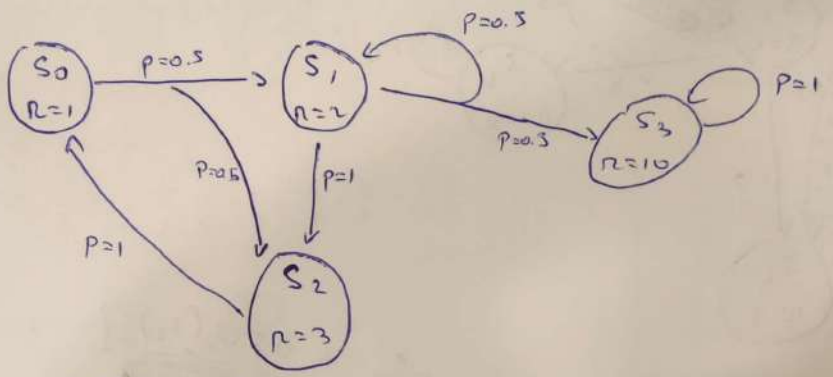


$\gamma = 1$. U_0 (fiecare stare) = 0

$t=1$: $U_1(S_0) = R(S_0) + 1 \cdot \max\left[0.5 \cdot U_0(S_1) + 0.5\, U_0(S_2)\right]$

$= 1 + \max\left[0.5 \cdot 0 + 0.5 \cdot 0\right) = 1$

$U_1(S_1) = 2 + \max\left[0.5 \cdot U_0(S_1) + 0.5\, U_0(S_3)\right.$ prima nemurită
$\left. 1 \cdot U_0(S_2)\right]$ a doua

$= 2 + \max[0, 0] = 2$

$U_1(S_2) = 3 + \max[1 \cdot U_0(S_0)] = 3$

$U_1(S_3) = 10 + \max[1 \cdot U_0(S_3)] = 10 + 0 = 10$

$t=2$: *foc direct pt S_1, nu mai foc pt S_0 că nu îl am printre rezultate*

$U_2(S_1) = R(S_1) + \max\left[0.5 \cdot U_1(S_1) + 0.5\, U_1(S_3)\right.$,
$\left. 1 \cdot U_1(S_2)\right]$

$= 2 + \max\left[0.5 \cdot 2 + 0.5 \cdot 10, 3\right] = 2 + \max\{1 + 5, 3\}$
$= 2 + 6 = 8$

6. (1.5p) Fie jocul din figura de mai jos. Ce strategie este dominată pentru Rose? Care este echilibrul Nash (pur sau mixt) al jocului? Cât câștigă Rose și Colin în situația de echilibru?

		Colin: P_D D	Colin: $1-P_D$ E
Rose: P_B	A	-2, 3	-4, 0
Rose: $1-P_B$	B	-1, 2	2, -1
	C	3, -2	-3, 1

A este strategie dominata

$$E_D = P_B \cdot 2 + (1-P_B)\cdot(-2)$$

$$E_E = P_B \cdot (-1) + (1-P_B)\cdot 1$$

$$4P_B - 2 = -2P_B + 1$$

$$P_B = \frac{1}{2}$$

$$1-P_B = \frac{1}{2}$$

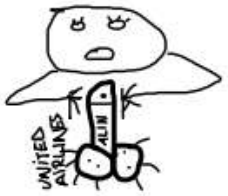
$$E_B = P_D \cdot (-1) + (1-P_D)\cdot 2$$

