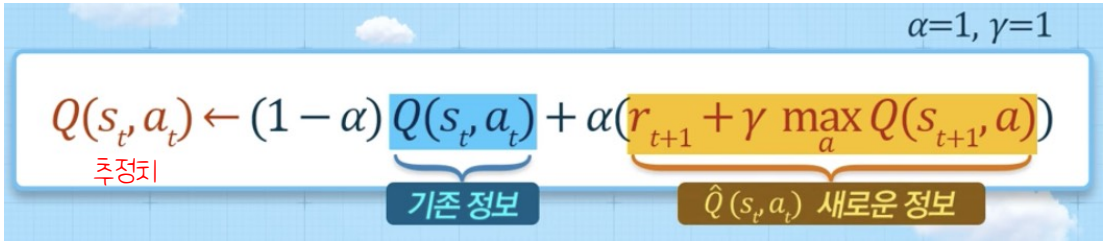
강화학습 맛보기

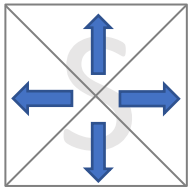
정 태 수
고려대학교 산업경영공학부
tcheong@korea.ac.kr


α=1, γ=1
$Q(s_t, a_t) \leftarrow (1-\alpha) Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a))$
추정치
기존 정보
$\hat{Q}(s_t, a_t)$ 새로운 정보

강화학습

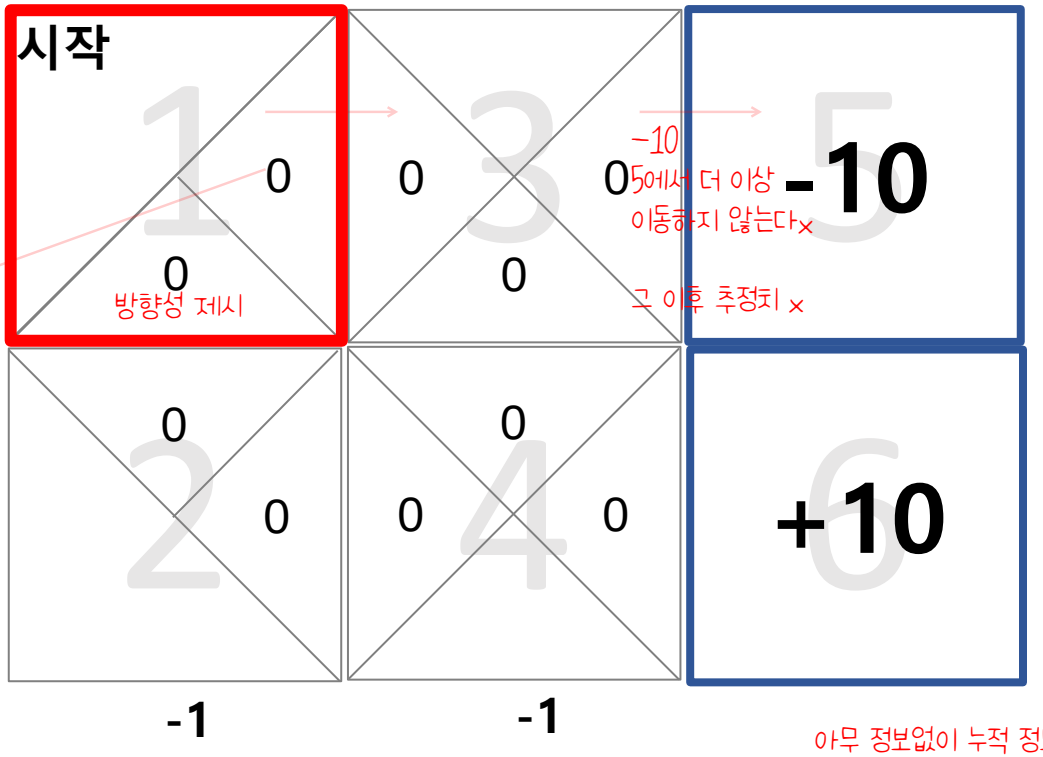
로봇의 움직임

$Q(s_t, a_t)$



$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(\underbrace{r_{t+1} + \gamma \max_a Q(s_{t+1}, a)}_{\hat{Q}(s_t, a_t)}\right) \qquad \alpha = 1, \gamma = 1$$

