강화 학습
Reinforcement Learning

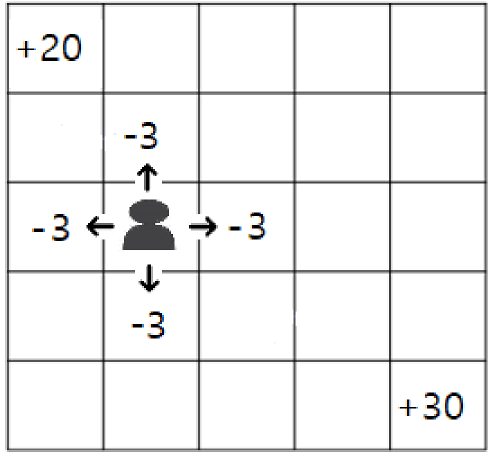
강화 학습이란

- 최적의 보상을 얻는 행동을 찾아가는 기계 학습 방법
  - Cart Pole
  - 알파고
  - 게임
- 참고 문헌
  - R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction

Example

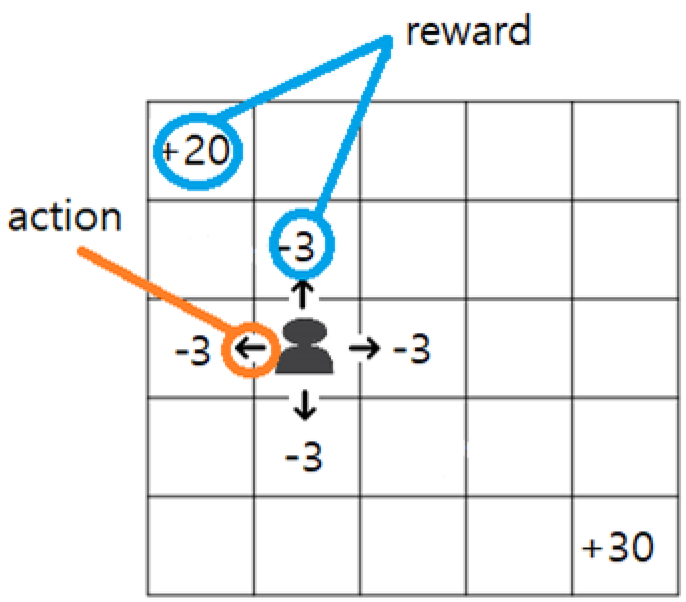
- Grid world
  - 이득이 가장 큰 방향은 어디인가?