$$E_C = P_D \cdot 3 + (1-P_D)\cdot(-3)$$

$$-3P_D + 2 = 6P_D - 3$$

$$P_D = \frac{5}{9}$$

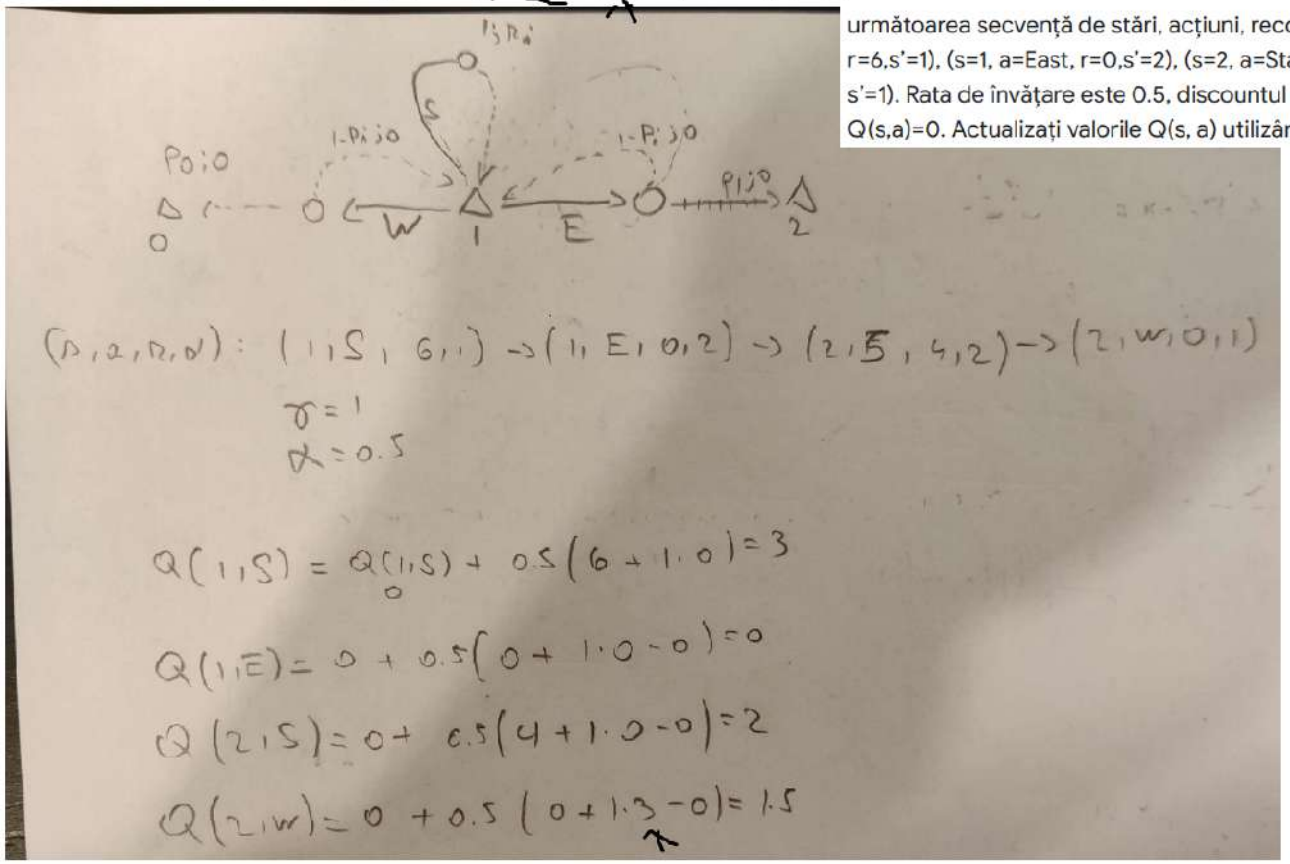
$$1-P_D = \frac{4}{9}$$



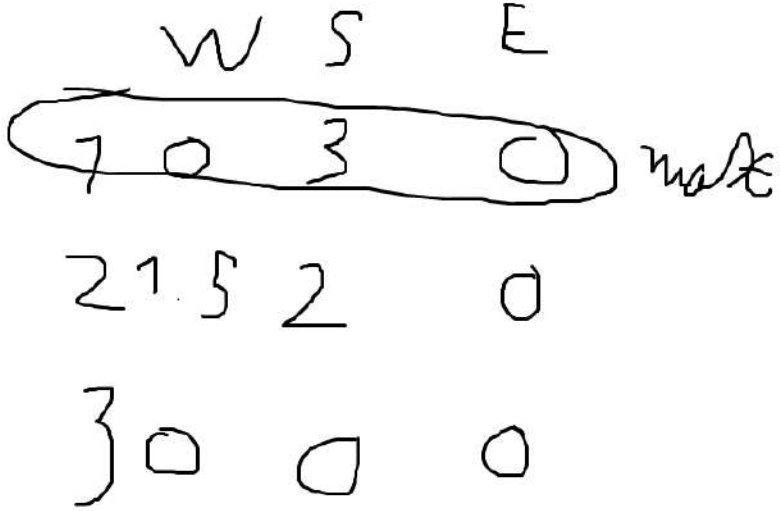
$$E_D = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot(-2) = 0$$

