Reinforcement learning

→ Agent : driverless car , robot

→ Environment: road , house floor

→ State : position, location, orientation, charge.

→ Action : Move left, apply breaks. Accelerate, go back to charging point.

Markov Decision Process

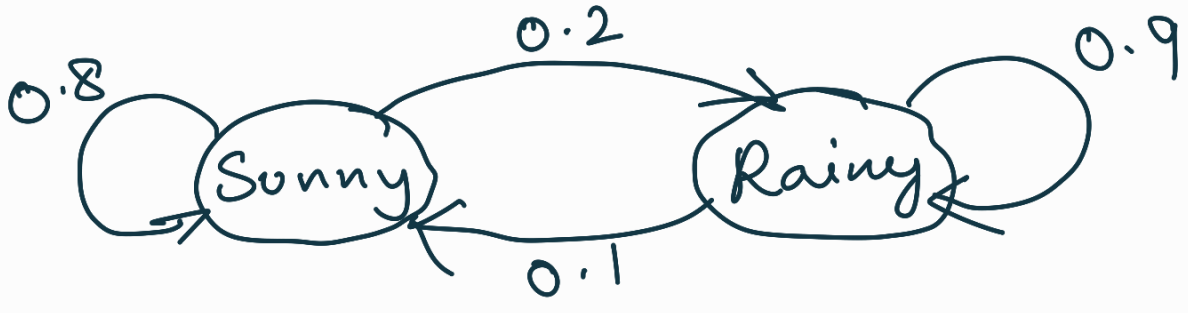
→ Decision making under uncertainity

Markov Process

State Transition Matrix

Today \ Tommorow	Sunny	Rainy
Sunny	0.8	0.2
Rainy	0.1	0.9

State Transition Diagram



time interv = 1day

init state = sunny

S S S S R R R R S S R
1 2 3 . . . . . t

$$\{X_t \mid t \geq 0\}$$

$$= \{X_0, X_1, \ldots X_t\}$$

X_0: S, R

Markov Property

The next state depends only on the current state and not on any of the previous state.

→ MP implies stationarity which means the transition matrix will remain same over the timesteps.