  - 어디에 투자할까?
    - 예금, 주식, 펀드



Variables

- State
  - 상태
- Action
  - 선택 가능한 행동
- Reward
  - 보상

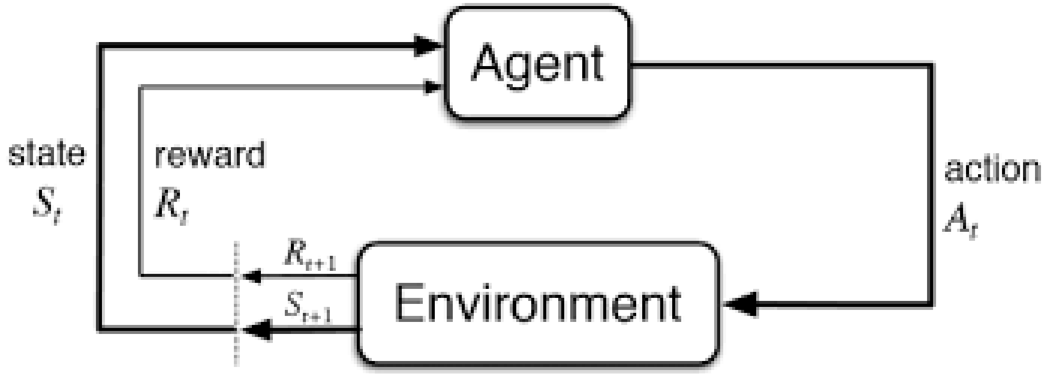


Agent-Environment Interface

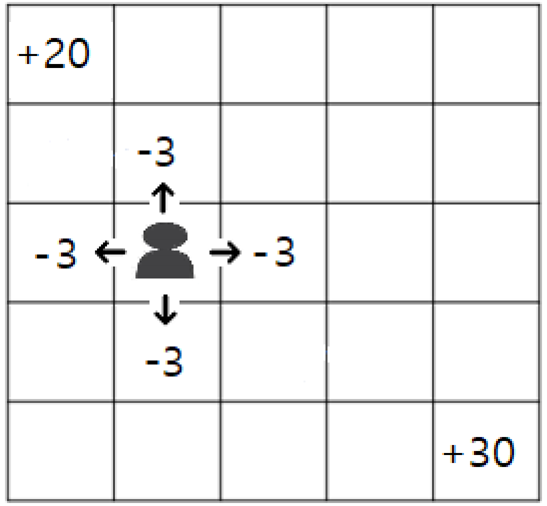
- Agent
  - 행위자
  - 각 state에서 action 선택
- Environment
  - 환경
  - Agent의 action에 대한 reward 제공
  - Agent의 action에 대하여 다음 state 결정

MDP – Markov Decision Process

- 의사결정을 위한 수학적 모델
- 직전의 state와 action 선택이 다음 state와 reward를 결정
- $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots$
- $S_0 \xrightarrow{A_0} R_1, S_1 \xrightarrow{A_1} R_2, S_2 \xrightarrow{A_2}$
  - S_t - state
  - A_t - action
  - R_t - reward

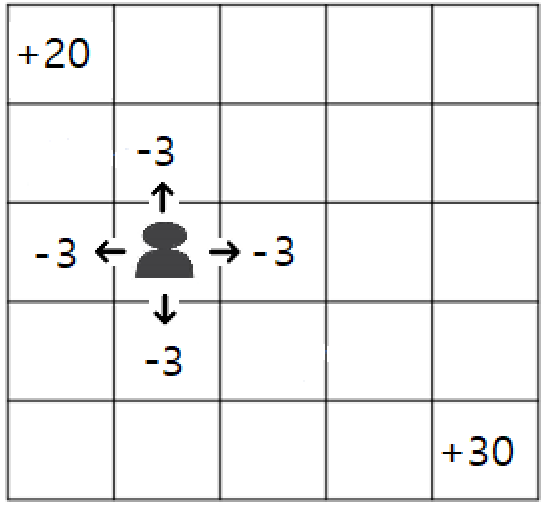


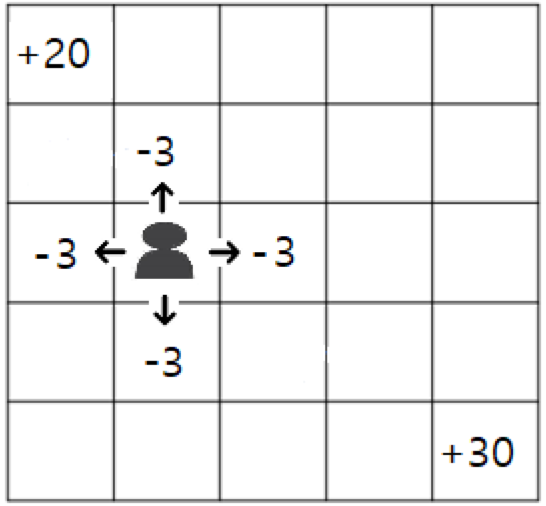
- Example 1
  - Valance(잔고): v
  - Action: 예금, 주식, 펀드
  - Reward: 수익
- Example 2
  - Grid world



- $\mathcal{S}$ - state의 집합
- $\mathcal{A}$ - action의 집합
  - $\mathcal{A}(s)$ - state s에서 취할 수 있는 action의 집합
- $\mathcal{R}$ - reward의 집합
  - $\mathcal{R}(s, a)$ - state s에서 action a를 취했을 때 얻는 reward의 집합

$$\mathcal{S} = \{(1,1), \dots, (5,5)\}$$
$$\mathcal{A} = \{\leftarrow, \uparrow, \rightarrow, \downarrow\}$$
$$\mathcal{R} = \{20, 30, -3\}$$



