

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

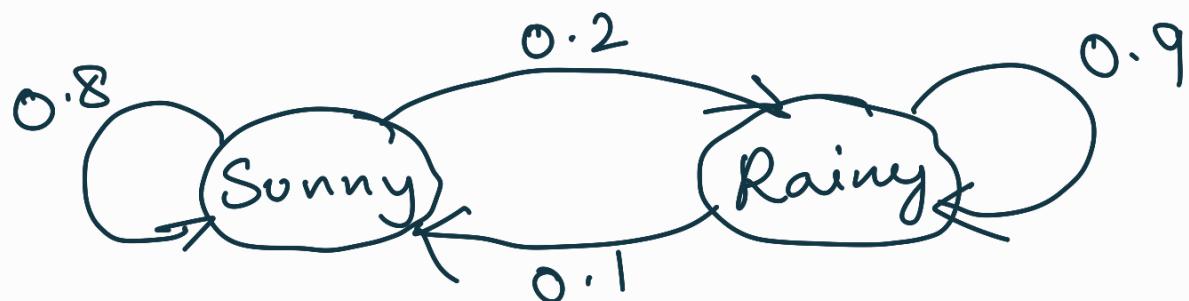
- Decision making under uncertainty

Markov Process

State Transition Matrix

		Tomorrow	
		Sunny	Rainy
Today	Sunny	0.8	0.2
	Rainy	0.1	0.9

State Transition Diagram



time interv = 1 day

init state = sunny

S S S S R ~~R R~~ R R S S R
1 2 3 t

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

$= \{ \underbrace{x_0}_S, x_1, \dots, x_t \}$

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

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

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

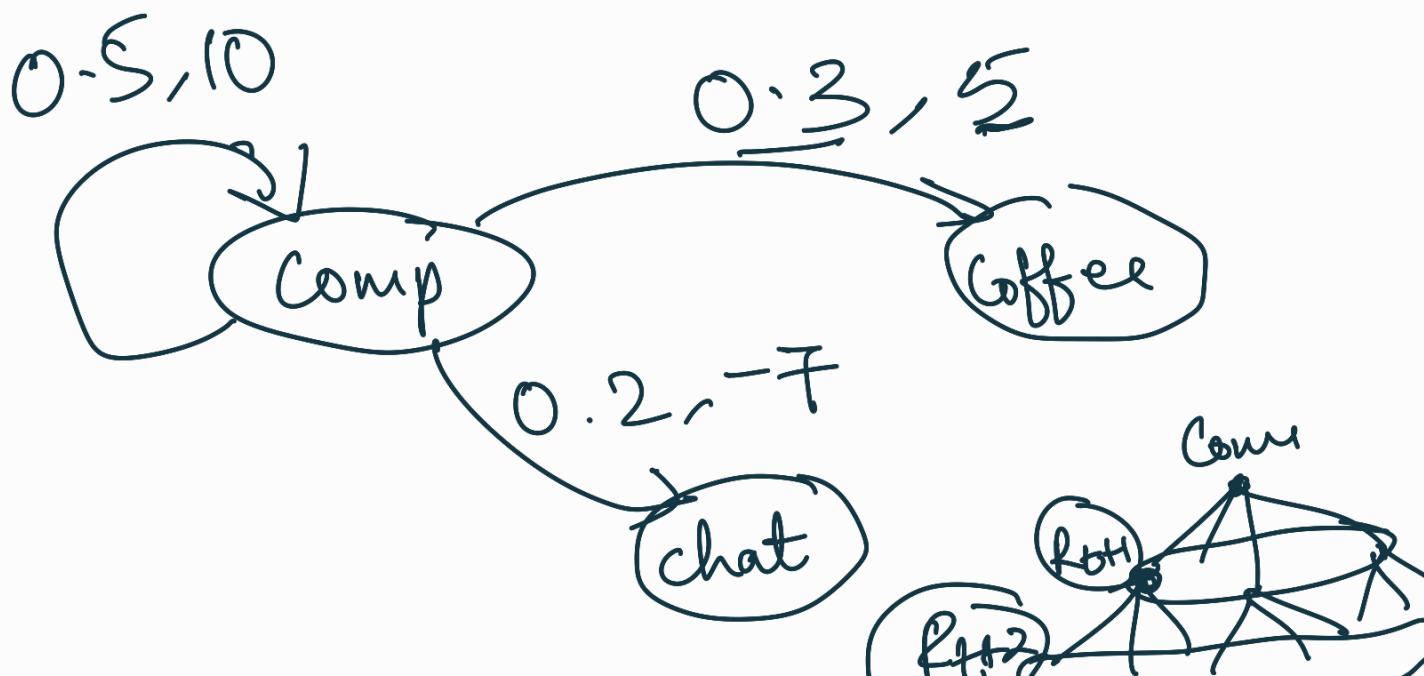
$$R(\underline{\text{comp}}, \underline{\text{coffee}}) = \underline{5}$$

