Introduction to Reinforcement Learning

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

RL의 구성

- Agent, Environment, State, Action, Reward

Objective: Maximize expected cumulative reward

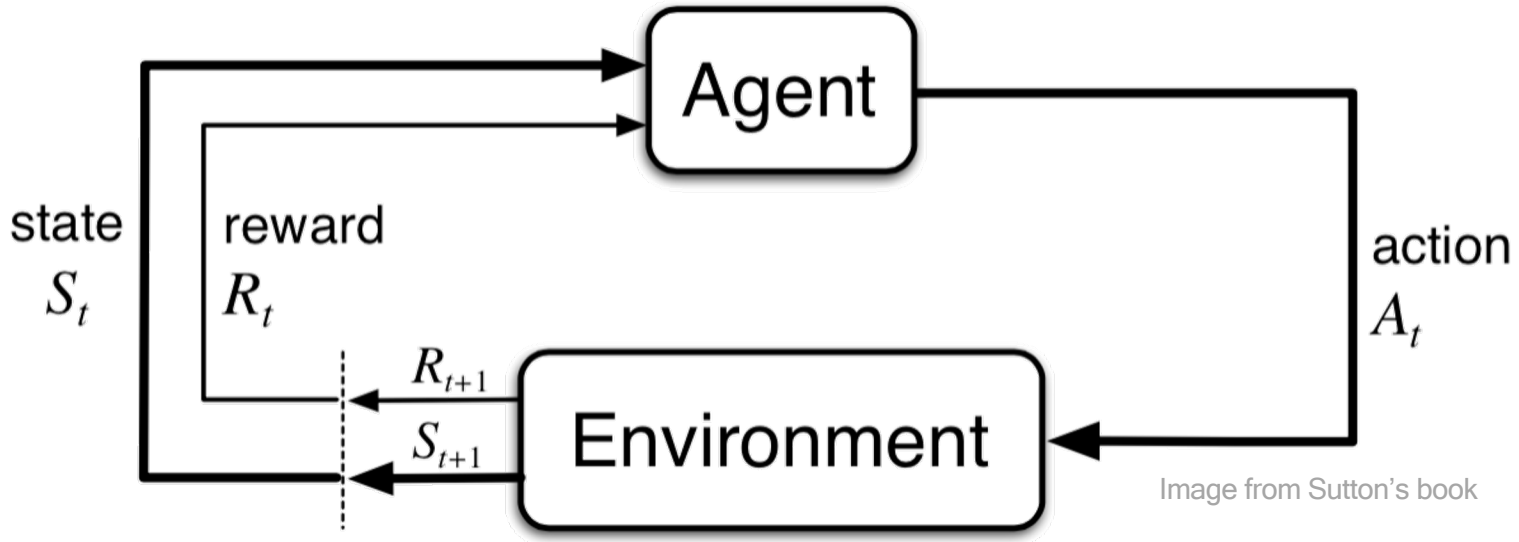


Image from Sutton's book

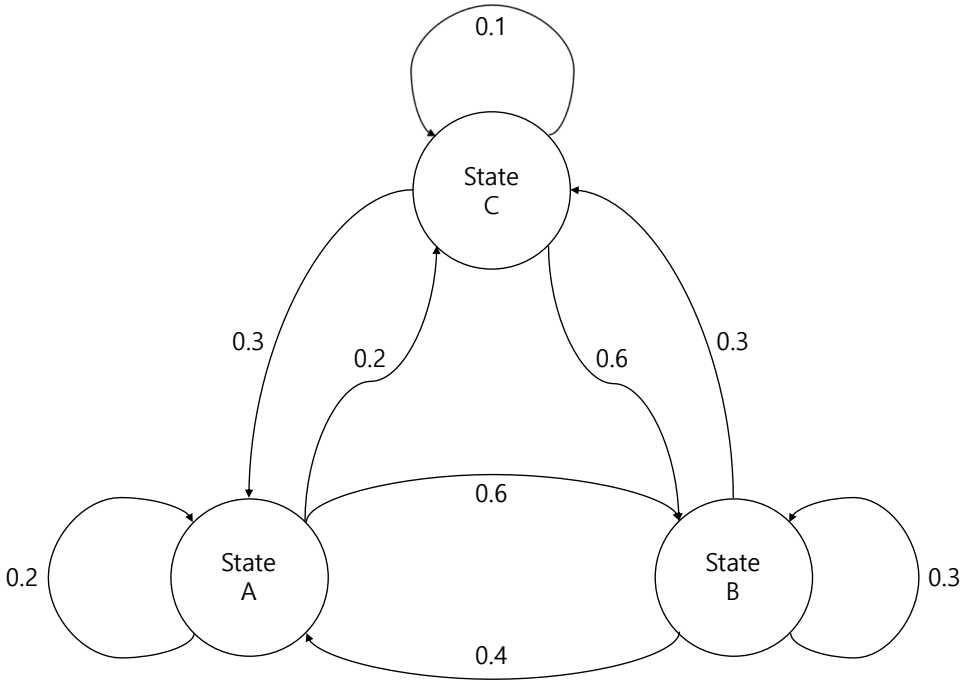
$$\text{episode} = \{s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, \cdots\}$$