1->3 : 다음 상태 3에서 큐값에 대한 추정치의 최대값 = 0



각 박스마다값이 있는데 의미하는 바는

1→3→5 : 누적보상 -11

1→2→4→6 : 누적보상 8

세모 안 숫자 : Q값

0이 최대값이인데

-1로 업데이트한다

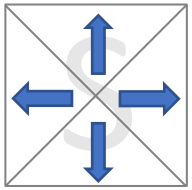
아무 정보없이 누적 정보가 최대가 될 수 있게끔

(1,→) (3,→) 5

강화학습

$Q(s_t, a_t)$ → State s_t에서 Action a_t선택 시 얻을 수 있는 최대 누적 보상의 기대치

$Q(s_t, a_t)$



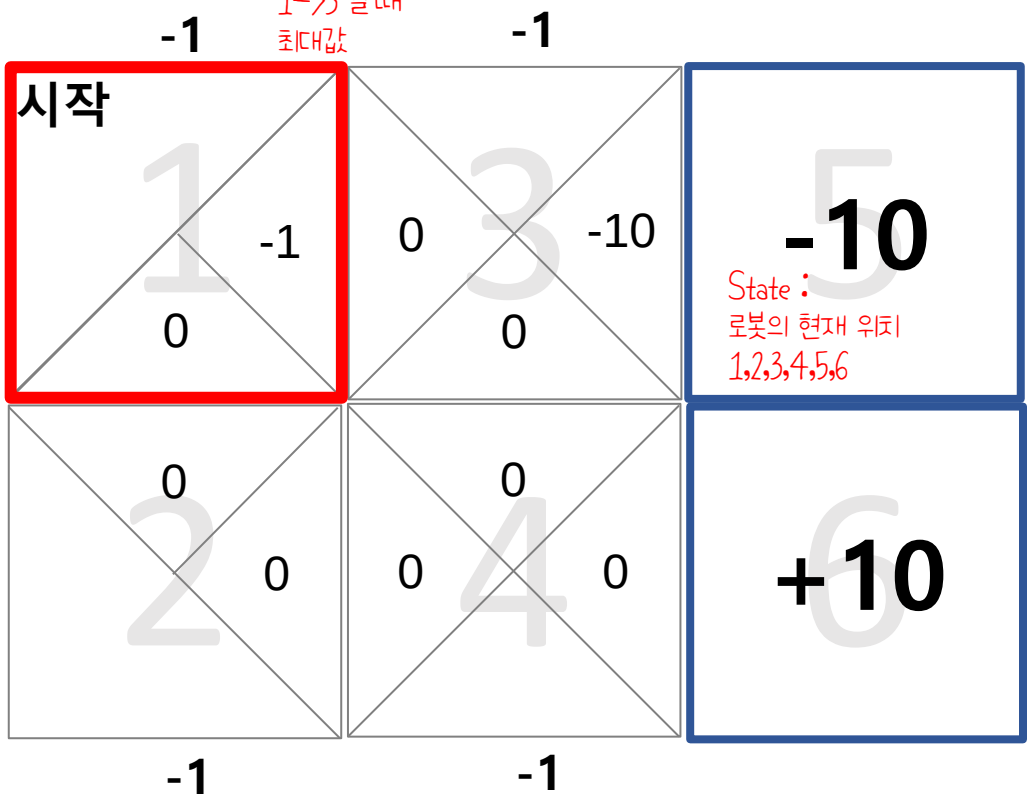
$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$\hat{Q}(s_t, a_t)$