$R(\text{coffee, comp})^5$

$f(\text{coffee}, \text{coffee}) = ?$

Fill up

State transition diagram



Return , G_t

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

↳ one-step reward

Discount factor γ - Gamma

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

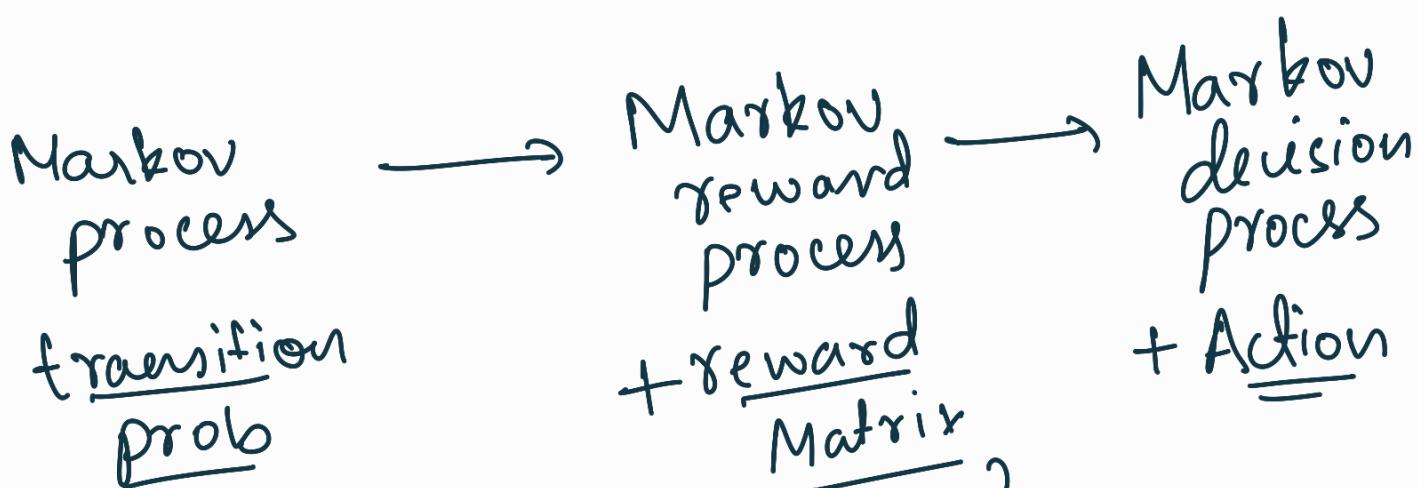
$0 \leq \gamma \leq 1$

$\text{Ex. } \gamma = 0.5$

Long term discounted return

Episodic task

→ the simulation will come to an end.



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

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

$$A = a_1$$

$P_{00}^{a_1}$	$P_{01}^{a_1}$	\dots
$P_{10}^{a_1}$	\dots	\dots
\dots	\dots	\dots

$$A = a_2$$

$P_{00}^{a_2}$	$P_{01}^{a_2}$	\dots
$P_{10}^{a_2}$	\dots	\dots
\dots	\dots	\dots

$$A = a_3$$

$P_{00}^{a_3}$	$P_{01}^{a_3}$	\dots
$P_{10}^{a_3}$	\dots	\dots
\dots	\dots	\dots

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.

