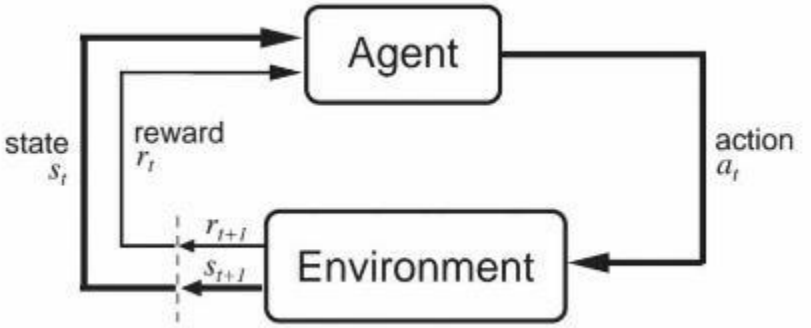
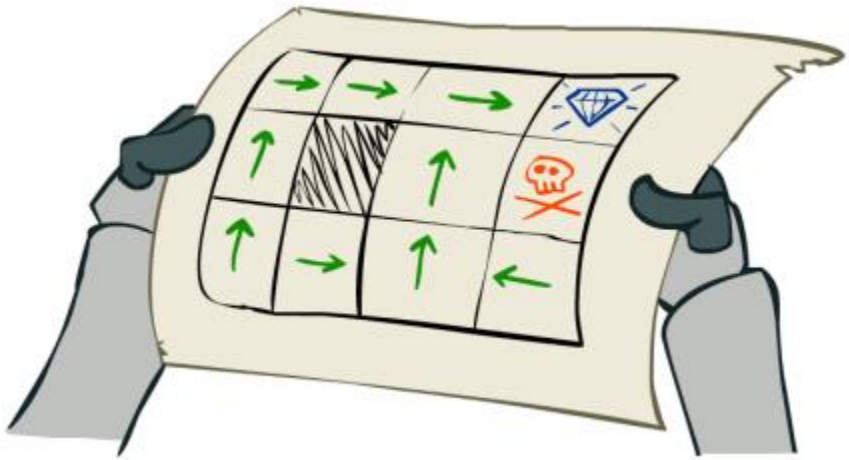
가치반복알고리즘
Value Interation





Infinite-horizon MDP 알고리즘

- **가치 반복 (value iteration) 알고리즘**

$$v^*(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\}$$

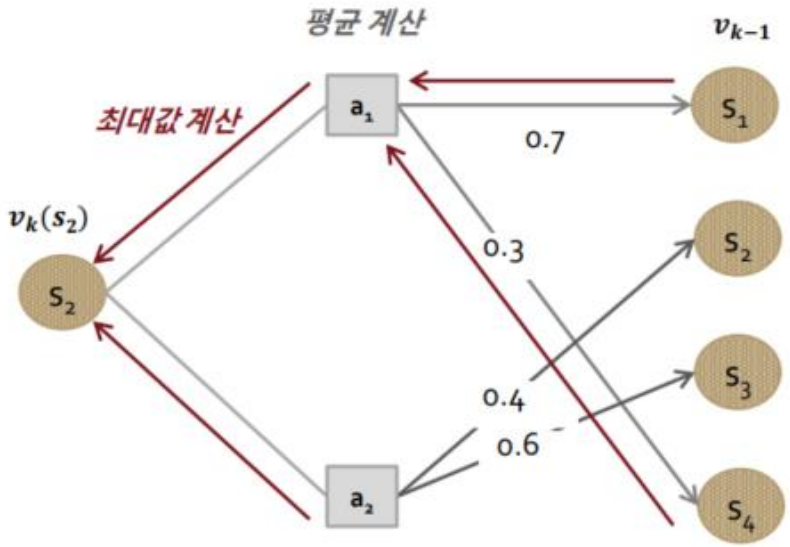
  - 초기화
    - $v_0(s) = 0$ for all s
    - $k = 1$
  - 반복 $(v_{k-1} \rightarrow v_k)$
    - $v_k(s) = \max_a \{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v_{k-1}(s') \}$
    - 만약 $\|v_k - v_{k-1}\| < \epsilon$, 종료. 아니면 $k \leftarrow k+1$