$\alpha = 1, \gamma = 1$

감가율

1->3 갈때 최대값



(1,→) (3,→) 5

러닝을 통해 어떤 정보를 업데이트 할까

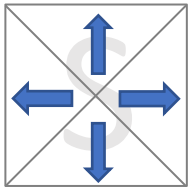
3번 상태에서 획득 가능한 누적 보상합의 최대 값을 활용

현재 1번 상태에서 우측으로 갔을 때 획득 가능한 누적 보상의 최대 값 추정

(1,↓) (2,→) (4,→) 6

강화학습

$Q(s_t, a_t)$

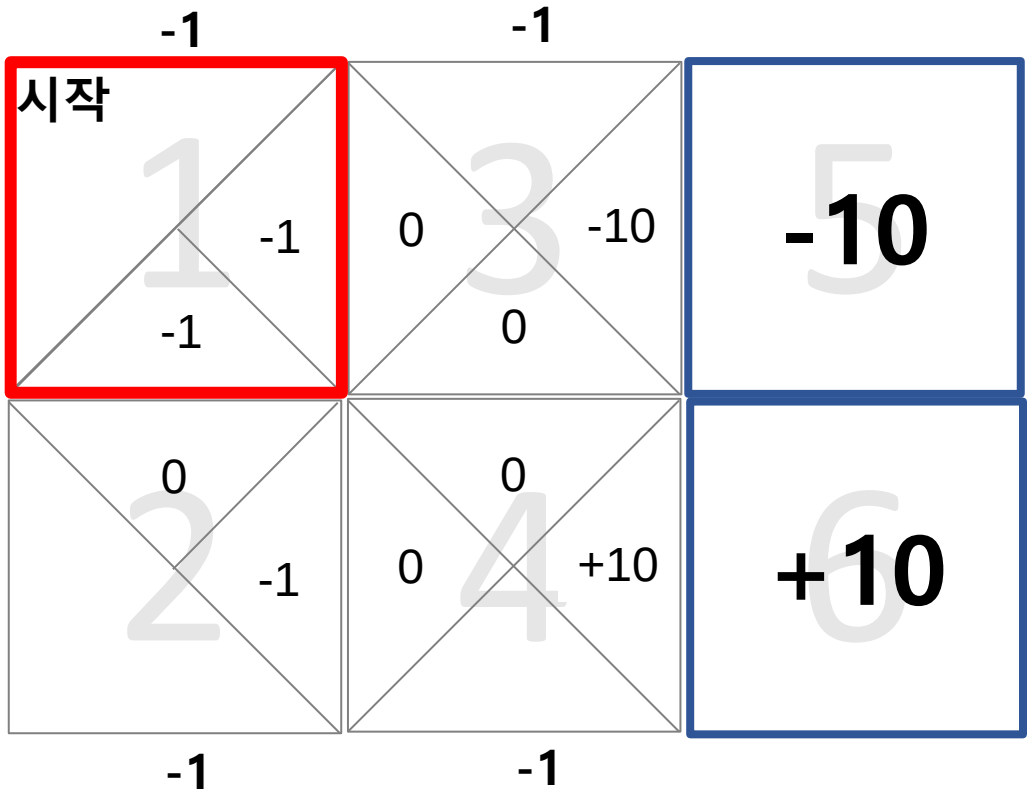


$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$\hat{Q}(s_t, a_t)$