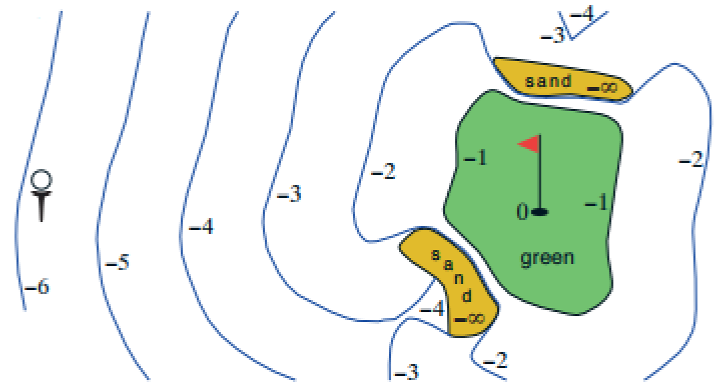


$$S_0 \xrightarrow{A_0} R_1, S_1 \xrightarrow{A_1} R_2, S_2 \xrightarrow{A_2}$$

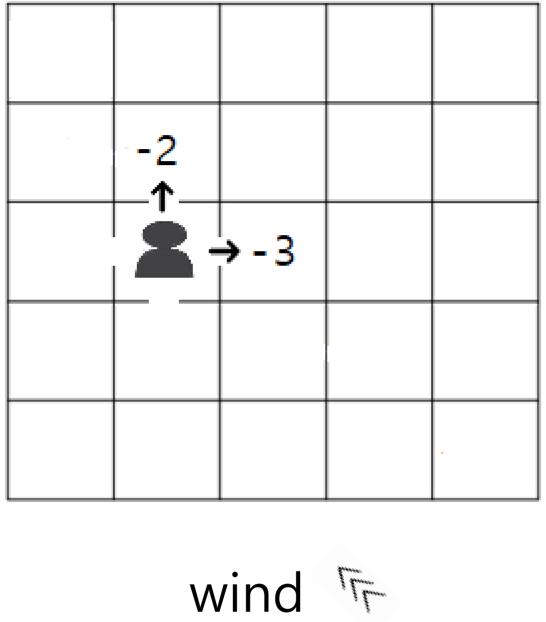
$$(2,3) \xrightarrow{\rightarrow} -3, (3,3) \xrightarrow{\rightarrow} -3, (4,3) \longrightarrow$$

- Action의 결과가 항상 같은 것은 아니다
    -

투자의 수익률은 일정하지 않다



다양한 요소의 영향을 받는다



$p((3,3)\,|(2,3),\rightarrow) \leq 1$

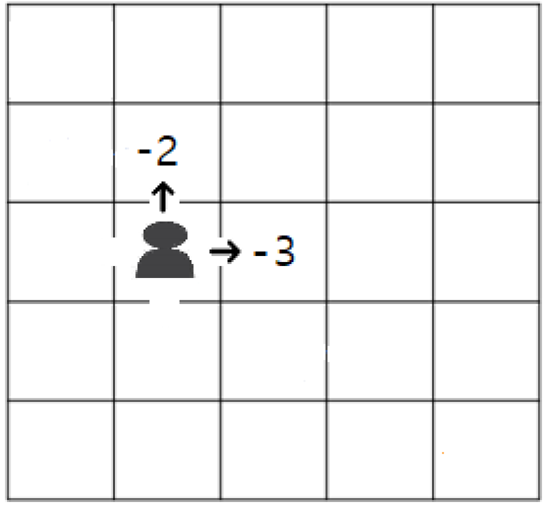
- state s에서 action a를 시행하여
  state s'이 되고 reward r을 얻을 확률