Infinite-horizon MDP 알고리즘

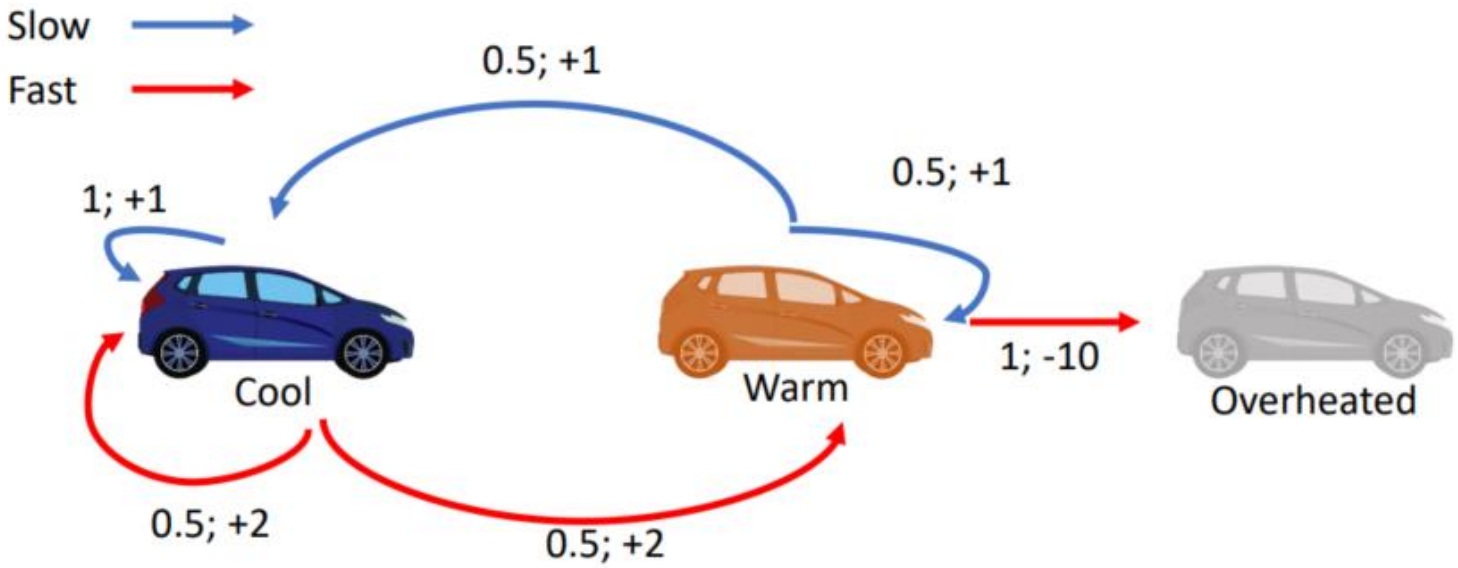
- 가치 반복 (value iteration) 알고리즘

$$v^*(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\}$$

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

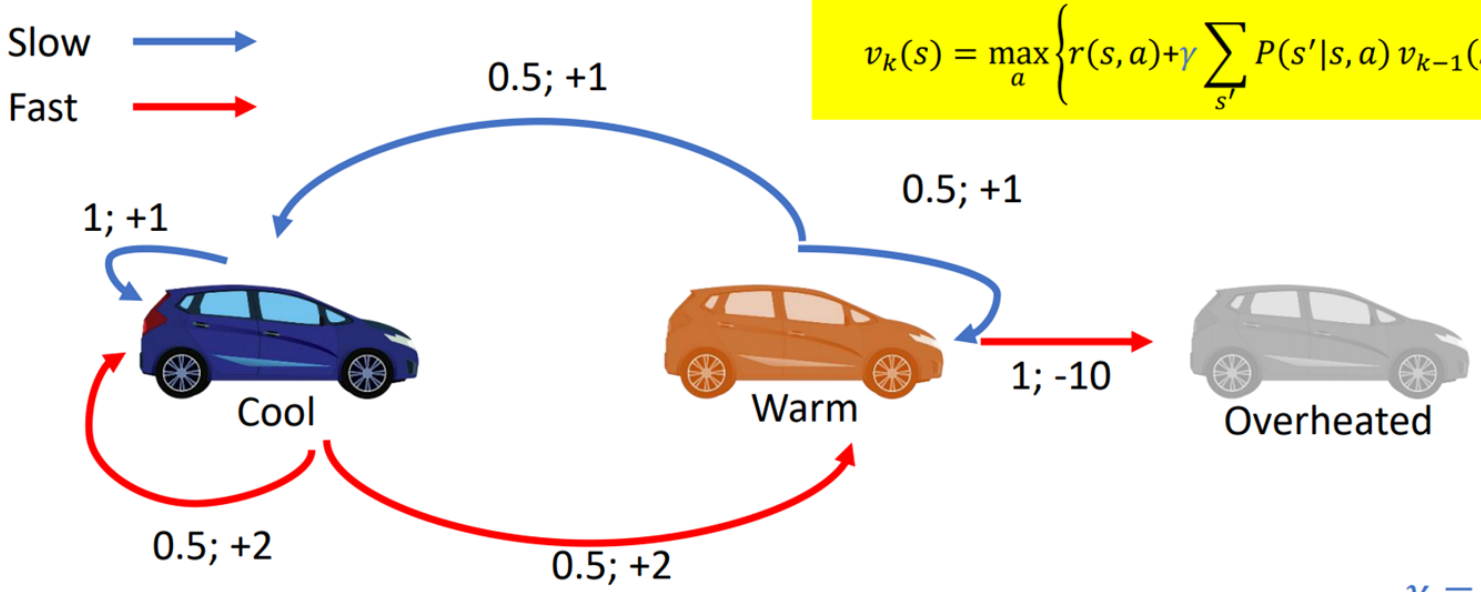


Infinite-horizon MDP 알고리즘



https://www.ccis.northeastern.edu/home/rplatt/cs5100_2015/slides/mdps.pdf

Infinite-horizon MDP 알고리즘



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

$\gamma = 0.8$

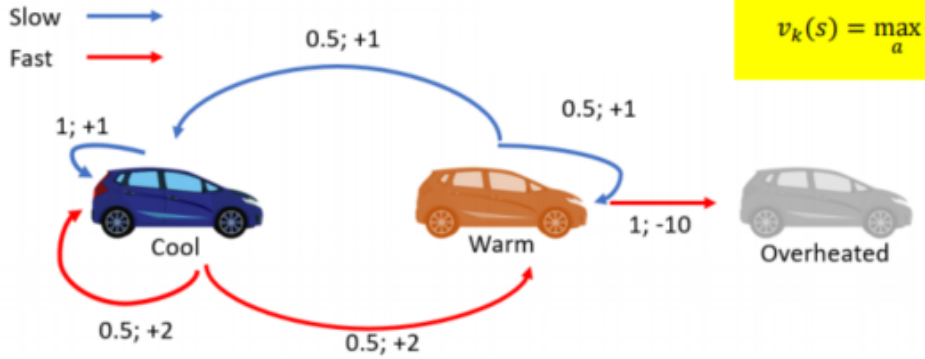
k	$v_k(Cool)$	$v_k(Warm)$	$v_k(Overheated)$
0	0	0	0

$$v_1(Cool) = \max_a \left\{ r(Cool,a) + \gamma \sum_{s'} P(s'|Cool,a)\, v_0(s') \right\} = \max\{r(Cool,Slow), r(Cool,Fast)\} = \max\{+1,+2\} = +2$$