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

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

Draw transition matrices:

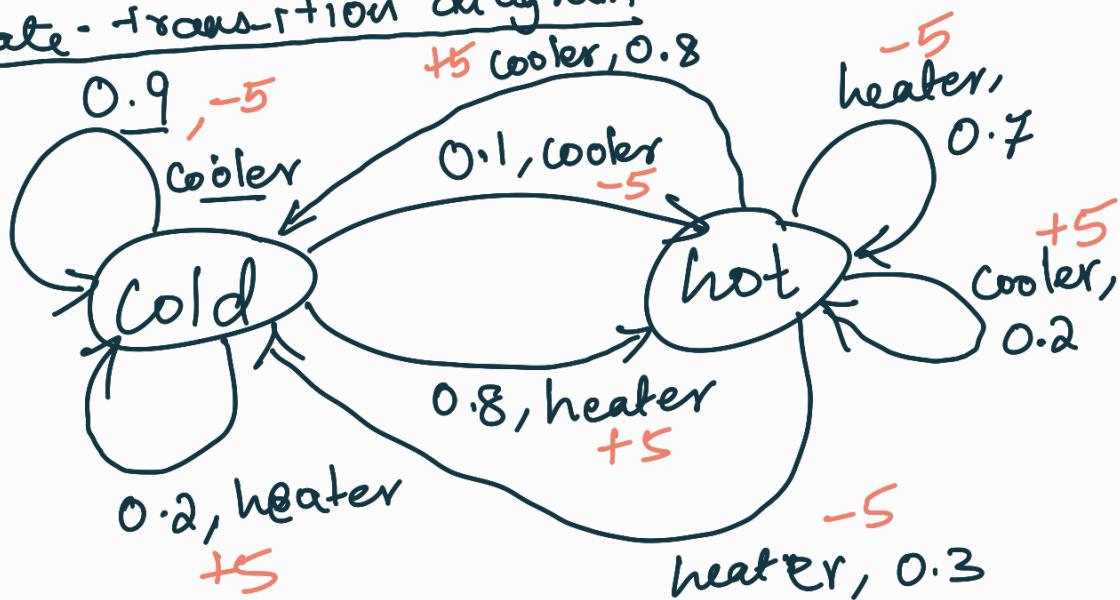
$A = \text{cooler}$

	cold	hot
cold	0.9	0.1
hot	0.8	0.2

$A = \text{heater}$

	cold	hot
cold	0.2	0.8
hot	0.3	0.7

State-transition diagram



Rewards cold \rightarrow turn on the cooler $\Rightarrow -5$

One-step rewards hot \rightarrow turn on the heater $\Rightarrow -5$

cold \rightarrow heater $\Rightarrow +5$
hot \rightarrow cooler $\Rightarrow +5$

* $P(\underbrace{\text{hot}}_{\text{hot}} | \text{hot, cooler}) = 0.2$

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

- $P(\text{hot} | \text{hot, heater}) = \frac{2}{3}$
 $P(\text{cold} | \text{hot, heater}) = ?$
 $P(\text{cold} | \text{cold, } \cancel{\text{heater}}) = ?$
 $P(\text{cold} | \text{hot, cooler}) = ?$
 $P(\text{cold} | \text{cold, cooler}) = ?$
 $P(\text{hot} | \text{cold, cooler}) = ?$
 $P(\text{hot} | \text{cold, heater}) = ?$
-

One step rewards :

* $\gamma(\underline{s_t}, \underline{a_t}, \underline{s_{t+1}})$
 $\gamma(\underline{\text{hot}}, \underline{\text{cooler}}, \underline{\text{hot}}) = +5$

⋮

Value functions

- State value function
- State-Action value function.

State - value function

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

$$V_{\pi}(s) = E[G_t \mid S_t = s]$$

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

State - Action value function

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

- 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!!...

$$* \quad v_{\pi}(s) = E[G_t | S_t = s]$$

Expectation using conditions

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

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

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

$$* \quad q_{\pi}(s, a) = E[G_t | S_t = s, A_t = a]$$

Linearity of Expectations

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

$$Q_{\pi}(s, a) = E \left[\underbrace{R_{t+1} + \gamma R_{t+2} + \dots}_{\gamma^2 R_{t+3}} \middle| S_t = s, A_t = a \right]$$

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

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

$$= \sum_{s' \in S} P(s' | s, a) \underbrace{\gamma(s, a, s')}_{\gamma(s, a, s')} + \gamma \sum_{s' \in S} P(s' | s, a) \times E \left[G_{t+1} \middle| \begin{array}{l} S_{t+1} = s' \\ S_t = s, A_t = a \end{array} \right]$$

