8|\mathcal{R}||\mathcal{A}|(n+1)).$$

A symmetric argument yields the same bound for $\mathbb{P}(X^k(s,a) \le -\varepsilon)$ so that $\mathbb{P}(|X^k(s,a)| \ge \varepsilon) \le \delta/(4|\mathcal{R}||\mathcal{A}|(n+1))$. Applying a union bound over all $(s,a) \in \mathcal{R} \times \mathcal{A}$ we get $\mathbb{P}(\|X^k\|_\infty \ge \varepsilon) \le \delta/(4(n+1))$, and then a second union bound over k and taking the complementary event gives $\mathbb{P}(\bigcap_{k=0}^n \{\|X^k\|_\infty \le \varepsilon\}) \ge 1 - \delta/4$.

Next, by the second statement of Lemma 2.3 and a union bound over k,

$$\mathbb{P}\left(\bigcap_{0\le k\le n} \{I^k_{s,a} = m_k \,\forall (s,a) \in \mathcal{R}\times\mathcal{A}\}\right) \ge 1 - \delta/4.$$

In this event, since $h^{-1} = 0$, by telescoping $\mathbf{T}Q^k = r + Ph^k = r + \sum_{i=0}^k Pd^i$ and we get

$$(T^k - \mathbf{T}Q^k)(s,a) = \sum_{i=0}^k (D^i - Pd^i)(s,a) = \sum_{i=0}^k Y^i(s,a) = X^k(s,a) \quad \forall (s,a) \in \mathcal{R}\times\mathcal{A}.$$

Then, the intersection of $\bigcap_{0\le k\le n}\{\|X^k\|_\infty \le \varepsilon\}$ and $\bigcap_{0\le k\le n}\{I^k_{s,a} = m_k \,\forall (s,a) \in \mathcal{R}\times\mathcal{A}\}$ gives

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{R}\times\mathcal{A}} |T^k - \mathbf{T}(Q^k)|(s,a) \le \epsilon \;\forall 0 \le k \le n\right) \ge 1 - \frac{\delta}{2}$$

The conclusion follows by a union bound and conditioning on the first sampling process. □

C.4 Proof of Theorem 3.3

We first introduce the following lemma.

Lemma C.2. *Consider a weakly communicating $MDP(\mathcal{S},\mathcal{A},P,r)$, and a restricted recurrent $MDP(\mathcal{R},\mathcal{A},P_\mathcal{R},r_{\mathcal{R}\times\mathcal{A}})$ where the transition probability $P_\mathcal{R} : \mathcal{R}\times\mathcal{A} \to \mathcal{M}(\mathcal{R})$ is defined by $P_\mathcal{R}(s' \mid s,a) = P(s' \mid s,a)$ for all $s,s' \in \mathcal{R}, a \in \mathcal{A}$ and $r_{\mathcal{R}\times\mathcal{A}}$ is defined by $r_\mathcal{R}(s,a) = r(s,a)$ for all $s \in \mathcal{R}, a \in \mathcal{A}$. Let $\pi^\star_\mathcal{R}$ and $Q^\star_\mathcal{R}$ be an optimal policy and the corresponding value function in the restricted MDP. Then, there exist $\pi^\star$ and corresponding value function $Q^\star(s,a)$ for $MDP(\mathcal{S},\mathcal{A},P,r)$ such that $\pi^\star_\mathcal{R}(a \mid s) = \pi^\star(a \mid s)$ and $Q^\star_\mathcal{R}(s,a) = Q^\star(s,a)$ for all $s,a \in \mathcal{R}\times\mathcal{A}$.*

