마르코프 결정 과정

Markov Decision Processes

- **A Markov decision process is an MRP with decisions:**
  $\langle S, A, P, R, \gamma \rangle$
  - A set of states $S = \{s_1, s_2, \dots, s_n\}$
  - A set of actions $A = \{a_1, a_2, \dots, a_m\}$
  - Transition function $P: S \times A \to S$, $P^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
  - Reward function $R: S \times A \to \mathbb{R}$, $R^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
  - Discount factor $\gamma \in [0,1]$

추가적으로 Action을 보고 결정

$$\pi(a|s) = P[A_t = a | S_t = s]$$

Policy : 어떤 state에서 어떤 action을 취할 확률(현재 state에만 의존), 시간에 따라 변화하지 않음.

Value Functions

$$V_\pi(s) = \mathbb{E}_\pi [G_t | S_t = s]$$

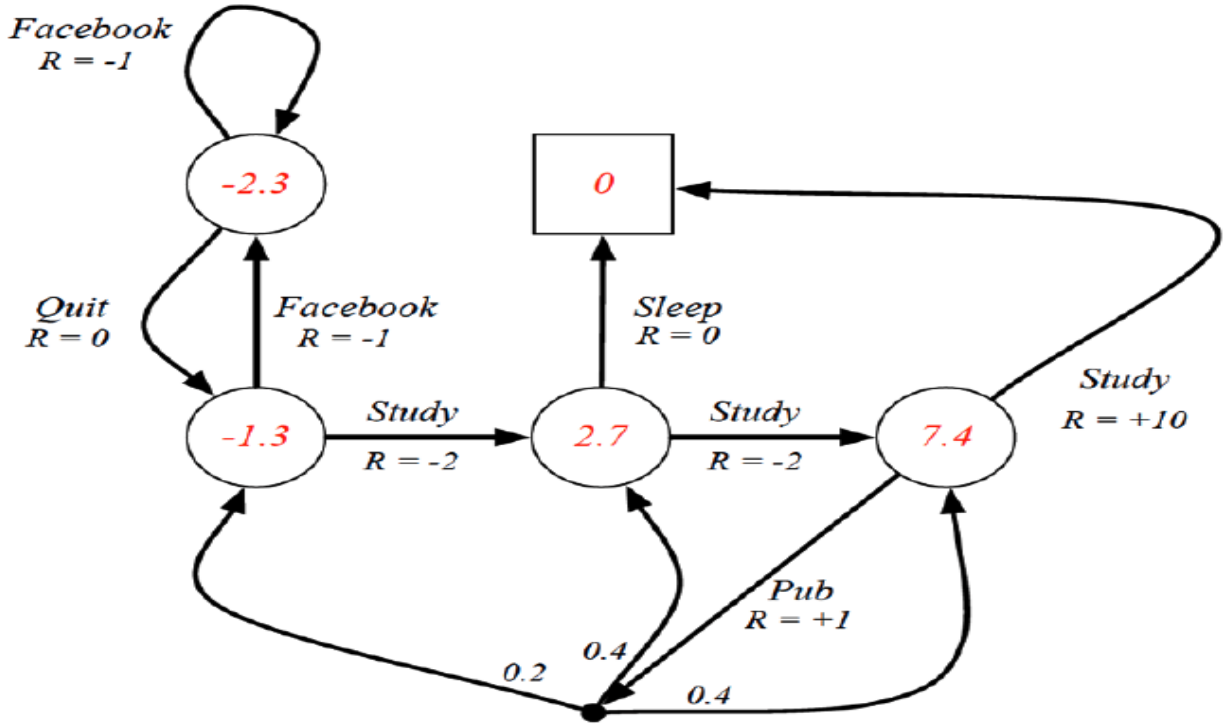
V : state value function
S에 대한 함수 (s만 주어지면 결정)
State가 주어졌을 때 return에 대한 expectation

$$Q_\pi(s, a) = \mathbb{E}_\pi [G_t | S_t = s, A_t = a]$$

Q : state action value function
S와 A에 대한 함수(s와 a가 주어지면 결정)
State가 주어졌을 때 어떤 action을 취했을 때 return에 대한 expectation

Example

- **Random policy with $\gamma = 1$**
  - $V_\pi(s)$ for $\pi(a|s) = 0.5$



검은색 동그라미 : action
Pub : 어떤 action을 취하면 보상을 받고(R) 다른 state로 갈 확률이 존재한다.