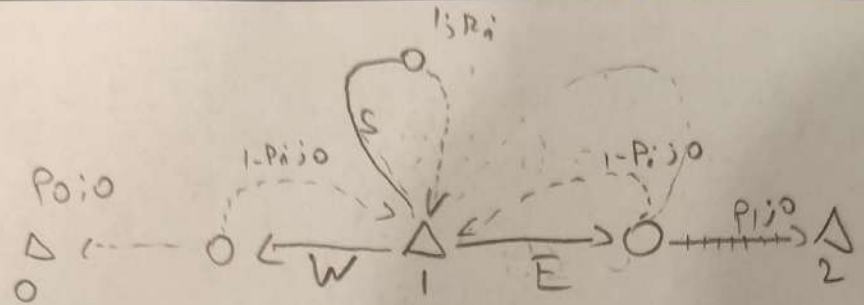
$$E_B = -\frac{5}{9} + \frac{8}{9} = \frac{3}{9} = \frac{1}{3}$$

$$Q(s,a) = Q(s,a) + \alpha(R(s) + \gamma max_{a'} Q(s',a') - Q(s,a))$$



următoarea secvență de stări, acțiuni, recompense (s, a, r, s') : (s=1, a=Stay, r=6,s'=1), (s=1, a=East, r=0,s'=2), (s=2, a=Stay, r=4, s'=2), (s=2, a=West, r=0, s'=1). Rata de învățare este 0.5, discountul este 1, iar valorile inițiale Q(s,a)=0. Actualizați valorile Q(s, a) utilizând algoritmul Q-learning. Care





(P, a, R, P'): $(1, S, 6, 1) \rightarrow (1, E, 0, 2) \rightarrow (2, S, 4, 2) \rightarrow (2, W, 0, 1)$

$\gamma = 1$

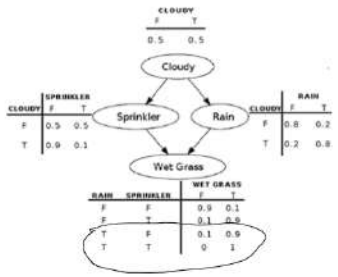
$\alpha = 0.5$

$Q(1,S) = Q_0(1,S) + 0.5(6 + 1 \cdot 0) = 3$

$Q(1,E) = 0 + 0.5(0 + 1 \cdot 0 - 0) = 0$

$Q(2,S) = 0 + 0.5(4 + 1 \cdot 0 - 0) = 2$

$Q(2,W) = 0 + 0.5(0 + 1 \cdot 3 - 0) = 1.5$



I $P(R \mid \neg C, W) = \alpha \sum_S P(\neg C, W, S)$

$= \alpha \sum_S P(\neg C) P(W \mid S, R) P(R \mid \neg C) P(S \mid \neg C)$

$P(\neg C) = 0.5$, $P(R \mid \neg C) = 0.2$

$= \alpha \left(0.5 \cdot 1 \cdot 0.2 \cdot 0.5 + 0.5 \cdot 0.9 \cdot 0.2 \cdot 0.5 \right)$

II $P(\neg R \mid \neg C, W) = \alpha \sum_S P(\neg C, W, S)$

$= \alpha \sum_S P(C) \cdot P(W \mid S, \neg R) \cdot P(\neg R \mid \neg C) \cdot P(S \mid \neg C)$

$= \alpha (0.5 \cdot 0.9 \cdot 0.8 \cdot 0.5 + 0.5 \cdot 0.1 \cdot 0.8 \cdot 0.5)$

$\alpha \cdot (0.295) = 1$

$$\alpha = \frac{1}{0.295} = 3.3898$$

rev la cerința

$\alpha \left(0.5 \cdot 1 \cdot 0.2 \cdot 0.5 + 0.5 \cdot 0.9 \cdot 0.2 \cdot 0.5 \right) = 0.322031$

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

New value of state t

Former estimation of value of state t

Learning Rate

Reward

Discounted value of next state

TD Target

Therefore, our $Q(S_t, A_t)$ **update formula goes like this:**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

New Q-value estimation

Former Q-value estimation

Learning Rate

Immediate Reward

Discounted Estimate optimal Q-value of next state

TD Target

Former Q-value estimation

TD Error