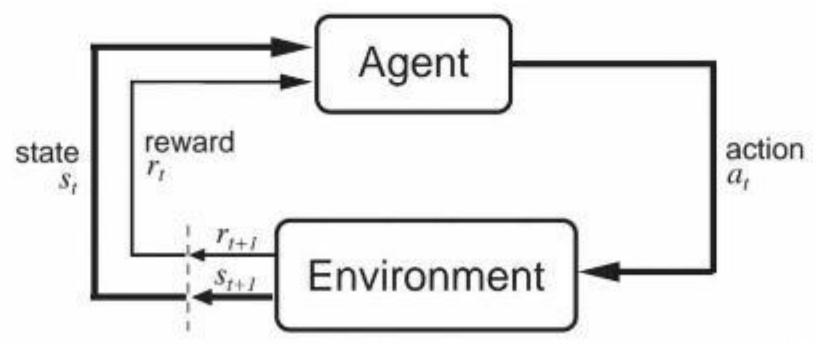
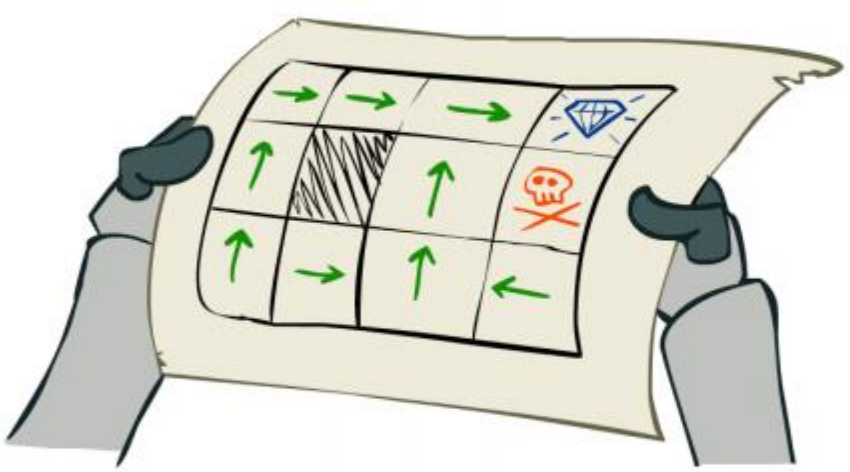
마르코브 의사결정 프로세스 가치함수 (Value function)





가치함수 (Value function)

- **상태-가치 함수 (State-value function)**
  - 의사결정 시점 t에서의 상태가 s일 때, t시점부터 정책 π를 따른 경우 기대 누적 보상합

$$v_t^{\pi}(s) = E^{\pi}[G_t | S_t = s] = E^{\pi}\left[\sum_{k=0} \gamma^k R_{t+k} \,|S_t = s\right]$$

- **행동-가치 함수 (Action-value function)**
  - 의사결정 시점 t에서의 상태가 s일 때 행동 a를 취한 후, 다음 시점 $t+1$ 이후부터 정책π를 따른 경우 기대 누적 보상합

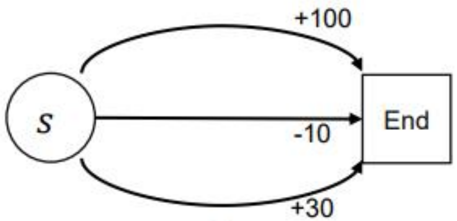
$$Q_t^{\pi}(s,a) = E^{\pi}[G_t | S_t = s, A_t = a] = E^{\pi}\left[\sum_{k=0} \gamma^k R_{t+k} \,|S_t = s, A_t = a\right]$$

가치함수 (Value function)

- **상태-가치 함수 (State-value function)**
  - 상태 s에서 정책 π를 따른 경우 기대 누적 보상합
  - 벨만 기대 방정식 (Bellman expectation equation)

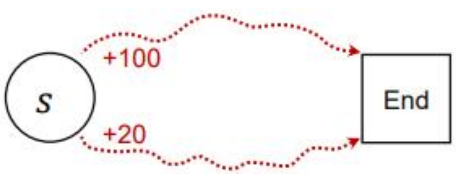
$$v_t^{\pi}(s) = E^{\pi}[G_t | S_t = s] = E^{\pi}\left[\sum_{k=0} \gamma^k R_{t+k} \,|S_t = s\right]$$

$$= E^{\pi}[\underbrace{R_t}_{r_t(s,\delta_t(s))} + \gamma v_{t+1}^{\pi}(s_{t+1}) | S_t = s]$$



