

# 기계 학습

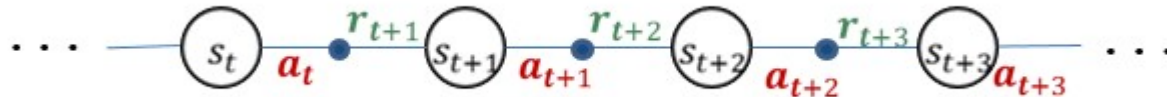
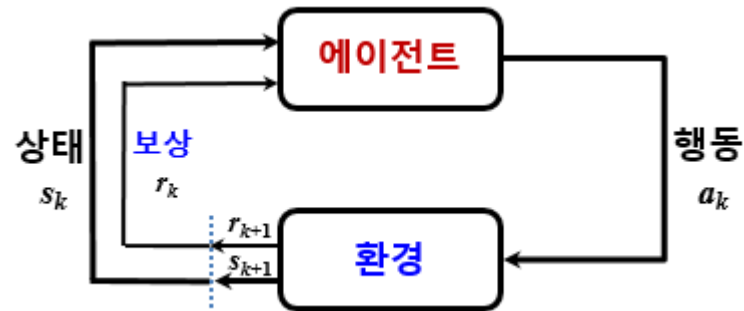
## Part V. 강화 학습

인공지능 : 튜링 테스트에서 딥러닝까지

# 11.1 강화 학습

## ❖ 강화 학습(reinforcement learning)

- 어떤 모르는 환경에서 동작하는 에이전트가 있을 때, 에이전트가 현재 **상태**(state)에서 향후 기대되는 **누적 보상값**(reward)이 **최대**가 되도록 **행동**(action)을 선택하는 **정책**(policy)을 찾는 것



- 강화학습 문제 표현
  - 마르코프 결정 과정(MDP) 사용 표현
    - 불확실성 반영

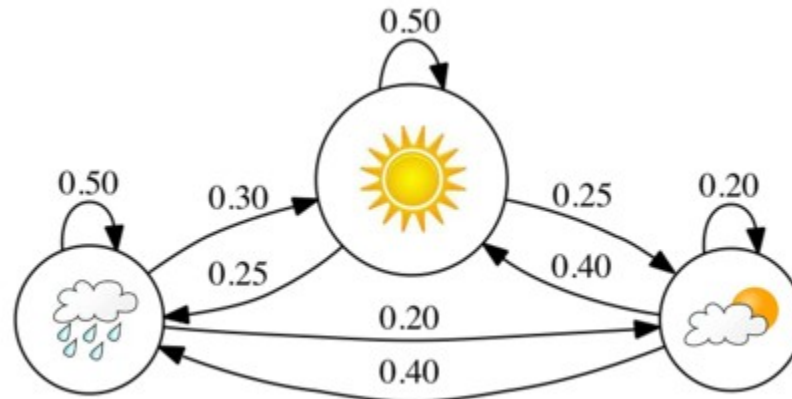
# 강화 학습

## ✧ 마르코프 결정 과정(Markov Decision Process, MDP)

- 상태 전이(state transition)가 현재 상태  $S_t$ 와 입력 (또는 행동)  $A_t$ 에 의해서 확률적으로 결정되는 마르코프 모델(Markov model)

- 마르코프 모델

- 미래의 상태  $S_{t+1}$ 는 현재 상태  $S_t$ 에 영향을 받고 과거 상태  $S_{t-1}, S_{t-2}, \dots$ 에는 영향을 받지 않는 시스템에 대한 확률 모델 (stochastic model)
- $P(S_{t+1}|S_t, S_{t-1}, \dots, S_0) = P(S_{t+1}|S_t)$



