note

Dongkun Zhang

Jan 2022

1 Markov Decision Process

(Definition) Return *Something*

$$g_{t_0} = \sum_{t=t_0}^{\infty} \gamma^{t-t_0} r(s_t, a_t)$$

(Definition) Index and Goal *the agent's goal is to obtain a policy which maximises the cumulative discounted reward from $t = 0$:*

$$J(\pi, \gamma, p_0, p) = \mathbb{E}[g_0; \pi, \gamma, p_0, p]$$
$$\max_{\pi} J(\pi, \gamma, p_0, p)$$

(Proposition) *Define $\Pr(s \to s', k, \pi)$ as the probability of transitioning from state s to state s' in k steps under policy π*

$$\begin{aligned} J(\pi, \gamma, p_0, p) &= J(\pi, \rho^{\pi}) \\ &= \mathbb{E}_{s \sim \rho^{\pi}(\cdot),\ a \sim \pi(\cdot|s)}[r(s, a)] \\ &= \sum_{s} \rho^{\pi}(s) \sum_{a} \pi(a|s) r(s, a) \end{aligned}$$

where $\rho^{\pi}(s) = \sum_{s_0} p_0(s_0) \sum_{t=0}^{\infty} \gamma^t \Pr(s_0 \to s, t, \pi)$ is the (improper) discounted state distribution.

Proof.
Some definitions:

$$\tau = (s_0, a_0, s_1, a_1, ...)$$

$$
\begin{aligned}
J(\pi, \gamma, p_0, p) &= \mathbb{E}[g_0; \pi, \gamma, p_0, p] \\
&= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t); \pi, p_0, p] \\
&= \mathbb{E}_{\tau \sim p_\tau(\cdot)}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \\
&= \sum_{\tau} p_\tau(\tau)(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)) \\
&= \sum_{s_0} p_0(s_0) \sum_{t=0}^{\infty} \gamma^t \sum_{s} \sum_{a} \pi(a|s) \Pr(s_0 \to s, t, \pi) r(s, a) \\
&= \sum_{s} \rho^\pi(s) \sum_{a} \pi(a|s) r(s, a) \\
&= \mathbb{E}_{s \sim \rho^\pi(\cdot),\ a \sim \pi(\cdot|s)}[r(s, a)]
\end{aligned}
$$

After. ■

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}[g_t | s_t = s; \pi, \gamma, p] \\
Q^\pi(s, a) &= \mathbb{E}[g_t | s_t = s, a_t = a; \pi, \gamma, p]
\end{aligned}
$$

2 Value Function

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a)] \\
Q^\pi(s, a) &= r(s, a) + \mathbb{E}_{s' \sim p(\cdot|s,a)}[V^\pi(s')]
\end{aligned}
$$

$$
\begin{aligned}
V(s) &= V^*(s) = \max_\pi V^\pi(s) \\
Q(s, a) &= Q^*(s, a) = \max_\pi Q^\pi(s, a)
\end{aligned}
$$

$$
V(s) = \max_a Q^\pi(s, a)
$$

Value Function *Something*

$$
\begin{aligned}
Q(s, a) &= r(s, a) + \mathbb{E}_{s' \sim p(\cdot|s,a)}[V(s')] \\
&= r(s, a) + \mathbb{E}_{s' \sim p(\cdot|s,a)}[\max_{a'} Q(s', a')]
\end{aligned}
$$

Proof.
Before

$$
\begin{aligned}
Q(s,a) &= \max_{\pi} Q^{\pi}(s,a) \\
&= r(s,a) + \max_{\pi} \mathbb{E}_{s' \sim p(\cdot|s,a)}[V(s')]
\end{aligned}
$$

After. ■

(Definition) Advantage Function *Something*

$$
\begin{aligned}
A^{\pi}(s,a) &= Q^{\pi}(s,a) - V^{\pi}(s) \\
\mathbb{E}_{a \sim \pi(\cdot|s)}[A^{\pi}(s,a)] &= 0
\end{aligned}
$$

Optimal Advantage Function *Something*

$$
\begin{aligned}
A(s,a) &= Q(s,a) - V(s) \\
A(s,a^*) = 0, \quad a^* &= \arg\max_{a} Q(s,a)
\end{aligned}
$$

Proof.
Before

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^{\pi}(s,a)] \\
&= \mathbb{E}_{a \sim \pi(\cdot|s)}[A^{\pi}(s,a) + V^{\pi}(s)] \\
&= \mathbb{E}_{a \sim \pi(\cdot|s)}[A^{\pi}(s,a)] + V^{\pi}(s)
\end{aligned}
$$

After. ■