Proof. First, we note that $P_\mathcal{R}$ is indeed a transition probability since $P(s' \mid s,a) = 0$ for $s' \in \mathcal{R}^\complement, s \in \mathcal{R}$ as shown in the proof of Lemma C.1. This implies that $P^\pi((s',a'),(s,a)) = P^{\pi_\mathcal{R}}_\mathcal{R}((s',a'),(s,a))$ and $P^\pi_\star((s',a'),(s,a)) = P^{\pi_\mathcal{R}}_{\mathcal{R},\star}((s',a'),(s,a))$ for all $s,s' \in \mathcal{R}, a,a' \in \mathcal{A}$ by property of product and sum of block matrix $\begin{bmatrix} A & \mathbf{0} \\ B & C \end{bmatrix}$ where A, C are square matrices. So, $J^{\pi_\mathcal{R}}_\mathcal{R}(s,a) = J^\pi(s,a)$ for any π satisfying $\pi_\mathcal{R}(a \mid s) = \pi(a \mid s)$ for all $s,a \in \mathcal{R}\times\mathcal{A}$ and this implies $J^\star(s,a) = J^{\pi^\star_\mathcal{R}}(s,a) = J^{\pi^\star}(s,a)$ for a policy $\pi^\star$ satisfying $\pi^\star_\mathcal{R}(a \mid s) = \pi^\star(a \mid s)$ for all $s,a \in \mathcal{R}\times\mathcal{A}$. Since for any policy $\pi \in \Pi$ (not Π_+), $\mathcal{T}\times\mathcal{A}$ for P^π continues to be a transient set through the same argument in Lemma C.2 with Fact 1 and using the fact that the optimal average reward is constant in a weakly communicating MDP, we can show that $J^{\pi^\star}(s,a)$ is a constant vector and $\pi^\star$ is an optimal policy in the original MDP. Lastly, by property of inverse and product of block matrix $\begin{bmatrix} A & \mathbf{0} \\ B & C \end{bmatrix}$ where A, C are square matrices, we also have $Q^\star_\mathcal{R}(s,a) = Q^\star(s,a)$ for all $s,a \in \mathcal{R}\times\mathcal{A}$. □

We now introduce the following key lemma.

Lemma C.3. *With the coefficients of Proposition 3.2, let (Q^n, T^n, π_n) be the iterates generated by SAVIC$(n_{\mathrm{esc}}, n_{\mathrm{rec}}, n, \{t_k\}_{k=0}^n, \varepsilon, \delta)$. Then, with probability at least $1-\delta$, we have*

$$\|Q^n - \mathbf{T}Q^n\|_{\mathrm{sp}} \le \frac{8\|Q^0 - Q^\star_\mathcal{R}\|_{\mathrm{sp}}}{n+2} + 4\varepsilon.$$

Proof. As the proof of Proposition 3.2 shows, with probability at least $1-\delta$, $R = \mathcal{R}$ and $I^k_{s,a} = m_k$ for all $(s,a) \in \mathcal{R}\times\mathcal{A}$. On the event of Proposition 3.2, combining these with the proof of Theorem 3.2 of (Lee et al., 2025), we have an upper bound on the Bellman error in the restricted recurrent $MDP(\mathcal{R},\mathcal{A},P_\mathcal{R},r_{\mathcal{R}\times\mathcal{A}})$ where the transition probability $P_\mathcal{R} : \mathcal{R}\times\mathcal{A} \to \mathcal{M}(\mathcal{R})$ is defined by $P_\mathcal{R}(s' \mid s,a) = P(s' \mid s,a)$ for all $s,s' \in \mathcal{R}, a \in \mathcal{A}$ and $r_{\mathcal{R}\times\mathcal{A}}$ is defined by $r_\mathcal{R}(s,a) = r(s,a)$ for all $s \in \mathcal{R}, a \in \mathcal{A}$. □

Now we are ready to prove Theorem 3.3.

Proof. With the coefficients in Theorem 3.3, we have $\|Q^n - \mathbf{T}(Q^n)\|_{\text{sp}} \leq \frac{\varepsilon}{2}$ by Lemmas C.2 and C.3, and we get $\left\|J^\star - J^{\pi^n_{\varepsilon/16}}\right\|_\infty \leq \varepsilon$ by Lemma 3.1.

Now, to estimate the total number of samples $\sum_{k=0}^n t_k$ we recall that $m_k = \max\{\lceil \eta c_k \|d^k\|_{\text{sp}}^2/\varepsilon^2 \rceil, 1\}$ which can be bounded as $t_k \leq 1 + e^2(\ln|\mathcal{R}||\mathcal{A}| + e)(t_{\text{cov}} + (m_k - 1)/d^{\pi_b}_{\min}) \log\left(\frac{4(n+1)}{\delta}\right)$. Then

$$\begin{aligned}\sum_{k=0}^n t_k &\leq n + 1 + \log\left(\frac{4(n+1)}{\delta}\right)\left(e^2(\ln|\mathcal{R}||\mathcal{A}| + e)(n+1)t_{\text{cov}} + 1/d^{\pi_b}_{\min}\sum_{k=0}^n m_k\right) \\ &= \widetilde{O}\left(t_{\text{cov}}\|Q^\star\|_{\text{sp}}/\varepsilon + 1/d^{\pi_b}_{\min}\|Q^\star\|^2_{\text{sp}}/\varepsilon^2\right)\end{aligned}$$

since $\sum_{k=0}^n m_k = O(\eta \ln^3(n+2)\|Q^\star_{\mathcal{R}}\|^2_{\text{sp}}/\varepsilon^2 + \eta\, n^2 \ln^2(n+2))$ from the proof of Theorem 3.2 of (Lee et al., 2025), and the fact that $1/d^{\pi_b}_{\min}, t_{\text{cov}} \geq |\mathcal{R}||\mathcal{A}|$. Lastly, by adding $n_{\text{hit}} + n_{\text{rec}}$, we obtain the desired result.

