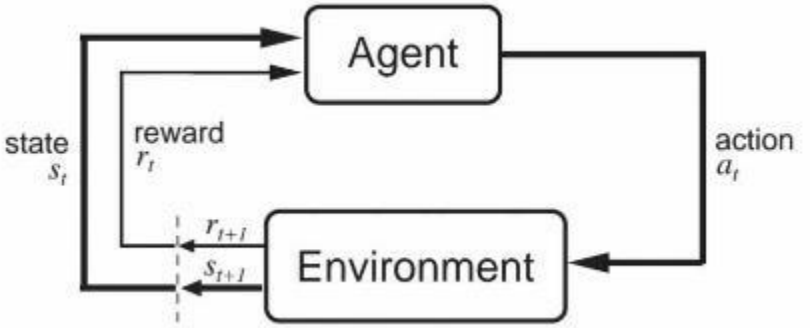
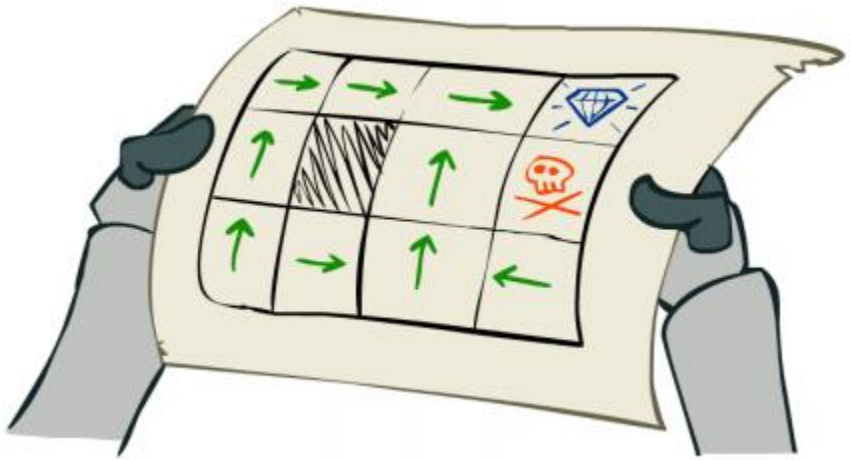
정책반복알고리즘
Policy Interation





Infinite-horizon MDP 알고리즘

- **정책 반복 (Policy iteration) 알고리즘**
  - 초기화
    - 임의의 정책 π를 선택
  - 반복 (정책의 변화가 없을 때까지)
    - 정책 평가 (policy evaluation)
      - 현 정책 π를 평가
        - $v^{\pi}(s) = r(s, \delta(s)) + \gamma \sum_{s'} P\left(s' | s, \delta(s)\right) v^{\pi}(s')$

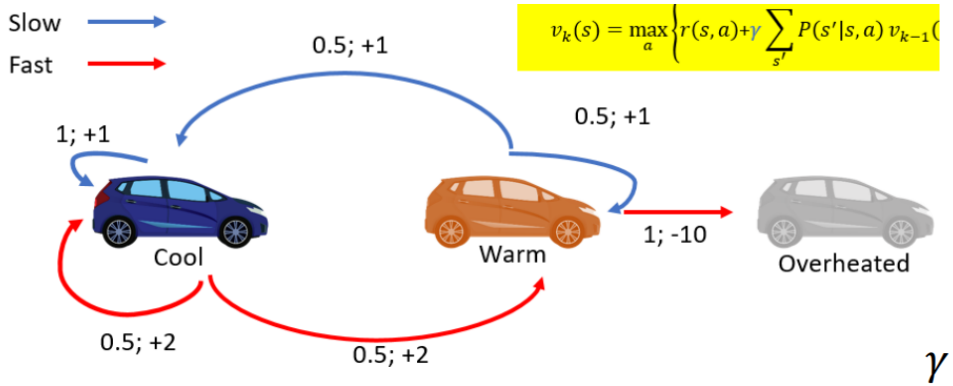
$$V^{\pi} = (I - \gamma P_{\pi})^{-1} R_{\pi}$$

    - 정책 개선 (policy improvement)
      - 모든 s에 대해, $\pi'(s) = \underset{a}{\operatorname{argmax}} \{r(s, a) + \gamma \sum_{s'} P(s' | s, a)\, v^{\pi}(s')\}$ 계산

$$v^{\pi'} \geq v^{\pi}$$

    - 정책 업데이트: $\pi \leftarrow \pi'$

Infinite-horizon MDP 알고리즘



$$v_k(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v_{k-1}(\right.$$

$$\gamma = 0.8$$

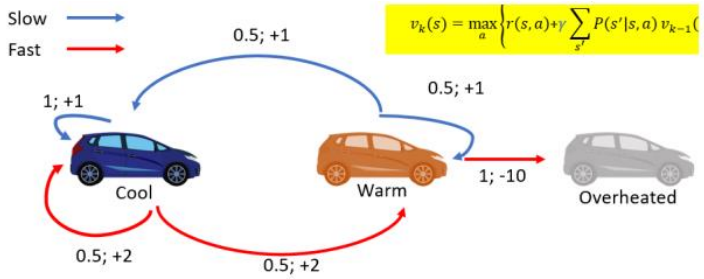
$$\pi_1 = (S, S, -) \qquad P_{\pi_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad R_{\pi_1} = \begin{bmatrix} +1 \\ +1 \\ 0 \end{bmatrix}$$

$$\pi_2 = (S, F, -) \qquad P_{\pi_2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad R_{\pi_2} = \begin{bmatrix} +1 \\ -10 \\ 0 \end{bmatrix}$$

$$\pi_3 = (F, S, -) \qquad P_{\pi_3} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad R_{\pi_3} = \begin{bmatrix} +2 \\ +1 \\ 0 \end{bmatrix}$$