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

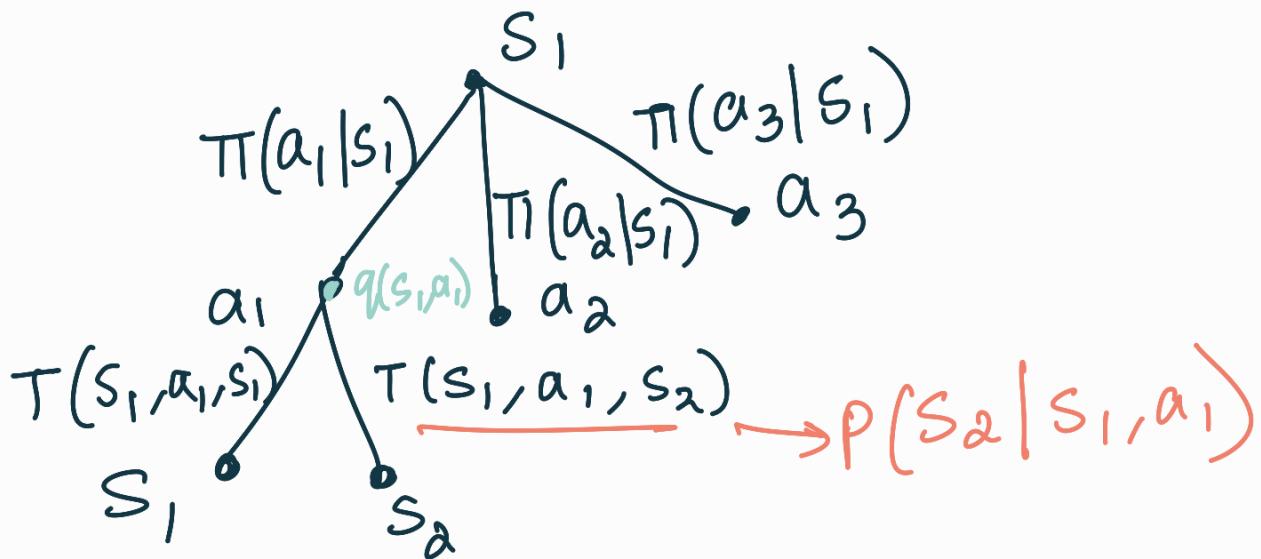
$$Q_{\pi}(s, a) = \sum_{s' \in S} P(s' | s, a) \left[\gamma(s, a, s') + \gamma V_{\pi}(s') \right]$$

\downarrow \downarrow
 $T(s, a, s')$ one-step reward

* Back up diagram

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

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

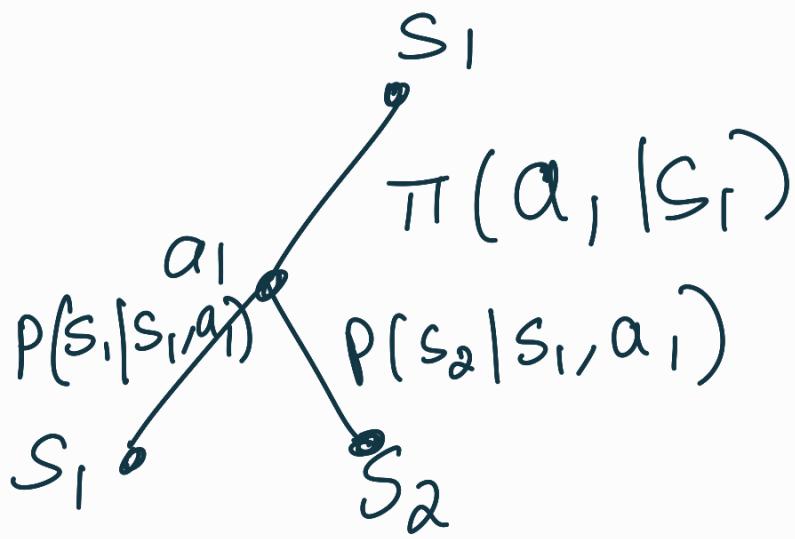


$$v_{\pi}(s_1) = \pi(a_1|s_1) q_{\pi}(s_1, a_1)$$

$$+ \pi(a_2|s_1) q_{\pi}(s_1, a_2)$$

$$+ \pi(a_3|s_1) q_{\pi}(s_1, a_3)$$

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



$$q_{\pi}(s_1, a_1) = P(s_1|s_1, a_1) [r(s_1, a_1, s_1) + \gamma v_{\pi}(s_1)] + P(s_2|s_1, a_1) [r(s_1, a_1, s_2) + \gamma v_{\pi}(s_2)]$$

$$q_{\pi}(s_1, a_1) \leq \sum_{s' \in S} P(s'|s_1, a_1) [r(s_1, a_1, s') + \gamma v_{\pi}(s')]$$

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

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

Agent decides

Bellman equation for state value function.