π_1

$$v_t^{\pi_1}(s) = \frac{100 - 10 + 30}{3} = +40$$



π_2

$$v_t^{\pi_2}(s) = \frac{100 + 20}{2} = +60$$

가치함수 (Value function)

- **행동-가치 함수 (Action-value function)**
  - 상태 s에서 행동 a를 취한 후, 이후 상태들에 대해서 정책π를 따른 경우 기대 누적 보상합

$$Q_t^{\pi}(s,a) = E^{\pi}[G_t | S_t = s, A_t = a] = E^{\pi}\left[\sum_{k=0} \gamma^k R_{t+k} \,| S_t = s, A_t = a\right]$$

$$= E^{\pi}[\underset{r_t(s,a)}{R_t} + \gamma Q_{t+1}^{\pi}(s_{t+1}, \delta_{t+1}(s_{t+1})) | S_t = s, A_t = a]$$

$$Q_t^{\pi}(\boldsymbol{s}, \boldsymbol{Right}) = \mathbf{+100}$$

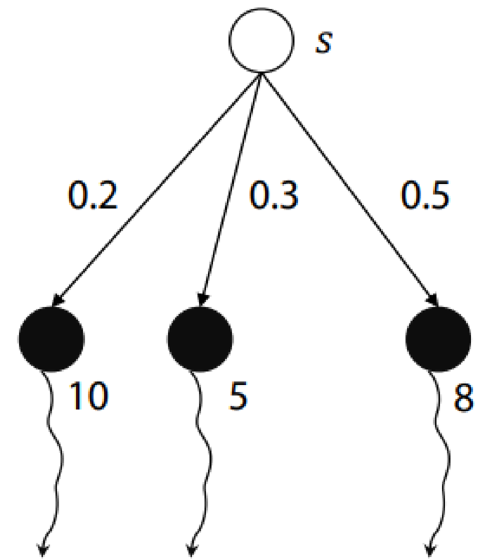
$$Q_t^{\pi}(\boldsymbol{s}, \boldsymbol{Left}) = \mathbf{+20}$$

가치함수 (Value function)

- $v_t^\pi(s)$와 $Q_t^\pi(s,a)$간의 관계 ($\pi = \{\delta_t\}_{\forall t}$)

$$v_t^\pi(s) = Q_t^\pi(s, \delta_t(s))$$

$$v_t^\pi(s) = \sum_{a \in A_s} \delta_t(a \mid s) Q_t^\pi(s,a)$$

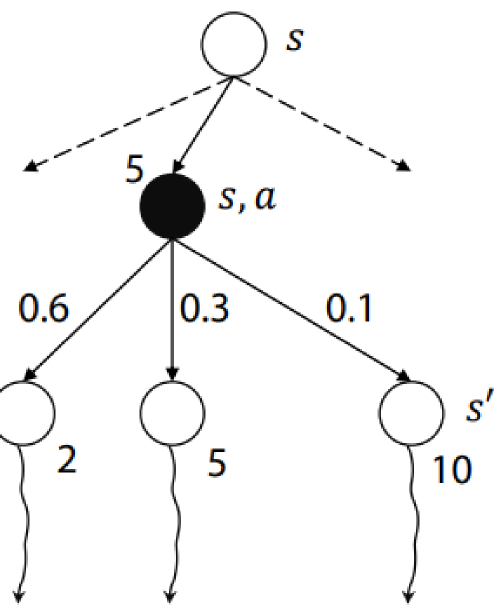


$$v_t^\pi(s) = 0.2 \times 10 + 0.3 \times 5 + 0.5 \times 8 = 7.5$$

가치함수 (Value function)

- $v_t^\pi(s)$와 $Q_t^\pi(s,a)$간의 관계 ($\pi = \{\delta_t\}_{\forall t}$)

$$Q_t^\pi(s,a) = r_t(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) v_{t+1}^\pi(s')$$



$$Q^\pi(s,a) \\ = 5 + \gamma \times (0.6 \times 2 + 0.3 \times 5 + 0.1 \times 10)$$