□

C.5 Proof of Theorem 3.4

Proof. For the proof, we again follow the proof framework of Lee et al. (2025).

The stopping time of SAVIC**+** is the random variable

$$I = \inf\{i \in \mathbb{N} : \|Q^{n_i} - T^{n_i}\|_{\text{sp}} \leq 14\,\varepsilon\}$$

where T^{n_i} and Q^{n_i} are the iterates generated in each loop of SAVIC**+**$(Q^0, \varepsilon, \delta)$.

We let $i_0 \in \mathbb{N}$ be the smallest integer satisfying $n_{i_0} \geq \|Q^*\|_{\text{sp}}/\varepsilon$, so that either $i_0 = 0$ and $n_{i_0} = 1$ or $n_{i_0-1} = n_{i_0}/2 < \|Q^*\|_{\text{sp}}/\varepsilon$, which combined imply $n_{i_0} \leq 2(1 + \|Q^*\|_{\text{sp}}/\varepsilon)$.

Lemma C.4. *Consider a communicating MDP. Let (Q^N, T^N, π^N) be the output of* SAVIC**+**$(\varepsilon/16, \delta/2)$ *with $\varepsilon \leq 1 \leq \|Q^\star\|_{\text{sp}}$. Then, with probability at least $1 - \delta/2$, we have $I \leq i_0$ and*

$$\|J^\star - J^{\pi_N}\|_\infty \leq \|Q^N - \mathbf{T}(Q^N)\|_{\text{sp}} \leq \varepsilon.$$

Proof. First, by Fact 4, every state is recurrent in communicating MDPs, and thus SAMPLE**+**(m_k, d^k) almost surely provides i.i.d. m_k samples for each state and action. Then, the proof of Corollary 3.5 of Lee et al. (2025) can be directly applied to SAVIC**+**$(\varepsilon/16, \delta/2)$ and the desired result follows. □

Due to Lemma C.4, we directly obtain the convergence result in Theorem 3.4.

We denote the number of samples used during the execution of SAVIC$(n_{\text{esc}}, n_{\text{rec}}, n_i, \emptyset, \varepsilon, \delta_i)$ in the i-th loop of SAVIC**+** by M_i, so that the total sample complexity is $M = \sum_{i=0}^I M_i$.

Note that by Lemma 2.3, with probability at least $1 - \delta_i$, the length of the trajectory sampled from SAMPLE**+**(m_k, d^k) in the i-th loop of SAVIC**+** is smaller than $t^i_k = 1 + e^2(\ln|\mathcal{S}||\mathcal{A}| + e)(t_{\text{cov}} + (m_k - 1)/d^{\pi_b}_{\min}) \log\left(\frac{4(n_i+1)}{\delta_i}\right)$ for $0 \leq k \leq n_i$ and $s, a \in \mathcal{S} \times \mathcal{A}$. Then, based on the sample complexity analysis in the proof of the Theorem 3.3 and Corollary 3.5 of Lee et al. (2025) and a

union bound, with probability at least $1-\delta/2$,

$$\begin{aligned}
M &\le \sum_{i=0}^{i_0}\sum_{k=0}^{n_i} t_k^i. \\
&= \sum_{i=0}^{i_0}(i+1) + \log\left(\frac{4(n_i+1)}{\delta_i}\right)\Bigg(\sum_{i=0}^{i_0} e^2(\ln|\mathcal{S}||\mathcal{A}|+e)(n+1)t_{\mathrm{cov}} \\
&+ 1/d_{\min}^{\pi_b}\sum_{i=0}^{i_0}\left(\eta_i \ln^3(n_i+2)\|Q^\star\|_{\mathrm{sp}}^2/\varepsilon^2 + \eta_i\, n_i^2 \ln^2(n_i+2)\right)\Bigg) \\
&= \widetilde{O}\left(t_{\mathrm{cov}}\|Q^*\|_{\mathrm{sp}}/\varepsilon + 1/d_{\min}^{\pi_b}\|Q^\star\|_{\mathrm{sp}}^2/\varepsilon^2\right)
\end{aligned}$$

where $\eta_i = \ln(8|\mathcal{S}||\mathcal{A}|(n_i+1)c_i/\delta)$ is the parameter used in the i-th internal cycle of SAVIC**+**. Lastly, by a union bound over the sample complexity bound and convergence results, we have Theorem 3.4.