$\alpha = 1, \gamma = 1$

(1,↓) (2,→) (4,→) 6



(1,↓) (2,→) (4,↑) (3,→) 5

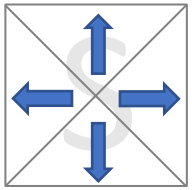
강화학습

$Q(s_t, a_t)$

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$$\hat{Q}(s_t, a_t)$$

$\alpha = 1, \gamma = 1$

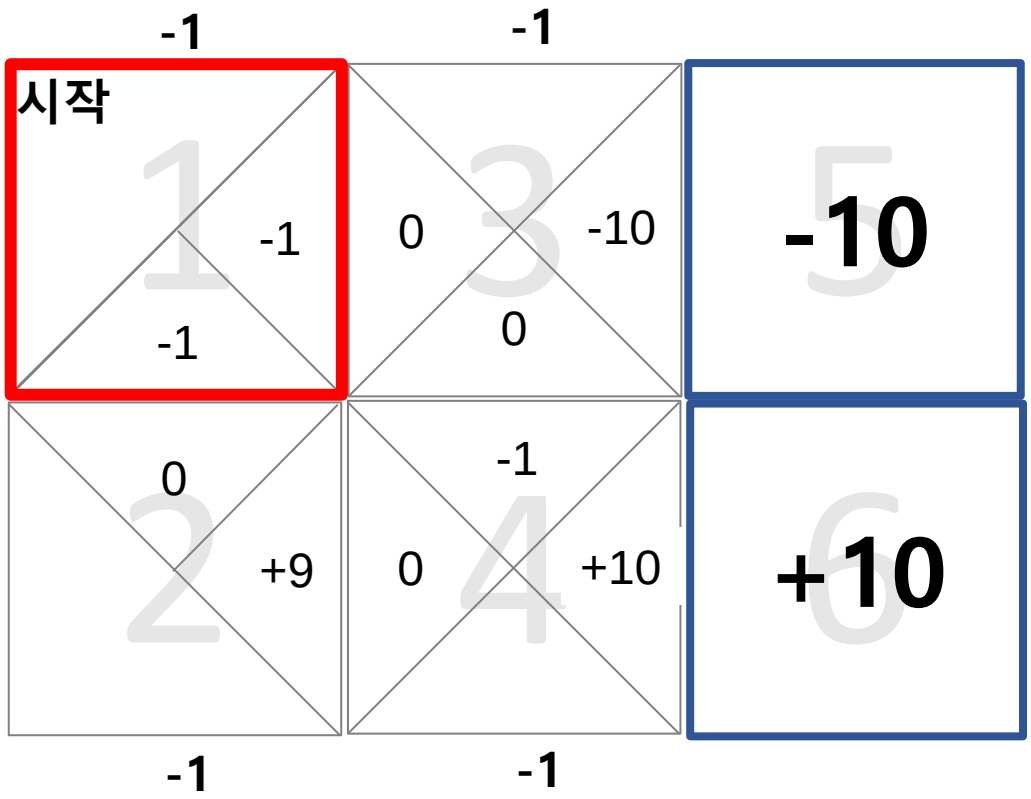


-1업뎃 +0업뎃

(1,↓) (2,→) (4,↑)