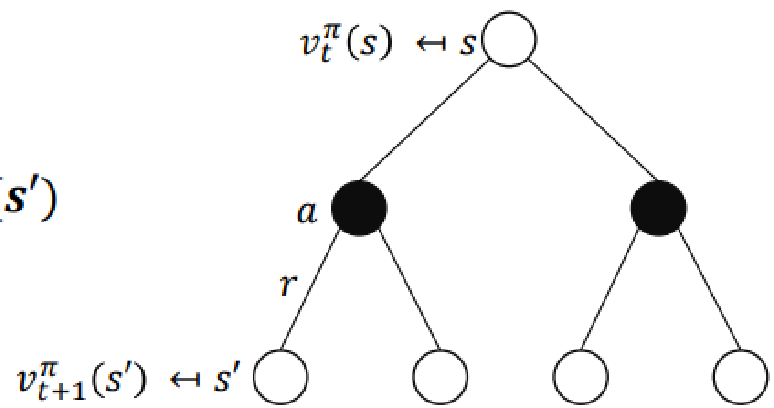
가치함수 (Value function)

$$v_t^\pi(s) = Q_t^\pi(s, \delta_t(s))$$

$$v_t^\pi(s) = \sum_{a \in A_s} \delta_t(a \mid s) Q_t^\pi(s, a)$$

$$Q_t^\pi(s, a) = r_t(s, a) + \gamma \sum_{s' \in S} P(s' \mid s, a) v_{t+1}^\pi(s')$$

$$\begin{aligned} v_t^\pi(s) &= Q_t^\pi\big(s, \delta_t(s)\big) \\ &= r_t\big(s, \delta_t(s)\big) + \gamma \sum_{s' \in S} P(s' \mid s, \delta_t(s)) v_{t+1}^\pi(s') \end{aligned}$$



$$v_t^\pi(s) = \sum_{a \in A_s} \delta_t(a \mid s) \left(r_t(s, a) + \gamma \sum_{s' \in S} P(s' \mid s, a) v_{t+1}^\pi(s') \right)$$

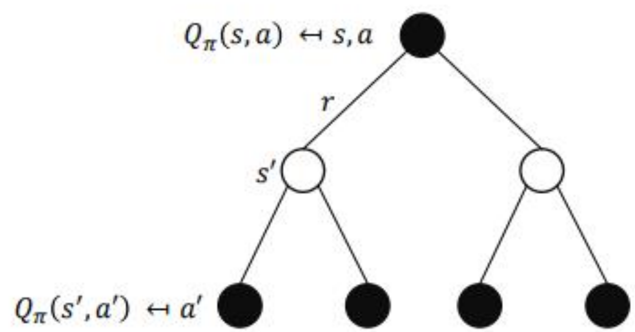
가치함수 (Value function)

$$Q_t^\pi(s,a) = r_t(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) v_{t+1}^\pi(s')$$

$$v_t^\pi(s) = Q_t^\pi(s, \delta_t(s))$$

$$v_t^\pi(s) = \sum_{a \in A_s} \delta_t(a \mid s) Q_t^\pi(s,a)$$

$$Q_t^\pi(s,a)$$
$$= r_t(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) Q_t^\pi(s', \delta_{t+1}(s'))$$



$$Q_t^\pi(s,a)$$
$$= r_t(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) \left(\sum_{a' \in A_{s'}} \delta_{t+1}(a' \mid s') Q_{t+1}^\pi(s',a') \right)$$

가치함수 (Value function)

- **최적 정책 (optimal policy)** π^*
  - 모든 $s \in S$와 모든 π 에 대해, $v_1^{\pi^*}(s) \geq v_1^{\pi}(s)$

- **(최적) 가치함수 ((optimal) value function)**
  - $v_1^*(s) = \max_{\pi} v_1^{\pi}(s)$
    - (감가율이 반영된) 기대 누적보상합의 최대값
  - π^*가 최적정책 $\Leftrightarrow$ 모든 $s \in S$에 대해, $v_1^{\pi^*}(s) = v_1^*(s)$

가치함수 (Value function)

- **(최적) 가치함수 $v_t(s_t)$**
  - 의사결정 시점 t의 상태 s_t에서 이후 남은 기간 동안의 총 예상 누적 보상합의 최대값

- **벨만 최적 방정식(Bellman optimality equation)**

$$v_t^*(s_t) = \max_{a_t \in A_{s_t}} \{r_t(s_t, a_t) + \gamma E[v_{t+1}^*(s_{t+1})]\}$$