□

D Omitted proofs in Section 4

We note that although SCPMA optimizes the average reward over the clipped policy set Π_α, the following fact guarantees that the gap between $J^\star$ and J^{π_α} can be made arbitrarily small by properly choosing α.

Fact 7 (Lemma 25, (Lee & Ryu, 2025a)). Consider a weakly communicating MDP. For $\pi \in \Pi_+$,

$$J^\star - J^\pi = \sum_{s\in\mathcal{R}}\sum_{a\in\mathcal{A}} d^\pi(s)\big(\pi_\star(a \mid s) - \pi(a \mid s)\big)Q^{\pi_\star}(s,a).$$

As $\alpha \to 0$, we can find $\pi \in \Pi_\alpha$ such that $\pi_\star(a \mid s) - \pi(a \mid s)$ becomes arbitrarily small, while $\|Q^{\pi_\star}\|_\infty$ and $\|d^\pi\|_1 (= 1)$ remain constant.

D.1 Proof of Proposition 4.1

In this section, we suppose that $\pi \in \Pi_+$ and the MDP is weakly communicating. We consider a restricted recurrent $MDP(\mathcal{R}, \mathcal{A}, P_{\mathcal{R}}, r_{\mathcal{R}\times\mathcal{A}})$ where the transition probability $P_{\mathcal{R}} : \mathcal{R}\times\mathcal{A} \to \mathcal{M}(\mathcal{R})$ is defined by $P_{\mathcal{R}}(s' \mid s,a) = P(s' \mid s,a)$ for all $s, s' \in \mathcal{R}, a \in \mathcal{A}$ and $r_{\mathcal{R}\times\mathcal{A}}$ is defined by $r_{\mathcal{R}}(s,a) = r(s,a)$ for all $s \in \mathcal{R}, a \in \mathcal{A}$. Let $\pi_{\mathcal{R}}$ and $Q_{\mathcal{R}}^\pi$ be a policy and the corresponding state-action value function, respectively. Let $d_{\mathcal{R}}^\pi (= P_{\mathcal{R}\star}^\pi)$ be the unique stationary distribution of $P_{\mathcal{R}}^\pi$.

We first introduce a lemma related to target time.

Lemma D.1. *Consider a weakly communicating MDP. Then, we have*

$$\left\| \frac{1}{k+1}\sum_{j=0}^{k}\left(P_{\mathcal{R}}^\pi\right)^j(\cdot,(s,a)) \;-\; d^\pi(\cdot)\right\|_1 \;\le\; \frac{2t_{\mathrm{tar}}^\pi}{k+1} \qquad \forall s,a \in \mathcal{R}\times\mathcal{A}.$$

Proof. Combining the proof of Lemma C.2 and Corollary 3 of Roberts & Rosenthal (1997), we have the desired result. □

Consider Anc-QI with the Bellman consistency operator :

$$Q^k = \beta_k \mathbf{T}^\pi Q^{k-1} + (1-\beta_k)Q^0.$$

We now present the convergence result of Anc-QI.

Lemma D.2. *Anc-QI with the Bellman consistency operator, $\beta_k = \frac{k}{k+1}$, and $Q^0 = 0$ exhibits the rate*

$$\|Q^k - Q_{\mathcal{R}}^\pi\|_{\mathrm{sp}} \le \frac{4t_{\mathrm{tar}}^\pi \|Q_{\mathcal{R}}^\pi\|_{\mathrm{sp}}}{k+1}.$$

Proof. By definition of Anc-QI, $Q^k = \frac{1}{k+1}\sum_{i=0}^{k}(\mathbf{T}^\pi)^i Q^0$, and

$$\begin{aligned}(\mathbf{T}^\pi)^i Q^0 &= \sum_{j=0}^{i-1}(P_\mathcal{R}^\pi)^j r_\mathcal{R} \\ &= \sum_{j=0}^{i-1}(P_\mathcal{R}^\pi)^j (J^\pi + (I - P_\mathcal{R}^\pi)Q_\mathcal{R}^\pi) \\ &= iJ^\pi + Q_\mathcal{R}^\pi - (P_\mathcal{R}^\pi)^i Q_\mathcal{R}^\pi,\end{aligned}$$

where the second and third equalities come from the Bellman equations. Utilizing the fact that $P_{\mathcal{R},\star}^\pi Q_\mathcal{R}^\pi = 0$ from Section A.5 of (Puterman, 2014) and the proof of Lemma C.2, we have

$$Q^k - Q_\mathcal{R}^\pi - \frac{k}{2}J^\pi = \left(P_{\mathcal{R}\star}^\pi - \frac{\sum_{i=0}^{k}(P_\mathcal{R}^\pi)^i}{k+1}\right)Q_\mathcal{R}^\pi.$$

