Róbert Leó Jónsson (robertt20@ru.is))

RL notes

Markov Decision Processes (MDP)

The at time t the state is $S_t \in \mathcal{S}$, the action is $A_t \in \mathcal{A}$, and the reward (received before seeing the state and doing the action) is $R_t \in \mathcal{R}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\mathcal{R} \subset \mathbb{R}$ is the reward space.

The **trajectory** is

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots \tag{1}$$

The probability of getting to state s' and gettign reward r after taking action a in state s is well defined, and given by

$$p(s', r \mid s, a) = \Pr\left\{S_t = s', R_t = r \mid S_{t-1}, A_{t-1} = a\right\} \tag{2}$$

where the function p defines **dynamics** of the MDP, and is called the **dynamics function**. The probabilities given p completely characterize the environment's dynamics. This also confirms that an MDP has the Markov property.

We can obtain the **state-transition probabilities** with

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

And the expected rewards for a state-action pair:

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

and the expected reward given the next state:

$$r(s, a, s') = \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'\right] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)} \tag{5}$$

Returns and episodes

The **return** is the discounted sum of rewards (if we are using discounting).

In an episodic task, this is

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{T-t-1} R_T \tag{6}$$

where T is the final time step. We think of each episode ending in the **same** terminal state - can be though of as an artificial state that occurs right *after* the real terminal state of the episode.

The final reward is given in this final terminal state.

For continuing tasks, the return is

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

If we define the reward to be zero after the final state, this also holds for episodic tasks.

Theree is a recursive relationship between G_t and G_{t+1}:

$$
\begin{align}
G_t &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{8}\\
&= R_t + \sum_{k=1}^{\infty} \gamma^k R_{t+k+1} \tag{9}\\
&= R_t + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} \tag{10}\\
&= R_t + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \tag{11}\\
&= R_t + \gamma G_{t+1} \tag{12}\\
& \tag{13}
\end{align}
$$

Policies and value function

The **value** function is defined to be

$$
\begin{align}
v_\pi(s) &= \mathbb{E}_\pi \left[G_t \mid S_t = s \right] \tag{14}\\
&= \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t + s \right] \tag{15}
\end{align}
$$

and the **action-value** function is

$$
\begin{align}
q_\pi(s, a) &= \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right] \tag{16}\\
&= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a \right] \tag{17}\\
& \tag{18}
\end{align}
$$

The **Bellman equation** for the value function is

$$
\begin{align}
v_\pi(s) &= \mathbb{E}_\pi \left[G_t \mid S_t = s \right] \tag{19}\\
&= \mathbb{E}_\pi \left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right] \tag{20}\\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma \mathbb{E}_\pi \left[G_{t+1} \mid S_{t+1} = s' \right] \right] \tag{21}\\
&= \sum_a \pi(a \mid s) \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right] \tag{22}\\
& \tag{23}
\end{align}
$$

And the Bellman equation for the action-value function is

$$
\begin{aligned}
q(s,a) &= \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a\right] && (24)\\
&= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a && (25)\\
&= \sum_{s'} \sum_{r} p(s', r \mid s, a)(r + \gamma \mathbb{E}_\pi \left[G_{t+1} \mid S_{t+1} = s'\right] && (26)\\
&= \sum_{s', r} p(s', r \mid s, a)(r + \gamma \sum_{a'} \pi(a' \mid s') \mathbb{E}\left[G_{t+1} \mid s', a'\right]) && (27)\\
&= \sum_{s', r} p(s', r \mid s, a)\Big(r + \gamma \sum_{a'} \pi(a' \mid s) q(a', s')\Big) && (28)\\
& && (29)
\end{aligned}
$$

We can write v_π in terms of q_π :

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left[q_\pi(a, s) \mid S_t = s\right] && (30)\\
&= \sum_a \pi(a \mid s) q_\pi(a, s) && (31)\\
& && (32)
\end{aligned}
$$

Or we could write q_π in terms of v_π:

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right] && (33)\\
&= \sum_{s', r'} p(s', r \mid s, a)(r + \gamma v_\pi(s')) && (34)\\
& && (35)
\end{aligned}
$$

Optimal policies and optimal value function

The **optimal policies** are denoted π_*, and they define the optimal value function

$$v_*(s) = \max_\pi v_\pi(s) = v_{\pi_*}(s), \text{ for all } s \in \mathcal{S} \tag{36}$$

They also define the optimal action-value function

$$q_*(s,a) = \max_\pi q_\pi(s,a) = q_{\pi_*}(s,a), \tag{37}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$.

$q_*(s,a)$ follows an optimal policy after the action a, so we have

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right] \tag{38}$$

Now, the **Bellman optimality equation** for v_* is

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) && (39)\\
&= \max_a \mathbb{E}_{\pi_*} \left[G_t \mid S_t = s, A_t = a\right] && (40)\\
&= \max_a \mathbb{E}_{\pi_*} \left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] && (41)\\
&= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right] && (42)\\
&= \max_a \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_*(s')\right] && (43)\\
& && (44)
\end{aligned}
$$

The corresponding equation for q_* is

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\middle|\; S_t = s, A_t = a\right] \tag{45}$$

$$= \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right] \tag{46}$$