Near-Optimal Sample Complexity for MDPs via Anchoring

Jongmin Lee,[1], Mario Bravo [2], Roberto Cominetti [3]

ICML 2025

[1]Department of Mathematical Sciences, Seoul National University
[2]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez
[3]Institute for Mathematical & Computational Engineering and Department of Industrial and Systems Engineering, Pontificia Universidad Católica de Chile

Average-Reward Markov Decision Process

Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$.

- $\mathcal{S}$, State space
- $\mathcal{A}$, Action space
- $\mathcal{P} \colon \mathcal{S} \times \mathcal{A} \to \mathcal{M}(\mathcal{S})$, Transition probability
- $r \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, Reward
- $\pi \colon \mathcal{S} \to \mathcal{M}(\mathcal{A})$, Policy

Define average-reward of a given policy as

$$g^\pi(s) = \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} r(s_t, a_t) \,|\, s_0 = s \right]$$

and Bellman operator as

$$TV(s) = \sup_{a\in\mathcal{A}} \left\{ r(s,a) + \mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a)} \left[V(s') \right] \right\}.$$

Weakly communicating MDP

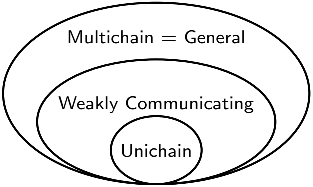

Figure: Unichain ⊂ Weakly Communicating ⊂ Multichain

In weakly communicating MDP,[4] Bellman equation is defined as

$$\max_{a \in \mathcal{A}} \left\{ r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) h(s') \right\} = h(s) + g^{\star}.$$

[4] The MDP is said to be weakly communicating if there is a set of states where each state in the set is accessible from every other state in that set under some policy, plus a possibly empty set of states that are transient for all policies.

Solving MDP with Generative model

Generative model provides independent samples of the next state for any given initial state and action (Reward is known).[5]

In average-reward MDP setup, model-free method exhibits a gap with respect to the lower bound. Furthermore, most methods require a priori bound on the span seminorm of the bias vector $h^\star$.

[5] Kearns & Singh, 1998

Framework I: Anc-VI with span seminorm

The *Anchored Value Iteration* is

$$Q^k = (1 - \beta_k)Q^0 + \beta_k T Q^{k-1} \qquad \text{(Anc-VI)}$$

We call the $(1 - \beta_k)Q^0$ term the *anchor term* since it serves to pull the iterates toward the starting point Q_0.

In weakly communicating MDP, we can show that Anc-VI exhibits

$$\|g^\star - g^{\pi_k}\|_\infty \le \|\mathcal{T}(Q^k) - Q^k\|_{\mathrm{sp}} \le \tfrac{4}{k+1}\|Q^0 - Q^*\|_{\mathrm{sp}}.$$

Framework II: Estimating $\mathcal{T}(Q^k)$ by recursive sampling[6]

To approximate $T^k \approx \mathcal{T}(Q^k)$, one can use naive sampling by collecting samples $\{s_j\}_{j=1}^{m_k} \sim \mathcal{P}(\cdot|s,a)$:

$$T^k(s,a) = r(s,a) + \frac{1}{m_k} \sum_{j=1}^{m_k} \max_{a' \in \mathcal{A}} Q^k(s_j, a').$$

Instead, we use *recursive sampling* by approximating the difference $\mathcal{T}(Q^k) - \mathcal{T}(Q^{k-1})$ and adding it T^{k-1}:

$$T^k(s,a) = T^{k-1}(s,a) + \frac{1}{m_k} \sum_{j=1}^{m_k} (\max_{a' \in \mathcal{A}} Q^k(s_j, a') - \max_{a' \in \mathcal{A}} Q^{k-1}(s_j, a')).$$

[6] Jin et al, 2024, Nguyen et al, 2017

Stochastic Anchored Value Iteration

Algorithm 1 $\text{SAVIA}(Q^0, n, \varepsilon, \delta)$

Input: $Q^0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}\,; n \in \mathbb{N}\,; \varepsilon > 0\,; \delta \in (0,1)$

$\alpha \;= \ln(2|\mathcal{S}||\mathcal{A}|(n{+}1)/\delta)$

$c_k = 5(k+2)\ln^2(k+2)\,;\ \beta_k = k/(k+2)$

$T^{-1} = r\,; h^{-1} = 0$

for $k = 0, \ldots, n$ **do**

$\quad Q^k = (1{-}\beta_k)\,Q^0 + \beta_k\,T^{k-1}$

$\quad h^k = \max_{\mathcal{A}}(Q^k)$

$\quad d^k = h^k - h^{k-1}$

$\quad m_k = \max\{\lceil \alpha\, c_k \|d^k\|^2_{\text{sp}}/\varepsilon^2 \rceil, 1\}$

$\quad D^k = \textsc{sample}(d^k, m_k)$

$\quad T^k = T^{k-1} + D^k$

end for

$\pi^n(s) \in \operatorname{argmax}_{a\in\mathcal{A}} Q^n(s,a) \quad (\forall s \in \mathcal{S})$

Output: (Q^n, T^n, π^n)

SAVIA+

Algorithm 2 $\mathrm{SAVIA+}(Q^0, \varepsilon, \delta)$

Input: $Q^0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}\,; \varepsilon > 0\,; \delta \in (0,1)$
for $i = 0, 1, \ldots$ **do**
  Set $n_i = 2^i, \delta_i = \delta/c_i$.
  $(Q^{n_i}, T^{n_i}, \pi^{n_i}) = \mathrm{SAVIA}(Q^0, n_i, \varepsilon, \delta_i)$
until $\|T^{n_i} - Q^{n_i}\|_{\mathrm{sp}} \leq 14\,\varepsilon$
Output: $Q^{n_i}, T^{n_i}, \pi^{n_i}$

We use doubling trick[7] and stopping rule based on the empirical Bellman error.

[7] Auer et al, 1995; Besson & Kaufmann, 2018

Sample Complexity of SAVIA+

Corollary

Assume $r(s,a) \in [0,1]$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*, and* $\|h^*\|_{\mathrm{sp}} \geq 1$*. Let* (Q^N, T^N, π^N) *be the output of* $\mathrm{SAVIA+}(Q^0, \varepsilon/16, \delta)$ *with* $Q^0 = 0$ *and* $\varepsilon \leq 1$*. Then, with probability at least* $1-\delta$ *we have*

$$\|g^* - g^{\pi_N}\|_\infty \leq \|\mathcal{T}(Q^N) - Q^N\|_{\mathrm{sp}} \leq \varepsilon,$$

with sample and time complexity $\widetilde{O}\big(|\mathcal{S}||\mathcal{A}|\|h^*\|_{\mathrm{sp}}^2/\varepsilon^2\big)$*.*

Summary

Our model-free algorithm $\mathrm{SAVIA+}$ achieve sample and time complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\|h^*\|_{\mathrm{sp}}^2/\varepsilon^2)$ which match the lower bound up to a factor $\|h^*\|_{\mathrm{sp}}$.

To the best of our knowledge, $\mathrm{SAVIA+}$ attains the best complexity among model-free methods, and furthermore, it requries no prior knowledge in weakly communicating MDP.

We also study expected sample complexity and extended this framework to discounted MDPs.