By taking $\|\cdot\|_{\text{sp}}$ and using properties of span seminorm $\|v + c\mathbf{1}\|_{\text{sp}} = \|v\|_{\text{sp}}$ for $c \in \mathbb{R}$ and $\|v\|_{\text{sp}} \leq 2\|v\|_\infty$, Lemma D.1, and the fact that J^π is constant vector, we have the desired result. □

We introduce the following lemma, which is analogous to Proposition 3.2. Its proof follows directly from that of Proposition 3.2 after appropriately adapting the coefficients (note that unlike SAVIC, $m_k = \max\{\lceil \eta\, c_k \|d^k\|_{\text{sp}}^2 (n/\varepsilon)^2 \rceil, 1\}$ in SAVIPE).

Lemma D.3. *Consider a weakly communicating MDP. Let* $n_{\text{esc}} \geq \lceil et_{\text{hit}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil, n_{\text{rec}} \geq \lceil et_{\text{cov}} \rceil \lceil \ln(\frac{4}{\delta}) \rceil$, $n = \left\lceil \frac{8t_{\text{tar}}^{\pi_k}\|Q^\pi\|_{\text{sp}}}{\varepsilon} \right\rceil$, *and let* T^k, Q^k *be the iterates generated by* SAVIPE$(\pi, n, \varepsilon/2, \delta, \eta, R)$. *Then, with probability at least* $1-\delta$, $R = \mathcal{R}$, $I_{s,a}^k = m_k$, *and* $\left\|T^k - \mathbf{T}^\pi(Q^k)\right\|_\infty \leq \varepsilon/(2n)$ *for all* $k = 0, \ldots, n-1$ *and* $(s,a) \in \mathcal{R} \times \mathcal{A}$.

Now, consider inexact Anc-QI with the Bellman consistency operator:

$$Q^k = \frac{k}{k+1}(\mathbf{T}^\pi Q^{k-1} + \epsilon_k) + \frac{1}{k+1}Q^0$$

where $\epsilon_k \in \mathbb{R}^{\mathcal{R}\times\mathcal{A}}$ represents the approximation error of the Bellman operator. Then, by direct calculation,

$$Q^k = \frac{1}{k+1}\sum_{i=0}^{k}(\mathbf{T}^\pi)^i Q^0 + \frac{k+1-i}{k+1}\sum_{i=1}^{k}(P^\pi)^{i-1}\epsilon_{k-i+1}.$$

Under the condition of Lemma D.3, $\left\|T^k - \mathbf{T}^\pi(Q^k)\right\|_\infty = \|\epsilon_k\|_\infty \leq \varepsilon/(2n)$ for all k. This implies

$$\|Q^k - \frac{1}{k+1}\sum_{i=0}^{k}(\mathbf{T}^\pi)^i Q^0\|_{\text{sp}} \leq \frac{\varepsilon}{2}. \tag{3}$$

Now we are ready to prove Proposition 4.1.

Proof. With the coefficients in Proposition 4.1, by equation 3, Lemma D.3 and D.2, and a union bound over $0 \leq k \leq K$, with probability at least $1-\delta$,

$$\begin{aligned}\|\hat{Q}^{\pi_k} - Q_\mathcal{R}^\pi\|_{\text{sp}} &\leq \|\hat{Q}^{\pi_k} - \frac{1}{n_k+1}\sum_{i=0}^{n_k}(\mathbf{T}^\pi)^i Q^0\|_{\text{sp}} + \|Q_\mathcal{R}^\pi - \frac{1}{n_k+1}\sum_{i=0}^{n_k}(\mathbf{T}^\pi)^i Q^0\|_{\text{sp}} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &\leq \varepsilon\end{aligned}$$

for all $0 \leq k \leq K$. □

Additionally, the following result will be needed for computing sample complexity in next subsection.

We follow the proof framework of (Lee et al., 2025). We keep the notation for all the sequences generated by SAVIC and we assume throughout that

$$\left\|T^k - \mathbf{T}^\pi(Q^k)\right\|_\infty = \|\epsilon_k\|_\infty \le \varepsilon \qquad \text{for all } k = 0, \dots, n-1, \tag{4}$$

which, in view of Lemma D.3, holds with probability at least $(1-\delta)$. Also, we have

$$\|T^k - T^{k-1}\|_{\text{sp}} = \|D^k\|_{\text{sp}} \le \|d^k\|_{\text{sp}} = \|h^k - h^{k-1}\|_{\text{sp}} \le \|Q^k - Q^{k-1}\|_{\text{sp}}. \tag{5}$$

We first establish two preliminary technical lemmas.