$$v_1(Warm) = \max_a \left\{ r(Warm,a) + \gamma \sum_{s'} P(s'|Warm,a)\, v_0(s') \right\} = \max\{r(Warm,Slow), r(Warm,Fast)\} = \max\{+1,-10\} = +1$$

$$v_1(Overheated) = 0$$

Infinite-horizon MDP 알고리즘



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

$\gamma = 0.8$

k	$v_k(Cool)$	$v_k(Warm)$	$v_k(Overheated)$
0	0	0	0
1	2	1	0

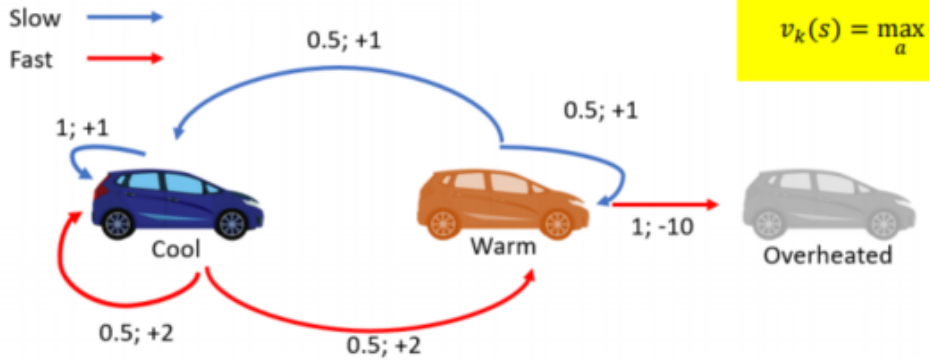
$$v_2(Cool) = \max_a \left\{ r(Cool,a) + \gamma \sum_{s'} P(s'|Cool,a)\, v_1(s') \right\} = \max \begin{Bmatrix} r(Cool,Slow) + 0.8 v_1(Cool) \\ r(Cool,Fast) + 0.8(0.5 v_1(Cool) + 0.5 v_1(Warm)) \end{Bmatrix} = \max \begin{Bmatrix} 1 + 0.8 \times 2 \\ 2 + 0.8(0.5 \times 2 + 0.5 \times 1) \end{Bmatrix} = 3.2$$

