we find value iteratively till it convergence. γ = discount factor.

i.e. $v_{11} = 0.9$
$v_{12} = 0.9$
$v_{13} = 0.9$ } thus no change, so stop.

Environment

↙ ↘

model-based methods → Dynamic Programming

model-free methods

Dynamic Programming ↙ ↓

Value Iteration

Policy Iteration ↙ ↘

Policy Evaluation

Policy Improvement

Deterministic

$p(r, s'|s, a) = 1$

Initially assume values of all states as zero (0)
↓
(can be any but mostly take 0.)

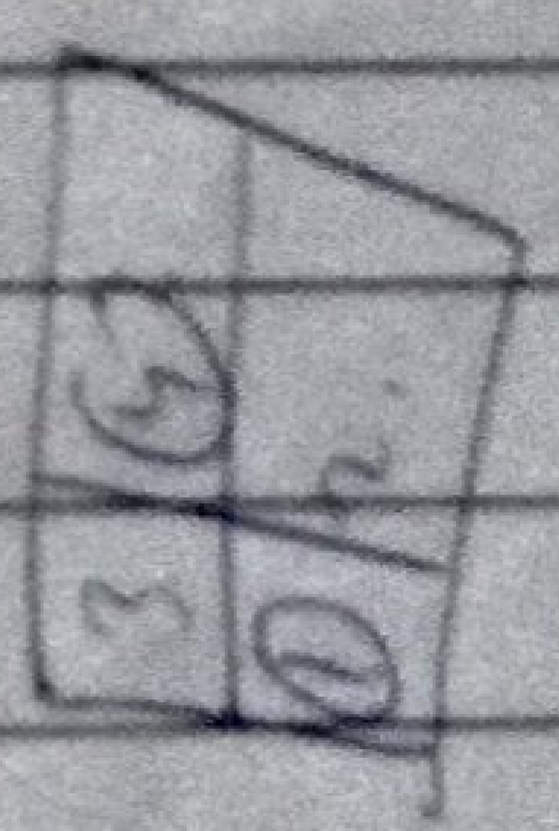
(off-policy basis working)

Q. Why Bellman eq are recursive in nature?

Q. Grid problem (ISE-2 - 5 marks) (ESE - 10 marks)
↳ reward → take minimum steps to reach the goal.

Q. Give a 2×2 matrix:- ($\gamma = 0.9$)

States	Rewards
S_1	-1
S_2	-1
S_3	-1
S_4	10



Sol:- Policy is Deterministic
∴ Let us assume the following action (a) and (s)

Current state	Action	Next state
S_1	a_1	S_2
S_2	a_2	S_3
S_3	a_3	S_4
S_4	~~a4~~ Terminal	

Scanned by TapScanner

∴ Formula :-

$$V_\pi(s) = \sum_a (r + \gamma V_\pi(s'))$$

Let Assume the value of all states as 0 ∴ $V_\pi(s_1) = 0$

$V_\pi(s_2) = 0$

$V_\pi(s_3) = 0$

$V_\pi(s_4) = 0$

∴ Iteration-1

$V_\pi(s_4) = r + \gamma V_\pi(s')$

∴ $V_\pi(s_1) = -1 + 0.9 \times 0 = -1$

$V_\pi(s_2) = -1 + 0.9 \times 0 = -1$

$V_\pi(s_3) = -1 + 0.9 \times 0 = -1$

$V_\pi(s_4) = 10 = 10$

I - 2

$V_\pi(s_1) = -1 + 0.9 \times (-1) = -1.9$

$V_\pi(s_2) = -1 + 0.9 \times (-1) = -1.9$

$V_\pi(s_3) = -1 + 0.9 \times 10 = 8$

$V_\pi(s_4) = 10$

I - 3

$V_\pi(s_1) = -1 + 0.9 \times (-1.9) = -2.71$

$V_\pi(s_2) = -1 + 0.9 \times (8) = 6.2$

$V_\pi(s_3) = -1 + 0.9 \times 10 = 8$

$V_\pi(s_4) = 10$

I - 4

$V_\pi(s_1) = -1 + 0.9 \times 6.2 = 5.58$

$V_\pi(s_2) = -1 + 0.9 \times 8 = 6.2$

$V_\pi(s_3) = -1 + 0.9 \times 10 = 8$

$V_\pi(s_4) = 10$

I - 5

$V_\pi(s_1) = -1 + 0.9(6.2) = 5.58$

$V_\pi(s_2) = -1 + 0.9 \times 8 = 6.2$

$V_\pi(s_3) = -1 + 0.9 \times 10 = 8$

$V_\pi(s_4) = 10$

SC

Delta Rule:-

$$\Delta w = \eta (t - y) x_i$$

- η → learning rate
- t → (observed)
- y → target (expected)
- x_i → (input value)

1) Objective
2) Formula
3) Initialized policy π
4) Initial set value of all states be zero(0) i.e $v(s)=0$

⇒ RL

→ Generalised Policy Iteration (GPI):-

Q.

