Bellman equation

24.01.11

정상혁

Value functions

- State-value function

$v_\pi(s)$

: state 's' 가 **얼마나 좋은지**를 나타내는 함수.

- Action-value function

$q_\pi(s, a)$

: state 's' 에서 action 'a'를 선택하는 것이 **얼마나 좋은지**를 나타내는 함수

'**좋다**' 의 기준

[Reward Hypothesis]

All goals can be described by the **maximization** of the **expected value** of the <u>cumulative sum of rewards</u> (=return)

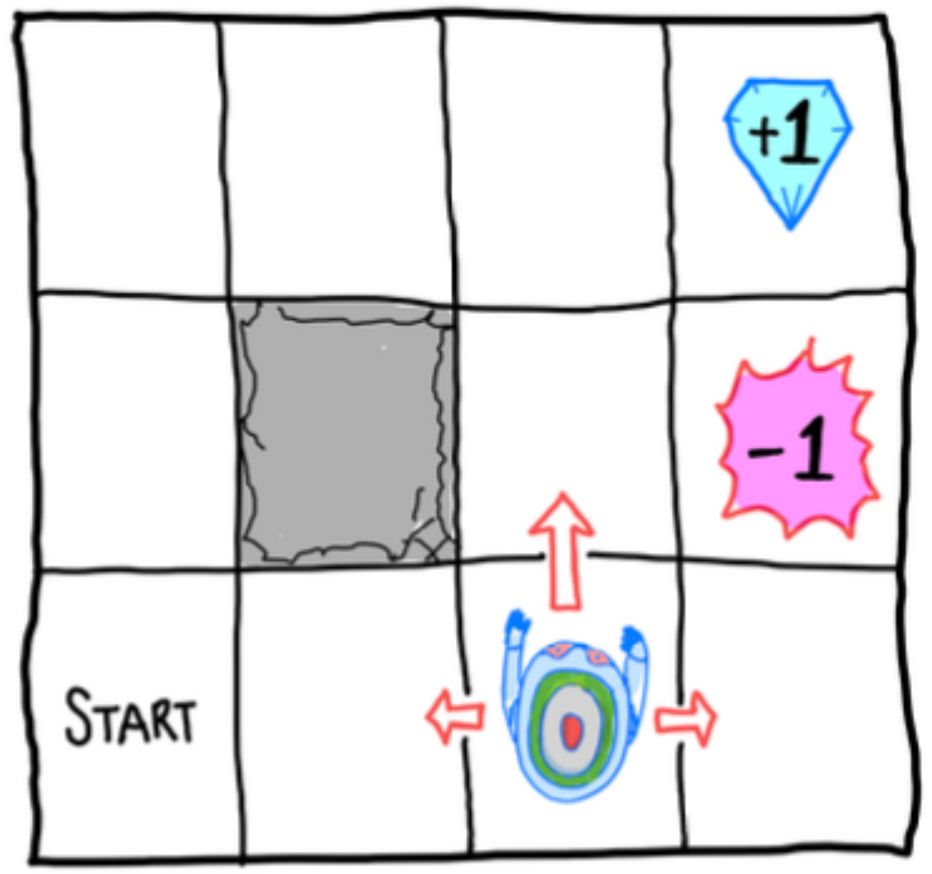
Value functions

- State-value function $v_\pi(s) = E_\pi \left[G_t \mid S_t = s \right]$

- Action-value function $q_\pi(s,a) = E_\pi \left[G_t \mid S_t = s, A_t = a \right]$

Advantage function

$$A_\pi(s,a) = q_\pi(s,a) \; - v_\pi(s)$$



Episode $(1,1) \xrightarrow{\text{north}} (1,2) \xrightarrow[\text{wall}]{\text{east}} (1,2) \xrightarrow{\text{north}} (1,3) \xrightarrow[\text{east } 10\%]{\text{south}} (2,3) \xrightarrow{\text{east}} (3,3) \xrightarrow{\text{east}} (4,3)$

[Law of total probability]

- $P(A) = \sum_n P(A \mid B_n) \cdot P(B_n)$
- $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$
- $\mathbb{E}[X \mid Z = z] = \sum_y \mathbb{E}[X \mid Y = y, Z = z] \cdot P(Y = y \mid Z = z)$ $\longrightarrow$