$$v_2(Warm) = \max_a \left\{ r(Warm,a) + \gamma \sum_{s'} P(s'|Warm,a)\, v_1(s') \right\} = \max \begin{Bmatrix} r(Warm,Slow) + 0.8(0.5 v_1(Cool) + 0.5 v_1(Warm)) \\ r(Warm,Fast) + 0.8 v_1(Overheated) \end{Bmatrix}$$

$$= \max \begin{Bmatrix} 1 + 0.8(0.5 \times 2 + 0.5 \times 1) \\ -10 + 0.8 \times 0 \end{Bmatrix} = 2.2$$

$$v_2(Overheated) = 0$$

Infinite-horizon MDP 알고리즘



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

$\gamma = 0.8$

k	$v_k(Cool)$	$v_k(Warm)$	$v_k(Overheated)$
0	0	0	0
1	2	1	0
2	3.2	2.2	0

$$v_3(Cool) = \max_a \left\{ r(Cool,a) + \gamma \sum_{s'} P(s'|Cool,a)\, v_2(s') \right\} = \max \begin{Bmatrix} r(Cool,Slow) + 0.8 v_2(Cool) \\ r(Cool,Fast) + 0.8(0.5 v_2(Cool) + 0.5 v_2(Warm)) \end{Bmatrix}$$

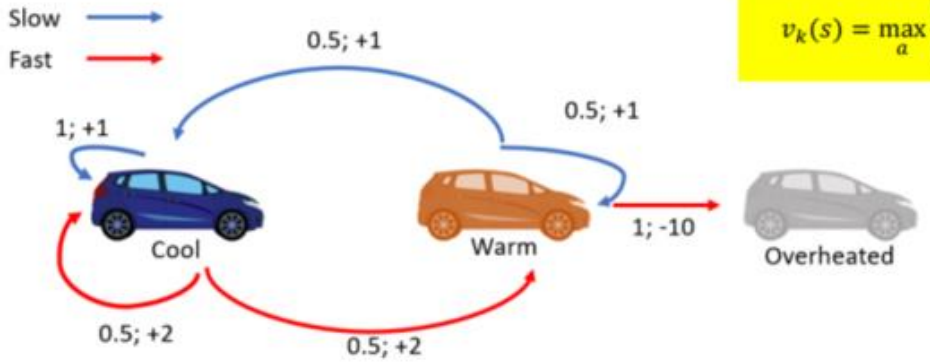
$$= \max \begin{Bmatrix} 1 + 0.8 \times 3.2 \\ 2 + 0.8(0.5 \times 3.2 + 0.5 \times 2.2) \end{Bmatrix} = 4.16$$

$$v_3(Warm) = \max_a \left\{ r(Warm,a) + \gamma \sum_{s'} P(s'|Warm,a)\, v_2(s') \right\} = \max \begin{Bmatrix} r(Warm,Slow) + 0.8(0.5 v_2(Cool) + 0.5 v_2(Warm)) \\ r(Warm,Fast) + 0.8 v_2(Overheated) \end{Bmatrix}$$

$$= \max \begin{Bmatrix} 1 + 0.8(0.5 \times 3.2 + 0.5 \times 2.2) \\ -10 + 0.8 \times 0 \end{Bmatrix} = 3.16$$

