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}:

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

Policies and value function

The **value** function is defined to be

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

and the **action-value** function is

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

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

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

And the Bellman equation for the action-value function is

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

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

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right] && (45) \\ &= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right] && (46) \end{aligned}$$

Policy evaluation

If the environment's dynamics are completely known, we can approximate v_π by starting with an arbitrary value function v_0 (but with the value of the terminal state equal to zero), and continuously perform the following iteration:

$$\begin{aligned} v_{k+1} &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s\right] && (47) \\ &= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_k(s')\right] && (48) \end{aligned}$$

for all $s \in \mathcal{S}$. In this case, $v_k \to v_\pi$ as $k \to \infty$. This is called **iterative policy evaluation**.

The analogous iteration for $q_\pi(s,a)$ is

$$\begin{aligned} q_{k+1}(s,a) &= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] && (49) \\ &= \sum_{s',r} p(s', r \mid s, a)\left(r + \gamma \sum_{a'} \pi(a' \mid s) q_k(a', s)\right) && (50) \\ & && (51) \end{aligned}$$

Policy Improvement

If we have a deterministic policy π and its value function v_π, then suppose π' is such that in state s, we choose the next action a greedily with respect to v_π, then the value of this behavior is

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1} \mid S_t = s, A_t = a\right] && (52) \\ &= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right] && (53) \\ & && (54) \end{aligned}$$

Then π' is better than π overall.

The **policy improvement theorem** states that if π and π' is a pair of deterministic policies such that

$$\forall s \in \mathcal{S} : q_\pi(s, \pi'(s)) \geq v_\pi(s) \tag{55}$$

Then π' is as good or better than π, that is

$$\forall s \in \mathcal{S} : v_{\pi'}(s) \geq v_\pi(s) \tag{56}$$

And if there is a strict inequality in (55) in s, then there is also a strict inequality in (56), in s.

Now, if we have a policy π, and $q_\pi(s,a)$, then we can construct the new *greedy* policy

$$\begin{align}
\pi'(s) &= \operatorname{argmax}_a q_\pi(s,a) \tag{57}\\
&= \operatorname{argmax}_a \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1} \mid S_t = s, A_t = a\right] \tag{58}\\
&= \operatorname{argmax}_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right] \tag{59}\\
& \tag{60}
\end{align}$$

This policy satisfies (55), so it is as good as or better than πi. This process is called *policy improvement*. If π' is not better than π but exactly as good, then we have

$$\begin{align}
v_{\pi'}(s) &= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a\right] \tag{61}\\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_{\pi'}(s')\right] \tag{62}\\
& \tag{63}
\end{align}$$

which is the Bellman optimality equation, so $v_{\pi'} = v_*$, and π and π' are optimal policies.

For stochastic policies, this also works - as long as wee assign a zero probability to non-optimal actions, all probability assignments are allowed.

Policy Iteration

Policy iteration is the process of taking a policy, calculating its value function, improving the policy, calculating the value function, and so on:

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} v_* \tag{64}$$

Value Iteration

The **value iteration** algorithm can be written as a simple update operation that combines the policy improvement and truncated policy evaluation steps:

$$\begin{align}
v_{k+1}(s) &= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a\right] \tag{65}\\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_k(s')\right] \tag{66}\\
& \tag{67}
\end{align}$$

for all $s \in \mathcal{S}$.

Off-policy prediction via Importance Sampling

Given a starting state S_t, the probability of the subsequent state-action trajectory $A_t, S_{t+1}, A_{t+1}, \ldots, S_T$ occurring under any policy π is

$$P(A, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T_1} \sim \pi) \tag{68}$$
$$= \pi(A_t \mid S_t) p(S_{t+1} \mid S_t, A_t) \pi(S_{t+1} \mid S_{t+1}) \cdots p(S_T \mid S_{T-1}, A_{T-1}) \tag{69}$$
$$= \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k) \tag{70}$$

This gives the importance-sampling ratio:

$$\rho_{t:T-1} = \frac{\prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)} \tag{71}$$
$$= \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)} \tag{72}$$

Thus, if we want to approximate v_π, but we only have episodes from the policy b, we can fix the expected value of the return with

$$\mathbb{E}\left[\rho_{t:T-1} G_t \mid S_t = s\right] = v_\pi(s) \tag{73}$$

To make the estimate, we can use either **ordinary importance sampling**, with

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|} \tag{74}$$

where $\mathcal{T}(s)$ is the set of time points where $S_t = s$ (only the first state in each episode if we are using first-visit sampling), and $T(t)$ is the end time of the episode.

We can also use **weighted importance sampling**, with

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}} \tag{75}$$

For action-values, this becomes

$$Q(s, a) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t+1:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t+1:T(t)-1}} \tag{76}$$

or

$$Q(s, a) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t+1:T(t)-1} G_t}{|\mathcal{T}(s)|} \tag{77}$$

for the ordinary importance sampling case.