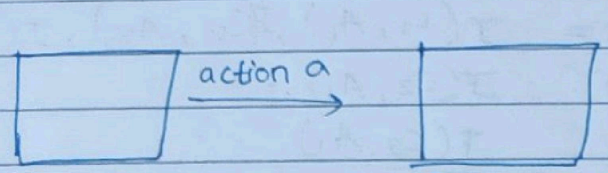
MOD 6 : Reinforcement Learning

M T W T F S S | Page No.: | Date: | YOUVA

5/04/2024 Monday

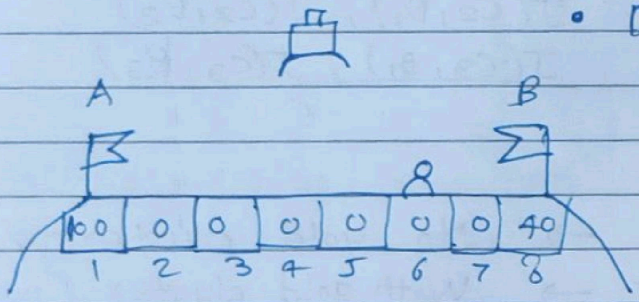
- Reinforcement → Strengthening something by adding some additional materials.

- No x, y



State s — new state s' + reward (feedback)

Mars Rover Example



- Discount factor γ (Gamma)

$0 \longleftrightarrow 1$

↑ lesser cost of travelling.

- Return

$$\text{Return} = 0 + \gamma \times 0 + \gamma^2 \times 0 + \gamma^3 \times 100 + \cdots$$

$$= (0.5)^3 \times 100 \quad \cdots \gamma = 0.5$$

$$= 12.5$$

$$\text{Return} = (0.99)^3 \times 100$$

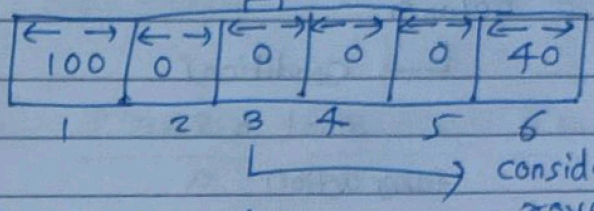
$$= 97.02 \quad \cdots \gamma = 0.99$$

γ	return	Dir^n
0.5	12.5	always left
0.99	97.02	always left.
0.5	20	always right
0.99	39.6	always right

} Return = $40 \times \gamma$

ex:

goes right & reaches → terminal state.



consider the state where rover goes as well.

1, 6 → terminal state.

$\gamma = 0.5$

For state 3 & right.

$$\text{Return} = \gamma^0 \times 0 + \gamma^1 \times 0 + \gamma^2 \times 40$$
$$= 0.25 \times 40$$
$$= 2.5 \times 4$$
$$= 10$$

For state 2 & right.

$$\text{Return} = \gamma^3 \times 40$$
$$= 0.125 \times 40$$
$$= 1.25 \times 4$$
$$= 5$$

For state 3 &

0 ← γ → 1

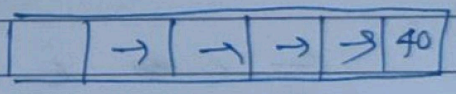
get closer reward tendency.

get farther reward tendency.

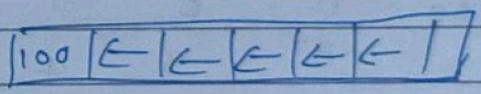
16/04/24 Tuesday

Policies (π)

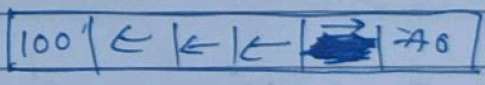
lower (right)



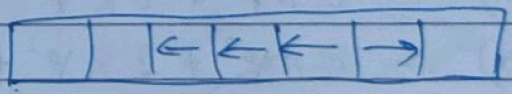
higher (left)



Neader



higher unless lower is 1 step away



(optimal)

$\pi(s) = a$

s → state

a → action

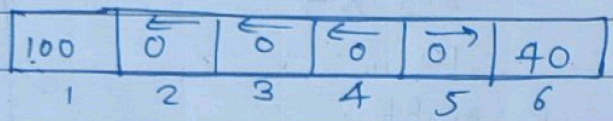
$Q(s, a)$ = Return
↑ ↳ conditions
state-action start in s
value function taken action a
behaves optimally

(eg)

$Q(3, \rightarrow) = ?$

$\gamma = 0.5$

first move right then follow optimal path.



$Q(3, \rightarrow) = \underbrace{0}_{(3 \to 4} + \underbrace{\gamma \times 0}_{\to 3} + \underbrace{\gamma^2 \times 0}_{\to 2} + \underbrace{\gamma^3 \times 0}_{\to 1)} + \gamma^4 \times 100$

$Q(5, \leftarrow) = 0 + \gamma \times 0 + \gamma^2 \times 0 + \gamma^3 \times 0 + \gamma^4 \times 100$

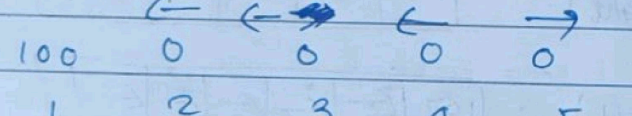
$= (0.5)^4 \times 100$

$= 0.25 \times 25$

$= 6.25$

$(5 \to 4 \to 3 \to 2 \to 1)$

(eg.2)



