Reinforcement Learning Equations

Aaron Hao Tan

Fininte Markov Decision Processes
Components of MDP

$$\{T, S, A_s, p_t(\cdot|s,a), r_t(s,a)\} \tag{1}$$

A state is *Markov* if and only if

$$\mathbb{P}\left[S_{t+1}|S_t\right] = \mathbb{P}\left[S_{t+1}|S_1,\ldots,S_t\right] \tag{2}$$

State-transition probabilities

$$p(s'|s,a) \doteq PrS_t = s'|S_{t-1} = s, A_{t-1} = a = \sum_{r\in\mathcal{R}} p(s',r|s,a) \tag{3}$$

Expected rewards for state-action pairs

$$r(s,a) \doteq \mathbb{E}\left[R_t|S_{t-1}=s, A_{t-1}=a\right] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p\left(s',r|s,a\right) \tag{4}$$

Policy

$$\pi(a|s) = Pr(A_t = a|S_t = s) \tag{5}$$

Returns

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{6}$$

Discounted Returns

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{7}$$

State Value Function and its Bellman Equations

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] \quad \forall s \in S \tag{8}$$

$$\begin{aligned} v_\pi(s) &= E_\pi[R_{t+1} + \gamma \cdot G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)[r + \gamma E_\pi[G_{t+1}|S_{t+1} = s']] \\ &= \sum_a \pi(a|s) \sum_{s',r} p\left(s',r|s,a\right)\left[r + \gamma v_\pi\left(s'\right)\right] \quad \forall s \in S \end{aligned} \tag{9}$$

Action Value Function and its Bellman Equations

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t|S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \tag{10}$$

$$q_\pi(s,a) = \sum_{s',r} p\left(s',r|s,a\right)\left[r + \gamma \sum_{a'} \pi\left(a'|s'\right) q_\pi\left(s',a'\right)\right] \tag{11}$$

Optimal State Value Function and Action Value Function

$$v_*(s) \doteq \max_\pi v_\pi(s) \tag{12}$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a) \tag{13}$$

Relationship between q and v

$$
\begin{aligned}
v_*(s) &= \max_{a \in A(s)} q_*(s, a) \\
&= \max_a q_*(s, a) \\
&= \max_a \mathbb{E}_{\pi_*}[G_t | S_t = s, A_t = a] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_{s', r} p(s', r | s, a)[r + \gamma v_*(s')]
\end{aligned} \tag{14}
$$

Bellman Optimality Equations

$$
\begin{aligned}
v_*(s) &= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*\left(S_{t+1}\right) | S_t = s, A_t = a\right] \\
&= \max_a \sum_{s', r} p\left(s', r | s, a\right)\left[r + \gamma v_*\left(s'\right)\right]
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
q_*(s, a) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*\left(S_{t+1}, a'\right) | S_t = s, A_t = a\right] \\
&= \sum_{s', r} p\left(s', r | s, a\right)\left[r + \gamma \max_{a'} q_*\left(s', a'\right)\right]
\end{aligned} \tag{16}
$$

Policy and Value Iteration