Optimal Non-asymptotic Rates of Value Iteration for Average-Reward MDPs

Jongmin Lee[1], Ernest K.Ryu [2]

ICLR 2025

[1]Department of Mathematical Sciences, Seoul National University
[2]Department of Mathematics, UCLA

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

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

Asymptotic convergence of Value Iteration in multichain MDP

Value Iteration provide asymptotic convergence in multichain MDP[3]. However, the *non-asymptotic* rate in multichain MDP remains unknown.

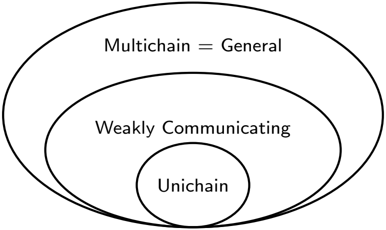

Figure: Unichain $\subset$ Weakly Communicating $\subset$ Multichain

[3] Schweitzer & Federgruen, 1977; Schweitzer & Federgruen, 1979; Puterman, 2014

Anchored Value Iteration

The *Anchored Value Iteration* is

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

We call the $\lambda_k V^0$ term the *anchor term* since it serves to pull the iterates toward the starting point V_0.

Non-asymptotic rates in multichain MDP

Theorem

Consider a general (multichain) MDP. Let $(g^\star, h^\star)$ be a solution of the modified Bellman equations. For $k > K$, the Bellman and policy errors of Anc-VI with $\lambda_k = \frac{2}{k+2}$[4] *exhibits the rate*

$$\|g^\star - g^{\pi_k}\|_\infty \leq \left\|TV^k - V^k - g^\star\right\|_\infty \leq \frac{8}{k+1}\left\|V^0 - h^\star\right\|_\infty + \frac{K}{k+1}\|g^\star\|_\infty,$$

where $K = \left(3\|r\|_\infty + 12\left\|V^0 - h^\star\right\|_\infty + 3\|g^\star\|_\infty\right)/\epsilon$,

$$0 < \epsilon = \inf_{\pi \in S \setminus \{\pi \mid \mathcal{P}^\pi g^\star = g^\star\}} \|\mathcal{P}^\pi g^\star - g^\star\|_\infty,$$

and S is the set of all deterministic policies.

[4] Sabach & Shtern; 2017 Contreras & Cominetti, 2022

Non-asymptotic rates in weakly communicating MDP

Corollary

Consider a general (multichain) MDP satsifying $\mathcal{P}^{\pi} g^{\star} = g^{\star}$ for any policy π. Let $(g^{\star}, h^{\star})$ be a solution of the Bellman equations. For $k \geq 1$, the Bellman and policy errors of Anc-VI with $\lambda_k = \frac{2}{k+2}$ exhibits the rate

$$\|g^{\star} - g^{\pi_k}\|_{\infty} \leq \left\|TV^k - V^k - g^{\star}\right\|_{\infty} \leq \frac{8}{k+1} \left\|V^0 - h^{\star}\right\|_{\infty}.$$

Note that in weakly communicating MDP, $\mathcal{P}^{\pi} g^{\star} = g^{\star}$ for all π.

Complexity lower bound

Theorem

Let $k \geq 0$, $n \geq k+2$, and $V^0 \in \mathbb{R}^n$. Then there exists a unichain MDP with $|\mathcal{S}| = n$ and $|\mathcal{A}| = 1$ such that its Bellman equations has a solution $(g^\star, h^\star)$ satisfying

$$\left\| \sum_{i=0}^{k} a_i (TV^i - V^i) - g^\star \right\|_\infty \geq \frac{1}{k+1} \left\| V^0 - h^\star \right\|_\infty$$

for any iterates $\{V^i\}_{i=0}^k$ satisfying the span condition and any choice of real numbers $\{a_i\}_{i=0}^k$ such that $\sum_{i=0}^k a_i = 1$.

Standard VI and Anc-VI all satisfy the span condition.

Optimality of Anc-VI

By previous two theorems,

$$\underbrace{\frac{1}{k+1}}_{\text{lower bound}} \leq \underbrace{\frac{8}{k+1}}_{\text{upper bound (rate of Anc-VI)}}.$$

This implies that Anc-VI is optimal up to a constant of factor 8 in weakly communicating MDPs.

Summary

Anc-VI first shows $\mathcal{O}(1/k)$ non-asymptotic rates on the Bellman and policy errors for multichain MDP.

Furthermore, Anc-VI is opitmal method up to a constant factor 8 in the weakly communicating and unichain setups.

In the paper, we provided non-asymptotic convergence rate of RX-VI and optimality of normalized iterate.