PROBLEM 6

Problem 5 Consider a robot that can either stand still and CHARGE (using its solar panels) or it can scurry around and EXPLORE. The robot's state (as we measure it) represents only how charged its battery is and can be EMPTY, LOW or HIGH. The robot is very eager to explore and this is how the rewards are set. The MDP transition probabilities and rewards are specified as shown below.

s	a	s'	T(s,a,s')	R(s,a)
HIGH	EXPLORE	LOW	1.0	+2
LOW	EXPLORE	EMPTY	1.0	+2
EMPTY	EXPLORE	EMPTY	1.0	-10
HIGH	CHARGE	HIGH	1.0	0
LOW	CHARGE	HIGH	1.0	-1
EMPTY	CHARGE	HIGH	1.0	-10

Note that the reward only depends on the robot's current state and action, not the state that it transitions to.

(5.1) **(3 points)** Based on the transitions and rewards (without further calculation), what is the optimal policy for this robot if we set the discount factor $\gamma = 0$?

$$\begin{aligned} \pi_0^*(HIGH) &= \\ \pi_0^*(LOW) &= \\ \pi_0^*(EMPTY) &= \end{aligned}$$

(5.2) **(4 points)** Could changing the discount factor γ change the optimal action to take in any state? (check all that apply)

( ) when s =HIGH?

( ) when s =LOW?

( ) when s =EMPTY?

(5.3) **(4 points)** Let's see how the robot values its states, and recovers those values through value iteration, when the discount factor is set to $\gamma = 0.5$. We start with all zero values as shown in the first value column. Please fill out the table

s	$V_0(s)$	$V_1(s)$	$V_2(s)$
EMPTY	0		
LOW	0		
HIGH	0		

(5.4) **(4 points)** If the robot uses values $V_2(s)$ as the true (converged) values, which action would it take in state $s = LOW$? (Show your calculation)