MDP with an Example

There is an agent that is training to regulate the temperature of a room. The room can either be cold or hot. The agent (thermostat) can either decide to turn on the cooler or the heater.

$S = \{ \text{cold}, \text{hot} \}$

$A = \{ \text{cooler}, \text{heater} \}$

* Given that the room is cold, by turning on the cooler there is a 90% chance of room remaining cold. However, if heater is turned on, there is 80% chance that the room gets hot.

* Given that the room is hot, by turning on the cooler there is a 80% chance of room becoming cold. However, if heater is turned on, there is 70% chance that the room gets hot.

$s = \{ \text{cold}, \text{hot} \}$

$A = \{ \text{cooler}, \text{heater} \}$

Draw transition matrices:

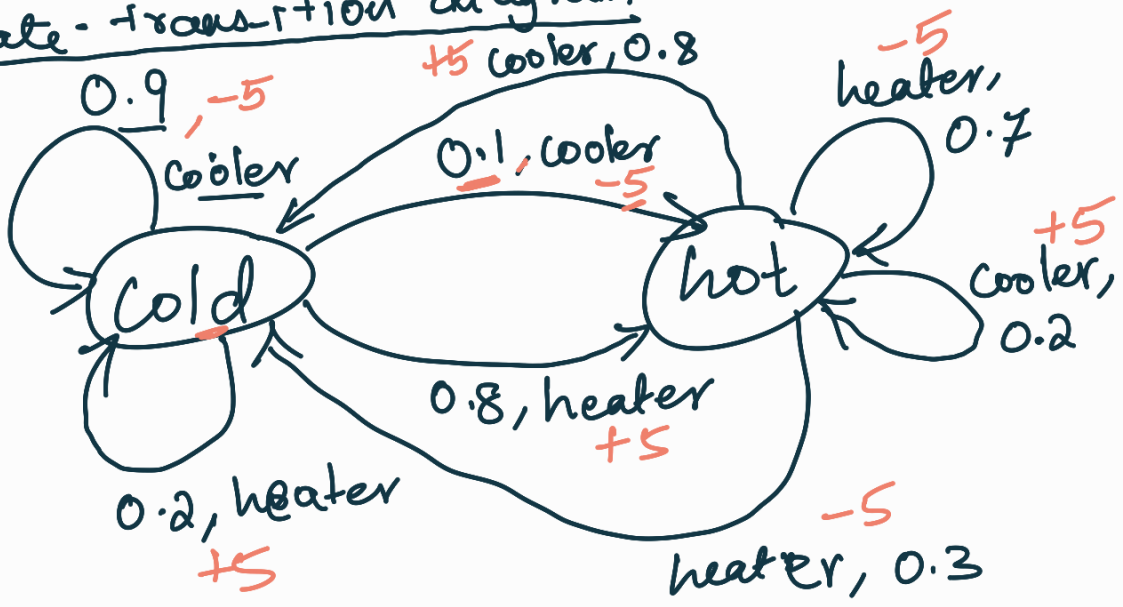
A = cooler

	cold	hot
cold	0.9	0.1
→ hot	0.8	0.2

A = heater

	cold	hot
cold	0.2	0.8
hot	0.3	0.7

State-transition diagram



Rewards

One-step rewards

cold → turn on the cooler ⇒ -5

hot → turn on the heater ⇒ -5

cold → heater ⇒ +5

hot → coole ⇒ +5

* $P(\text{hot} \mid \text{hot}, \text{cooler}) = 0.2$

$P(S_{t+1} = \text{hot} \mid S_t = \text{hot}, A_t = \text{cooler})$

$= T(\text{hot}, \text{cooler}, \text{hot})$

$p(\text{hot} \mid \text{hot}, \text{heater}) = 2_0$

$p(\text{cold} \mid \text{hot}, \text{heater}) = ?_0$

$p(\text{cold} \mid \text{cold}, \text{heater}) = ?_0$

$p(\text{cold} \mid \text{hot}, \text{cooler}) = ?_0$

$p(\text{cold} \mid \text{cold}, \text{cooler}) = ?_b$

$p(\text{hot} \mid \text{cold}, \text{cooler}) = ?_0$

$p(\text{hot} \mid \text{cold}, \text{heater}) = ?_*$

One step rewards:

* $r(S_t, A_t, S_{t+1})$

$r(\text{hot}, \text{cooler}, \text{hot}) = +5$

⋮

Discounted return

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= R_{t+1} + \gamma [R_{t+2} + \gamma R_{t+3} + \cdots]$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$\gamma \in [0, 1]$

$= 0 \leq \gamma \leq 1$

$G_{t+2} = 10$

$R_{t+1} = 5$

$\gamma = 0.9$

$G_t = 5 + 0.9(10)$

chess $\Rightarrow \gamma = ?$

$\gamma = 1$

$S = \{ \text{Sunny}, \text{Rainy} \}$

$A = \{ \underset{U}{\text{umbrella}}, \underset{N.U}{\text{no umbrella}} \}$

Given that today is rainy there is 10% chance that tommor is sunny.

Given that today is sunny there is 20% chance that tommorow is rainy.

If you carry an umbrella on a rainy day ⇒ +5 reward

Sunny day ⇒ −5 reward

No umbrella, rainy day ⇒ −10

No umbrella, sunny day ⇒ +10

* Draw state transition diagram.
* write the probability transition matrix
* $P(S_{t+1} \mid S_t, A_t)$
* $r(S_t, A_t, S_{t+1})$

G_t R_{t+1}

Exercise 3.8 Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, \ldots, G_5$? Hint: Work backwards. □

