# Value-based, Policy-based

## Value-based

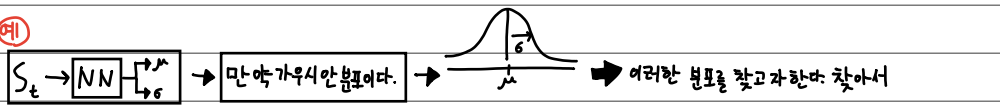
- 가치함수를 Maximization 하는 action을 고르는것
- ㉠ DQN

  greedy ➡ $\arg\max_{a_t} Q(s_t, a_t)$ ➡ 가치값이 큰 action을 고른다.

  $\varepsilon$-greedy ➡ greedy 방식에서 $\varepsilon$를 추가한것

## Policy-based

- 강화학습 자체가 좋은 policy [$P(a_t|s_t)$]를 찾는건데 Policy-based는 그 확률 분포를 찾는 것이다.
- ㉠




➡ 여러한 분포를 찾고자 한다: 찾아서 action별 continuce한 policy를 찾는다.

- Policy-based를 많이 쓰는 이유

1) continuous action을 고를때 좋다.

➡ Value-based 기반의 DQN을 보면 CNN을 거쳐서 $\begin{bmatrix} Q(s,\text{오}) \\ Q(s,\text{왼}) \end{bmatrix}$ 이런식으로 discrete 하게뽑아야한다.

그러다보니 continuous한 action (자동차 핸들)을 control 하기 힘들다.

2) stochastic policy를 찾을수있다.

➡ ㉠

| $s_0$ | $s_1$ | $s_2$ | $s_1$ | $s_3$ |
|---|---|---|---|---|
| Die | ① | Goal | ② | Die |

**Value-based 입장**

- $s_0, s_2, s_3$ 에서는 문제가 생기지 않는다.
- $s_1$ 상태에서는 ①위치에서 컴퓨터입장에서 보았을때 우측으로가는것이 value값이 크다:

  ➡ $Q(s,\text{왼}) < Q(s,\text{오})$

  하지만 ②위치 에서는 왼쪽으로 가는 것이 value 값이 크다.

  ➡ $Q(s,\text{왼}) > Q(s,\text{오})$

  → 이런 상황이 계속 반복된다.

**Policy-based 입장**

- 오른쪽 왼쪽 확률을 $\frac{1}{2}, \frac{1}{2}$로 만들면 충분히가능하다.