(3,→) 5

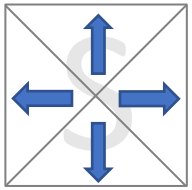
-1 업뎃



(1,→) (3,↓) (4,↑) (3,→) 5

강화학습

$Q(s_t, a_t)$

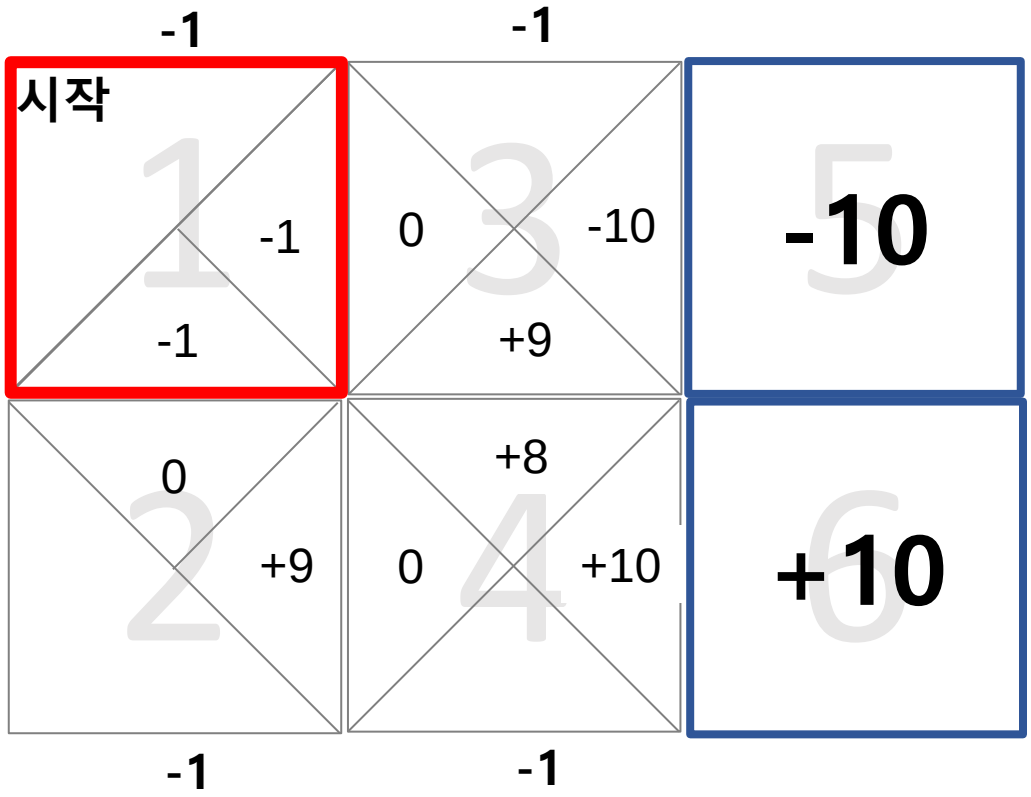


$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$$\hat{Q}(s_t, a_t)$$

$\alpha = 1, \gamma = 1$

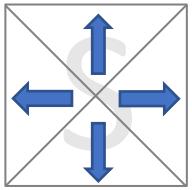
(1,→) (3,↓) (4,↑)
(3,→) 5



(1,↓) (2,→) (4,←) (2,↑) (1,→) (3,→) 5

강화학습

$Q(s_t, a_t)$

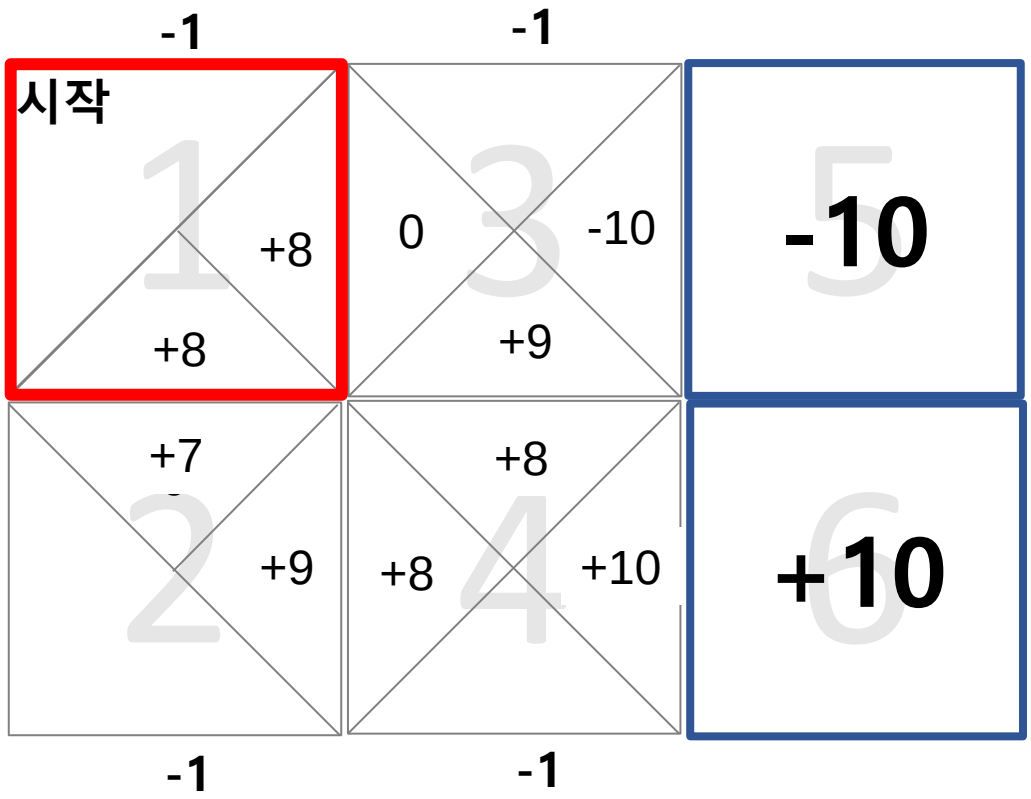


$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$$\hat{Q}(s_t, a_t)$$

$\alpha = 1, \gamma = 1$

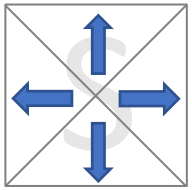
(1,↓) (2,→) (4,←)
(2,↑) (1,→) (3,→) 5



.....