Bellman Expectation Equation for MDPs

- **The value function can be decomposed into two parts:**
  - Immediate reward R_{t+1}
  - Discounted value of successor state $\gamma V(s_{t+1})$
- **The state-value function can be decomposed**

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[G_t|s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots|s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)|s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(s_{t+1})|s_t = s] \end{aligned}$$

- **The state-value function can be decomposed**

$$V_\pi(s) = \mathbb{E}_\pi\,[R_{t+1} + \gamma V(S_{t+1})|S_t = s]$$

- **The action-value function can be decomposed**

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

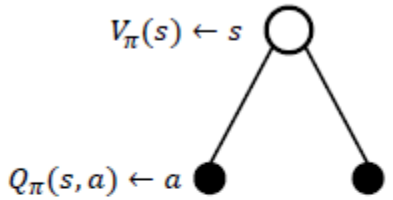
우리가 정의한 value function에 대해서
현재 state의 value function과
그 다음 state의 value function의
상관관계를 정의

Value function
- Immediate reward expectation
- 다음 state로 갔을 때의 reward expectation

Bellman Equation

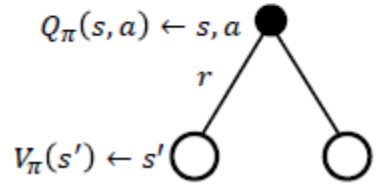
- **Bellman expectation equation for V_π**

$$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s,a)$$



- **Bellman expectation equation for Q_π**

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s')$$



1. 어떤 state에서 다른 action
2. Action에서 다른 state

Policy : 어떤 state에서 action을 선택할 확률

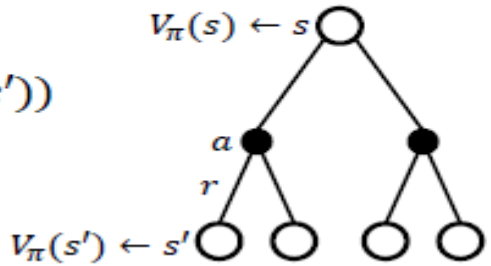
V : state에서 취하게 될 모든 action에 대해서 확률 분포를 다 구한다음 그 확률 분포대로 Q 값을 곱한다.

Q : 내가 state S에서 A를 취했을 때의 Reward + 누적합[Pss(모든 action 확률)*State Value값]

Bellman Equation

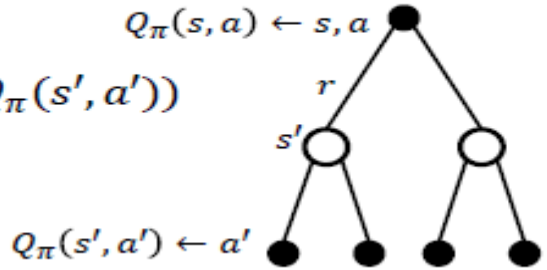
- **Bellman expectation equation for V_π (2)**

$$V_\pi(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s'))$$



- **Bellman expectation equation for Q_π (2)**

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a (\sum_{a' \in A} \pi(a'|s') Q_\pi(s',a'))$$

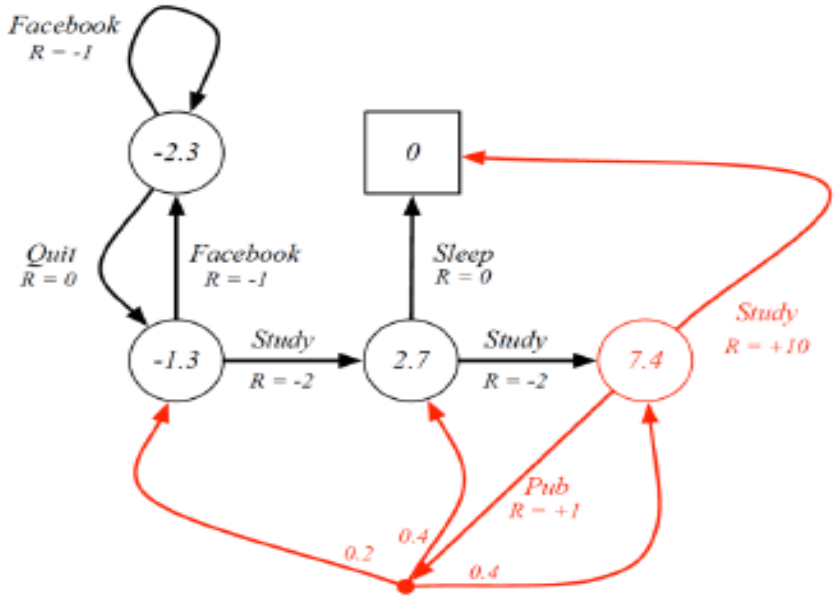


V는 V에 대한 equation으로만,
Q는 Q에 대한 equation으로만
표현

Example

$$V_\pi(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s'))$$

$$V_\pi(s) = 0.5 * (1 + 0.2 * (-1.3) + 0.4 * 2.7 + 0.4 * 7.4) + 0.5 * 10 = 7.4$$



Random : 0.5고정(어떤 action을 취할 확률)
Study선택 : 0.5*10
Pub 선택 : 0.5*(1+0.2*
(-1.3)+0.4*2.7+0.4*7.4)
(C1으로 갈 확률*보상, C2로 갈 확률*보상,
C3로 갈 확률*보상)