Andrei Andreyevich Markov



1856-1922  
러시아, 수학자

- 마르코프 결정과정

- $P(S_{t+1}|S_t, S_{t-1}, \dots, S_0, A_t) = P(S_{t+1}|S_t, A_t)$

# 강화 학습

## ❖ 마르코프 결정 과정(Markov Decision Process, MDP) – cont.

- **상태**의 집합  $S = \{s_1, s_2, \dots, s_N\}$
- **행동**의 집합  $A = \{a_1, a_2, \dots, a_M\}$
- **상태 전이(state transition)** 결정 확률분포
  - $t$  시점의 상태  $s_t$ 에서 행동  $a_t$ 를 취할 때 도달하는 다음 상태  $s_{t+1}$ 를 결정하는 것  $(s_t, a_t, s_{t+1})$ 
    - $P(s_{t+1} = s' | s_t = s, a_t = a) = T(s, a, s')$
- 상태 전이가 일어날 때 **즉시 보상값(immediate reward)**
  - 상태 전이  $(s_t, a_t, s_{t+1})$ 에서 받는 즉시 보상값  $r_{t+1}$ 
    - $R(s_t, a_t, s_{t+1}) = R(s_{t+1}) = r_{t+1}$

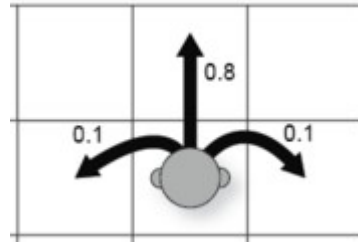
## ❖ 강화학습의 목적

- 기대 누적 보상값(expected accumulated reward)이 최대가 되도록 하는 **정책(policy)**을 찾는 것
  - 정책 : 각 상태에서 선택할 행동 지정

# 강화 학습

## 예. 강화 학습 문제

- $S = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)$   
 $(3,1), (3,2), (3,3), (4,1), (4,2), (4,3)\}$
- $A = \{east, west, south, north\}$
- 상태전이확률



$$T((3,1), north, (3,2)) = 0.8$$

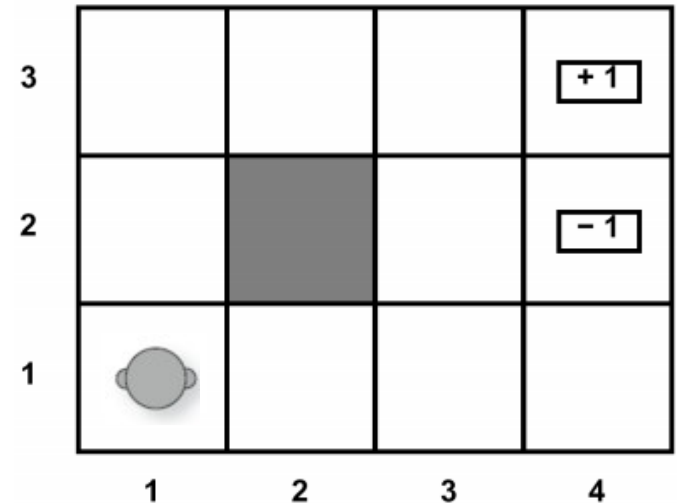
$$T((3,1), north, (2,1)) = 0.1$$

$$T((3,1), north, (4,1)) = 0.1$$

- 보상(Reward)

$$R((4,3)) = +1, R((4,2)) = -1$$

$$R((x,y)) = c \quad (x,y) \neq (4,1) \text{ or } (4,3)$$

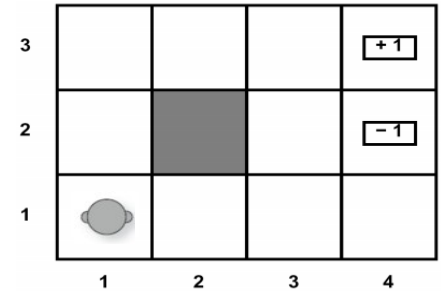


# 강화 학습

## ❖ 예. 강화 학습 문제

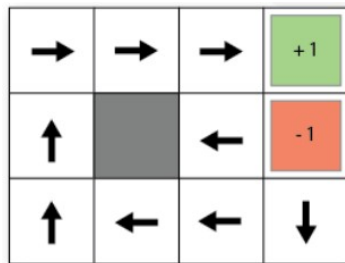
### ▪ 정책(policy)

