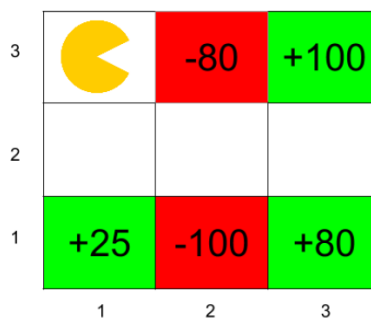


CAP 6629 Homework 1

Due: Feb-11, 2022 (11:59PM)

Consider the grid-world given below and Pacman who is trying to learn the optimal policy. If an action results in landing into one of the shaded states the corresponding reward is awarded during that transition. All shaded states are terminal states, i.e., the MDP terminates once arrived in a shaded state. The other states have the *North*, *East*, *South*, *West* actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grid). Assume the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state (1, 3).



Hint: Q-values obtained by Q-learning updates - $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma \max_{a'} Q(s', a'))$.

(a) What is the value of optimal value function V^* at the following states:

$V^*(3,2) =$

$V^*(2,2) =$

$V^*(1,3) =$

(b) 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 is an Episode is a tuple containing (s, a, s', r) .

Hint: N, E, S, W refer to the moving directions.

Episode 1	Episode 2	Episode 3
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

Using Q-learning updates, what the following Q-values after the above three episodes:

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

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

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