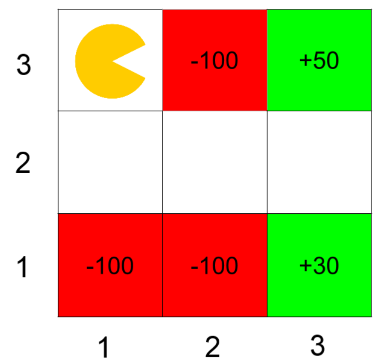
CS 6300 | **Q-Learning Practice** | **March 9, 2021**

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) = 50 $\qquad$ Q((2,3), S) = 30 $\qquad$ Q((2,2), E) = 40

Direct evaluation is just averaging the discounted reward after performing action a in state s.

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 $(\cdots, s_t, a_t, s_{t+1}, r_t, \cdots)$. 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. If the specified Q value never becomes non-zero, write *never*.

Particularize the Q-learning equation for this problem.

$$\begin{aligned}
Q((2,1), E) &= \frac{1}{2}Q((2,1), E) + \frac{1}{2}\max\{Q((2,2), E), Q((2,2), S)\} \\
Q((2,2), E) &= \frac{1}{2}Q((2,2), E) + \frac{1}{2}\max\{Q((2,3), N), Q((2,3), S)\} \\
Q((2,2), S) &= \frac{1}{2}Q((2,2), S) - 50 \\
Q((2,3), N) &= \frac{1}{2}Q((2,3), N) + 25 \\
Q((2,3), S) &= \frac{1}{2}Q((2,3), S) + 15 \\
Q((3,1), S) &= \frac{1}{2}Q((3,1), S) + \frac{1}{2}Q((2,1), E)
\end{aligned}$$

Initially all Q-values are zero. Starting with episode 1 trial 1,

Episode	Trial	Update
1	1	$Q((3,1),S) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{0,0\} = 0$
	3	$Q((2,2),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{0,0\} = 0$
	4	$Q((2,3),N) = \frac{1}{2} \cdot 0 + 25 = 25$
2	1	$Q((3,1),S) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{0,0\} = 0$
	3	$Q((2,2),S) = \frac{1}{2} \cdot 0 - 50 = -50$
3	1	$Q((3,1),S) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{0,-50\} = 0$
	3	$Q((2,2),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{0,25\} = 12.5$
	4	$Q((2,3),S) = \frac{1}{2} \cdot 0 + 15 = 15$
4	1	$Q((3,1),S) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2} \cdot 0 + \frac{1}{2} \max\{12.5,-50\} = 6.25$

The answer is:

Q((2,1), E) = (4,2) Q((2,2), E) = (3,3) Q((2,3), S) = (3,4)

3. Repeat with SARSA. The update equation is:

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

Particularize the Q-learning equation for this problem.

$$
\begin{aligned}
Q((2,1),E) &= \frac{1}{2}Q((2,1),E) + \frac{1}{2}\begin{cases} Q((2,2),E) & \text{action } E \\ Q((2,2),S) & \text{action } S \end{cases} \\
Q((2,2),E) &= \frac{1}{2}Q((2,2),E) + \frac{1}{2}\begin{cases} Q((2,3),N) & \text{action } N \\ Q((2,3),S) & \text{action } S \end{cases} \\
Q((2,2),S) &= \frac{1}{2}Q((2,2),S) - 50 \\
Q((2,3),N) &= \frac{1}{2}Q((2,3),N) + 25 \\
Q((2,3),S) &= \frac{1}{2}Q((2,3),S) + 15 \\
Q((3,1),S) &= \frac{1}{2}Q((3,1),S) + \frac{1}{2}Q((2,1),E)
\end{aligned}
$$

Initially all Q-values are zero. Starting with episode 1 trial 1,

Episode	**Trial**	**Update**
1	1	$Q((3,1),S) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	3	$Q((2,2),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	4	$Q((2,3),N) = \frac{1}{2}\cdot 0 + 25 = 25$
2	1	$Q((3,1),S) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	3	$Q((2,2),S) = \frac{1}{2}\cdot 0 - 50 = -50$
3	1	$Q((3,1),S) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	3	$Q((2,2),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	4	$Q((2,3),S) = \frac{1}{2}\cdot 0 + 15 = 15$
4	1	$Q((3,1),S) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	3	$Q((2,2),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 25 = 12.5$
	4	$Q((2,3),N) = \frac{1}{2}\cdot 25 + 25 = 37.5$
5	1	$Q((3,1),S) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$
	2	$Q((2,1),E) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 12.5 = 6.25$

The answer for Q-state $Q((2,1),E)$ has changed.

Q((2,1), E) = (5,2) Q((2,2), E) = (3,3) Q((2,3), S) = (3,4)