Dynamic Programming

Shangtong Zhang

University of Virginia

Prediction

Bellman operator

$$\begin{aligned} \mathcal{T}_\pi v &\doteq r_\pi + \gamma P_\pi v, \\ v_{k+1} &\doteq \mathcal{T}_\pi v_k \end{aligned}$$

Sample complexity of applying Bellman operator

$$\left\| \mathcal{T}_\pi^{(n)} v_k - v_\pi \right\|_\infty \leq \gamma^n \| v_0 - v_\pi \|_\infty$$

Control: Value Iteration

Bellman optimality operator

$$\mathcal{T}_* v \doteq \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) v(s') \right\},$$

$$v_{k+1} \doteq \mathcal{T}_* v_k$$

Sample complexity of value iteration

$$\left\| \mathcal{T}_*^{(n)} v_k - v_* \right\|_\infty \leq \gamma^n \| v_0 - v_* \|_\infty$$

Control: Async Value Iteration

$$v_{k+1}(s) \doteq \begin{cases} (\mathcal{T}_* v_k)(s), & \text{if } s = s_k \\ v_k(s), & \text{otherwise} \end{cases}$$

Convergence of Async Value Iteration

When $\mathcal{T}_* v_0 \geq v_0$:

$$\mathcal{T}_* v_k \geq v_0, v_k \geq v_0$$
$$\implies \mathcal{T}_* v_{k+1} \geq v_0, v_{k+1} \geq v_0$$

$$v_{k+1}(s) = \begin{cases} v_k(s) \geq v_0(s) \\ (\mathcal{T}_* v_k)(s) \geq (\mathcal{T}_* v_0)(s) \geq v_0(s) \end{cases}$$

$$\mathcal{T}_* v_{k+1} \geq \mathcal{T}_* v_0 \geq v_0$$

Convergence of Async Value Iteration

Let $k_0 = 0$ and k_m be the first time that each state has been updated at least once after k_{m-1}.

$$
\begin{aligned}
&\forall k \geq k_1, \exists k' < k \quad \text{such that} \\
&v_k(s) = (\mathcal{T}_* v_{k'})(s) \geq (\mathcal{T}_* v_0)(s) \\
&v_k \geq \mathcal{T}_* v_0 \\
&\forall k \geq k_2, \exists k' \in [k_1, k_2) \quad \text{such that} \\
&v_k(s) = (\mathcal{T}_* v_{k'})(s) \geq (\mathcal{T}_* \mathcal{T}_* v_0)(s) \\
&v_k \geq \mathcal{T}_*^{(2)} v_0 \\
&\ldots \\
&\forall k > k_m, v_* \geq v_k \geq \mathcal{T}_*^{(m)} v_0 \\
&v_k \to v_*
\end{aligned}
$$

Convergence of Async Value Iteration

When $\mathcal{T}_* v_0 \leq v_0$:

$$v_k \geq \mathcal{T}_*^{k+1} v_0$$

$$v_{k+1}(s) = \begin{cases} v_k(s) \geq (\mathcal{T}_*^{k+1} v_0)(s) \geq (\mathcal{T}_*^{k+2} v_0)(s) \\ (\mathcal{T}_* v_k)(s) \geq (\mathcal{T}_*^{k+2} v_0)(s) \end{cases}$$

$$v_k \to v_*$$

Convergence of Async Value Iteration

$$
\begin{aligned}
v_0^+ &\doteq v_0 + c\mathbb{1} \\
\mathcal{T}_* v_0^+ &= \mathcal{T}_* v_0 + c\gamma\mathbb{1} \\
v_0^- &\doteq v_0 - c\mathbb{1} \\
\mathcal{T}_* v_0^- &= \mathcal{T}_* v_0 - c\gamma\mathbb{1}
\end{aligned}
$$