- 각 상태  $s$ 에서 취할 행동  $a$ 을 결정해 둔 것

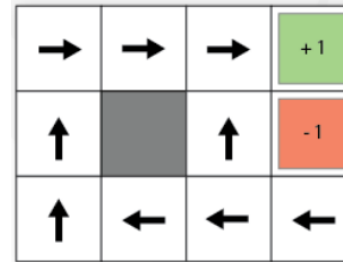


$$R((4,3)) = +1, R((4,2)) = -1, R((x,y)) = c \quad (x,y) \neq (4,1) \text{ or } (4,3)$$

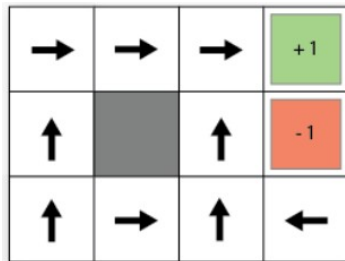
$$c = -0.01$$



$$c = -0.04$$



$$c = -0.09$$



$$c = -2.00$$



## 11.2 누적 보상치

### ❖ 누적 보상치의 계산 방법

- 단순 합계

- $V(s_0, s_1, \dots) = r(s_0) + r(s_1) + r(s_2) + \dots$
- 연속해서 보상치가 더해지면 지속적으로 커질 수 있음

- 할인 누적 합계 (sum of discounted reward)

- $V(s_0, s_1, \dots) = r(s_0) + \gamma * r(s_1) + \gamma^2 * r(s_2) + \dots$ 
  - 할인율 (discount factor)  $\gamma : 0 < \gamma < 1$
  - 가까운 보상이 먼 미래의 보상보다 가치가 있음

## 11.3 가치 함수

### ❖ 가치 함수 (value function)

#### ■ 상태 가치 함수(state value function) $V^\pi(s)$

- 상태  $s$ 에서 시작하여 정책  $\pi$ 에 따라 행동을 할 때 얻게 되는 기대 보상(expected reward)

$$\begin{aligned} V^\pi(s) &= \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots | s_t = s, \pi] \\ &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, \pi\right] \end{aligned}$$

#### ■ 상태-행동 가치 함수(state-action value function) $Q^\pi(s, a)$

- 상태  $s$ 에서 행동  $a$ 를 한 다음, 정책  $\pi$ 에 따라 행동을 할 때 얻게 되는 기대 보상

$$\begin{aligned} Q^\pi(s, a) &= \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots | s_t = s, a_t = a, \pi] \\ &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a, \pi\right] \end{aligned}$$

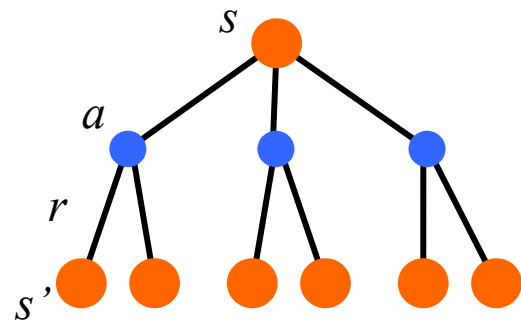


# 가치 함수

## ❖ Bellman 방정식

- 상태 가치 함수와 상태-행동 가치 함수의 관계

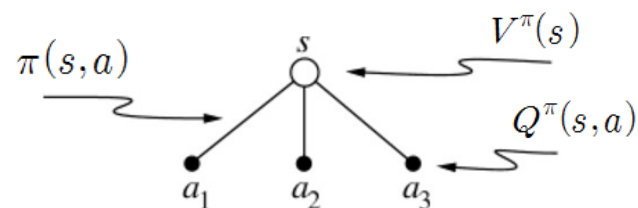
$$\begin{aligned}
 V^\pi(s) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, \pi\right] \\
 &= \mathbb{E}\left[r_{t+1} + \gamma \sum_{s'} \gamma^k r_{t+k+2} \mid s_t = s, \pi\right] \\
 &= \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a \left[ r_{ss'}^a + \gamma \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s'\right] \right] \\
 &= \sum_a \pi(s, a) \left[ \sum_{s'} P_{ss'}^a \left[ r_{ss'}^a + \gamma V^\pi(s') \right] \right] \\
 &= \sum_a \pi(s, a) Q^\pi(s, a)
 \end{aligned}$$