Lemma D.4. *Assuming equation 4, we have*

$$\|Q^k - Q^\pi_{\mathcal{R}}\|_{\text{sp}} \le \|Q^0 - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + k\varepsilon \quad \textit{for all } k = 0, \dots, n.$$

Proof. From the iteration $Q^k = (1-\beta_k)Q^0 + \beta_k\, T^{k-1}$ with $\beta_k = \frac{k}{k+1}$ we get

$$\|Q^k - Q^\pi_{\mathcal{R}}\|_{\text{sp}} \le \tfrac{1}{k+1}\|Q^0 - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + \tfrac{k}{k+1}\|T^{k-1} - Q^\pi_{\mathcal{R}}\|_{\text{sp}}. \tag{6}$$

Using the invariance of $\|\cdot\|_{\text{sp}}$ by addition of constants and the nonexpansivity of $\mathbf{T}^\pi$ for this seminorm, a triangle inequality together with $\|\cdot\|_{\text{sp}} \le 2\|\cdot\|_\infty$ and the bound equation 4 imply

$$\|T^{k-1} - Q^\pi_{\mathcal{R}}\|_{\text{sp}} = \|T^{k-1} - \mathbf{T}^\pi(Q^\pi_{\mathcal{R}})\|_{\text{sp}} \le 2\varepsilon + \|Q^{k-1} - Q^\pi_{\mathcal{R}}\|_{\text{sp}}$$

which plugged back into equation 6 yields

$$\|Q^k - Q^\pi_{\mathcal{R}}\|_{\text{sp}} \le \tfrac{1}{k+1}\|Q^0 - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + \tfrac{2}{k+1}k\varepsilon + \tfrac{k}{k+1}\|Q^{k-1} - Q^\pi_{\mathcal{R}}\|_{\text{sp}}.$$

By induction, the conclusion follows. □

Lemma D.5. *Assume equation 4 and let $\rho_k = 2\|Q^0 - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + k\varepsilon$. Then*

$$\|Q^k - Q^{k-1}\|_{\text{sp}} \le \tfrac{1}{k}\textstyle\sum_{i=1}^k \frac{\rho_{i+1}}{i+1} \quad \textit{for all } k = 1, \dots, n.$$

Proof. From $Q^k = \frac{1}{k+1}Q^0 + \frac{k}{k+1}T^{k-1}$ and $Q^{k-1} = \frac{1}{k}Q^0 + \frac{k-1}{k}T^{k-2}$ we derive

$$Q^k - Q^{k-1} = \tfrac{1}{k(k+1)}(T^{k-1} - Q^0) + \tfrac{k-1}{k}(T^{k-1} - T^{k-2}).$$

Now, from the previous proof of Lemma D.4, we have $\|T^{k-1} - Q^0\|_{\text{sp}} \le \|T^{k-1} - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + \|Q^\pi_{\mathcal{R}} - Q^0\|_{\text{sp}} \le \rho_{k+1}$, and

$$\|Q^k - Q^{k-1}\|_{\text{sp}} \le \tfrac{1}{k(k+1)}\rho_{k+1} + \tfrac{k-1}{k}\|Q^{k-1} - Q^{k-2}\|_{\text{sp}}.$$

The conclusion follows from induction. □

By Lemma D.5, we have

$$\begin{aligned}\|d^k\|_{\text{sp}} &\le \tfrac{1}{k}\textstyle\sum_{i=1}^k \frac{\rho_{i+1}}{i+1} \\ &\le \tfrac{2\ln(k+1)}{k}\|Q^0 - Q^\pi_{\mathcal{R}}\|_{\text{sp}} + \varepsilon.\end{aligned}$$

D.2 Proof of Theorem 4.2

Following the framework of (Lee & Ryu, 2025a), consider *inexact policy mirror ascent*

$$\pi_{k+1}(\cdot \mid s) = \operatorname*{arg\,max}_{p \in \mathcal{M}_\alpha(\mathcal{A})} \left\{ b_k \sum_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(s,a)p(a) - D(p(\cdot), \pi_k(\cdot \mid s)) \right\}, \qquad \forall s \in \mathcal{S},$$

where $\hat{Q}^{\pi_k}$ is an inexact evaluation of Q^{π_k}. We analyze the convergence of inexact policy mirror ascent under the following assumption.