Markov Decision Process (MDP)

Markov Process

- 상태의 이동 제약 조건에 이전 상태만 영향을 받는다.

$$P(s_t|s_{t-1})$$

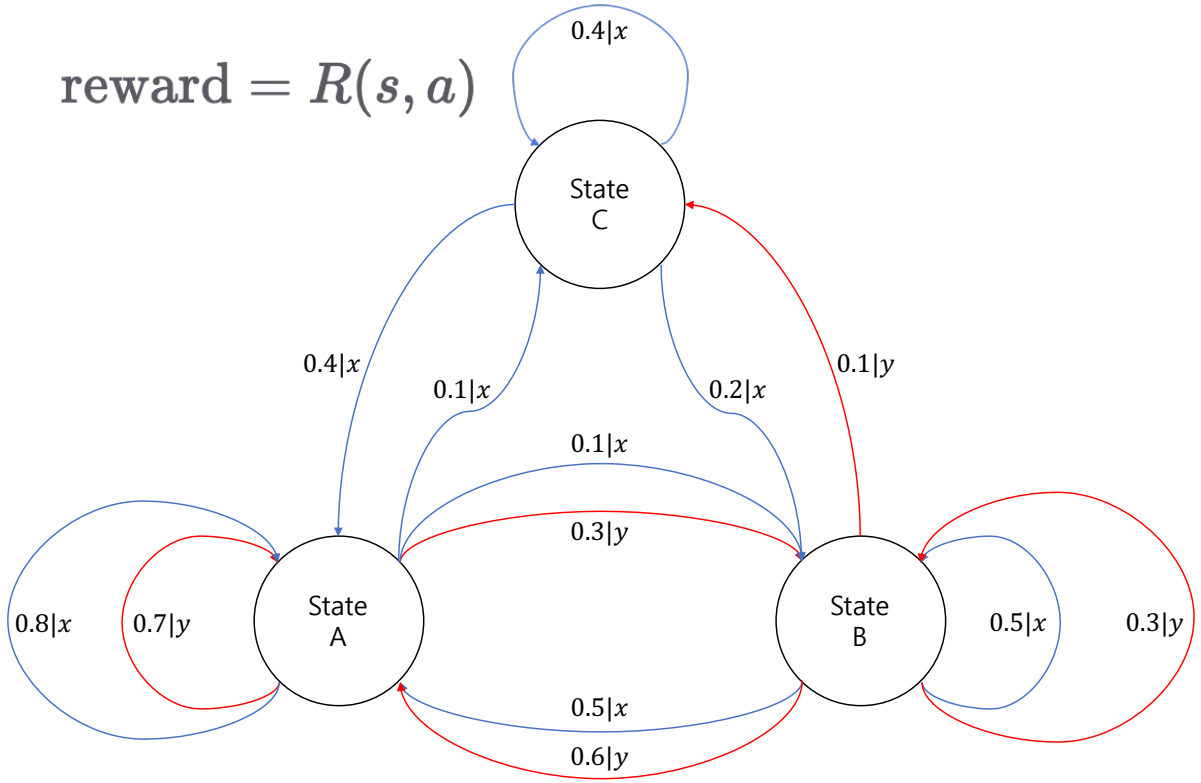


Markov Decision Process

- 상태의 이동 제약 조건에 이전 상태와 행해진 액션에 영향을 받는다.