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

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r | s, a) = 1$$

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

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

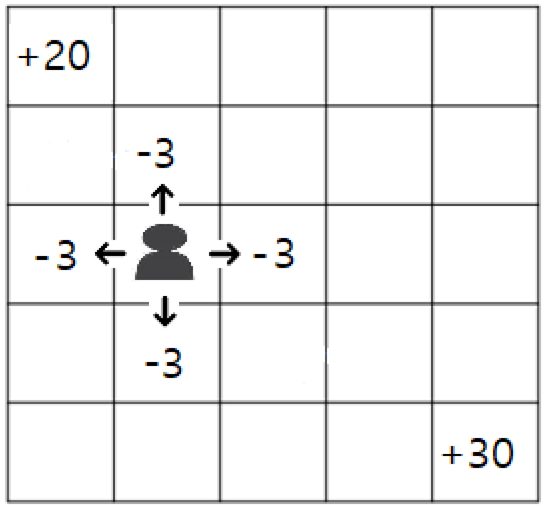


wind

$p((3,3)\ |(2,3),\rightarrow) = 0.8$
$p((1,1)\ |(2,3),\rightarrow) = 0.2$

$$r\big((2,3),\rightarrow\big) = -3 \times 0.8 - 2 \times 0.2$$

- Goal: the maximization of the expected value of the cumulative sum of rewards



Episodes and Returns

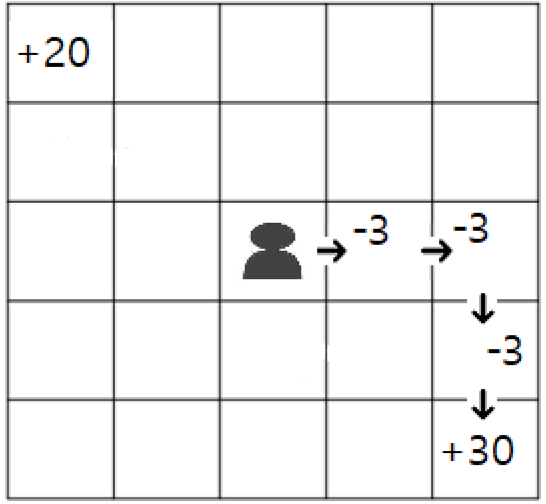
- Episode
  - agent와 environment의 상호작용이 완료되는 과정
  - $S_0 \xrightarrow{A_0} R_1, S_1 \xrightarrow{A_1} R_2, S_2 \xrightarrow{A_2} \cdots \xrightarrow{A_{T-1}} R_T, S_T$
- Return
  - Time t 에서 에피소드가 종료될 때까지 얻는 보상의 총합
  - $G_t = R_{t+1} + R_{t+2} + \cdots + R_T$

Return and discount rate

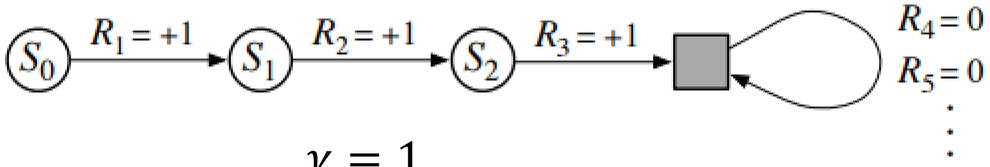
- Discount rate, 할인율
  - 미래에 얻을 이익을 현재 가치로 환산하기 위한 비율
- 할인율 γ를 적용한 return
  - $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$
    $= \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

Example

- episode
  - $S_0 \xrightarrow{A_0} R_1, S_1 \xrightarrow{A_1} R_2, S_2 \xrightarrow{A_2} R_3, S_3 \xrightarrow{A_3} R_4, S_4$
- discount rate
  - $\gamma = 0.9$
- return
  - $G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4$
    $= -3 + 0.9 \cdot (-3) + 0.81 \cdot (-3) + 0.729 \cdot 30$
    $= 13.74$
  - $G_1 = 18.6$



Example



$\gamma = 1$

$$G_0 = 3$$
$$G_1 = 2$$
$$G_2 = 1$$
$$G_3 = 0$$

Policy

- agent의 정책
- action을 선택하는 방법
- $\pi(a|s) = \Pr(A_t = a \mid S_t = s)$

deterministic policy

각 state에서 action이 결정되어 있음

stochastic policy

각 state에서 확률에 따라 action을 선택

Example

deterministic policy

$\pi(a|s) = 0$ or 1

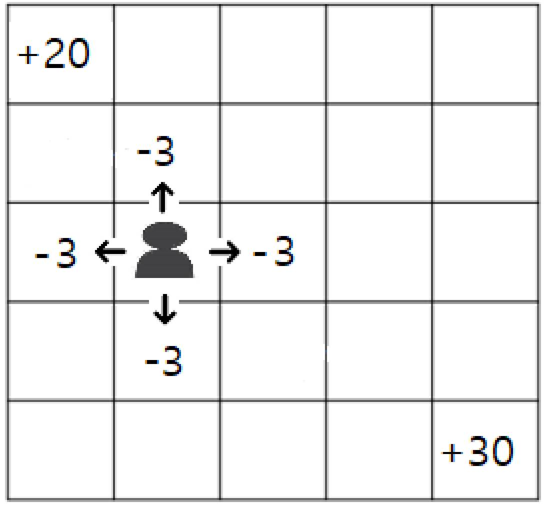
random policy

$\pi(a|s) = \pi(a'|s)$ for all actions a and a'