To find an action a_t that maximizes

Expected immediate reward in period t

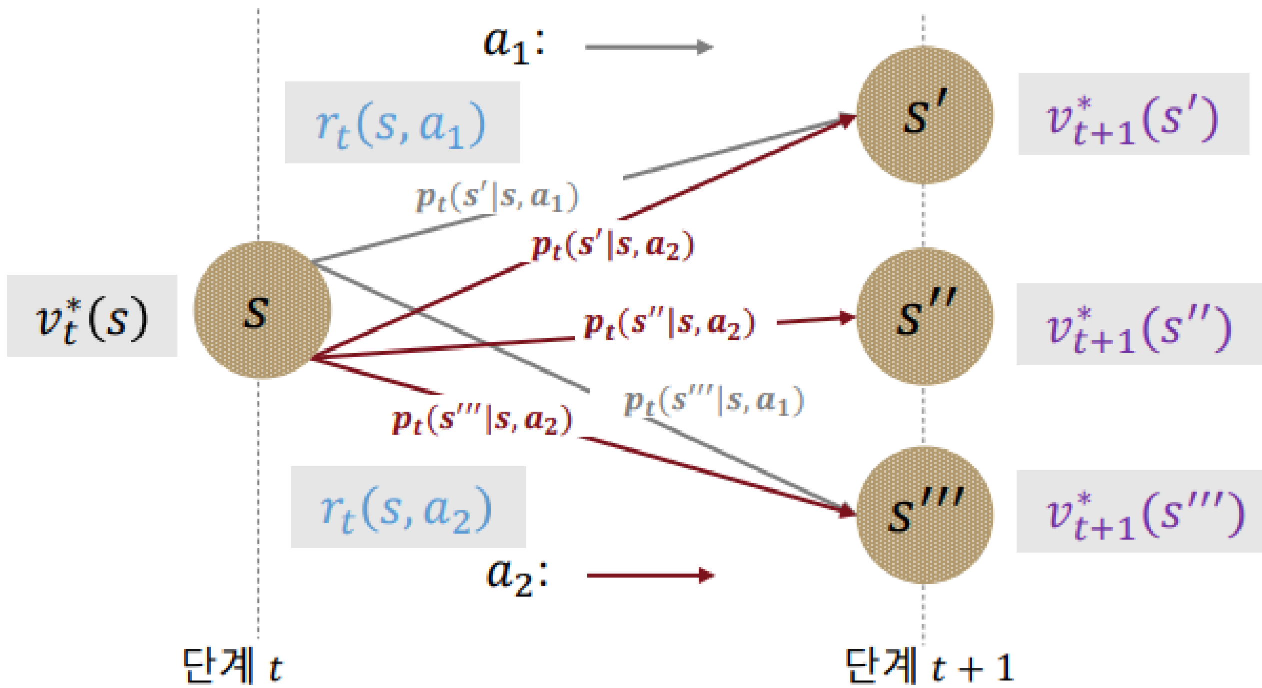
+

Expected maximum total remaining rewards in periods $t+1, t+2, \ldots$ (expected maximum total reward-to-go)

v.s. $v_t^*(s_t) = \max_{a_t}\{r_t(s_t, a_t) + v_{t+1}^*(f_t(s_t, a_t))\}$

가치함수 (Value function)

$$v_t^*(s_t) = \max_{a_t \in A_{s_t}} \left\{ r_t(s_t, a_t) + \gamma \sum_{j \in S} p(j|s_t, a_t) v_{t+1}^*(j) \right\}$$



가치함수 (Value function)

- **Finite-horizon MDP 해법**
  - 역진 귀납법 (backward induction)

1. $t = N$ 설정. 모든 $s \in S$에 대해 $v_N^*(s) = r_N(s)$
2. $t \leftarrow t-1$ 로 설정 후, 모든 $s \in S$에 대해 하기 문제 해결

$$v_t^*(s) = \max_{a \in A_s} \left\{ r_t(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_{t+1}^*(s') \right\}$$

$$A_{s,t}^* = \operatorname*{argmax}_{a \in A_s} \left\{ r_t(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_{t+1}^*(s') \right\}$$

3. $t = 1$이면 종료. 아니면 단계 2로 이동

$\boldsymbol{\pi}^* = (\boldsymbol{\delta}_1^*, \boldsymbol{\delta}_2^*, \dots, \boldsymbol{\delta}_{N-1}^*)$ **with** $\boldsymbol{\delta}_t^*(s) \in \boldsymbol{A}_{s,t}^*$ **for every** $\boldsymbol{t}$ **and** $\boldsymbol{s}$