Markov reward Process

$S = \{ \text{Computer}, \text{Coffee}, \text{chat} \}$

state transition matrix

	Comp	coffee	chat
Comp	0.5	0.3	0.2
Coffee	0.5	0.2	0.3
chat	0.7	0.1	0.2

time = 1 hr

Reward matrix

	Comp	coffee	chat
Comp	10	5	-7
Coffee	5	0	-3
chat	5	-1	-3

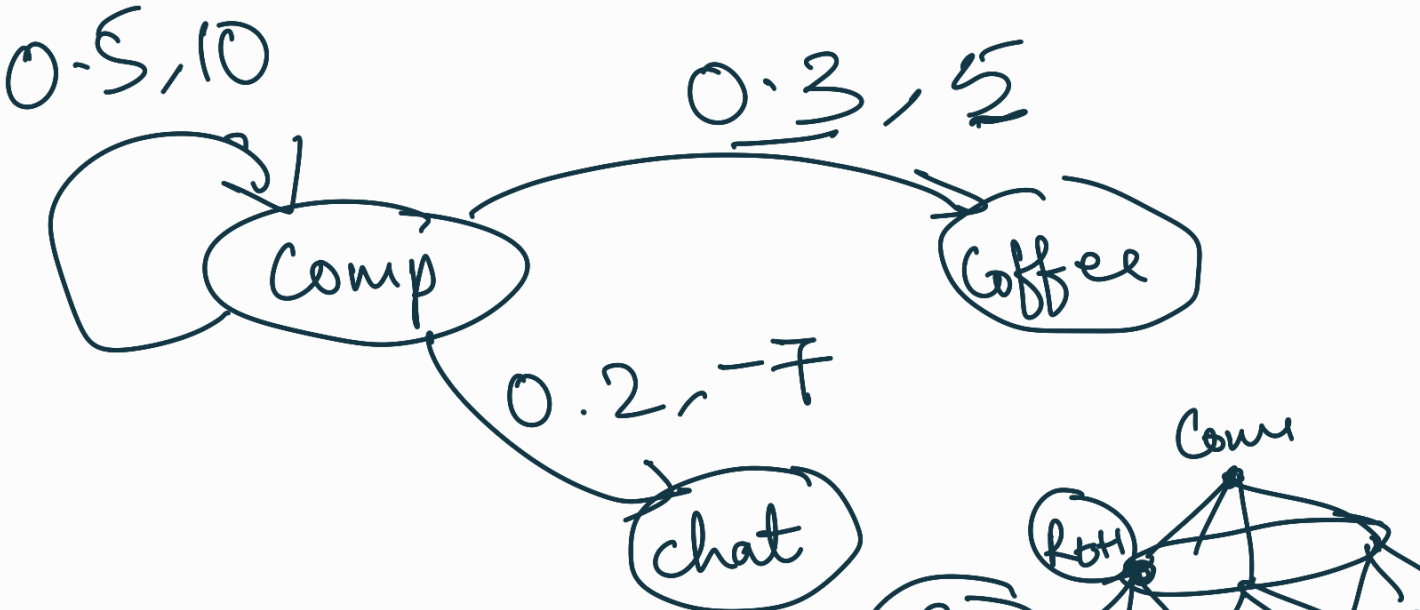
R (comp , coffee) = 5

R (coffee , comp) = 5

R (coffee , coffee) = 0

⋮ Fill up ⋮

state transition diagram



Return , G_t

$$G_t = R_{t+1} + R_{t+2} \cdots\cdots + R_T$$

↳ one-step reward

Discount factor γ - Gamma

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

Ex. $\gamma = 0.5$

$$0 \leq \gamma \leq 1$$

↳ Long term discounted return

Episodic task

→ the simulation will come to an end.

Markov process → Markov reward process → Markov decision process

transition prob + reward Matrix + Action

$$S = \{ S_1, S_2, S_3 \}$$

$$A = \{ a_1, a_2, a_3 \}$$

$A = a_1$

$P_{00}^{a_1}$	$P_{01}^{a_1}$	. . .
$P_{10}^{a_1}$	.	.
.	.	.

$A = a_2$

$P_{00}^{a_2}$	$P_{01}^{a_2}$	. .
$P_{10}^{a_2}$	. .	. .
.	.	. .

$A = a_3$

$P_{00}^{a_3}$	$P_{01}^{a_3}$	. . .
$P_{10}^{a_3}$	.	. .
.	. .	. .

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.

* 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:

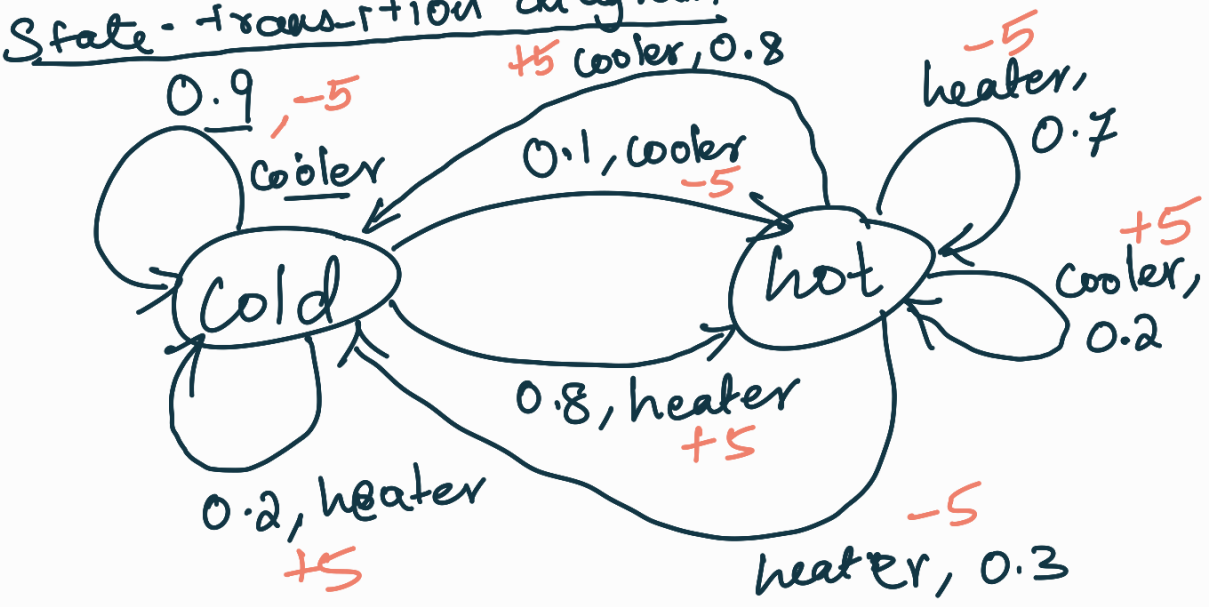
A = cooler

	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

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

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

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

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

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

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

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

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

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

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

One step rewards:

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

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

⋮

Value functions

→ State value function
→ state-Action value function.

State-value function

It tells how valuable it is to start from the state s.

$$V_\pi(s) = E\left[G_t \mid S_t = s \right]$$

It is the expected long term discounted return for starting from state s.

State-Action value function

$$q_\pi(s,a) = E\left[G_t \mid S_t = s, A_t = a \right]$$

→ how valuable it is to start from state s and take an action a.

→ Exp long term discounted return for starting from s and taking an action a.

SESSIONAL - 1

Syllabus for s-1 ends here

All the best!!...

* $v_\pi(s) = E[G_t \mid S_t = s]$

Expectation using conditions

$$E[X] = \sum_{y \in Y} E[X \mid y]\, P(y)$$

$$v_\pi(s) = \sum_{a \in A} E[G_t \mid S_t = s, A_t = a] * P(A_t = a \mid S_t = s)$$

$$\boxed{v_\pi(s) = \sum_{a \in A} q_\pi(s,a)\, \pi(a \mid s)}$$

* $q_\pi(s,a) = E[G_t \mid S_t = s, A_t = a]$

Linearity of Expections

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$$

$$q_\pi(s,a) = E\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s, A_t = a \right]$$

$$= E\left[R_{t+1} \mid S_t = s, A_t = a \right] + \gamma E\left[R_{t+2} + \gamma R_{t+3} + \cdots \mid S_t = s, A_t = a \right]$$

$$= E\left[R_{t+1} \mid S_t = s, A_t = a \right] + \gamma E\left[G_{t+1} \mid S_t = s, A_t = a \right]$$

$$= \sum_{s' \in S} p(s' \mid s,a)\, r(s,a,s') + \gamma \sum_{s' \in S} p(s' \mid s,a) \cdot E\left[G_{t+1} \mid S_{t+1} = s', \cancel{S_t = s, A_t = a} \right]$$

$$= \sum_{s' \in S} p(s' \mid s,a) \left[r(s,a,s') + \gamma E\left[G_{t+1} \mid S_{t+1} = s' \right] \right]$$

$$\boxed{q_\pi(s,a) = \sum_{s' \in S} p(s' \mid s,a) \left[r(s,a,s') + \gamma v_\pi(s') \right]}$$

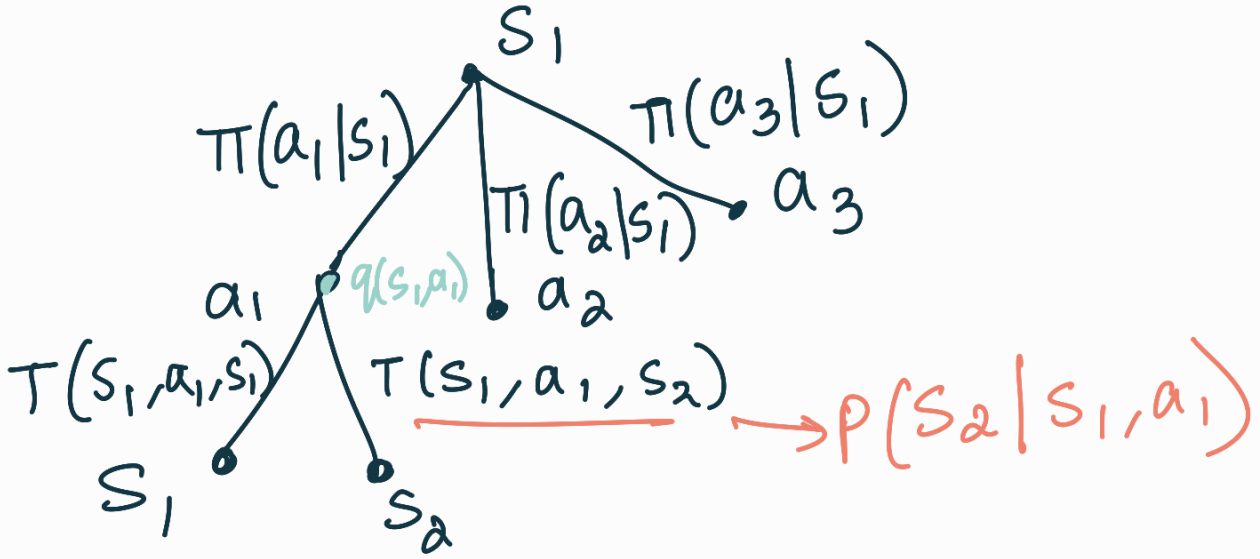
$p(s' \mid s,a)$: $T(s,a,s')$ — one-step reward

