Reinforcement Learning Ch.3

2018.10.02
발표자 장유환

CONTENTS

01

MDP
MDP Constituents

02

State-value Func.
Action-value Func.
Bellman Expect Eqn.

03

Optimal Policy
Optimal Value Func.
Bellman Optimal Eqn.

01. Markov Decision Process

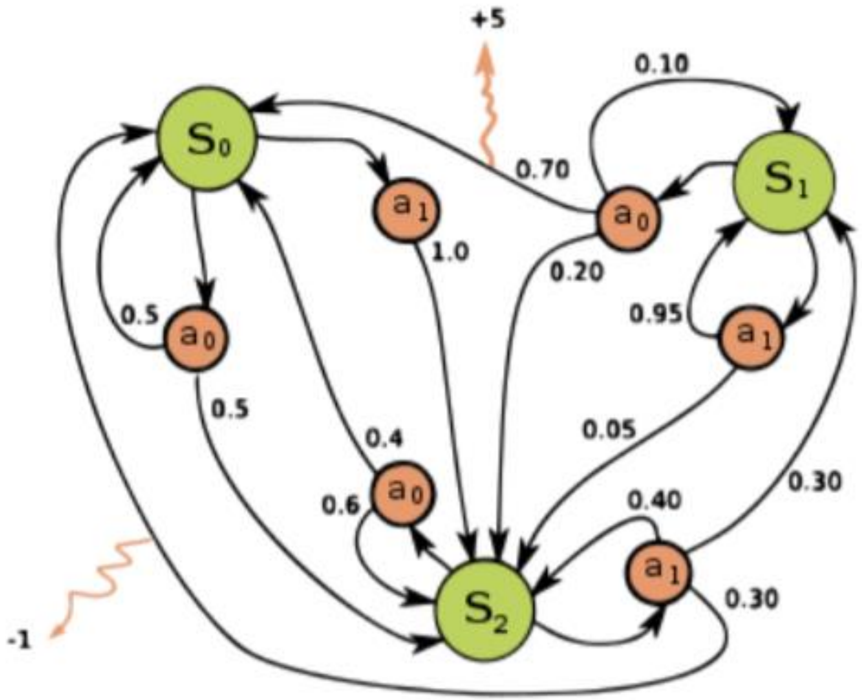
- **시간에 따라 상태 변화 / 상태 공간 안에서 움직이는 Agent**

Agent의 action 선택

action에 따른 다음 state, reward

>>확률적으로 Modelize : MDP

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$



01. MDP의 구성 요소

- **MDP는 5가지 요소로 구성**

$$\mathrm{MDP} = \{S,\ A,\ R,\ P_{ss'}^{a},\ \gamma\}$$

S : State (상태)

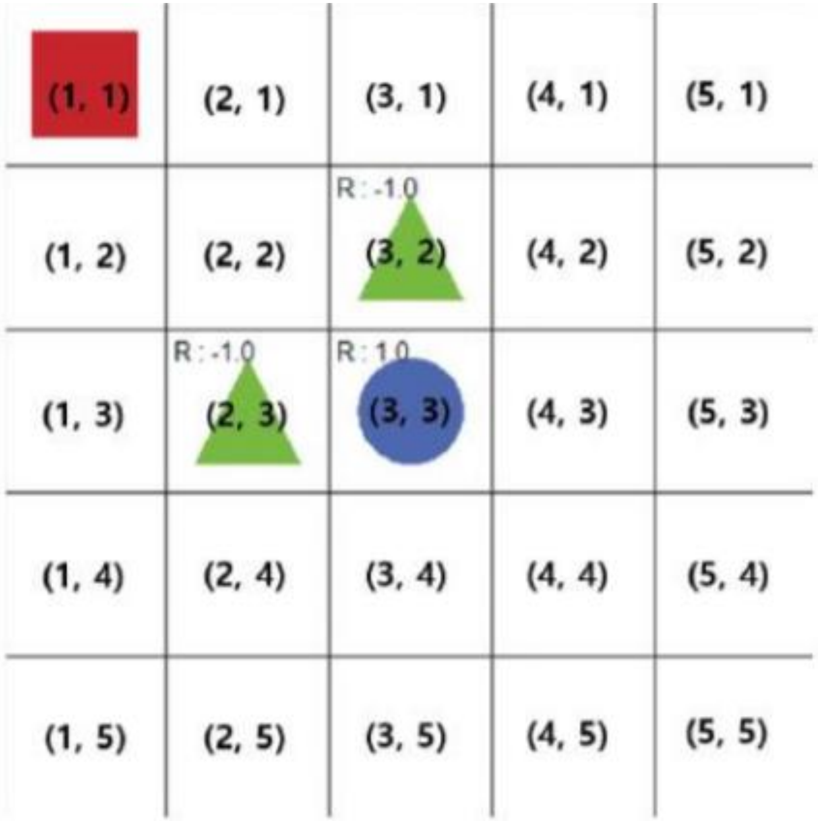
A : Action (행동)