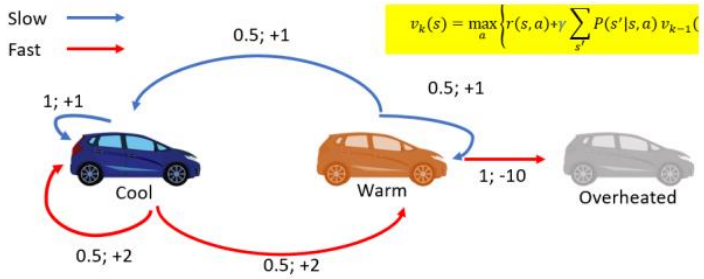
$$\pi_4 = (F, F, -) \qquad P_{\pi_4} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad R_{\pi_4} = \begin{bmatrix} +2 \\ -10 \\ 0 \end{bmatrix}$$

Infinite-horizon MDP 알고리즘



- **초기화**
  - $\pi = (S, S, -)$
- **반복(1)**
  - 정책 평가
    - $v^{\pi}(C) = r(C,S) + 0.8\sum_{s'} P(s'|C,S)v^{\pi}(s')$
      $\rightarrow v^{\pi}(C) = 1 + 0.8(1 \times v^{\pi}(C)) \rightarrow v^{\pi}(C) = 5$
    - $v^{\pi}(W) = r(W,S) + 0.8\sum_{s'} P(s'|W,S)v^{\pi}(s')$
      $\rightarrow v^{\pi}(W) = 1 + 0.8(0.5v^{\pi}(C) + 0.5v^{\pi}(W)) \rightarrow v^{\pi}(W) = 5$
    - $v^{\pi}(O) = r(O,-) + 0.8\sum_{s'} P(s'|O,-)v^{\pi}(s')$
      $\rightarrow v^{\pi}(O) = 0 + 0.8(1 \times v^{\pi}(O)) \rightarrow v^{\pi}(O) = 0$
  - 정책 평가
    - $\pi'(C) = \underset{a}{argmax}\{r(C,a) + \gamma\sum_{s'} P(s'|C,a)v^{\pi}(s')\}$
      $= argmax\{1 + 0.8 \times v^{\pi}(C), 2 + 0.8 \times (0.5v^{\pi}(C) + 0.5v^{\pi}(W))\}$
      $= argmax\{5,6\} = F$
    - $\pi'(W) = \underset{a}{argmax}\{r(W,a) + \gamma\sum_{s'} P(s'|W,a)v^{\pi}(s')\}$
      $= argmax\{1 + 0.8 \times (0.5v^{\pi}(C) + 0.5v^{\pi}(W)), -10 + 0.8 \times v^{\pi}(O)\}$
      $= argmax\{5, -10\} = S$
    - $\pi'(O) = -$

Infinite-horizon MDP 알고리즘



- **반복(2)** $\pi = (F, S, -)$
  - 정책 평가
    - $v^{\pi}(C) = r(C,F) + 0.8\sum_{s'} P(s'|C,F)v^{\pi}(s')$,
      $\rightarrow v^{\pi}(C) = 2 + 0.8(0.5v^{\pi}(C) + 0.5v^{\pi}(W)) \rightarrow v^{\pi}(C) = 8$
    - $v^{\pi}(W) = r(W,S) + 0.8\sum_{s'} P(s'|W,S)v^{\pi}(s')$
      $\rightarrow v^{\pi}(W) = 1 + 0.8(0.5v^{\pi}(C) + 0.5v^{\pi}(W)) \rightarrow v^{\pi}(W) = 7$
    - $v^{\pi}(O) = r(O,-) + 0.8\sum_{s'} P(s'|O,-)v^{\pi}(s')$
      $\rightarrow v^{\pi}(O) = 0 + 0.8(1 \times v^{\pi}(O)) \rightarrow v^{\pi}(O) = 0$
  - 정책 평가
    - $\pi'(C) = \underset{a}{argmax}\{r(C,a) + \gamma\sum_{s'} P(s'|C,a)v^{\pi}(s')\}$
      $= argmax\{1 + 0.8 \times v^{\pi}(C), 2 + 0.8 \times (0.5v^{\pi}(C) + 0.5v^{\pi}(W))\}$
      $= argmax\{7.4, 8\} = F$
    - $\pi'(W) = \underset{a}{argmax}\{r(W,a) + \gamma\sum_{s'} P(s'|W,a)v^{\pi}(s')\}$
      $= argmax\{1 + 0.8 \times (0.5v^{\pi}(C) + 0.5v^{\pi}(W)), -10 + 0.8 \times v^{\pi}(O)\}$
      $= argmax\{7, -10\} = S$
    - $\pi'(O) = -$

$\pi = \pi'$