	0	1	2	3
0				+10
1		Wall		-10
2				
3	(agent)			

Terminal State:-
S(3,3)
S(2,3)

Rewards:-
1) Target state (3,3) = +10
2) Terminal state (2,3) = -10
3) Rest of the state = -1

Policy is Stochastic
Greedy action probability = 0.8
Non-Greedy action probability = 0.1
Discount factor = 0.6

S.T. Formula:-

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)[R + \gamma V_\pi(s')]$$

- $V_\pi(s)$ → value function
- $\pi(a|s)$ → probability of taking an action given a state
- $P(s',r|s,a)$ → environment
- R → Reward
- γ → Discount factor
- $V_\pi(s')$ → Value generated from next state

{ If [illegible] then explain this equation i.e tell what each [illegible] }

{ Objective:- To reach to the goal (target state) in minimum number of steps.

∴ Greedy Action
Following will be the optimal policy

S \ A	0	1	2	3
0	→	→	→	+10
1	↑	[illegible]	↑	-10
2	↑	→	↑	←
3	↑	↑	↑	↑

Let initially value of all states be zero

$\therefore V(s) = 0$

~~Hardcoded~~ ∴ The environment is deterministic

$\therefore P(s', r | s, a) = 1$

Iteration-1 :-

Lets consider ->

$V(0,1) = 0.8 \times (1 \times (-1 + 0.6 \times 0))$ (Right)
$+ 0.1 \times (1 \times (-1 + 0.6 \times 0))$ (left)
$= -0.9$

(non greedy) left 50 0.1 ← [0,1] → (Right (greedy) 50 0.8)

No up, down because wall

Discount factor (γ) = 1

Episode 1 :- $(S_1, a_1, 5)$ $(S_5, a_3, 5)$ $(S_7, A_{10}, 10)$ $(S_4, L)5, 3)$ $(S_{10}, A_5, 6)$

Formula :-

$q_\pi(s, a) = \sigma + \gamma V_\pi(s')$

$\therefore q_\pi(s, A_1) = 5 + 5 + 10 + 3 + 6 = 29$
$q_\pi(s_7, A_{10}) = 10 + 3 + 6 = 19$

→ $q_\pi(s_1, a_1) = 5 + \gamma(1 \times V_\pi(s_5))$
$= 5 + (1 \times 5)$
$= 5 + 5 = 10$
$V_\pi(s_5) = 5$
$q_\pi(s_5, a_3) = 5 + (1 \times V_\pi(s_7))$
$V_\pi(s_7) = 10$
$q_\pi(s_7$

when episodes are expected :-

Ex :-
Episode - $(S_1, a_1, 5)$ $(S_5, A_3, 5)$ $(S_4, A_{10}, 10)$ $(S_1, A_1, 6)$ $(S_{10}, A_5, 6)$

$\therefore q_\pi(s_1, a_1)$

ϵ -> It is a hyperpeter which keeps balance between exploration & exploitation.

high ↓ explore low ↓ exploit

avg => total / no. of episodes.

ϵ - greedy :-

Episode - 3 :- $(S_2, A_3, 6)\ (S_1, a, 5)\ (S_2, a_3, 3)\ (S_1, A_1, 2)\ (S_7, A_2, 4)$

↳

First visit	Every Visit
$q_\pi(S_2, a_1) = 5+3+2+4$	$q_\pi(S_1, a_1) = (5+3+2+4)+(2+4)$
$= 14$	$= 20$
mostly used	
because in every visit,	
relation of every ~~state~~ state-action pair that is repeated	
can add biasness to the solution.	

Total reward

For (S_1, a_1) :-

	First visit	Every visit
Ep-1 :-	$5+5+10+3+6 = 29$	$= 29$
2 :-	$5+6+4+3 = 18$	$= 18$
3 :-	$5+3+2+4 = 14$	$(5+3+2+4)+(2+4) = 20$
$\therefore q_\pi(S_1, a_1)$	$= 20.33$	$\therefore q_\pi(S_1, a_1) = 22.33$

Q- Monte-carlo (ab Diver Question ($\gamma = 1$)

reward = -1

$q(S=(0,0), a = Right)$

Formula :-

$q_\pi(s, a) = \gamma + \gamma V_\pi(s')$

$V_\pi(s) = \Sigma \pi(a/s) \cdot q_\pi(s, a)$

$\pi(a/s)$:- 1 because deterministic

$q_\pi(S=(0,0), Right) = \gamma + \gamma(V(S=(0,1))$

$V(S=(0,1)) = 1 \times q_\pi(S=(0,1), Right)$

$q_\pi(S=(0,1), Right) = \gamma + \gamma(V(S=(0,2))$

$V(S=(0,2)) = 1 \times q_\pi(S=(1,2), Down)$

~~$q(S=(1,2)) = -1 + \gamma$~~

$q_\pi(S=(1,2), Down) = \gamma + \gamma(V(S=(2,2))$

$V_\pi(S=(2,2)) = 1 \times q_\pi(S=(2,3), Right)$

$q_\pi(S=(2,3), Right) = \gamma + \gamma($