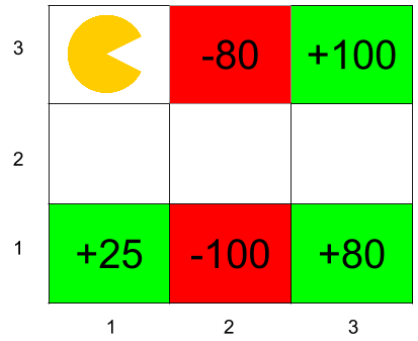
1 Approximate Q-Learning

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).



1. What is the value of the optimal value function V^* at the following states:

State	Optimal value
$V^*(3,2)$	100
$V^*(2,2)$	50
$V^*(1,2)$	25

2. 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
(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

Use Q-Learning to update the Q-values you can.

The formula for Q-learning particularized to this problem is:

$$Q(s,a) \leftarrow \frac{1}{2}Q(s,a) + \frac{1}{2}(R(s,a,s') + \frac{1}{2}\max_{a'} Q(s',a'))$$

Non-zero updates to the following Q-values take place.

$$\begin{aligned}
Q((2,2),S) &= \frac{1}{2}\cdot 0 + \frac{1}{2}(-100 + \frac{1}{2}\cdot 0) = -50 \\
Q((3,2),N) &= \frac{1}{2}\cdot 0 + \frac{1}{2}(100 + \frac{1}{2}\cdot 0) = 50 \\
Q((2,2),E) &= \frac{1}{2}\cdot 0 + \frac{1}{2}(0 + \frac{1}{2}\cdot 50) = 12.5 \\
Q((3,2),S) &= \frac{1}{2}\cdot 0 + \frac{1}{2}(80 + \frac{1}{2}\cdot 0) = 40
\end{aligned}$$

3. Consider a feature based representation of the Q-value function:

$$\hat{Q}(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$f_1(s)$: The x coordinate of the state $\qquad$ $f_2(s)$: The y coordinate of the state

$$f_3(N) = 1,\ f_3(S) = 2,\ f_3(E) = 3,\ f_3(W) = 4$$

(a) Given that all w_i are initially 0, what are their values after the first episode using approximate Q-learning weight updates? For a particular trial, update the weights independently.
The weight update rule is:

$$w_i \leftarrow w_i + \alpha([R(s,a,s') + \gamma \max_{a'} Q(s',a')] - \hat{Q}(s,a))f_i(s,a)$$

For this problem, the weights are updated in the last action of episode 1. All features have the value 2.

$$w_i = 0 + \frac{1}{2}([-100 + \frac{1}{2}\cdot 0] - 0)\cdot 2 = -100$$

Weight	**Value**
w_1	-100
w_2	-100
w_3	-100

(b) Assume the weight vector w is equal to $(1, 1, 1)$. What is the action prescribed by the Q-function in state $(2, 2)$?

The Q function for state (2,2) under action a is:

$$Q((2, 2), a) = 1 \cdot 2 + 1 \cdot 2 + 1 \cdot f_3(a)$$

Action $a = west$ gives the largest Q-state value for state (2,2).