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

Bellman 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^*
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.

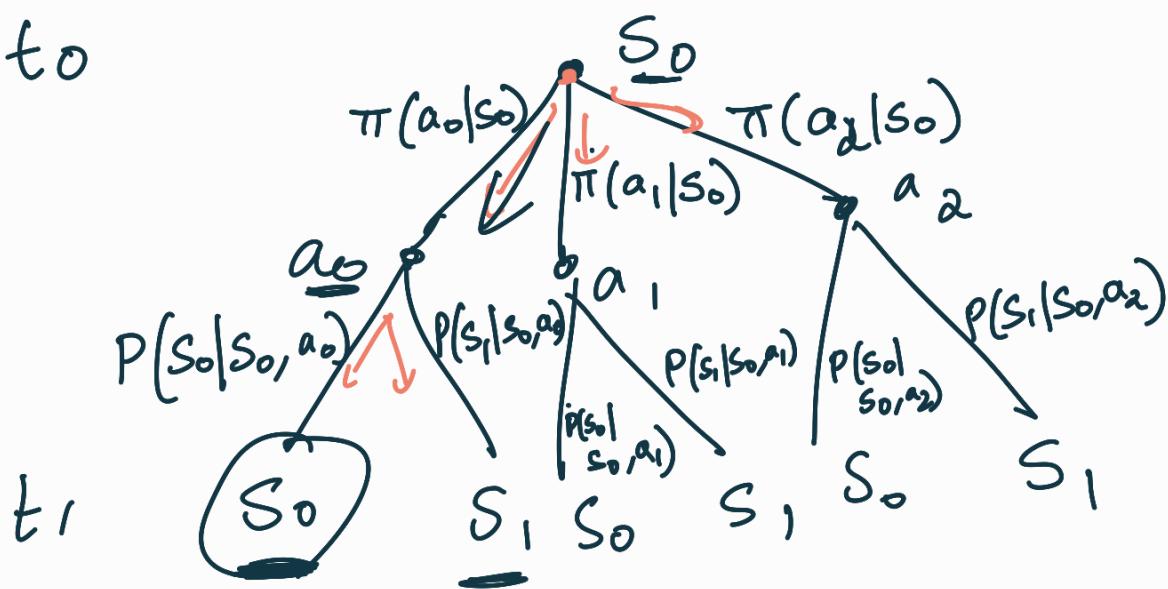
two-level backup diagrams

① State value function

$$G_n, \quad S = \{s_0, s_1\}$$

$$A = \{a_0, a_1, a_2\}$$

to



$$V_{\pi}(s_0) = \pi(a_0|s_0) \left[\underbrace{P(s_0|s_0, a_0) [\gamma(s_0, a_0, s_0) + \gamma V_{\pi}(s_0)]}_{+ P(s_1|s_0, a_0) [\gamma(s_0, a_0, s_1) + \gamma V_{\pi}(s_1)]} \right]$$

$$+ \pi(a_1|s_0) \left[\underbrace{P(s_0|s_0, a_1) [\gamma(s_0, a_1, s_0) + \gamma V_{\pi}(s_0)]}_{+ P(s_1|s_0, a_1) [\gamma(s_0, a_1, s_1) + \gamma V_{\pi}(s_1)]} \right]$$

$$+ \pi(a_2|s_0) \left[\underbrace{P(s_0|s_0, a_2) [\gamma(s_0, a_2, s_0) + \gamma V_{\pi}(s_0)]}_{+ P(s_1|s_0, a_2) [\gamma(s_0, a_2, s_1) + \gamma V_{\pi}(s_1)]} \right]$$

$$= \pi(a_0|s_0) \left[\sum_{s' \in S} P(s'|s_0, a_0) [\gamma(s_0, a_0, s') + \gamma V_{\pi}(s')] \right]$$

