Reinforcement Learning

Let’s start a study for reinforcement learning by sutton.

3. Finite Markov Decision Process

3.1 The Agent-Environment Interface

Agent

The learner and decision maker is called the agent.

Environment

The thing it interacts with, comprising everything outside the agent, is called the environment.

Agent는 action들을 선택하고 Environment는 이러한 action들에 반응하고 agent에게 새로운 state를 나타내면서 계속해서 상호작용한다.

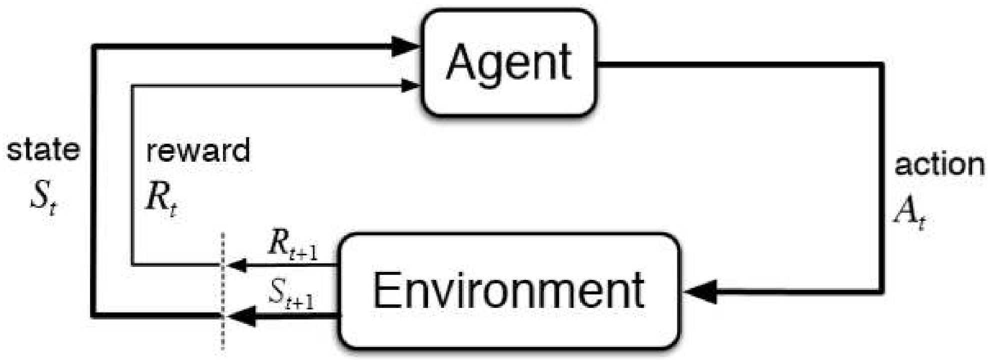

Figure 3.1: The agent–environment interaction in a Markov decision process.

Return G_t

The return G_t is the total discounted reward from time-step t

$$G_t = r_{t+1} + \gamma r_{t+2} + +\gamma^2 r_{t+3} \cdots$$
$$= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

3.5 Policies and Value Functions

Value Function

Value Functions estimate how good it is for the agent to be in a given state (or how good it is to perform a given action in a given state).

Policy

A policy is a mapping from states to probabilities of selecting each possible action.
If the agent is following policy π at time t

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

Reinforcement learning methods specify how the agent's policy is changed as a result of its experience.
(강화학습은 어떻게 Agent의 policy가 그 실험의 결과로써 변화하는지 명시한다.)

The state-value function for policy π : $v_\pi(s)$

- The expected return when starting in s and following policy π thereafter.
($v_\pi(s)$는 상태 s에서 시작하여 그 후 policy π를 따를 때 기대되는 반환 값이다.)

- For MDPs, we can define $v_\pi(s)$ formally by

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

The action-value function for policy π : $q_\pi(s, a)$

- The expected return starting from s, taking the action a, and thereafter following policy π:
($q_\pi(s, a)$ 는 상태 s에서 시작하여 액션 a를 취한 후에 정책 π를 따를 때 기대되는 반환 값이다.)

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

Bellman equation for v_π

- It expresses a relationship between the value of a state and the values of its successor states.
(이것은 한 상태의 가치와 그 후속 상태의 가치들과의 관계를 나타낸다.)

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma(R_{t+2}+\gamma R_{t+3}+\cdots)|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma G_{t+1}|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}|S_t=s]+\mathbb{E}_\pi[\gamma G_{t+1}|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}|S_t=s]+\mathbb{E}_\pi[\gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s']|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s']|S_t=s] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma v_\pi(s')|S_t=s] \\
&= \sum_r p(r|s)\big[r+\gamma v_\pi(s')\big] \\
&= \sum_{s'}\sum_r p(s',r|s)\big[r+\gamma v_\pi(s')\big] \\
&= \sum_a \pi(a|s)\sum_{s'}\sum_r p(s',r|s,a)\big[r+\gamma v_\pi(s')\big] \\
&= \sum_a \pi(a|s)\sum_{s',r} p(s',r|s,a)\big[r+\gamma v_\pi(s')\big]
\end{aligned}
$$

Bellman equation for q_π

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[G_t|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma(R_{t+2}+\gamma R_{t+3}+\cdots)|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}|S_t=s,A_t=a]+\mathbb{E}_\pi[\gamma G_{t+1}|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}|S_t=s,A_t=a]+\mathbb{E}_\pi[\gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s',A_{t+1}=a']|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s',A_{t+1}=a']|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma q_\pi(s',a')|S_t=s,A_t=a] \\
&= \sum_r p(r|s,a)\big[r+\gamma q_\pi(s',a')\big] \\
&= \sum_{s'}\sum_r p(s',r|s,a)\big[r+\gamma q_\pi(s',a')\big] \\
&= \sum_{s',r} p(s',r|s,a)\big[r+\gamma q_\pi(s',a')\big]
\end{aligned}
$$

3.6 Optimal Policies and Optimal Value Functions.

Optimal state-value function

$$
v_*(s) = \max_\pi v_\pi(s)
$$

for all $s \in S$

Optimal action-value function

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

for all $s \in S$ and $a \in A(s)$

For the state-action pair (s,a), this function gives the expected return for taking action a in state s and thereafter following an optimal policy.
(state-action (s,a)에 대해 이 함수는 상태 s에서 액션 a)를 취하고 그 후에 최적의 정책을 따르는 것에 대한 기대되는 반환 값을 준다.)

$\therefore$

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

Bellman optimality equation (Bellman equation for v_*)

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