Q1. Reinforcement Learning Background

1. Each True/False question is worth 1 points. *Briefly justify* your answers.

   (i) [*true* or *false*] Temporal difference learning is a model-based learning method.

   (ii) [*true* or *false*] Using an optimal exploration function for Q-learning will not have any regret while learning the optimal policy.

   (iii) [*true* or *false*] In a deterministic MDP, Q-learning with a learning rate of $\alpha = 1$ cannot learn the optimal q-values.

   (iv) [*true* or *false*] A large γ (around 1) incentivizes greedy behavior.

   (v) [*true* or *false*] A living reward less than 0 encourages policies that maximize the probability of reaching terminal states.

   (vi) [*true* or *false*] Every $\gamma < 1$ can be rewritten using a mathematically equivalent negative living reward.

2. Properties of reinforcement learning algorithms.

   (a) Which algorithm, assuming ample exploration, does not provide enough information to obtain an optimal policy at the time of convergence? (Select all that apply)

   □ Temporal Difference learning to estimate $U(s)$.

   □ Direct Evaluation to estimate $U(s)$.

   □ Q-Learning to estimate $Q(s, a)$.

   □ Model-based learning of $T(s, a, s')$ and $R(s, a, s')$.

   (b) Assuming we run for infinitely many steps, for which exploration policies is Q-learning guaranteed to converge to the optimal Q-values for all state-action pairs. Assume we chose reasonable values for α and all states of the MDP are connected via some path. (Select all that apply)

   □ A fixed optimal policy.

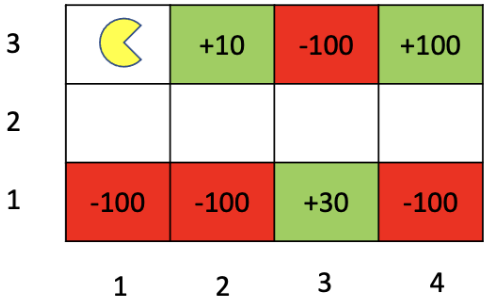
   □ A fixed policy taking actions uniformly at random.

   □ An ϵ-greedy policy

   □ A greedy policy.

Q2. Grid-World Reinforcement

(a) 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 will take the exit action and collect the corresponding reward once it arrives 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). For this question, Pacman **does not** have to learn the values for the terminal (shaded) states, these are given to him and remain fixed.



3		+10	-100	+100
2				
1	-100	-100	+30	-100
	1	2	3	4

1. 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	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), S, (1,1), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(1,1), Exit, D, -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), N, (2,3), 0	(2,2), E, (3,2), 0
	(3,2), E, (4,2), 0	(3,2), S, (3,1), 0	(2,3), Exit, D, +10	(3,2), E, (4,2), 0
	(4,2), N, (4,3), 0	(3,1), Exit, D, +30		(4,2), N, (4,3), 0
	(4,3), Exit, D, +100			(4,3), Exit, D, +100

Fill in the following Q-values obtained from direct evaluation from the samples (round to three decimal places):

Q((4,2), N) = ________ Q((1,2), E) = ________ Q((2,2), E) = ________

2. Q-learning is an online algorithm to learn optimal Q-values in an MDP with unknown rewards and transition function. The update equation is:

$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

where γ is the discount factor, α is the learning rate and the sequence of observations are $(\cdots, s_t, a_t, s_{t+1}, r_t, \cdots)$. Given the episodes in (1), fill in the episode at which the following Q values first become non-zero. If the specified Q value never becomes non-zero, write *never*.

Q((1,3), S) = ________ Q((2,2), E) = ________ Q((3,2), E) = ________

3. What is the value of the optimal value function V^* at the following states: (Unrelated to answers from previous parts)

$V^*(1,3) =$ ________________ $V^*(2,2) =$ ________________ $V^*(3,2) =$ ________________

4. Using Q-Learning updates, what are the following Q-values after the above five episodes:

   Q((3,2),N) = __________________ Q((1,2),S) =__________________ $Q((2,2),E)$ =__________________

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

   $Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$

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

   $f_3(N) = 1,\ f_3(W) = 2,\ f_3(S) = 3,\ f_3(E) = 4,\ f_3(Exit) = 5$

   (a) Given that all w_i are initially 0, what are their values after the first episode:

   w_1 = __________________ w_2 =__________________ w_3 =__________________

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

   ____

Q3. Simple HMM

Consider a Markov Model with a binary state X (i.e., X_t is either 0 or 1). The transition probabilities are given as follows:

X_t	X_{t+1}	$P(X_{t+1} \mid X_t)$
0	0	0.7
0	1	0.3
1	0	0.6
1	1	0.4

1. The prior belief distribution over the initial state X_0 is uniform, i.e., $P(X_0 = 0) = P(X_0 = 1) = 0.5$. After one timestep, what is the new belief distribution, $P(X_1)$?

X_1	$P(X_1)$
0	
1	

2. Now, we incorporate sensor readings. For all parts, answer using a decimal rounded to three places.

   (a) The sensor model is parameterized as is shown in the table below. Notice the sensor is better at detecting $X_1 = 0$ than $X_1 = 1$:

X_t	E_t	$P(E_t \mid X_t)$
0	0	0.9
0	1	0.1
1	0	0.3
1	1	0.7

   At $t = 1$, we get the first sensor reading, $E_1 = 0$. Use your answer from part (1) to compute $P(X_1 = 1 \mid E_1 = 0)$:

   (b) The sensor model is parameterized by a number $\beta \in [0, 1]$:

X_t	E_t	$P(E_t \mid X_t)$
0	0	$(1 - \beta)$
0	1	β
1	0	$(1 - \beta)$
1	1	β

At $t = 1$, we get the first sensor reading, $E_1 = 0$. Please simplify your answer when possible. Use your answer from part (1) to compute $P(X_1 = 0 \mid E_1 = 0)$:

(c) For what range of values, if any, of β will a sensor reading $E_1 = 0$ increase our belief that $X_1 = 0$? That is, what is the range of β for which $P(X_1 = 0 \mid E_1 = 0) > P(X_1 = 0)$? (Use β as defined in part (b))

3. Unfortunately, the sensor breaks after just one reading, and we receive no further sensor information. Compute $P(X_\infty \mid E_1 = 0)$, the stationary distribution *very many* timesteps from now.

X_∞	$P(X_\infty \mid E_1 = 0)$
0	
1	

4. How would your answer to part (3) change if we never received the sensor reading E_1, i.e. what is $P(X_\infty)$ given no sensor information?

X_∞	$P(X_\infty)$
0	
1	

Q4. Hours Worked

(a) How many hours did you spend on this homework? Any reasonable answer (number greater than zero) will receive credit. This will not affect your score on any other problem.