# Tic Tac Toe Value function

# Problem Set up > MDP

Environment : Initial state(비어 있는 보드)
(current) State: 보드의 현재 상태
Action: Agent가 board에 O나 X를 두는 행동
Reward: win -> +1, lost -> -1, draw -> 0

State Set: the location of X and O

➔매 타임 t마다 agent는 t 번째 state에 존재
➔Time t에서 action $a_t$수행 -> t+1 번째 state로 이동 -> reward 받음

# 초기 Value function

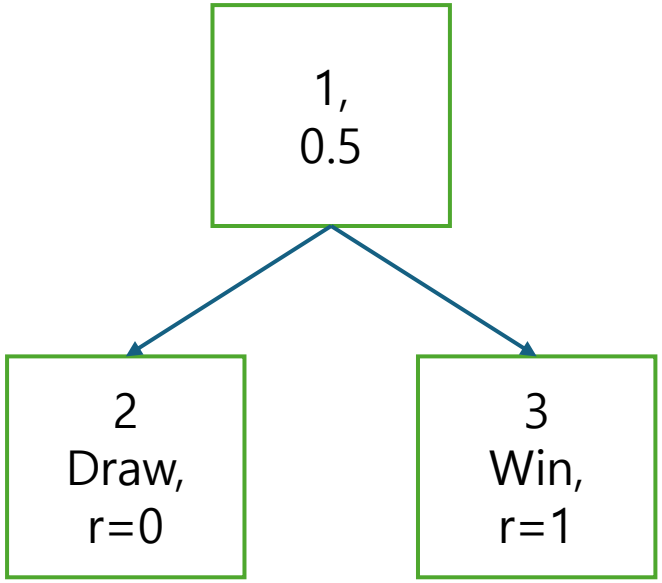
1. Leaf node state(경기가 끝난 final state) – agent가 승리한 경우 V(S) = 1
   agent가 패배한 경우 V(S) = -1
   무승부인 경우 V(S) = 0

2. Final state가 아닌 State의 초기 value – 0.5(50%의 승률)

# Value 값 계산 방식

$$V(S_t) \leftarrow V(S_t) + \alpha\left[V(S_{t+1}) - V(S_t)\right]$$

각 수에 대한 가중치를 점점 갱신해 나간다 => 즉 각 state 마다 가중치를 갖고,
이를 value function을 통해 계산한다.




State 1의 value: V(1), TD(0) 방식에서 구할 때는 최대값을 선택하거나, Weighted Average 이용

최대값 선택: V(1) ← 0.5 + $\alpha([\max(0,1) - 0,5]$
V(1)의 value가 0.5 + 0.5$\alpha$로 update 된다.

⇒ 각각의 중간 state들은 leaf node의 reward 값의 영향을 받아 계산된다.

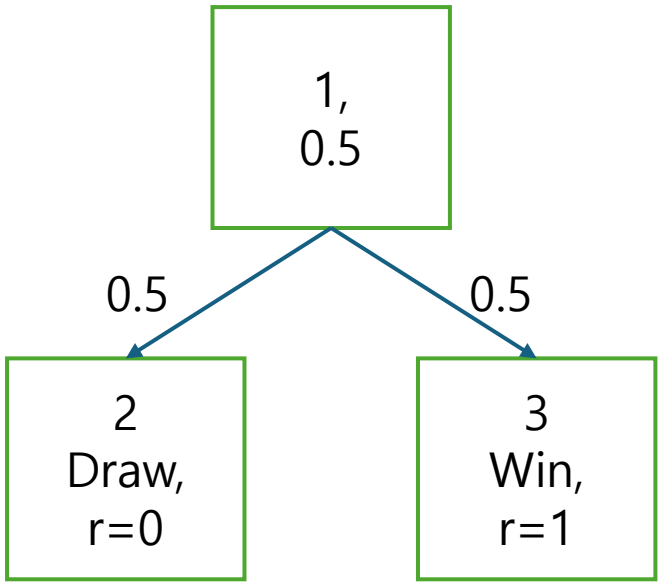
다음 State의 value를 알아야 값을 update 할 수 있기 때문에, 게임이 종료된 후에만 update 된다.

게임이 끝난 후 역방향으로 value를 계산한다.

# Value 값 계산 방식

$$V(S_t) \leftarrow V(S_t) + \alpha [V(S_{t+1}) - V(S_t)]$$

Weighted averag를 이용할 경우 천이 확률 필요




Explore만 수행하는 경우 -> Uniform distribution
Explore-Exploit을 수행하는 경우 -> $\varepsilon - greedy$를 사용해, $\varepsilon$ 비율로 무작위 탐험을 진행한 후 1- $\varepsilon$ 비율로 가장 좋은 행동 선택