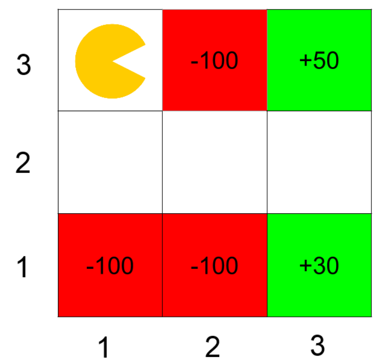
CS 6300 | # Q-Learning Practice | **March 9, 2021**

Consider the grid-world given below and an agent who is trying to learn the optimal policy. Rewards are only awarded for taking the *Exit* action from one of the shaded states. Taking this action moves the agent to the Done state, and the MDP terminates. Assume $\gamma = 1$ and $\alpha = 0.5$ for all calculations. All equations need to explicitly mention γ and α if necessary.



1. The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r).

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(3,1), S, (2,1), 0	(3,1), S, (2,1), 0	(3,1), S, (2,1), 0	(3,1), S, (2,1), 0	(3,1), S, (2,1), 0
(2,1), E, (2,2), 0	(2,1), E, (2,2), 0	(2,1), E, (2,2), 0	(2,1), E, (2,2), 0	(2,1), E, (2,2), 0
(2,2), E, (2,3), 0	(2,2), S, (1,2), -100	(2,2), E, (2,3), 0	(2,2), E, (2,3), 0	(2,2), E, (2,3), 0
(2,3), N, (3,3), +50		(2,3), S, (1,3), +30	(2,3), N, (3,3), +50	(2,3), S, (1,3), +30

Fill in the following Q-values obtained from direct evaluation from the samples:

Q((2,3), N) = __________ Q((2,3), S) = __________ Q((2,2), E) = __________

2. Q-learning is an online algorithm to learn optimal Q-values in an MDP with unknown rewards and transition function. The update equation is:

$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(R(s_t, a_t, s_{t+1}) + \gamma \max_{a'} Q(s_{t+1}, a'))$$

where γ is the discount factor, α is the learning rate and the sequence of observations are $(\cdots, s_t, a_t, s_{t+1}, r_t, \cdots)$.

(a) Particularize the Q-learning equation for this problem.

$$Q((2,1),E) =$$

$$Q((2,2),E) =$$

$$Q((2,2),S) =$$

$$Q((2,3),N) =$$

$$Q((2,3),S) =$$

$$Q((3,1),S) =$$

(b) Given the episodes in part 1, fill in the time at which the following Q values first become non-zero. Your answer should be of the form (**episode#,iter#**) where **iter#** is the Q-learning update iteration in that episode.

Q((2,1), E) = _________ Q((2,2), E) = _________ Q((2,3), S) = _________

3. Repeat with SARSA. The update equation is:

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha(R(s_t, a_t, s_{t+1}) + \gamma Q(s_{t+1}, a_{t+1}))$$

(a) Particularize the SARSA equation for this problem.

$$\begin{aligned}
Q((2,1),E) &= \\
Q((2,2),E) &= \\
Q((2,2),S) &= \\
Q((2,3),N) &= \\
Q((2,3),S) &= \\
Q((3,1),S) &=
\end{aligned}$$

(b) Given the episodes in part 1, fill in the time at which the following Q values first become non-zero.

Q((2,1), E) = __________ Q((2,2), E) = __________ Q((2,3), S) = __________