- $\pi(s, a)$ : 정책  $\pi$ 가 상태  $s$ 에서 행동  $a$ 를 선택할 확률
- $P_{ss'}^a$ : 상태  $s$ 에서 행동  $a$ 를 할 때, 상태  $s'$ 이 될 확률
- $r_{ss'}^a$ : 상태  $s$ 에서 행동  $a$ 를 할 때, 보상값
- $\gamma$ : 할인율

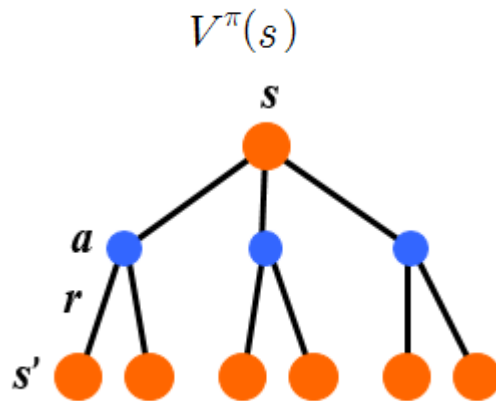
$$V^\pi(s) = \sum_a \pi(s, a) Q^\pi(s, a)$$

$$Q^\pi(s, a) = \sum_{s'} P_{ss'}^a \left[ r_{ss'}^a + \gamma V^\pi(s') \right]$$

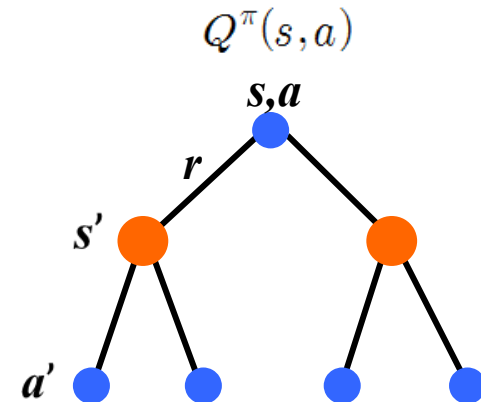


# 가치 함수

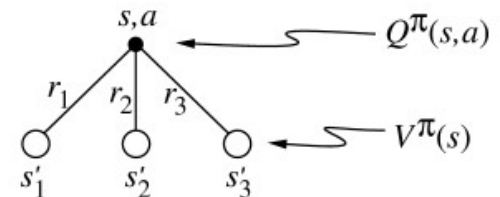
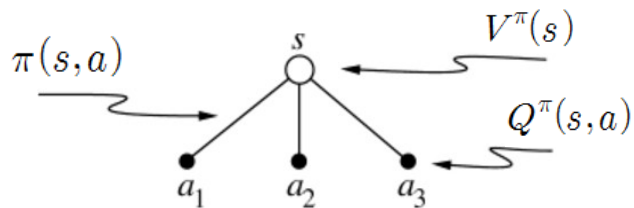
## ❖ 가치 함수 (value function)



$$V^\pi(s) = \sum_a \pi(s, a) Q^\pi(s, a)$$



$$Q^\pi(s, a) = \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V^\pi(s')]$$



# 가치 함수

## ❖ 가치 함수 계산 방법

- 동적계획법 방법 (dynamic programming, DP)
  - 모든 상태에 대한 섭렵하면서 **Bellman 최적 방정식** 성질을 이용하여 가치 함수 계산
  - 정책반복 학습, 값반복 학습 알고리즘
- 몬테 카를로 방법(Monte Carlo method)
  - 주어진 정책  $\pi$ 에 따라 에이전트가 행동을 하여 상태와 행동에 따른 보상값을 기록하여 상태 가치 함수 또는 상태-행동 가치 함수 추정
- 모수적 함수(parameterized function) 학습 방법
  - 상태의 개수의 매우 많은 경우 각 상태에 대한 보상값 관리 곤란
  - 가치 함수의 역할을 하는 모수적 함수를 학습하여 사용

# 11.4 최적 정책

❖ **최적 정책**(optimal policy)  $\pi^*$ 과 최적 상태 가치 함수  $V^*$

$$\pi^* = \arg \max_{\pi} V^{\pi}(s), (\forall s)$$

$$V^*(s) = V^{\pi^*}(s)$$

❖ **Bellman 최적 방정식**(optimality equation)