For sufficiently large c,

$$
\begin{aligned}
& v_0^+ \geq \mathcal{T}_* v_0^+ \\
& v_0^- \leq \mathcal{T}_* v_0^- \\
& v_k^+ \to v_*, \; v_k^- \to v_*
\end{aligned}
$$

Convergence of Async Value Iteration

$$v_k \leq v_k^+ \implies v_{k+1} \leq v_{k+1}^+$$

$$v_k \geq v_k^- \implies v_{k+1} \geq v_{k+1}^-$$

$$v_k^- \leq v_k \leq v_k^+$$

Policy Improvement Theorem

$$\forall s, \sum_a \pi'(a|s) q_\pi(s,a) \geq v_\pi(s) \implies \forall s, v_{\pi'}(s) \geq v_\pi(s)$$

$$\begin{aligned}
& v_\pi(s) \\
\leq & \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t \sim \pi'(\cdot|S_t)\right] \\
\leq & \mathbb{E}\left[R_{t+1} + \gamma \mathbb{E}\left[R_{t+2} + v_\pi(S_{t+2}) | S_{t+1}, A_{t+1} \sim \pi'\right] | S_t = s, A_t \sim \pi'(\cdot|S_t)\right] \\
\\
\leq & v_{\pi'}(s)
\end{aligned}$$

Control: Policy Iteration

For $k = 0, 1, \ldots$

- Policy evaluation $\pi_k \to v_{\pi_k}$
- Policy Improvement $v_{\pi_k} \to \pi_{k+1}$

$$\lim_{k \to \infty} v_{\pi_k} = v_*$$

Control: Policy Iteration

$$\text{Finite MDP} \quad \Longrightarrow \quad v_{\pi_{k+1}} = v_{\pi_k}$$

$$v_{\pi_{k+1}}(s) = \sum_a \pi_{k+1}(a|s) \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_{\pi_{k+1}}(s') \right)$$

$$v_{\pi_k}(s) = \sum_a \pi_{k+1}(a|s) \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}(s') \right)$$

$$v_{\pi_k}(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}(s') \right)$$

Control: Async Policy Iteration

At $(k+1)$-the iteration, either

$$v_{k+1}(s) = \begin{cases} \left(\mathcal{T}_{\pi_k} v_k\right)(s) & \text{if} \quad s \in \mathcal{S}_k \\ v_k(s) & \text{otherwise} \end{cases}$$

or

$$\pi_{k+1}(s) = \begin{cases} \text{greedy}(v_k)(s) & \text{if} \quad s \in \mathcal{S}_k \\ \pi_k(s) & \text{otherwise} \end{cases}$$

If $\mathcal{T}_{\pi_0} v_0 \geq v_0$, then

$$\lim_{k \to \infty} v_k = v_*.$$

If $\mathcal{T}_{\pi_0} v_0 \geq v_0$ does not hold, no convergence guarantee.

Convergence of Async Policy Iteration

$\mathcal{T}_{\pi_k} v_k \geq v_k \implies \mathcal{T}_{\pi_{k+1}} v_{k+1} \geq v_{k+1} \geq v_k$
In case of value update

$$v_{k+1}(s) = \begin{cases} (\mathcal{T}_{v_k} v_k)(s) \geq v_k(s) & \text{if} \quad s \in \mathcal{S}_k \\ v_k(s) & \text{otherwise} \end{cases}$$

$$\left(\mathcal{T}_{\pi_{k+1}} v_{k+1}\right)(s) = (\mathcal{T}_{\pi_k} v_{k+1})(s) \geq (\mathcal{T}_{\pi_k} v_k)(s) \begin{cases} = v_{k+1}(s) \\ \geq v_k(s) = v_{k+1}(s) \end{cases}$$