$$G_5 = R_{5+1} = R_6 = 0$$

$$G_4 = R_5 + \gamma G_6$$

$$= 2$$

$$G_3 = R_4 + \gamma^2 G_4$$

G_2

G_1

G_0

---

* $P(S_{t+1} \mid S_t, A_t)$ or $T(S_t, A_t, S_{t+1})$

$P(S_{t+1} = \text{cold} \mid S_t = \text{hot}, A_t = \text{heater})$

$= T(S_t = \text{hot}, A_t = \text{heater}, S_{t+1} = \text{cold})$

$= 0.3$

* $R(S_t, A_t, S_{t+1})$

$= R(S_t = \text{hot}, A_t = \text{heater}, S_{t+1} = \text{cold})$

$= -5$

Backup diagrams

Policy :-

$$\pi(A_t \mid S_t)$$

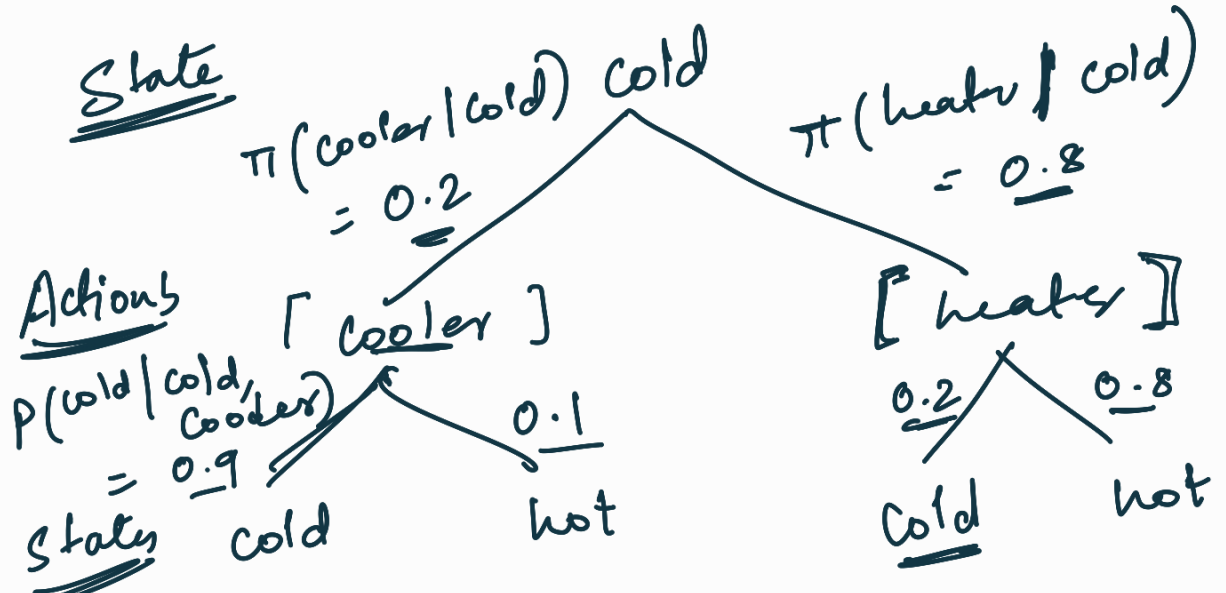
$$= P(A_t \mid S_t)$$

$$\pi(\text{cooler} \mid \text{cold}) = 0.2$$

$$\pi(\text{heater} \mid \text{hot}) = 0.3$$

$$\pi(\text{heater} \mid \text{cold}) = 1 - 0.2 = 0.8$$

$$\pi(\text{cooler} \mid \text{hot}) = 1 - 0.3 = 0.7$$



$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \left[\sum_{s' \in S} P(s' \mid s, a) \left(\gamma(s, a, s') + \gamma v_\pi(s') \right) \right]$$

* Policy evaluation
* Policy improvement / value iteration

Policy evaluation

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \left[\sum_{s' \in S} P(s'|s,a) \left(r(s,a,s') + \gamma v_\pi(s') \right) \right]$$

* Given $\pi(a|s)$

$$V(s) \approx v_\pi(s)$$

1. init $V(s) = 0$
2. $V_{New}(s) = \underbrace{\underbrace{\quad\quad V_{OLD}(s)}_{\text{Bellmans eqn.}}}$

Iteration 1

$V(cold) = 0 \qquad V(hot) = 0$ ←

Gm. $\gamma = 1$

$$V(cold) = 0.2\left[\begin{matrix}0.9\,(-5 + 1(0)) \\ +0.1\,(-5 + 1(0))\end{matrix}\right] + 0.8\left[\begin{matrix}0.2\,(5 + 1(0)) \\ +0.8\,(5 + 1(0))\end{matrix}\right]$$

$$= 0.2\,(-4.5 - 0.5) + 0.8\,(1 + 4)$$

$$= 0.2\,(-5) + 3.2$$

$$= -1 + 3.2 = 2.2$$

$V(cold) = 2.2$ ←

$V(hot) = ?$

Iteration 2

$V(cold) = 2.2 \qquad V(hot) = ?$

$V(cold) =$ _____

$V(\text{hot}) =$ ———

Policy improvement / value iteration

Bellmans optimality equation.

$$v_*(s) = \max_{a \in A} \left[\sum_{s' \in S} P(s'|s,a) \left(r(s,a,s') + \gamma v_*(s) \right) \right]$$

$v_{\pi^*}(s) \rightarrow v_*(s)$

$\pi^* \rightarrow$ optimal policy.

1. init $V(s) = 0$
2. $V(s) =$ ——— $V(s)$

Bellmans optimality equation.

Iteration 1

$V(\text{cold}) = 0$ $\quad$ $V(\text{hot}) = 0$

$$V(cold) = \max\left(\begin{bmatrix} 0.9\,(-5 + 1(0)) \\ +0.1\,(-5 + 1(0)) \end{bmatrix}, \begin{bmatrix} 0.2\,(5 + 1(0)) \\ +0.8\,(5 + 1(0)) \end{bmatrix} \right)$$

$$= \max\left(\begin{matrix} (-4.5 & -0.5) \\ (1 & +\ 4) \end{matrix} \right)$$

$$= \max(-5, 5)$$

$$= \underline{5}$$

$$V(hot) : \quad =$$