EX)

$$v_\pi(s) = E_\pi \left[G_t \mid S_t = s \right]$$

$$q_\pi(s,a) = E_\pi \left[G_t \mid S_t = s, A_t = a \right]$$

$$\downarrow$$

$$v_\pi(s) = \sum_a \pi(a \mid s)\, q_\pi(s,a)$$

[Law of large numbers]

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

$X : X_1, \ldots, X_n$ i.i.d random samples (independent and identically distributed)

$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) \rightarrow \mathbb{E}[X]$ as $n \rightarrow \infty$

EX)

$$\frac{1}{n}(G_1 + \cdots + G_n) \longrightarrow q_\pi(s,a) = E_\pi \left[G_t \mid S_t = s, A_t = a \right] \quad \text{as } n \rightarrow \infty$$

Optimal value functions and policy

- Optimal state-value function $v_*(s) = \max_\pi v_\pi(s)$

- Optimal action-value function $q_*(s,a) = \max_\pi q_\pi(s,a)$

- Optimal policy $\pi_*(s)$ ($\pi_*(s) \geq \pi$ for all π)

(Define a partial policies $\pi' \geq \pi$ if $v_{\pi'}(s) \geq v_\pi(s)$ for all s)

[Theorem] Any MDP satisfies the followings.

- There exists an optimal policy $\pi_* \geq \pi$ for all π.
- All optimal policies achieve the optimal state-value function $\boxed{v_{\pi_*}(s) = v_*(s).}$
- All optimal policies achieve the optimal action-value funct. $\boxed{q_{\pi_*}(s,a) = q_*(s,a).}$

Optimal value functions and policy

- Optimal action-value function 을 알면 Optimal policy 를 구할 수 있다.

$$\pi_*(s) = arg\max_a q_*(s,a) \qquad \pi_*(a \mid s) = \begin{cases} 1 & if \;\; a = arg\max_a q_*(s,a) \\ 0 & \text{otherwise} \end{cases}$$

- Optimal state-value function 을 알면 Optimal policy 를 구할 수 있을까 ?

$$v_*(s) = \max_a q_*(s,a) \qquad q_*(s,a) = r(s,a) \; + \; \gamma \sum_{s'} \boxed{p(s' \mid s,a)} \, v_*(s')$$

Known MDP

$$\boxed{v_*(s)} \underset{O}{\overset{O}{\rightleftarrows}} q_*(s,a) \xrightarrow{O} \pi_*(s)$$

Unknown MDP

$$v_*(s) \underset{O}{\overset{X}{\rightleftarrows}} \boxed{q_*(s,a)} \xrightarrow{O} \pi_*(s)$$

Bellman equation

Bellman expectation equation **(Markov property)**

$$v_\pi(s) = E_\pi\ [G_t \mid S_t = s]$$
$$= E_\pi\ [R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

$$q_\pi(s,a) = E_\pi\ [G_t \mid S_t = s, A_t = a]$$
$$= E_\pi\ [R_{t+1} + \gamma\ q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

Bellman optimality equation

$$v_*(s) = \max_\pi v_\pi(s)$$
$$= \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$$

$$q_*(s) = \max_\pi q_\pi(s,a)$$
$$= E[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a]$$
$$= \sum_{s',r} p(s', r \mid s, a)[r + \gamma \max_{a'} q_*(s', a')]$$

Bellman equation

Known MDP

$$\boxed{v_*(s)} \underset{O}{\overset{O}{\rightleftarrows}} q_*(s,a) \xrightarrow{O} \pi_*(s)$$

Unknown MDP

$$v_*(s) \underset{O}{\overset{X}{\rightleftarrows}} \boxed{q_*(s,a)} \xrightarrow{O} \pi_*(s)$$

Dynamic Programming 사용 ← Tabular updating method → **Reinforcement Learning 사용**

- $V(s)$ (State-value function table)을 반복해서 update 하며, optimal policy를 구한다.
- 이때, $v_*(s)$ 를 구하기 위해서, Bellman equation 을 이용한다.

- $Q(s,a)$ (Action-value function table)을 반복해서 update 하며, optimal policy를 구한다.
- 이때, $q_*(s,a)$ 를 근사하기 위해서. random samples, Bellman equation 을 이용한다.