State-Value Function

- Values of states
- State-Value function for policy π
  - the expected value of returns for all possible episodes
  - $v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$

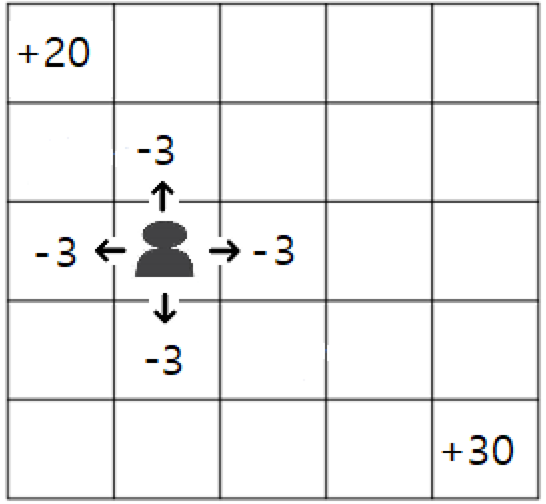
Value of the state $(2{,}3)$?
Determined by policy



Action-Value Function

- Values of actions
- Action-Value function for policy π
  - $q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$

Value of the action → at the state (2,3)?
Determined by policy



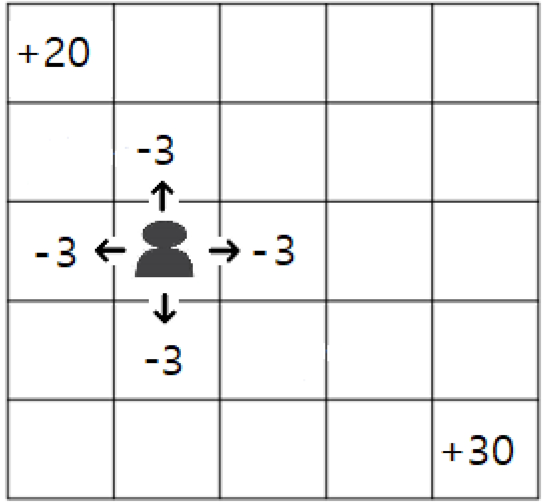
Bellman equation

- Bellman equation for v_π

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$$

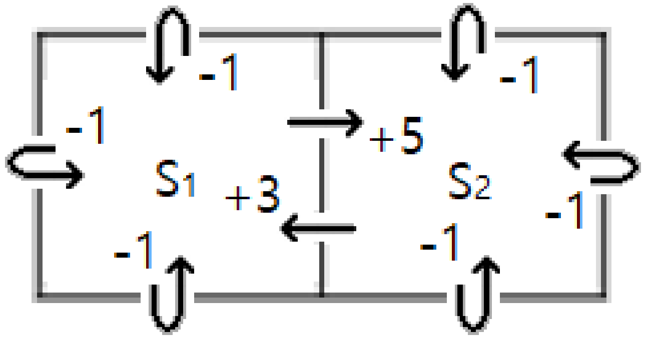
- Bellman equation for q_π

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$$



The value of (2,3) is determined by values of (1,3), (2,2), (3,3), (2,4)

Example - Tiny World



- $\pi(a|s) = 1/4$
- $p(s', r|s, a) = 1$
- Bellman equation
  - $v_\pi(S_1) = \frac{1}{4} \cdot 3 \cdot [-1 + \gamma v_\pi(S_1)] + \frac{1}{4} \cdot [5 + \gamma v_\pi(S_2)]$
  - $v_\pi(S_2) = \frac{1}{4} \cdot [3 + \gamma v_\pi(S_1)] + \frac{1}{4} \cdot 3 \cdot [-1 + \gamma v_\pi(S_2)]$
- Solution
  - $v_\pi(S_1) = \frac{4-3\gamma}{4(1-\gamma)(2-\gamma)}$, $v_\pi(S_2) = \frac{\gamma}{4(1-\gamma)(2-\gamma)}$
  - $\gamma = 0.9 \Longrightarrow v_\pi(S_1) = 2.95,\ v_\pi(S_2) = 2.05$

Bellman Equation

- Bellman equation을 풀어 v_π(s)를 모두 구할 수 있다
  - 형태는 $Av + b = 0$ 꼴
  - 시간 문제
  - 메모리 문제