R : Reward (보상)

$P_{ss'}^{a}$: State Transition Probaility (상태 변환 확률)

γ : Discount Factor (할인율)

01-1. **State (상태)**

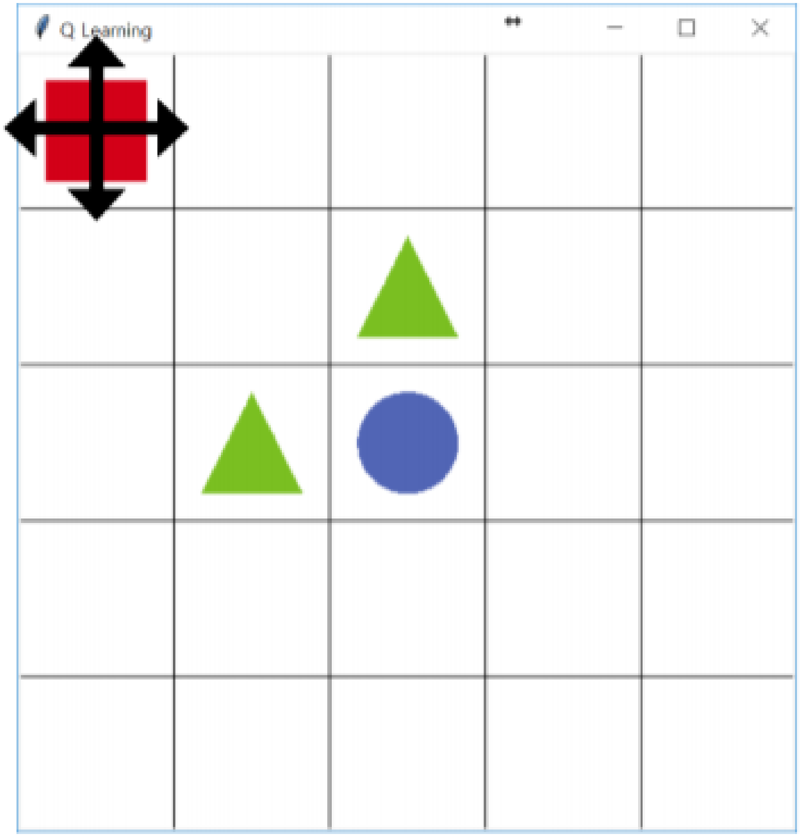


Agent에서 관찰되는 모든 상태 집합

$$\boldsymbol{S} = \{(1,1),(2,1),(1,2),\ \cdots\ ,\ (5,5)\}$$

특정 시간 t에서의 상태 : $S_t = (1,3)$ 와 같이 표현

01-2. **Action (행동)**



Agent가 할 수 있는 행동의 집합

A = { 상, 하, 좌, 우 }

특정 시간 t에서의 행동 : $A_t = a$ 와 같이 표현

확률에 따라 Action이 달라진다? → 상태 변환 확률