$\gamma = 0.5$

$Q(3, \leftarrow) = 0 \times \gamma^0 + 0 \times \gamma^1 + 100 \times \gamma^2 = 25$

$Q(3, \rightarrow) = 0 \times \gamma^0 + 0 \times \gamma^1 + 0 \times \gamma^2 + 40 \times \gamma^3$

$= 40 \times 0.5 \times 0.25$

$= 20 \times 0.25$

$= 2.5 \times 2$

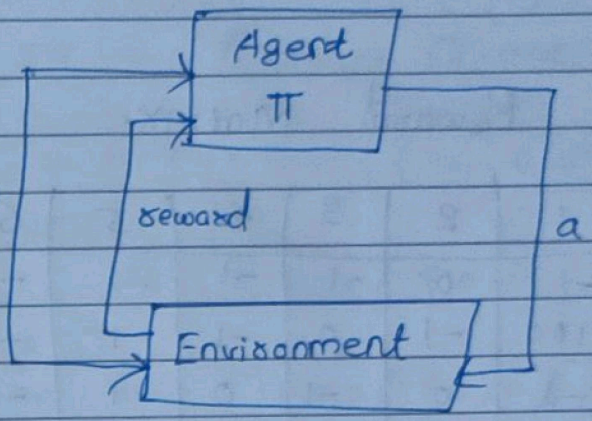
$0 \times \gamma^0 + 0 \times \gamma^1 + 0 \times \gamma^2 + 0 \times \gamma^3 + 100\,\gamma^4$

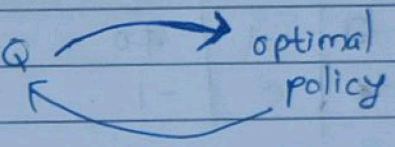
$Q(3, \rightarrow) = 6.25$

$\max(Q(s,a))$

18/04/24 Thursday

Markov Decision Process

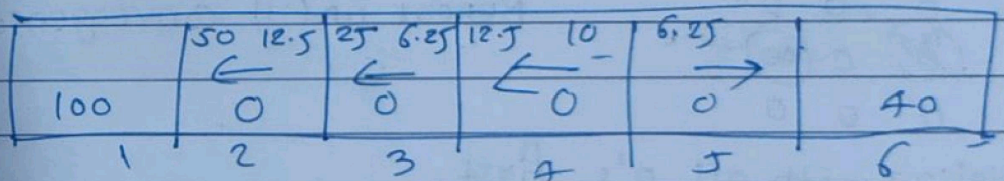




Bellman Equation (guides all ~~reinforcement~~ algos).

$$Q(s,a) = R(s) + \gamma \max Q(s',a')$$

- $Q(s,a)$: Current state, action
- $\max Q(s',a')$: over all actions; after going to s' whatever actions we can take is a'



$Q(4,\rightarrow) = 0 \times \gamma^0 + 0 \times \gamma^1 + 40 \times \gamma^2 = 40 \times 0.25$ ~~= 10~~

$= \sim\sim 20 \times 0.25 \sim\sim$

$= 5.10$

Using Bellman equation.

$Q(3,\rightarrow) = R(3) + (0.5) \max([12.5, 10])$

$Q(4,\rightarrow) = R(4) + (0.5) \max([6.25, 20])$

$Q(4,\leftarrow) = R(4) + (0.5) \max([25, 6.25]) = 0 + 12.5 = 12.5$

$$Q(s,a) = R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots$$
$$= R_1 + \gamma [R_2 + \gamma R_3 + \gamma^2 R_4 + \cdots]$$

22/04/2024 Monday

(e.g.

Reward Matrix.

	1	2	3	4	5	6
1	-1	0	-1	-1	-1	-1
2	100	-1	0	-1	-1	-1
3	-1	0	-1	0	-1	-1
4	-1	-1	0	-1	0	-1
5	-1	-1	-1	0	-1	40
6	-1	-1	-1	-1	0	-1

Q - Matrix (Based on Bellman Equation)

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	(100)	0	0	0	0	0
3	0	(50)	0	0	0	0
4	0	0	(25)	0	0	0
5	0	0	0	(12.5)	0	(40)
6	0	0	0	0	0	0

(e

Choose action
Perform action
Calculate Q-matrix (Bellman eqn)
Repeat until convergence.

Initialize with all 0's first.

$$Q(2,3) = R + \gamma \max_{\text{over all } a} Q(3,a)$$

$$= 0 + 0.5 \max([0,0])$$

$$= 0$$

→ take from Q-matrix.

$Q(2,1) = 100 + 0.5 \times \max(\{0\})$

$= 100$

(now Q matrix changes).

$Q(3,2) = 0 + 0.5 \times \max(\{100, 0\})$

$= 50$ (change Q-matrix).

$Q(4,3) = 0 + 0.5 \times \max(\{0, \ldots 0, 50\})$

$= 25$

$Q(5,4) = 0 + 0.5 \times \max(\{0, 0, \ldots 25\})$.

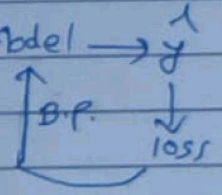
$= 12.5$

$Q(5,6) = 40 + 0.5 \max(\{0, 0\})$

$= 40$

Deep Reinforcement Learning

- Replay of data



- Exploration & Exploitation