CS 7545: Machine Learning Theory October 31, 2018

Lecture 19: Reinforcement Learning

Instructor: Jake Abernethy *Qian Shao, Yuyang Shi*

1 Markov Decision Process (MDP)

A typical MDP has the following elements:

- A set S of states;
- Initial state s_0;
- A set of actions A;
- Transition probability $\Pr[s'|s,a]$; If there is no randomness, then s' is a function of current state s and action a, i.e., $s' = \delta(s,a)$.
- Reward probability $\Pr[r'|s,a]$. If there is no randomness, then r' is a function of current state s and action a, i.e., $r' = r(s,a)$.

A policy $\pi : S \to A$ is a mapping from the state space S to the action space A.
For *finite horizon* setting, under policy π, the overall reward up to time T is

$$\sum_{t=0}^{T} r(s_t, \pi(s_t)).$$

For *infinite horizon* setting, the overall reward under policy π is defined as

$$\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)),$$

where γ is a parameter.

2 Policy Value

For finite horizon setting, the value of a policy π is defined as

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T} r(s_t, \pi(s_t))|s_0 = s\right].$$

For infinite horizon setting, the value of a policy π is

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))|s_0 = s\right].$$

3 Bellman Equation

Theorem 3.1 (Bellman Equation)**.**

$$V_\pi(S) = \mathbb{E}[r(s, \pi(s))] + \gamma \sum_{s'} \Pr[s'|s, \pi(s)] V_\pi(s'), \forall s' \in S.$$

Proof.

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,\pi(s_t))|s_0=s\right] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_{t+1},\pi(s_{t+1}))|s_0=s\right] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \sum_{s'} r(s_{t+1},\pi(s_{t+1}))|s_0=s, s_1=s'\right]\Pr[s_1=s'|s_0=s] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma\mathbb{E}\left[\sum_{s'}\sum_{t=0}^{\infty}\gamma^t r(s_{t+1},\pi(s_{t+1}))|s_1=s'\right]\Pr[s_1=s'|s_0=s] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma\sum_{s'} V_\pi(s')\Pr[s'|s,\pi(s)].
\end{aligned}$$

□

Theorem 3.2. *For a finite MDP, Bellman Equation has a unique solution.*

Sketch of proof. For finite MDP, Bellman Equation can be expressed as

$$V = R + \gamma P V,$$

where P is the transition probability matrix. Since $||P||_\infty = 1$, $||\gamma P||_\infty < 1$ as $\gamma < 1$, thus $I - \gamma P$ is invertible. The unique solution for V is

$$V = (I - \gamma P)^{-1} R.$$

□

4 Optimal Policy

π^* is optimal if it has maximal value of $V_\pi(s), \forall s \in S$, i.e., for any $s \in S$,

$$V_{\pi^*}(s) = \max_{\pi\in\Pi} V_\pi(s).$$

5 Value Iteration

We know transition $\Pr[s'|s,a], r(s,a)$, to optimize value:

1: $V \leftarrow V_0$
2: **while** $||V - \phi_V||_\infty \geq \frac{1-\gamma}{\gamma}\epsilon$ **do**
3: $\quad V \leftarrow \phi_V$, where $\phi_V(s) = \max_{a\in A}(\mathbb{E}[r(s,a)] + \gamma\sum_{s'\in S}\Pr[s'|s,a]V(s'))$
4: **end while**
5: return V

Claim 5.1. *If ϕ is γ-Lipschitz in $||\cdot||_\infty$, $||\phi_U - \phi_V||_\infty \leq \gamma||U - V||_\infty$, where ϕ is from state s, and $a^*(s)$, which is the best action of the state from current knowledge.*

Proof.

$$\begin{aligned}
\phi_V(s) - \phi_U(s) &\leq \phi_V(s) - (\mathbb{E}[r(s, a^*(s))] + \gamma \sum_{s' \in S} \Pr[s'|s, a^*(s)]U(s')) \\
&= \gamma \sum_{s' \in S} \Pr[s'|s, a^*(s)](V(s') - U(s')) \\
&\leq \gamma \sum_{s' \in S} \Pr[s'|s, a^*(s)]||V - U||_\infty \\
&\leq \gamma ||V - U||_\infty
\end{aligned}$$

□

Theorem 5.2. *For any V_0, value iteration converges to V^**

Proof. We know $V* = \phi(V*)$, and $V_{t+1} = \phi(V_t)$

$$\begin{aligned}
||V^* - V_{t+1}||_\infty = ||\phi(V^*) - \phi(V_t)||_\infty &\leq \gamma ||V^* - V_t||_\infty \\
&\leq \gamma^t ||V^* - V_0||_\infty
\end{aligned}$$

Since $\gamma < 1$

$$\lim_{t \to \infty} ||V^* - V_{t+1}||_\infty = 0$$

So we know it will always converge. Now we want to find the bound.

$$\begin{aligned}
||V^* - V_{t+1}||_\infty &= ||\phi(V^*) - \phi(V_{t+1}) + \phi(V_{t+1}) - \phi(V_t)||_\infty \\
&\leq ||\phi(V^*) - \phi(V_{t+1})||_\infty + ||\phi(V_{t+1}) - \phi(V_t)||_\infty \\
&\leq \gamma ||V^* - V_{t+1}||_\infty + \gamma ||V_{t+1} - V_t||_\infty \\
\frac{1-\gamma}{\gamma} ||V^* - V_{t+1}||_\infty &\leq ||V_{t+1} - V_t||_\infty
\end{aligned}$$

Assume $||V^* - V_{t+1}||_\infty \geq \epsilon$

$$\begin{aligned}
\frac{1-\gamma}{\gamma} \epsilon \leq \frac{1-\gamma}{\gamma} ||V^* - V_{t+1}||_\infty &\leq ||V_{t+1} - V_t||_\infty \\
&\leq \gamma^t ||\phi(V_0) - V_0||_\infty
\end{aligned}$$

Set $c = \frac{1-\gamma}{\gamma ||\phi(V_0) - V_0||_\infty}$, since it's not dependent on ϵ

$$\begin{aligned}
c\epsilon &\leq \gamma^t \\
(\frac{1}{\gamma})^t &\leq \frac{1}{c\epsilon} \\
t &\leq O(\log(\frac{1}{\epsilon}))
\end{aligned}$$

□