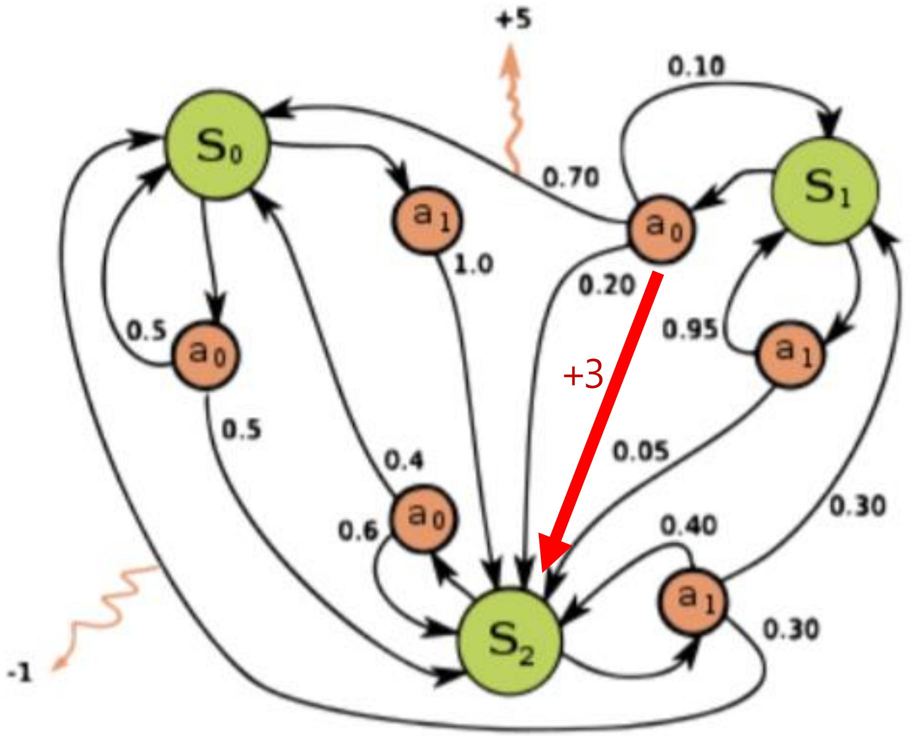
01-3. State Transition Probability (상태 변환 확률)

- t-1 시점에서의 상태는 s / 이때 취한 행동은 a
- t 시점에서 상태 s'로 변화 / reward는 r

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

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s).$$

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



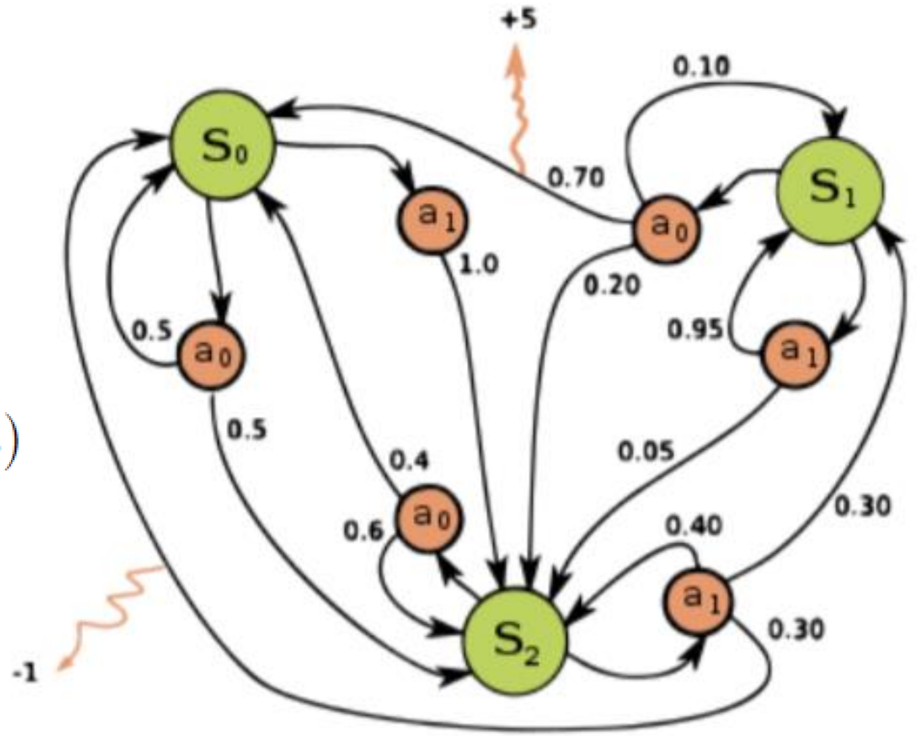
01-3. Expected Reward (보상의 기댓값)