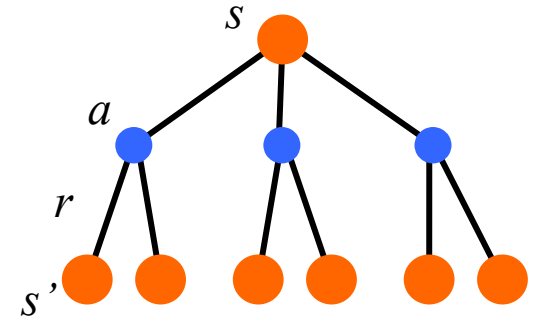
- 최적 정책에 따른 가치 함수들이 만족하는 성질
- 상태 가치 함수의 경우

$$V^*(s) = \max_a \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V^*(s')]$$

- 모든 가능한 **행동** 중에서 가장 큰 기대보상값을 주는 행동의 값

- 상태-행동 가치 함수의 경우

$$Q^*(s, a) = \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V^*(s')]$$



# 11.5 강화 학습 알고리즘

## ✧ 정책 평가 (policy evaluation) $\pi \rightarrow V^\pi$

- 주어진 정책  $\pi$ 을 따를 때, 각 상태에서 얻게 되는 기대보상 값  $V^\pi$  계산

$$V_{k+1}(s) = \sum_a \pi(s, a) \sum_{k'} P_{ss'}^a [r_{ss'}^a + \gamma V_k(s')]$$

- 임의의 가치 함수  $V_0$ 에서 시작하여,  $V_k$ 가 수렴할 때까지 반복

**Input :** 평가할 정책  $\pi$

$V(s) \leftarrow 0$  for each  $s \in S$

repeat

$\Delta \leftarrow 0$

    for each  $s \in S$

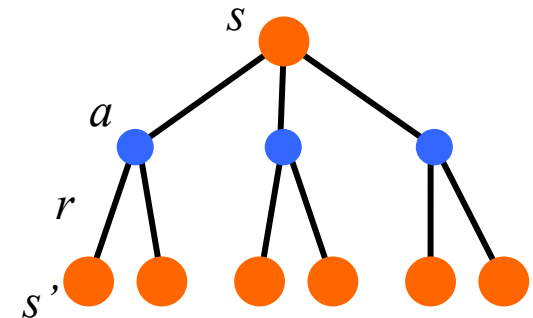
$temp \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(s, a) \sum_{k'} P_{ss'}^a [r_{ss'}^a + \gamma V_k(s')]$

$\Delta \leftarrow \max(\Delta, |temp - V(s)|)$

until  $\Delta < \theta$  (작은 양수)

**Output :**  $V \approx V^\pi$



# 강화 학습 알고리즘

✧ 정책 개선 (policy improvement)  $V^\pi \rightarrow \pi$

- 상태 가치 함수  $V(s)$  값으로 부터 정책  $\pi$  결정

$$\begin{aligned}\pi'(s) &= \arg \max_a Q^\pi(s, a) \\ &= \arg \max_a \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V^\pi(s')]\end{aligned}$$

**Input** : 상태가치 함수  $V$

for each  $s \in S$

$$\pi(s) \leftarrow \arg \max_a \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V^\pi(s')]$$