Optimal Value Functions

- The optimal state value function $V_*(s)$ is the max value function over all policies

$$V_*(s) = \max_\pi V_\pi(s)$$

- The optimal action-value function $Q_*(s,a)$ is the maximum action-value function over all policies

$$Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

- Theorems: For any MDP
  - There exists an optimal policy $\pi_* \geq \pi, \forall \pi$
  - All optimal policies achieve the optimal state-value, $V_{\pi_*}(s) = V_*(s)$
  - All optimal policies achieve the optimal action-value, $Q_{\pi_*}(s,a) = Q_*(s,a)$

주어진 문제에서 모든 가능한 policy를 고려했을 때 가장 최대가 되는 value function

모든 MDP에 항상 optimal policy가 존재함
Optimal policy = 최적의 value function

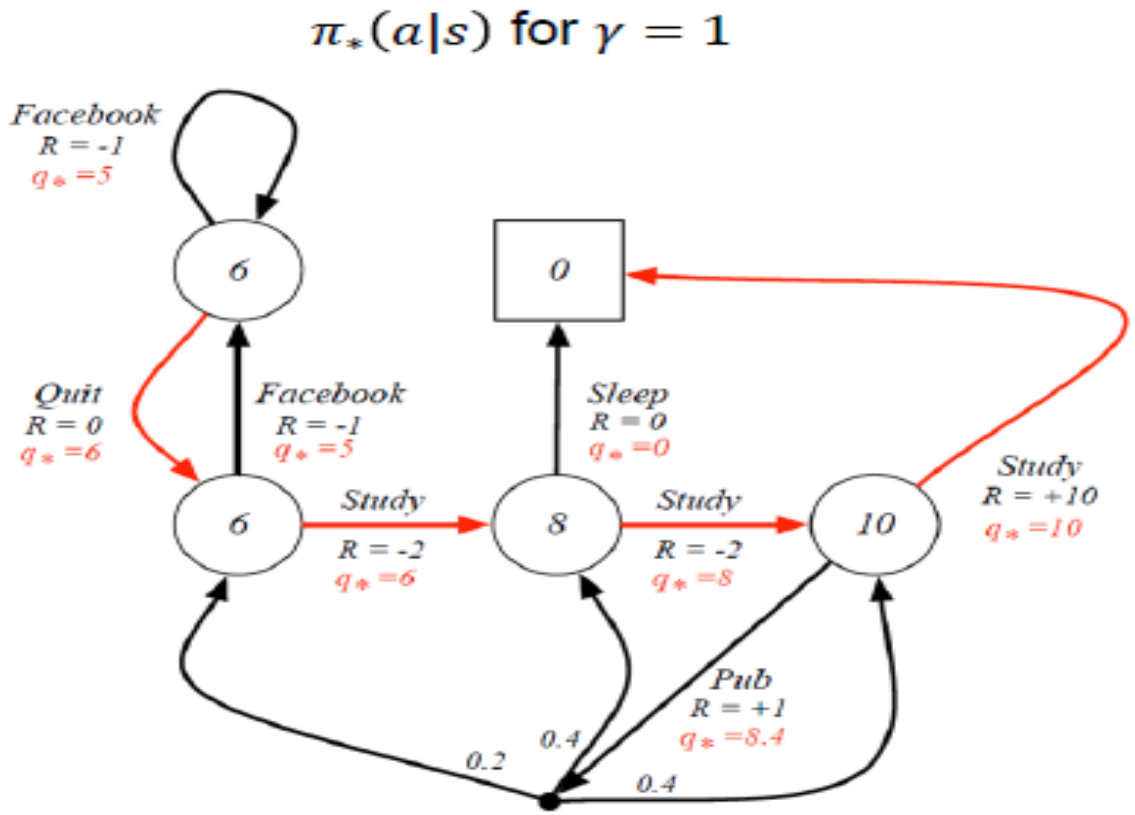
An optimal policy can be found by maximizing over $Q_*(s,a)$

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \underset{a \in A}{\operatorname{argmax}}\, Q_*(s,a) \\ 0 & \text{otherwise} \end{cases}$$

- If we know $Q_*(s,a)$, we immediately have the optimal policy

어떤 state가 있었을 때 그 state에서 취할 수 있는 action이 여러 개 있다고 할 때, 첫 번째 Q, 두 번째 Q … 값들을 따진 다음 항상 가장 큰 Q값만 가지는 action을 취한다.

Optimal Policy Example



C3에 있을 때 pub, study action을 취할 수 있다.
Study : 10
Pub : 8.4
Study가 더 크므로 Study action을 취한다.