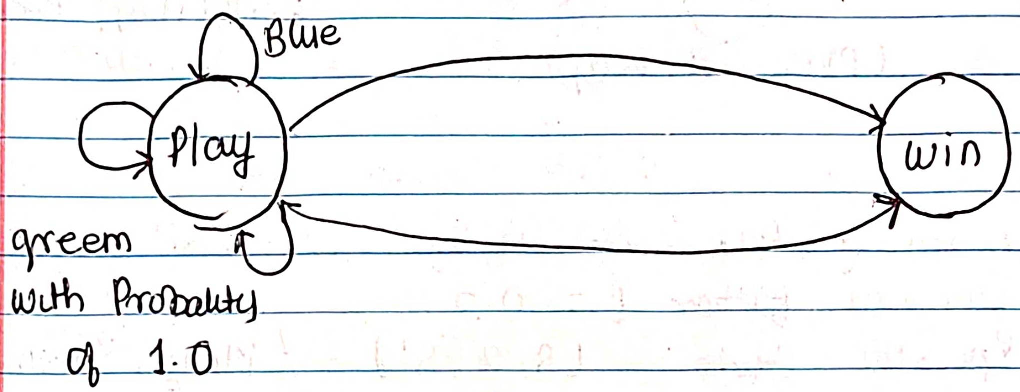
NAME : PAURAVI UDAY NAGARKAR.

ID: W1650209

Homework No.: 5

Problem 2: Game has two States Play & Win. There are three actions available at State Play: Play the Blue machine, Green machine, Red machine. The arrows represent the State transitions, and actions are labelled on arrows. however for Playing Blue & Red machine, the transition Probabilities are unknown. Once win State is reached, game is over.



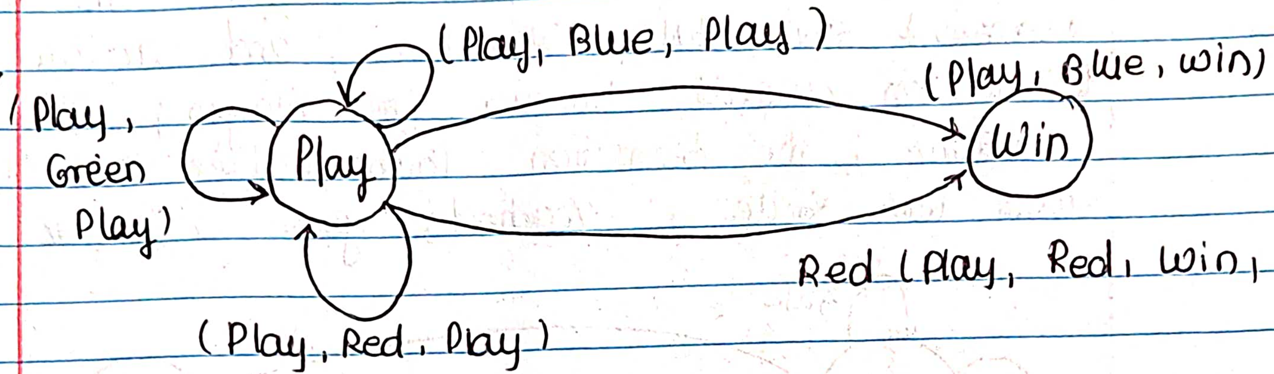
The reward are given:

s	a	s'	R(s, a, s')
Play	green	Play	2
Play	Blue	Play	4
Play	Blue	win	10
Play	Red	Play	0
Play	Red	win	50

Use temporal difference learning to learn the value of Play State, by Successively applying the two episode below. The initial values of all States from 0. With learning rate $\alpha = 0.5$ and discount factor $\gamma = 0.5$. what value of State Play do we learn?

s	a	s'
Play	Green	Play
Play	Red	Play.

→



→ Learning Rate $\alpha = 0.5$
Discount factor $\gamma = 0.5$
Episodes Given :- (S, a, S') = (Play, green, Play)
(Play, Red, Play)

using temporal difference learning to learn the value of the 'play state'

updated $V(S) \leftarrow V(S) + \alpha[\text{Sample} - V(S)]$ 4
$\text{Sample} = R(S, a, S') + \gamma V(S')$

Initalizing the State's value as 0
$V_0(S) = 0$, for S = Play win.

First Episode :

$\text{Sample} = R(\text{Play, Green, Play}) + \gamma V(\text{Play})$
$= 2 + 0.5 * 0 = 2$

updated Value

$V(\text{Play}) = V(\text{Play}) + \alpha(\text{Sample} - V(\text{play})]$
$= 0 + 0.5[2-0]$

$V(\text{Play}) = 1$ after Episode 1.

Second Episode :

Sample = R(Play, Red, Play) + γ V(Play)

= 0 + 0.5 * 1 = 0.5

updated

V(Play) = V(Play) + α[Sample - V(Play)]

= 1 + 0.5 * (0.5 - 1)

V(Play) = 0.75 (after 2nd episode)

Value of State Play for Episode 1 : 1

Value of State Play for Episode 2 : 0.75

Problem 2 for Same game shown in figure 1, we observed three episodes :

S	a	S'
Play	Green	Play
Play	Red	Play
Play	Blue	win

Use Q- learning to update the values of Q- State by applying the above three episodes one after another. use a learning of 0.5 and a discount of 0.5. Intialize all Q states values as 0

S	a	Q(S,a)
Play	Green	
Play	Red	
Play	Blue	

→ learning Rate $\alpha = 0.5$
Discount factor $\gamma = 0.5$
Episodes given - $(S, a, S') \rightarrow$ (Play, Green, Play)
(Play, Red, Play
(Play, Blue, win)

using Q learning to update the values of Q-states
Initializing Q states values as '0'
$Q_0(S) = 0$ for S = Play, win.
updated $Q(S, a) \leftarrow Q(S, a) + \alpha.(\text{Sample} - Q(S, a))$
where
$\text{Sample} = R(S, a, S') + \gamma \max_a Q'(S', a')$

Episode 1 :
Sample = R(Play, green, Play) + 0.5 * Max
(Q (Play a'))

$= 2 + 0 * 0.5$
$= 2$.

updated Play
$Q(S, a) = 0 + 0.5(2 - 0)$
$= 1$.

$Q(\text{Play}, \text{green}) = 1$

Episode 2 :
$\text{Sample} = R(\text{Play}, \text{Red}, \text{Play}) + 0.5 * \max_{a'} (Q(\text{Play}, a'))$

$\text{Sample} = 0 + 0.5 * 1$
$= 0.5$

updated
$Q(S, a) = Q(\text{Play}, \text{Red}) + 0.5 * (\text{Sample} - Q(\text{Play}, \text{Red}))$
$= 0 + 0.5 * (0.5 - 0)$
$Q(S, a) = 0.25$

$Q(\text{Play}, \text{Red}) = 0.25$

Episode 3

Sample = $R(\text{Play}, \text{Blue}, \text{win}) + \gamma \max_{a'} Q(\text{win}, a')$

$= 10 + 0.5 * 0$

$= 10$

updated

$Q(s, a) = Q(\text{Play}, \text{Blue})$

$= Q(\text{Play}, \text{Blue}) + \alpha(\text{Sample} - Q(\text{Play}, \text{Blue}))$

$= 0 + 0.5 * (10 - 0)$

$Q(\text{Play}, \text{Blue}) = 5$

S	a	Q(S, a)
Play	green	1
Play	Red	0.25
Play	Blue	5