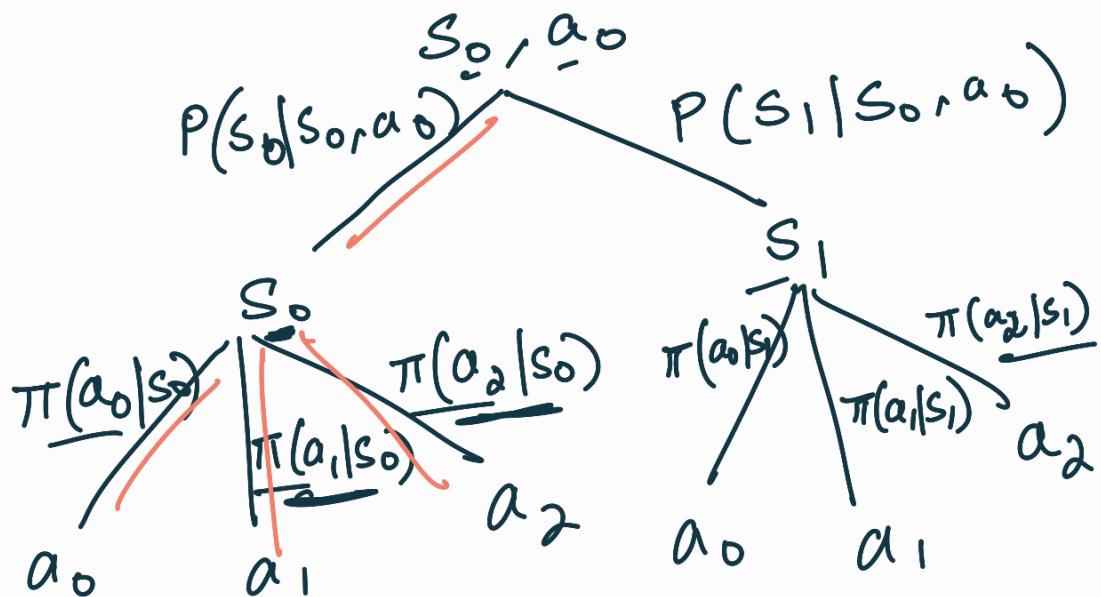
$$+ \pi(a_1|s_0) \left[\sum_{s' \in S} P(s'|s_0, a_1) [\gamma(s_0, a_1, s') + \gamma V_{\pi}(s')] \right]$$

$$+ \pi(a_2|s_0) \left[\sum_{s' \in S} P(s'|s_0, a_2) [\gamma(s_0, a_2, s') + \gamma V_{\pi}(s')] \right]$$

$$V_{\pi}(s) = \sum_{a \in A} \pi(a|s_0) \left[\sum_{s' \in S} \underline{P(s'|s_0, a)} \right. \\ \left. [r(s_0, a, s') + \gamma V_{\pi}(s')] \right]$$

Bellman's Equation for state value function.

2 Action Value function



$$q_{\pi}(s_0, a_0) = P(s_0|s_0, a_0) \left[r(s_0, a_0, s_0) + \gamma \left[\begin{array}{l} \pi(a_0|s_0) q_{\pi}(s_0, a_0) \\ + \pi(a_1|s_0) q_{\pi}(s_0, a_1) \\ + \pi(a_2|s_0) q_{\pi}(s_0, a_2) \end{array} \right] \right] \\ + P(s_1|s_0, a_0) \left[r(s_0, a_0, s_1) + \gamma \left[\begin{array}{l} \pi(a_0|s_1) q_{\pi}(s_1, a_0) \\ + \pi(a_1|s_1) q_{\pi}(s_1, a_1) \\ + \pi(a_2|s_1) q_{\pi}(s_1, a_2) \end{array} \right] \right]$$

$$= \left[\frac{P(S_0 | S_0, a_0)}{1 - P(S_1 | S_0, a_0)} \left[\gamma (S_0, a_0, S_0) + \gamma \sum_{a' \in A} \pi(a' | S_0) \frac{q_{\pi}(S_0, a')}{\pi(a' | S_0)} \right] \right. \\ \left. + P(S_1 | S_0, a_0) \left[\gamma (S_0, a_0, S_1) + \gamma \sum_{a' \in A} \pi(a' | S_1) \frac{q_{\pi}(S_1, a')}{\pi(a' | S_1)} \right] \right]$$

$$q_{\pi}(S_0, a_0)$$

$$= \sum_{S' \in S} P(S' | S_0, a_0) \left[\gamma (S_0, a_0, S') + \gamma \sum_{a' \in A} \pi(a' | S') \frac{q_{\pi}(S', a')}{\pi(a' | S')} \right]$$

Bellman's equation State - ~~action~~
 value function.