Assumption 1. The inexact evaluations $\hat{Q}^{\pi_k}$ satisfy

$$\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_{\text{sp}} \le \varepsilon, \qquad \forall k \in \mathbb{N}.$$

For $\rho \in \mathcal{M}(\mathcal{S})$, define the weighted divergence

$$D_\rho(\pi, \pi') \;=\; \sum_{s\in\mathcal{S}} \rho(s)\, D(\pi(\cdot \mid s), \pi'(\cdot \mid s)).$$

Then, the following fact holds. We briefly note that although the original assumption of Fact 8 requires a $\|\cdot\|_\infty$ error bound, $c\mathbf{1} + \hat{Q}^\pi$ and $\hat{Q}^\pi$ generate the same π_k in the policy mirror ascent framework.

Fact 8 (Appendix D.3, (Lee & Ryu, 2025a)). Consider a weakly communicating MDP. Under Assumption 1, for $\pi_0 \in \Pi_\alpha$, the inexact α-clipped policy mirror ascent with constant step size $b > 0$ generates a sequence of policies $\{\pi_k\}_{k=1}^\infty$ satisfying

$$\|J^{\pi_\alpha} - J^{\pi_k}\|_\infty \;\le\; \frac{1}{k+1}\left(\frac{D_{d^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{b} + C_\alpha \|J^{\pi_\alpha} - J^{\pi_0}\|_\infty\right) + (2C_\alpha + k + 2)\varepsilon.$$

and with adaptive stepsize $b_{k+1}(C_\alpha - 1) \ge b_k C_\alpha > 0$ generates a sequence of policies $\{\pi_k\}_{k=1}^\infty$ satisfying

$$\|J^{\pi_\alpha} - J^{\pi_k}\|_\infty \le \left(1 - \frac{1}{C_\alpha}\right)^k \left(\|J^{\pi_\alpha} - J^{\pi_0}\|_\infty + \frac{1}{b_0(C_\alpha - 1)} D_{d^{\pi_\alpha}}(\pi_\alpha, \pi_0)\right) + 4C_\alpha\varepsilon.$$

Now, we introduce lemmas for the last step of our algorithm.

Lemma D.6. *Given $\pi \in \Pi_+$ and $\hat{Q} \in \mathbb{R}^{|\mathcal{R}||\mathcal{A}|}$ such that $J^\star - J^\pi \le \epsilon_1$ and $\|\hat{Q} - Q^\pi\|_{\text{sp}} < \epsilon_2$, we have $\left\|J^\star - J^{\hat{\pi}}\right\|_\infty \le \epsilon_1 + \epsilon_2$, where $\hat{\pi}(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}(s,a)\, \forall s \in \mathcal{R}$.*

Proof. By definition, $Q^\pi(s, \hat{\pi}(s)) \ge \max_a Q^\pi(s,a) - \epsilon_2 \ge Q^\pi(s, \pi(s)) - \epsilon_2$ and this implies $r + P^{\hat{\pi}} Q^\pi - Q^\pi \ge J^\pi - \epsilon_2 \mathbf{1}$ by the Bellman equation. By multiplying both sides by $P_\star^{\hat{\pi}}$, we have $J^{\hat{\pi}} \ge J^\pi - \epsilon_2 \mathbf{1}$. By assumption, we finally have $\left\|J^\star - J^{\hat{\pi}}\right\|_\infty \le \epsilon_1 + \epsilon_2$. □

Now, we prove Theorem 4.2.

Proof. By Fact 7, there exists $\pi \in \Pi_\alpha$ such that $\|\pi^\star - \pi\|_1 \le 2|\mathcal{A}|\alpha$, and this implies $J^\star - J^{\pi_\alpha} \le \frac{G}{8}$.

First, for constant step size, with coefficients in Theorem 4.2, Assumption 1 is satisfied with $\frac{\varepsilon}{16(2C_\alpha + K + 2)}$. Then, by Proposition 4.1, $\|d^k\|_{\text{sp}} \le \frac{2\ln(k+1)}{k}\|Q^0 - Q_{\mathcal{R}}^\pi\|_{\text{sp}} + \varepsilon$, and Fact 8, with probability at least $1 - \delta$, $J^{\pi_\alpha} - J_\mu^{\pi_K} \le \frac{\varepsilon}{8}$ with sample complexity

