Reinforcement Learning

Ash Bellett

1 Markov processes

1.1 Markov process

A Markov process is described by a finite set of states $\mathcal{S}$ and a state transition probability matrix $P_{s,s'}$ where $s, s' \in \mathcal{S}$. A sequence of random variables $\{S_t \in \mathcal{S}\}$ follow the Markov property: the next state S_{t+1} depends only on the current state S_t:

$$\mathrm{P}(S_{t+1} = s' | S_t = s_t, S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, ...) = \mathrm{P}(S_{t+1} = s' | S_t = s_t)$$

The state transition probability matrix $P_{s,s'}$ gives the probability of transitioning from state s to state s':

$$P_{s,s'} = \mathrm{P}(S_{t+1} = s' | S_t = s)$$

1.2 Markov reward process

A Markov reward process is described by a finite set of states $\mathcal{S}$, a finite set of rewards $\mathcal{R} \subset \mathbb{R}$, a state transition probability matrix $P_{s,s'}$, a function $r : \mathcal{S} \to \mathcal{R}$ and a discount factor $\gamma \in [0, 1]$. The function r maps a state $s \in \mathcal{S}$ to a reward $R_t \in \mathcal{R}$:

$$R_t = r(s | S_t = s)$$

The next expected reward R_s at state s is defined as:

$$\begin{aligned} R_s &= \mathrm{E}(R_{t+1} | S_t = s) \\ &= \sum_{s' \in \mathcal{S}} r(s') P_{s,s'} \end{aligned}$$

The return $G_t \in \mathbb{R}$ is a sum of attenuated, future rewards represented as a geometric series:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \end{aligned}$$

The return can be rewritten as:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + ...) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

The state value function $v : S \to \mathbb{R}$ is the expected return at state s:

$$\begin{aligned} v(s) &= \mathrm{E}(G_t | S_t = s) \\ &= \mathrm{E}(R_{t+1} + \gamma G_{t+1} | S_t = s) \end{aligned}$$

From the law of total expectation, $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y))$:

$$\begin{aligned} \mathrm{E}(G_{t+1}) &= \mathrm{E}(\mathrm{E}(G_{t+1}|S_{t+1} = s')) \\ &= \mathrm{E}(v(s')) \end{aligned}$$

The Bellman equation for the state value function $v(s)$ is:

$$\begin{aligned} v(s) &= \mathrm{E}(R_{t+1} + \gamma v(s')|S_t = s) \\ &= \sum_{s' \in \mathcal{S}} P_{s,s'} r(s') + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'} v(s') \\ &= R_s + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'} v(s') \end{aligned}$$

1.3 Markov decision process

A Markov decision process is described by a finite set of states $\mathcal{S}$, a finite set of rewards $\mathcal{R} \subset \mathbb{R}$, a finite set of actions $\mathcal{A}$, a state transition probability matrix $P^a_{s,s'}$, a policy $\pi \in [0,1]$, a function $r : \mathcal{S} \to \mathcal{R}$ and a discount factor $\gamma \in [0,1]$. The state transition probability matrix $P^a_{s,s'}$ gives the probabilities of transitioning from state s to state s' if action $a \in \mathcal{A}$ is taken:

$$P^a_{s,s'} = \mathrm{P}(S_{t+1} = s'|S_t = s, A_t = a)$$

The policy π maps a state s to the probability of selecting action a:

$$\pi(a|s) = \mathrm{P}(A_t = a|S_t = s)$$

The action A_t is sampled from the policy:

$$A_t \sim \pi(\cdot|S_t = s)$$

While following policy π, the state transition probability matrix $P^\pi_{s,s'}$ is:

$$P^\pi_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) P^a_{s,s'}$$

The next expected reward R^a_s at state s given action a is taken is defined as:

$$\begin{aligned} R^a_s &= \mathrm{E}(R_{t+1}|S_t = s, A_t = a) \\ &= \sum_{s' \in \mathcal{S}} r(s') P^a_{s,s'} \end{aligned}$$

While following policy π, the next expected reward R^π_s is:

$$R^\pi_s = \sum_{a \in \mathcal{A}} \pi(a|s) R^a_s$$

The state value function $v_\pi : S \to \mathbb{R}$ is the expected return at state s under policy π:

$$\begin{aligned} v_\pi(s) &= \mathrm{E}_\pi(G_t | S_t = s) \\ &= \mathrm{E}_\pi(R_{t+1} + \gamma G_{t+1} | S_t = s) \end{aligned}$$

From the law of total expectation, $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y))$:

$$\begin{aligned} \mathrm{E}_\pi(G_{t+1}) &= \mathrm{E}_\pi(\mathrm{E}_\pi(G_{t+1} | S_{t+1} = s')) \\ &= \mathrm{E}_\pi(v_\pi(s')) \end{aligned}$$

The Bellman equation for the state value function $v_\pi(s)$ under policy π is:

$$\begin{aligned} v_\pi(s) &= \mathrm{E}_\pi(R_{t+1} + \gamma v_\pi(s') | S_t = s) \\ &= \mathrm{E}_\pi(R_{t+1} | S_t = s) + \gamma \mathrm{E}_\pi(v_\pi(s') | S_t = s) \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) R_s^a + \gamma \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s') \\ &= R_s^\pi + \gamma \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s') \end{aligned}$$

The action-value function $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected return at state s, taking action a and following policy π:

$$\begin{aligned} q_\pi(s,a) &= \mathrm{E}_\pi(G_t | S_t = s, A_t = a) \\ &= \mathrm{E}_\pi(R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a) \end{aligned}$$

From the law of total expectation, $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y,Z))$:

$$\begin{aligned} \mathrm{E}_\pi(G_{t+1}) &= \mathrm{E}_\pi(\mathrm{E}_\pi(G_{t+1} | S_{t+1} = s', A_{t+1} = a')) \\ &= \mathrm{E}_\pi(q_\pi(s',a')) \end{aligned}$$

The Bellman equation for the action-state value function $q_\pi(s,a)$ under policy π is:

$$\begin{aligned} q_\pi(s,a) &= \mathrm{E}_\pi(R_{t+1} + \gamma q_\pi(s',a') | S_t = s, A_t = a) \\ &= \mathrm{E}_\pi(R_{t+1} | S_t = s, A_t = a) + \gamma \mathrm{E}_\pi(q_\pi(s',a') | S_t = s, A_t = a) \\ &= R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a') \end{aligned}$$

The state value function v_π can be written in terms of the action-state value function q_π:

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$

Similarly, the action-state value function q_π can be written in terms of the state value function v_π:

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s)$$