$r(s,a,s')$

* Back up diagram

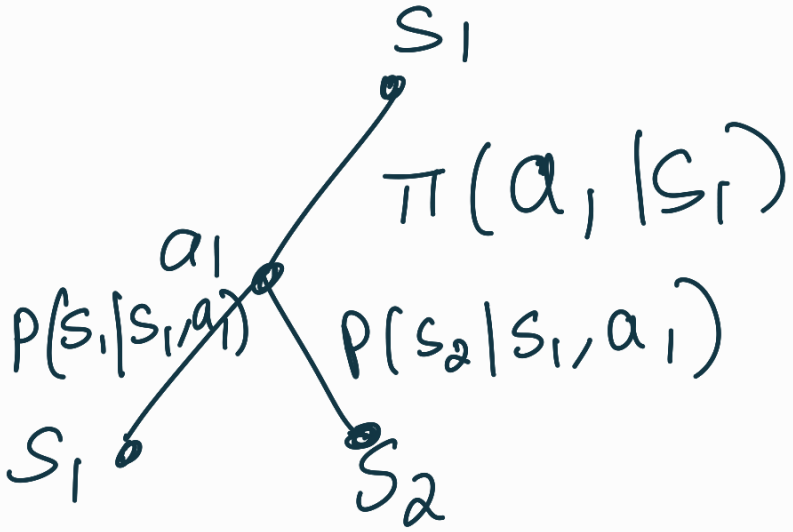
$$S = \{s_1, s_2\}$$

$$A = \{a_1, a_2, a_3\}$$



$$v_\pi(s_1) = \pi(a_1|s_1)\, q_\pi(s_1,a_1) + \pi(a_2|s_1)\, q_\pi(s_1,a_2) + \pi(a_3|s_1)\, q_\pi(s_1,a_3)$$

$$\boxed{V_\pi(s_1) = \sum_{a \in A} \pi(a|s_1)\, q_\pi(s_1,a)}$$



$$q_\pi(s_1,a_1) = p(s_1|s_1,a_1)\left[r(s_1,a_1,s_1) + \gamma v_\pi(s_1)\right]$$

$$+ p(s_2|s_1,a_1)\left[r(s_1,a_1,s_2) + \gamma v_\pi(s_2)\right]$$

$$\boxed{q_\pi(s_1,a_1) = \sum_{s' \in S} p(s'|s_1,a_1)\left[r(s_1,a_1,s') + \gamma v_\pi(s')\right]}$$

$$v_\pi(s) = \sum_{a \in A} \pi(a|s)\, q_\pi(s,a)$$

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

Agent decides

Bellman equation for state value function.

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

Bellmen equation for state-action value function.

Policy π

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

→ decided by the agent.

Optimal policy

for any MDP,

there exists a policy π^* (optimal policy)

such that.

$$V_{\pi^*}(s) \geq V_\pi(s)$$

$$q_{\pi^*}(s,a) \geq q_{\pi}(s,a)$$

under the optimal policy, the state value function and state-action value function will have max output.