$$v_3(Overheated) = 0$$

Infinite-horizon MDP 알고리즘



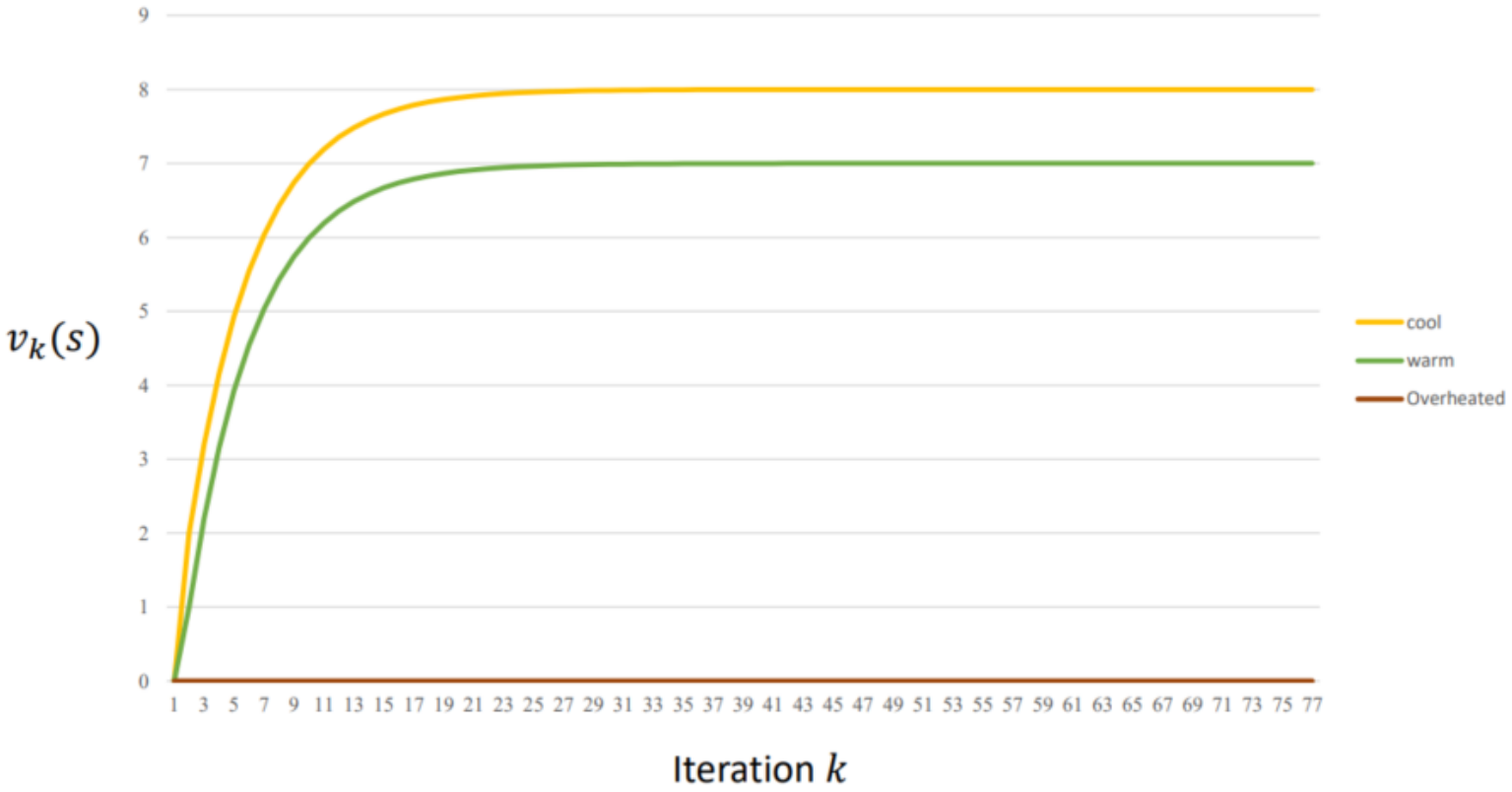
$$v_k(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v_{k-1}(s') \right\}$$

$\gamma = 0.8$

k	$v_k(Cool)$	$v_k(Warm)$	$v_k(Overheated)$
0	0	0	0
1	2	1	0
2	3.2	2.2	0
3	4.16	3.16	0

⋮

Infinite-horizon MDP 알고리즘



Infinite-horizon MDP 알고리즘

- 최적 정책은?
  - $\pi^*(s) = \underset{a}{\text{argmax}}\{r(s,a)+\gamma \sum_{s'} P(s'|s,a)\, v^*(s')\}$

- $\pi^*(Cool) = \text{argmax}\begin{Bmatrix} 1+0.8\times v^*(Cool) \\ 2+0.8\big(0.5v^*(Cool)+0.5v^*(Warm)\big) \end{Bmatrix}$
  $= \text{argmax}\begin{Bmatrix} 1+0.8\times 8 \\ 2+0.8(0.5\times 8+0.5\times 7) \end{Bmatrix} = Fast$
- $\pi^*(Warm) = \text{argmax}\begin{Bmatrix} 1+0.8\big(0.5v^*(Cool)+0.5v^*(Warm)\big) \\ -10+0.8v^*(Overheated) \end{Bmatrix}$
  $= \text{argmax}\begin{Bmatrix} 1+0.8(0.5\times 8+0.5\times 7) \\ -10+0.8\times 0 \end{Bmatrix} = Slow$