- **t-1 시점에서의 상태는 s / 이때 취한 행동은 a**
- **이때 t 시점에서 얻을 수 있는 보상 R_t 의 기댓값**

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

- 계산 예시

  r(s1, a0) = (0 * (-1)) + (0.1 * 0) + (0.2 * 0) + (0.7 * 5)

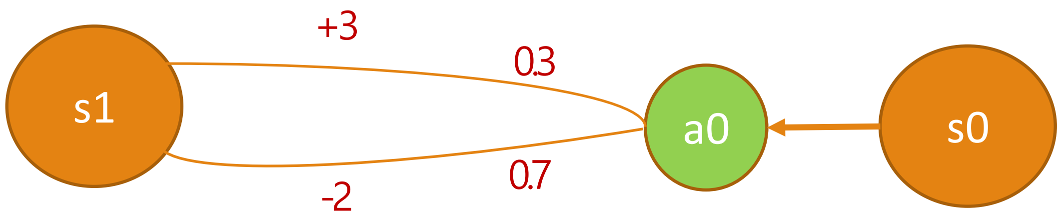


01-3. Expected Reward (보상의 기댓값)

- **t-1 시점에서의 상태는 s / 이때 취한 행동은 a / 다음 상태는 s′**
- **이때 t 시점에서 얻을 수 있는 보상 R_t 의 기댓값**

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

- 계산 예시 r(s0, a0, s1) = (0.3 * 3) + (0.7 * (-2))



01-4. Discount Factor (할인율)

- **미래에 받을 보상은 가치 ↓, 가까운 보상에 대해 가치 ↑**
- 0~1 사이의 실수 γ로 나타냄

$$R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

$$\Downarrow$$

$$R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

02-1. 반환값 (Return value)

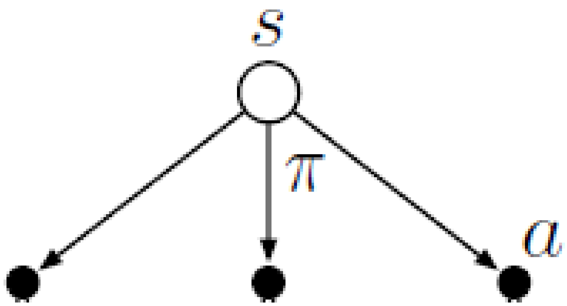
- 할인율 적용 시 앞으로 얻을 Reward의 합 : Return value

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

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

02-2. 정책 (Policy)

- Agent는 각 상태마다 행동을 선택함
- **각 상태에서 어떻게 행동할지? → 정책(Policy)**



상태 s에서 a를 선택할 확률 : $\pi(a|s)$ **로 표기**

02-3. 가치함수 (state-value function)

- **어떤 상태 s로 갈 때, 그 이후로 받게 되는 Return value의 기댓값**

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S},$$

- 상태 s에서 가능한 반환값들의 평균 : 정책에 따라 영향을 받음

02-3. Q함수 (action-value function)

- 어떤 상태 s 에서, 행동 a 를 했을 때 받게 되는 Return Value의 기댓값

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

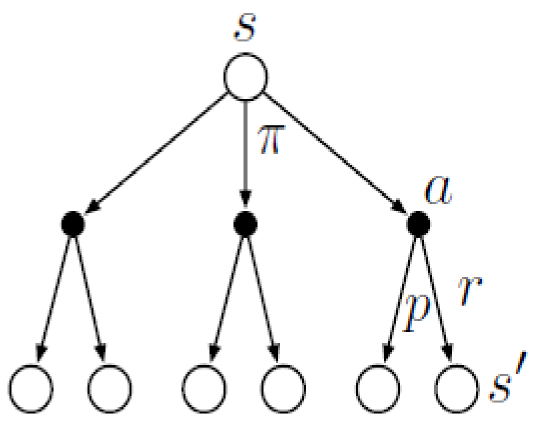
- 가치함수는 Q함수의 기댓값으로 표현 가능

$$v_\pi(s) = \boldsymbol{E}_{a\sim\pi}[q_\pi(s,a) | S_t = s]$$

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

02-3. Bellman Expectation Equation

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\Big[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\Big] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a)\Big[r + \gamma v_\pi(s')\Big], \quad \text{for all } s \in \mathcal{S}, \end{aligned}$$



