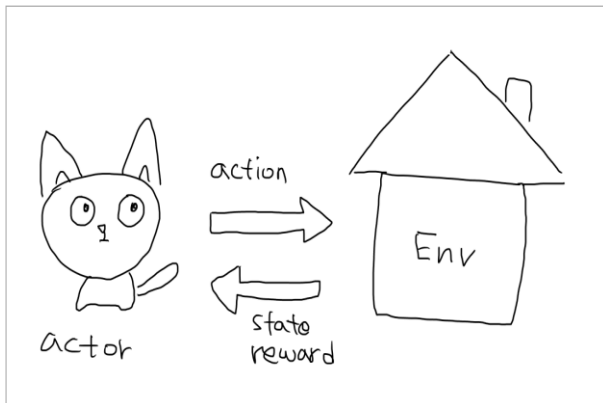
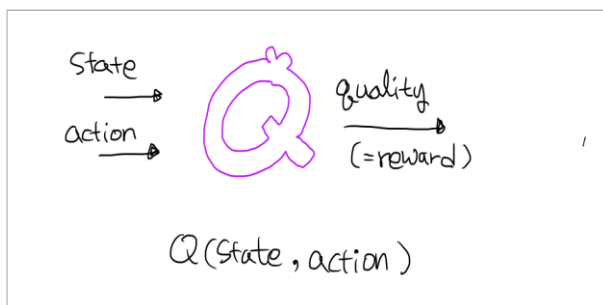


Reinforce Learning

- 강화학습은 state 와 reward 를 통해 학습한다.
- **actor** 가 주어진 **Environment** 에서 어떤 **action** 을 취할 때마다 변화된 **state** 와 **reward** 를 얻게 된다.



Q-learning



- Q 에 **State** 와 **action** 을 입력하면 **quality (reward)** 를 출력한다

Optimal policy (π), and Max Q

$$\begin{aligned} \text{Max } Q &= \max_{a'} Q(\text{state}, a') && = \text{현재 state에서 나올수 있는 가장 큰 Q의 값?} \\ \pi^*(s) &= \underset{a}{\operatorname{argmax}} Q(\text{state}, a) && = \text{현재 state에서 Q의 값을 가장 크게 만드는 a의 값?} \end{aligned}$$

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

↑
↑
↑

업데이트 될 Q값
현재 얻을 수 있는 reward
그 다음 단계에서 얻을 reward

Q-algorithm

- 1 $Q(s, a) \leftarrow 0$: Q의 모든 값을 0으로 초기화한다.
- 2 환경을 만들고 상태 (state) 를 가져온다.

(③ ~ ⑥ 을 무한 반복한다.)

- 1 어떤 행동 (action) 을 취한다.
- 2 행동에 대한 보상 (reward) 과 변화된 상태 (state') 를 받는다.
- 3 $\hat{Q}(s, a) \leftarrow r + \max_{a'} \hat{Q}(s', a')$
: 보상 (reward) 과 다음 상태 (state') 의 최대 Q 값을 더해서 Q 를 업데이트 한다.
- 4 상태 (state) 를 변화된 상태 (state') 로 바꾼다.