강화학습

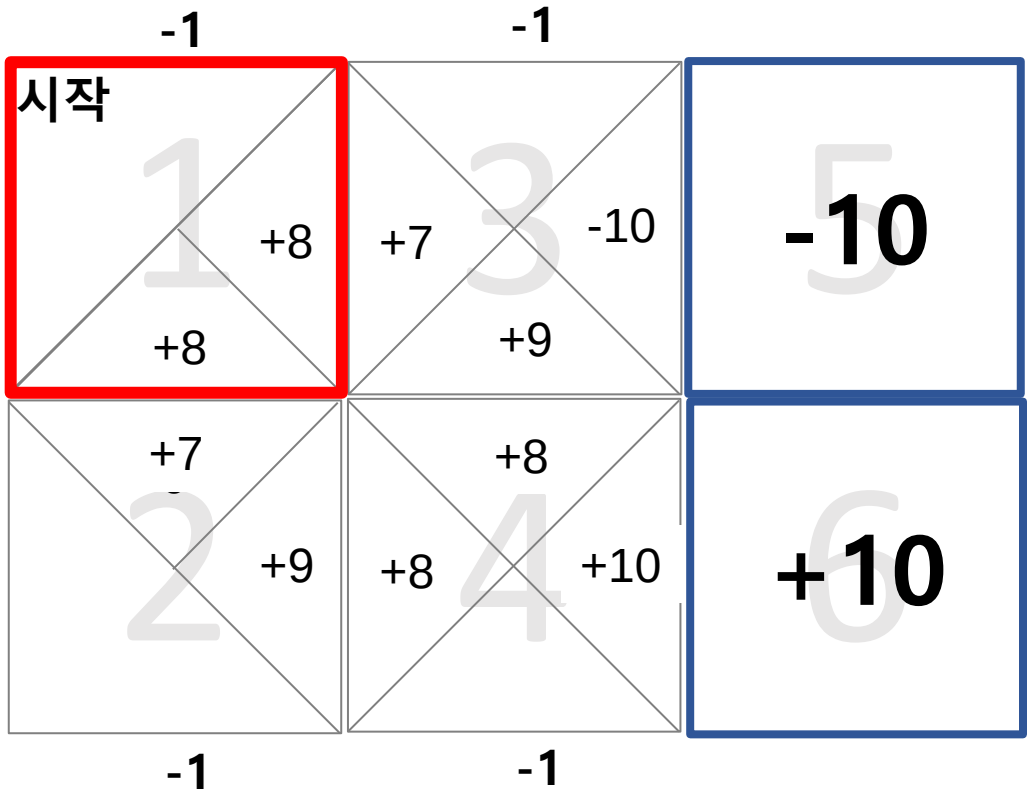
$Q(s_t, a_t)$

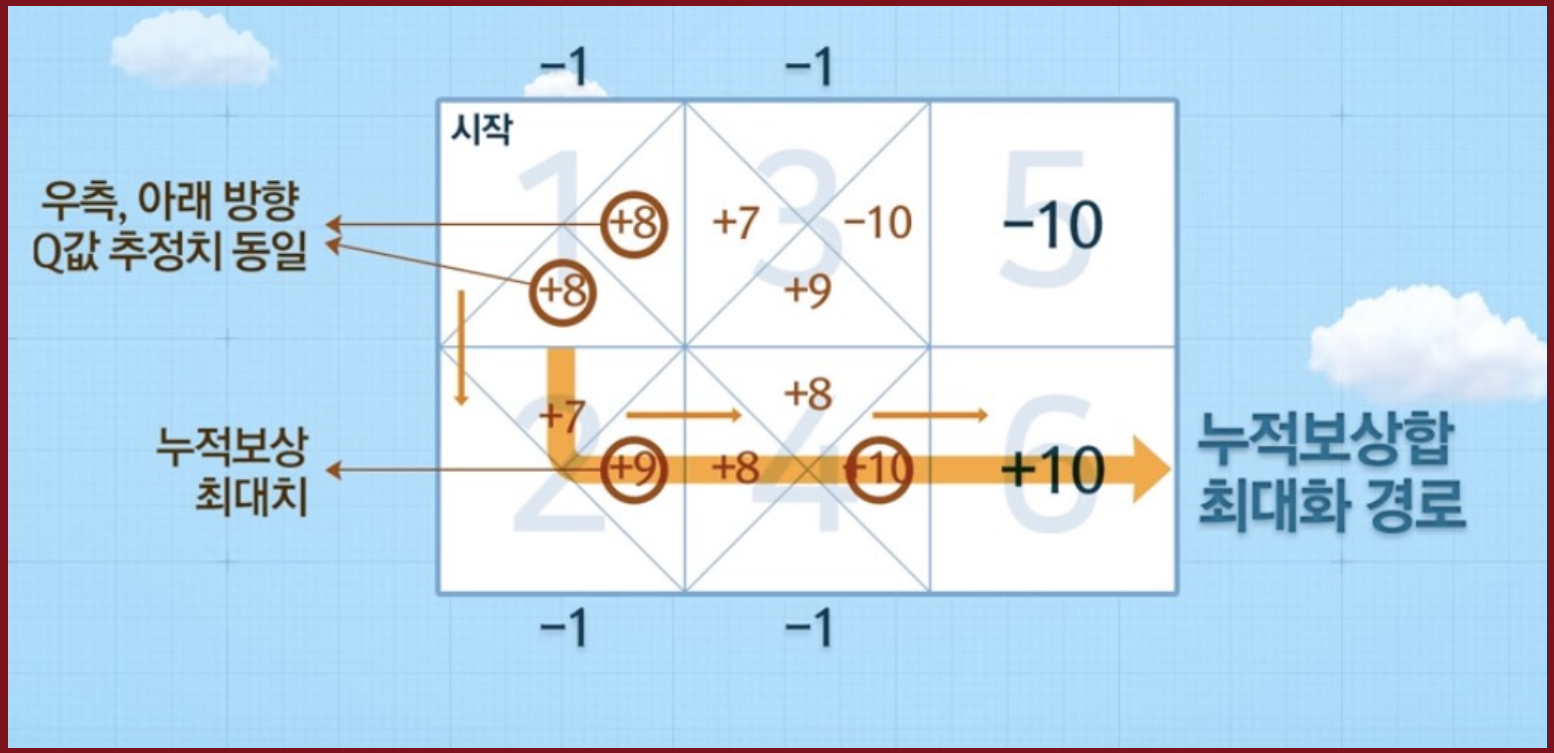


$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right)$$

$$\hat{Q}(s_t, a_t)$$

$\alpha = 1, \gamma = 1$




-1
-1
시작
우측, 아래 방향
Q값 추정치 동일
+8
+8
+7
-10
-10
+9
+8
+7
누적보상
최대치
+9
+8
-10
+10
누적보상합
최대화 경로
-1
-1