① policy based $\pi(a | s)$

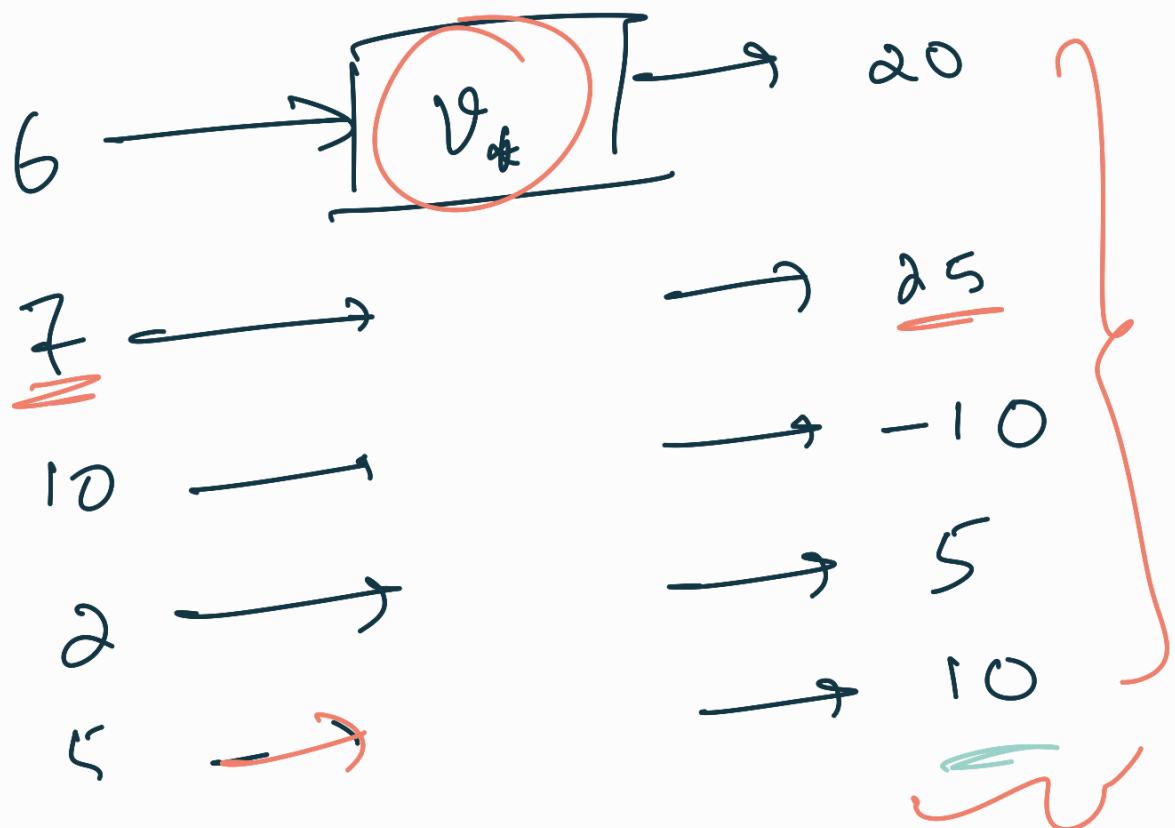
② value based $v_*(s) \xrightarrow{\text{action}}$

Methods to solve RL
 problems.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

$$\pi_{11}(\text{right}|6) = 1$$

Number



Bellmans optimality equation

the optimal policy satisfies,

$$v_{\pi^*}(s) = \max_{a \in A} \underline{q_{\pi}(s, a)}$$

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

RHS ↓
one-step
reward

long term
discounted
return

$$\underline{V}_A(s_0) = \max_{a \in A} \left[\sum_{s' \in S} P(s'|s_0, a_0) \left(\gamma (s_0, a_0, s') + \gamma \underline{V}_*(s') \right) \right]$$

$$\sum_{s' \in S} P(s'|s_0, a_1) \left(\gamma (s_0, a_1, s') + \gamma \underline{V}_*(s') \right)$$

$$\sum_{s' \in S} P(s'|s_0, a_2) \left(r(s_0, a_2, s') + \gamma \underline{V}_*(s') \right)$$

$$= \max_{a \in A} \left[q_{*}(s_0, a_0), q_{*}(s_0, a_1), q_{*}(s_0, a_2) \right]$$

Value iteration

* initialize $\underline{v}(s) = 0$,

* until $\|\underline{v}_{\text{new}} - \underline{v}_{\text{old}}\| < \text{tol}$:

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