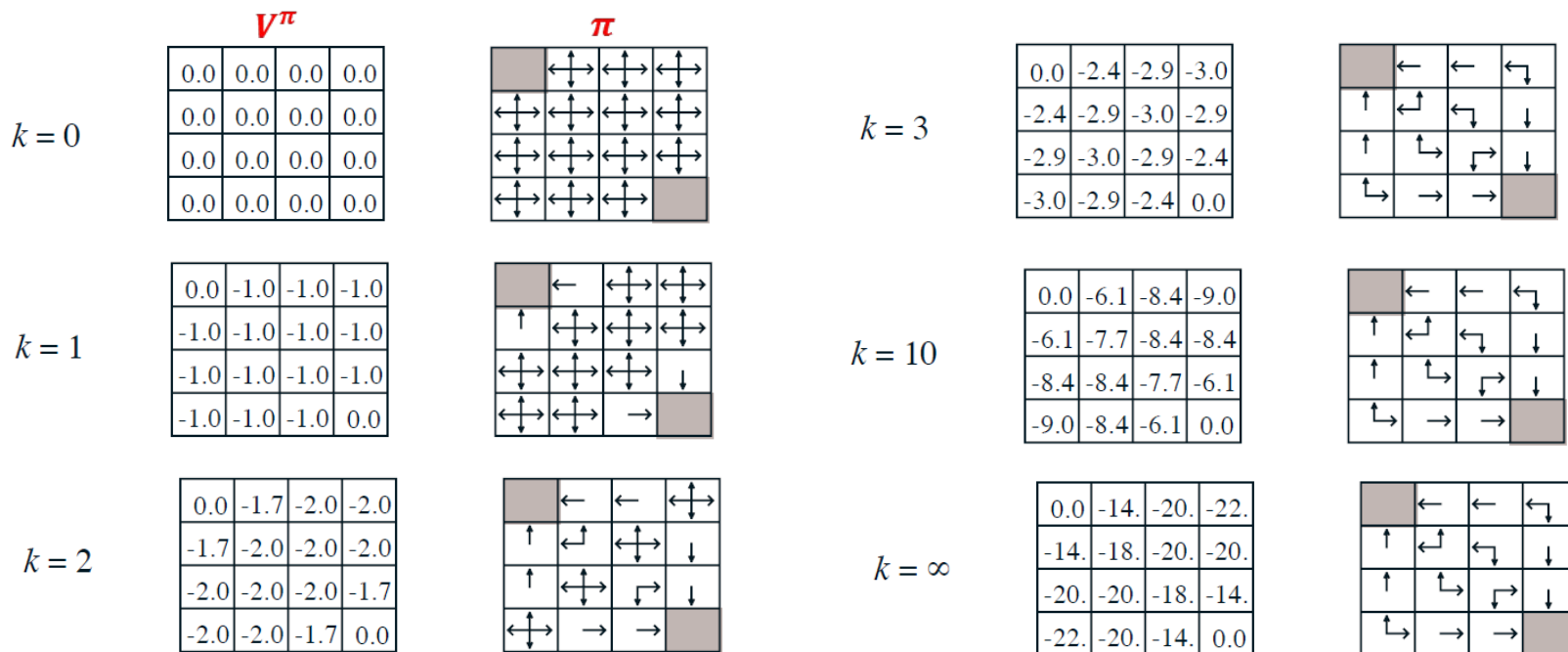
**Output** : 정책  $\pi$

# 정책 반복 학습 알고리즘

## ❖ 정책 반복(policy iteration) 학습



- 임의의 정책  $\pi$ 에서 시작하여,  $\pi$ 에 대해서 Bellman 방정식을 수렴할 때까지(즉, 바뀌지 않을 때까지) 적용하여  $V^\pi$ 를 계산하고,  $V^\pi$ 를 사용하여  $\pi$ 를 개선하는 과정을 정책  $\pi$ 가 수렴할 때까지 반복



# 값 반복 학습 알고리즘

## ❖ 값 반복(value iteration) 학습

$$V_{k+1}(s) = \max_a \sum_{s'} P_{ss'}^a [r_{ss'}^a + \gamma V_k(s')]$$

- 임의의 가치 함수  $V_0$ 에서 시작하여 정책은 계산하지 않고 **가치 함수가 수렴할 때까지 반복**
- 수렴한 가치 함수  $V^*$ 를 사용하여 **정책  $\pi$ 를 결정**

```
V(s) ← 0 for each s ∈ S
repeat
  Δ ← 0
  for each s ∈ S
    temp ← V(s)
    V(s) ← max_a ∑_{s'} P_{ss'}^a [r_{ss'}^a + γ V_k(s')]
    Δ ← max(Δ, |temp - V(s)|)
until Δ < θ (작은 양수)
```

