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 $\gamma = 0$?

$\pi_0^*(HIGH)$ = EXPLORE
$\pi_0^*(LOW)$ = EXPLORE
$\pi_0^*(EMPTY)$ = EXPLORE or CHARGE

(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? doesn't change if

(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	-10	-9
LOW	0	2	0
HIGH	0	2	3

(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)

$$\text{argmax} \begin{cases} +2 + \gamma V_2(\text{EMPTY}) & a = \text{EXPLORE} \\ -1 + \gamma V_2(\text{HIGH}) & a = \text{CHARGE} \end{cases}$$

change from CHARGE because, even though both CHARGE and EXPLORE appear to cost the same -10 in the table, performing a CHARGE leads to higher reward states.

c) Recall that $V(s) = \max_a Q(s,a)$. The possible s' in this model for s =EMPTY are EMPTY and HIGH. We determine $Q_2(EMPTY, EXPLORE)$ and $Q_2(EMPTY, CHARGE)$ and take the max of the two to be $V_2(EMPTY)$.

$$Q_2(EMPTY, EXPLORE) = -10 + 0.5 * V_1(EMPTY) = -15$$

$$Q_2(EMPTY, CHARGE) = -10 + 0.5 * V_1(HIGH) = -9$$

So $V_2(EMPTY) = -9$.

d) Solutions are fine.

7) a) Recall that the normal vector to a hyperplane is defined as the θ parameter. The hidden sign units in this network perform the same kind linear combination with an offset that our signed hyperplanes use.

b) Using the fact that the hidden units act as linear classifiers, we can draw decision boundaries that separate the signed points (orientation is important). Multiple solutions.

c) Solutions are fine.

d) Solutions are fine.

8) a) Solutions are fine.

b) We get the optimal policy from the largest $Q_1(s,a)$-values.

c) We can get the $V_1(s)$ values from the largest $Q(s,a)$ values.

9) a) Recall that the margin of a dataset with respect to a separator is the minimum margin of all points in the dataset with respect to the separator. Removing a single point can result in the margin of the dataset increasing (if the removed point had the minimum margin) or staying the same (if it did not).

b) If there were a separator for the dataset with some of the coordinates omitted, then the same separator (with 0 weights for the omitted coordinates) would still separate the original dataset.

c) The setup of this problem is unusual because we are given a fixed set of classifiers $\mathcal{H} = \{h_1, h_2, h_3\}$ and select one based on a training set error. The parameters of the classifiers do not actually depend on the training set, and the test set error is fixed for a particular choice of classifier. We are also told the relative test errors are $\mathcal{E}(h_1) < \mathcal{E}(h_2) < \mathcal{E}(h_3)$. The point of the problem is to see how the function we choose given a training set changes with the number of training points.