Reinforcement Learning CAP 6629, Homework 2

Professor Xiangnan Zhong

Matthew Acs, Fall 2023

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 grad). 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 - $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a'} Q(s',a'))$.

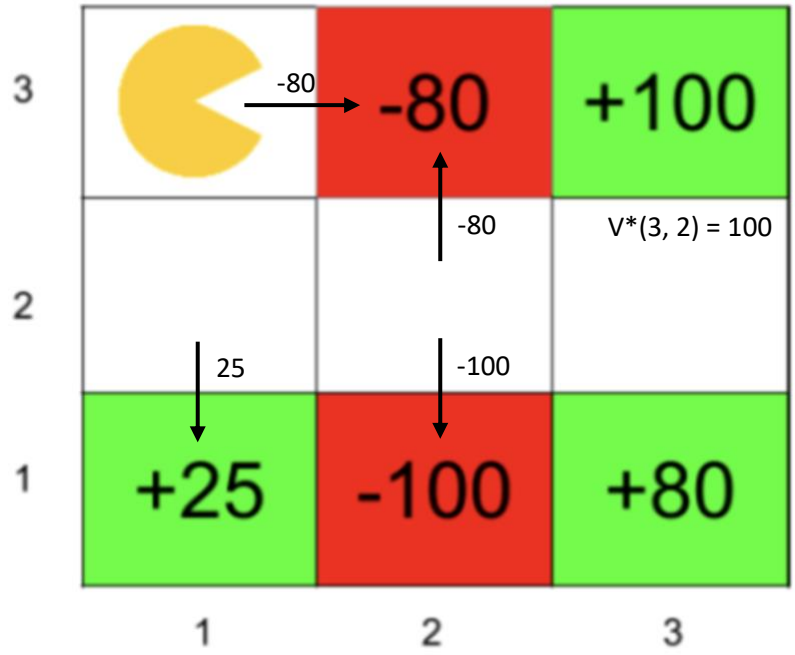
(a) What is the value of optimal value function V^* at the following states:
$V^*(3,2)$ =

$V^*(2,2)$=

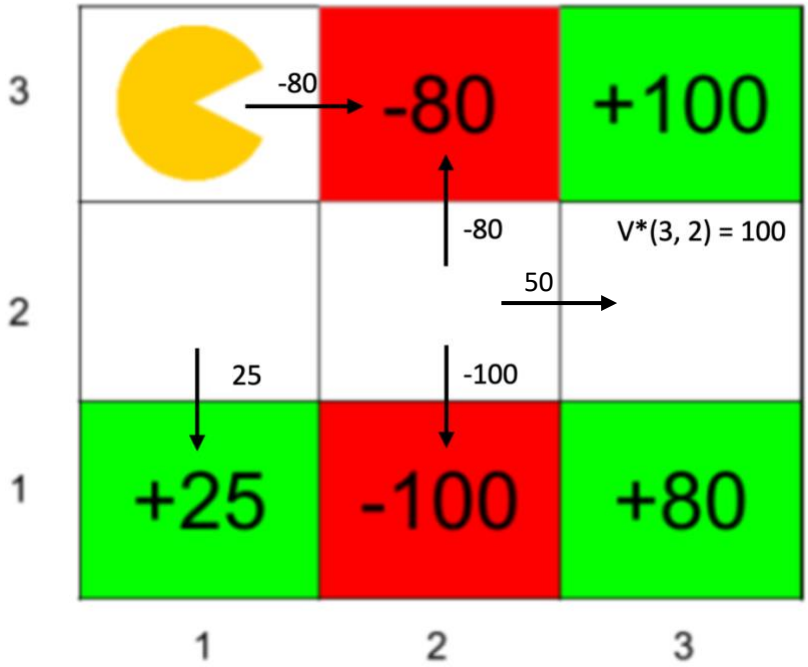
$V^*(1,3)$=

$V^*(s) = \max_{a'}(Q(s, a'))$
$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma \max_{a'} Q(s', a'))$
$Q(s, a) \leftarrow (0.5)Q(s, a) + 0.5(R(s, a, s') + 0.5 \max_{a'} Q(s', a'))$