### 정책결정

```
for each s ∈ S
  π(s) = arg max_a ∑_{s'} P_{ss'}^a [r_{ss'}^a + γ V^π(s')]
```



# Q-learning 알고리즘

## ❖ 정책 반복, 값 반복 학습 알고리즘

- 정확한 MDP 모델이 필요
- 실제 상황에서는 정확한 MDP 모델을 모르는 경우가 많음

## ❖ Q-learning 알고리즘

- 모델이 없이 학습하는 강화학습 알고리즘

```
for each  $s$  and  $a$   
     $\hat{Q}(s, a) \leftarrow 0$   
현재 상태  $s$  관찰  
repeat forever  
    행동  $a$ 를 선택하여 수행  
    즉시보상값  $r$ 를 관측  
    새로운 상태  $s'$  관찰  
     $\hat{Q}(s, a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s', a')$   
     $s \leftarrow s'$ 
```

# Q-learning 알고리즘

for each  $s$  and  $a$

$$\hat{Q}(s, a) \leftarrow 0$$

현재 상태  $s$  관찰

repeat forever

행동  $a$ 를 선택하여 수행

즉시보상값  $r$ 를 관측

새로운 상태  $s'$  관찰

$$\hat{Q}(s, a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s', a')$$

$$s \leftarrow s'$$

상태  $s = 1$ 에서 시작,  $\gamma = 0.8$

행동  $a = 5$  선택 수행

새로운 상태  $s' = 5$  관측

$$\hat{Q}(1, 5) \leftarrow$$

$$r(1, 5) + 0.8 * \max\{\hat{Q}(5, 1), \hat{Q}(5, 4), \hat{Q}(5, 5)\} \\ = 100 + 0.8 * 0 = 100$$

State	Action					
	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

State	Action					
	0	1	2	3	4	5
0	-1	-1	-1	-1	0	-1
1	-1	-1	-1	0	-1	100
2	-1	-1	-1	0	-1	-1
3	-1	0	0	-1	0	-1
4	0	-1	-1	0	-1	100
5	-1	0	-1	-1	0	100

State	Action					
	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	100
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

# 강화 학습 알고리즘

## ❖ 강화 학습 알고리즘

- 몬테카를로 방법(Monte Carlo method)
- 시간 차이 학습(temporal difference learning, TD-learning)
- 정책 그래디언트 알고리즘(policy gradient algorithm)
  - 연속구간 행동을 갖는 강화학습

## 11.6 역강화 학습

### ❖ 역강화 학습(inverse reinforcement learning)

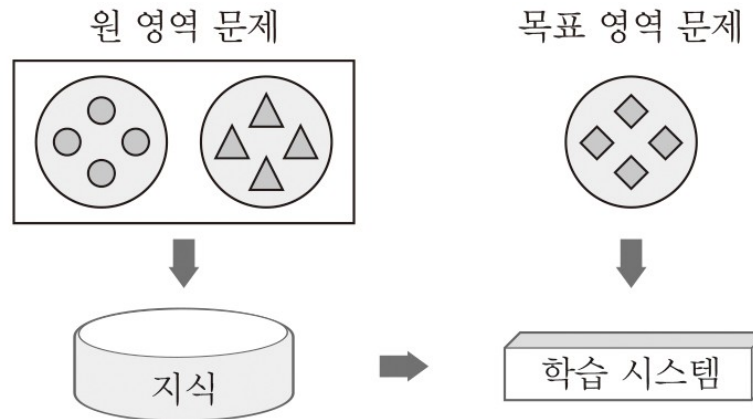
- 보상함수가 직접적으로 제공되지 않는 경우 적용
- 전문가의 바람직한 행동 시연이 가능한 상황
- 시연을 관측한 데이터로부터 보상함수 학습  
➔ 보상함수를 사용하여 가치함수를 함수를 학습하고 정책 결정
- 상태  $s$ 에 대한 전형적인 보상함수  $R(s)$ 의 표현
  - 상태  $s$ 의 특징  $\phi_i(s)$ 들에 대한 선형결합 표현

$$R(s) = \sum_{i=1}^N w_i \phi_i(s)$$

## 4.7 전이 학습

### ❖ 전이 학습(transfer learning)

- 특정 문제를 해결하는 데 사용되는 지식 또는 모델을 관련된 다른 문제의 학습에 이용하는 것



- 활용 가능한 학습 데이터가 부족할 때, 과거에 습득한 지식을 목표영역으로 이전시켜야 할 때 유용