Convergence of Async Policy Iteration

$\mathcal{T}_{\pi_k} v_k \geq v_k \implies \mathcal{T}_{\pi_{k+1}} v_{k+1} \geq v_{k+1} \geq v_k$
In case of policy update, for $s \in \mathcal{S}_k$,

$$\begin{aligned}\left(\mathcal{T}_{\pi_{k+1}} v_{k+1}\right)(s) = \left(\mathcal{T}_{\pi_{k+1}} v_k\right)(s) = \left(\mathcal{T}_* v_k\right)(s) \geq \left(\mathcal{T}_{\pi_k} v_k\right)(s) \\ \geq v_k(s) = v_{k+1}(s)\end{aligned}$$

For $s \notin \mathcal{S}_k$,

$$\left(\mathcal{T}_{\pi_{k+1}} v_{k+1}\right)(s) = \left(\mathcal{T}_{\pi_{k+1}} v_k\right)(s) = \left(\mathcal{T}_{\pi_k} v_k\right)(s) \geq v_k(s) = v_{k+1}(s)$$

Convergence of Async Policy Iteration

$v_k \leq v_{k+1},\ \mathcal{T}_* v_k \geq \mathcal{T}_{\pi_k} v_k \geq v_k$

$$\mathcal{T}_{\pi_k}^{(m)} v_k \geq v_k \implies v_{\pi_k} \geq v_k \implies v_* \geq v_k$$

$$\begin{aligned} \lim_{k\to\infty} v_k &= \bar{v}, \\ v_k &\leq \bar{v}, \\ \mathcal{T}_* \bar{v} &\geq \bar{v} \end{aligned}$$

Convergence of Async Policy Iteration

If $\exists s, (\mathcal{T}_* \bar{v})(s) > \bar{v}(s)$, then $\exists \bar{k}$ such that $\forall k \geq \bar{k}$,

$$(\mathcal{T}_* v_k)(s) > \bar{v}(s)$$

Let $k > \bar{k}$ be an iteration where policy update is done for s; let k' be an iteration of the first value update for s after k.

$$\begin{aligned}
v_{k'+1}(s) &= \left(\mathcal{T}_{\pi_{k'}} v_{k'}\right)(s) \\
&\geq \left(\mathcal{T}_{\pi_k} v_k\right)(s) \\
&\geq \left(\mathcal{T}_{\pi_k} v_{k-1}\right)(s) \\
&= \left(\mathcal{T}_* v_{k-1}\right)(s) \\
&> \bar{v}(s)
\end{aligned}$$

Contradiction!

Span Seminorm for Average Reward

$$\mathrm{sp}\left(v\right) = \max_s v(s) - \min_s v(s)$$

Contraction under Span Seminorm

$$\mathrm{sp}\left(P_\pi v\right) \leq \gamma_d \mathrm{sp}\left(v\right)$$

where

$$\gamma_d \doteq 1 - \min_{s,s'} \sum_j \min\left\{P_\pi(s,j), P_\pi(s',j)\right\}$$

Prediction

$$\mathcal{T}_\pi v \doteq r_\pi + P_\pi v$$

$$\lim_{k \to \infty} \mathrm{sp}\left(\mathcal{T}_\pi^{(k)} v - \bar{v}_\pi\right) = 0$$

Control

$$\mathcal{T}_* v \doteq \max_{\pi} \{ r_\pi + P_\pi v \}$$

If there exists an n such that $\forall \pi_1, \pi_2$,

$$\eta(\pi_1, \pi_2) \doteq \min_{s,s'} \sum_j \min \left\{ P^n_{\pi_1}(s, j), P^n_{\pi_2}(s', j) \right\} > 0,$$

then

$$\mathrm{sp} \left(\mathcal{T}^n_* v - \mathcal{T}^n_* v' \right) \leq \gamma' \mathrm{sp} \left(v - v' \right),$$

where

$$\gamma' \doteq 1 - \min_{\pi_1, \pi_2} \{ \eta(\pi_1, \pi_2) \}$$

References

- Markov Decision Processes: Discrete Stochastic Dynamic Programming by Martin Puterman
- Neuro-Dynamic Programming by Dimitri Bertsekas and John Tsitsiklis