$$\begin{aligned}
&\sum_{k=0}^{K} \sum_{i=0}^{n_k - 1} t_i^{\pi_k} \\
&= \sum_{k=0}^{K-1} \widetilde{O}\left((2C_\alpha + K + 2)^2 t_{\text{cov}} t_{\text{tar}}^{\pi_k} \|Q^{\pi_k}\|_{\text{sp}}^2/\varepsilon^2 + 1/d_{\min}^{\pi_b}(t_{\text{tar}}^{\pi_k})^2 \|Q^{\pi_k}\|_{\text{sp}}^4 (2C_\alpha + K + 2)^4/\varepsilon^4\right) \\
&\le \widetilde{O}\left(g t_{\text{cov}} t_\alpha C_\alpha^2 Q_\alpha^2/\varepsilon^3 + (t_{\text{cov}} t_\alpha g^3 Q_\alpha^2 + 1/d_{\min}^{\pi_b} g C_\alpha^4 t_\alpha^2 Q_\alpha^4)/\varepsilon^5 + 1/d_{\min}^{\pi_b} t_\alpha^2 Q_\alpha^4 g^5/\varepsilon^9\right).
\end{aligned}$$

Lastly, by adding $n_{\text{hit}} + n_{\text{rec}}$, we have the total sample complexity.

For adaptive step size $b_{k+1}(C_\alpha - 1) \ge b_k C_\alpha$, with coefficients in Theorem 4.2, Assumption 1 is satisfied with $\frac{\varepsilon}{64 C_\alpha}$. Then, by Proposition 4.1, $\|d^k\|_{\text{sp}} \le \frac{2\ln(k+1)}{k}\|Q^0 - Q_{\mathcal{R}}^\pi\|_{\text{sp}} + \varepsilon$, and Fact 8, with probability at least $1 - \delta$, $J^{\pi_\alpha} - J_\mu^{\pi_K} \le \frac{\varepsilon}{8}$ with sample complexity

$$\begin{aligned}
&n_{\text{hit}} + n_{\text{rec}} + \sum_{k=0}^{K-1} \sum_{i=0}^{n_k - 1} t_i^{\pi_k} \\
&= \widetilde{O}\left(t_{\text{hit}} + K(t_\alpha t_{\text{cov}} Q_\alpha^2 C_\alpha^2/\varepsilon^2 + 1/d_{\min}^{\pi_b} t_\alpha^2 Q_\alpha^4 C_\alpha^4/\varepsilon^4)\right).
\end{aligned}$$

Therefore, by combining results, we have $J^\star - J^{\pi_K} \leq (\varepsilon + G)/8$. Lastly, by Lemma D.6, we get $\|J^\star - J^\pi\|_\infty < (5\varepsilon + G)/8$. By the definition of G, if $\varepsilon < G$, then π is an optimal policy. Otherwise, $\|J^\star - J^\pi\|_\infty < \varepsilon$.

□

E Omitted Algorithms in Section C

Algorithm 6: SAMPLE+(d, m)

Input: $d \in \mathbb{R}^{\mathcal{S}}$; $m \in \mathbb{N}$
Initialize: $m_{s,a}, D_{s,a} = 0,\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
for $t = 0, 1, \ldots$ **do**
  $a_t \sim \pi_b(\cdot \mid s_t),\ s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  If $m_{s_t,a_t} < m$ **then**
    $D_{s_t,a_t} \leftarrow \frac{d(s_{t+1})}{m_{s_t,a_t}+1} + \frac{m_{s_t,a_t} D_{s_t,a_t}}{m_{s_t,a_t}+1}$
  EndIf
  $m_{s_t,a_t} \leftarrow m_{s_t,a_t} + 1$
until $m_{s,a} \geq m\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
Output: D

Algorithm 7: SAVIC+(ε, δ)

Input: $\varepsilon > 0,\ \delta \in (0,1)$
Initialize: $n_{\text{esc}}, n_{\text{rec}} = 0,\ n_i = 2^i,\ \delta_i = \delta/(5(i+2)\ln^2(i+2))\ \forall i \in \mathbb{N}$
for $i = 0, 1, \ldots$ **do**
  $\text{SAVIC}(n_{\text{esc}}, n_{\text{rec}}, n_i, \emptyset, \varepsilon, \delta_i)$
  $= (Q^{n_i}, T^{n_i}, \pi^{n_i})$
until $\|Q^{n_i} - T^{n_i}\|_{\text{sp}} \leq 14\,\varepsilon$
Output: $Q^{n_i}, T^{n_i}, \pi^{n_i}$

F AI use statement

In this work, we used generative AI tools for language editing and proofreading. We have not used generative AI tools for the development of the core methodology and theoretical results. We checked all grammar corrections and comments about the correctness of proofs proposed by generative AI. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.