$$P(s_t|s_{t-1}, a)$$

$\text{reward} = R(s, a)$



Cumulative Reward

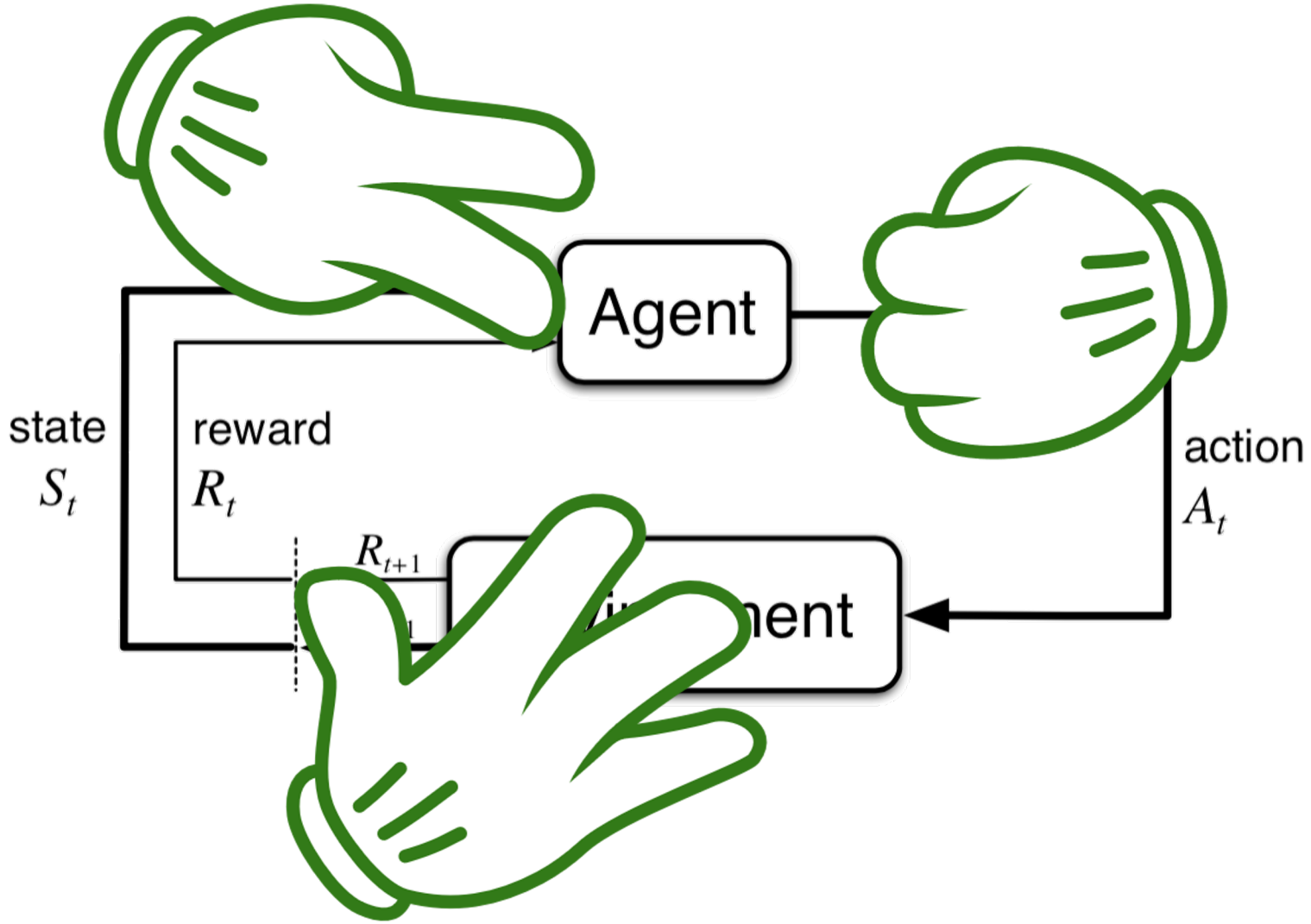
- Add all reward after t, until end of episode:

$$G_t = r_{t+1} + r_{t+2} + \cdots + r_T$$

- Apply discount factor:

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

Example: 가위, 바위, 보



Policy and Value Function, Action-Value Function

- Policy:

$$\pi(a|s) = P(A_t = a|S_t = s)$$

- Value Function:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|S_t = s\Big], \\ &\forall s \in \mathcal{S}. \end{aligned}$$

- Action-Value Function:

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|S_t = s, A_t = a\Big],$$
$$\forall s \in \mathcal{S} \text{ and } \forall a \in \mathcal{A}.$$

Summary

- RL Objective: Maximize expected cumulative reward, $\mathbb{E}_{\pi}[G_t]$
  - 현재 상황에서 얻을 수 있는 누적 보상의 기대 값을 최대화

- Environment에 대한 정보가 부족하므로,
  샘플링을 통해 value(or Q) function을 근사 하거나, policy 함수를 학습.
  - Value 기반의 방법과, Policy 기반의 방법으로 크게 나뉨

- 우리는 Policy 기반의 방법을 통해, NLG를 고도화 할 것