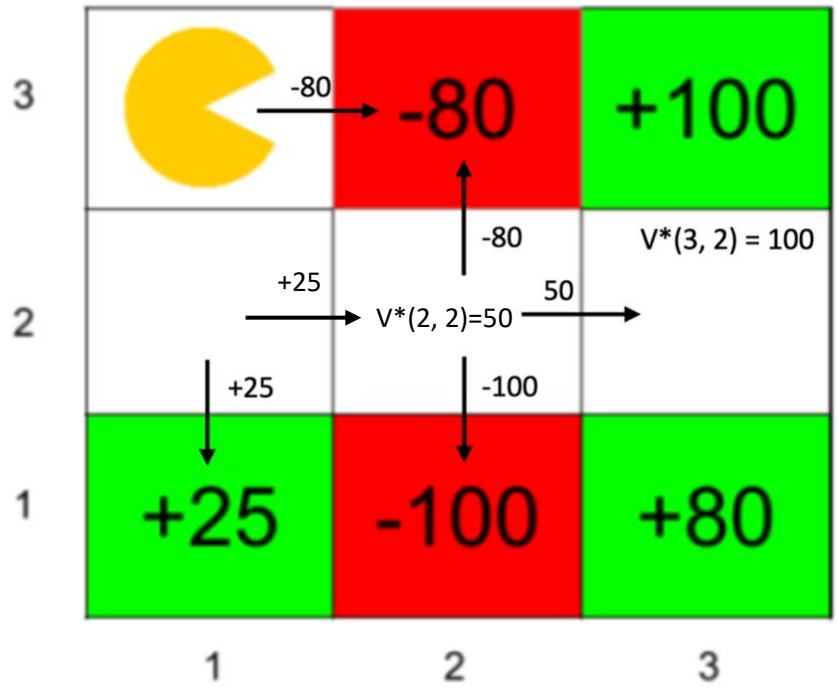
Based on the updating rules above, the state next to the terminal state with the greatest reward will converge to having a $V^*(s)$ of the reward. Thus, $V^*(3,2) = 100$. Additionally, the Q value for a state-action pair will converge to having the value of the reward if the state-action pair moves the agent to a terminal state. Thus, the following diagram can be drawn:



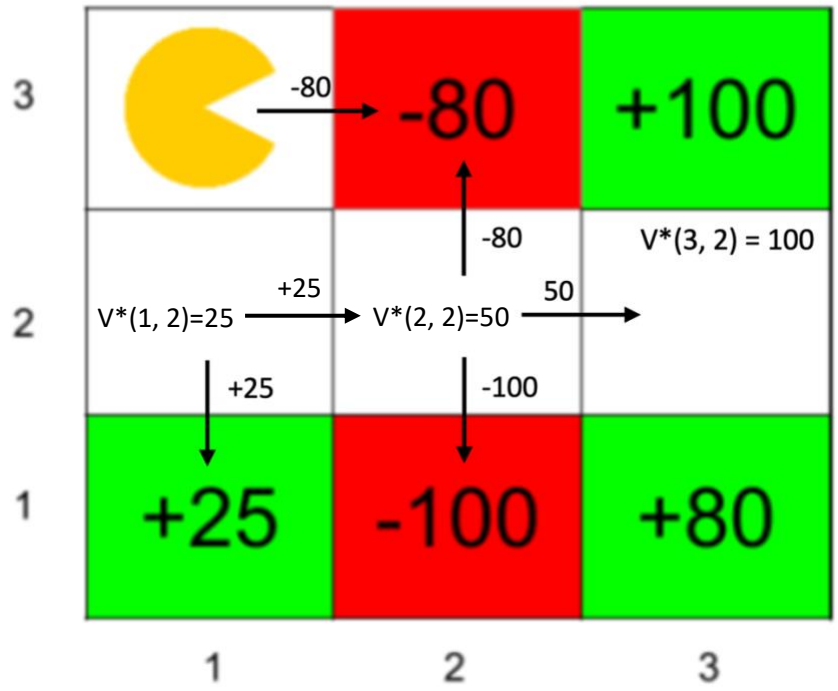
The value of a state-action pair will converge to half of the V* value of the next state if the transition provides no reward. Thus, the diagram can be updated further:

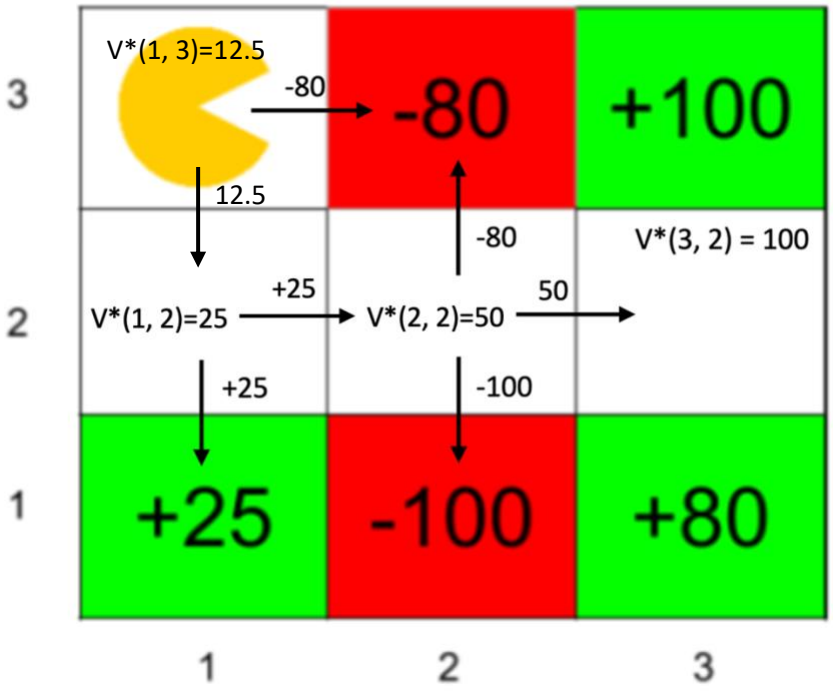


Finally, V*(2,2) = max(50, ½ (V*(1,2)) and V*(1,2) = max(25, ½ (V*(2,2)), which leads to V*(2,2) = max(50, ½ (max(25, ½ (V*(2,2))) → V*(2,2) = max(50, ½ (max(25, ½ (max(50, ½ (V*(1,2)))) → V*(2,2) = max(50, ½ (max(25, ½ (max(50, ½ (max(25, ½ (V*(2,2))))) → … → V*(2,2) = 50. Hence, V*(2,2) = 50.



Based on the V* value obtained for (2,2), we can finish updating the gridworld examples Q and V values.





Thus, the V* values for the states (3, 2), (2, 2), and (1, 3) are:

$V^*(3, 2) = 100$
$V^*(2, 2) = 50$
$V^*(1, 3) = 12.5$

(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)=

$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma \max_{a'} Q(s', a'))$

$\alpha = 0.5$
$\gamma = 0.5$

Q-tables for the three episodes:

$Q^{initial}$

	North	**South**	**East**	**West**
1, 2	0	0	0	0
1, 3	0	0	0	0
2, 2	0	0	0	0
3, 2	0	0	0	0

Q^{E1}

	North	**South**	**East**	**West**
1, 2	0	0	0	0
1, 3	0	0	0	0
2, 2	0	-50	0	0
3, 2	0	0	0	0

Q^{E2}

	North	**South**	**East**	**West**
1, 2	0	0	0	0
1, 3	0	0	0	0
2, 2	0	-50	0	0
3, 2	50	0	0	0

Q^{E3}

	North	South	East	West
1, 2	0	0	0	0
1, 3	0	0	0	0
2, 2	0	-50	12.5	0
3, 2	50	40	0	0

Q((3, 2), N) = 50
Q((1, 2), S) = 0
Q((2, 2), E) = 12.5