Backup diagram for v_π

$$v_\pi(s) = \boldsymbol{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$

$$q_\pi(s, a) = \boldsymbol{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

03-1. 최적 정책 (Optimal Policy)

- Optimal Policy – Return이 최대로 되도록 하는 최적 정책 π_*

- Optimal State-value Function – 최적 정책 하에서 나오는 최대치 함수 $v_*(s) \doteq \max_\pi v_\pi(s)$

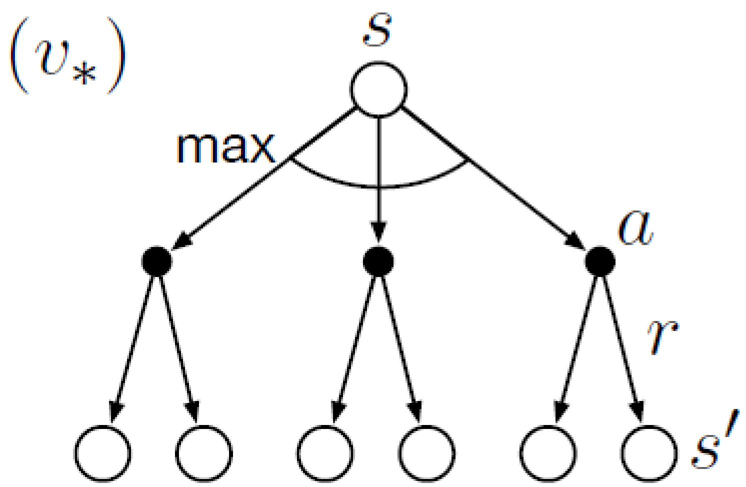
- Optimal Action-value Function $q_*(s,a) \doteq \max_\pi q_\pi(s,a)$

- 최적 가치함수를 찾은 이후 정책의 Update

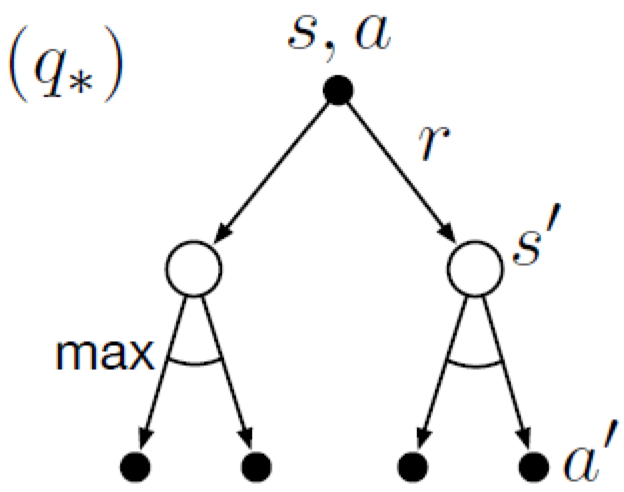
$$\pi_*(s,a) = \begin{cases} 1, & if \;\; a = argmax_a q_\pi(s,a) \\ 0, & otherwise \end{cases}$$

03-1. Bellman Optimality Equation

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



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



03-1. 요약

- 벨만 기대 방정식 (Bellman Expectation Equation)
  - $v_\pi(s) = \boldsymbol{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$
  - $q_\pi(s,a) = \boldsymbol{E}_{\boldsymbol{\pi}}[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$

- 벨만 최적 방정식 (Bellman Optimality Equation)
  - $v^*(s) = \max_a \boldsymbol{E}[R_{t+1} + \gamma v^*(S_{t+1})|S_t = s, A_t = a]$
  - $q^*(s,a) = \boldsymbol{E}[R_{t+1} + \gamma\, max_{a'}\, q^*(S_{t+1}, a')\,|S_t = s, A_t = a]$

Thank you