CSE 571 Homework 3

Name: Student ID:

Name: Student ID:

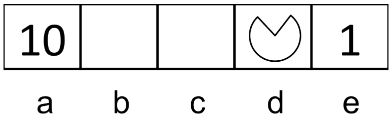
Due: October 13., 2016

For each homework we will state here if you have to work alone or if you can team up with another student.
For this homework you are allowed to work in **teams of two students** for all questions. Each group has to submit one **handwritten** (!) copy and state their names on the front page. Use the same groups as in Blackboard. Failure to do so could result in the allegation of plagiarism! Submission is possible at the end of a lecture or during an office hour till the due date of this homework (which you can find above).
Print this homework and write all your answers in the space below the questions. If you need additional space you might want to use the backside of the pages.
Also,

- unstapled homework will result in a decrease of at least 30% of the achieved points.
- handwritten text which is not readable will be graded with zero points.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7*	Q8*	Sum
Points	5	5	10	10	5	5	5	5	40 + 10 Bonus
Achieved									

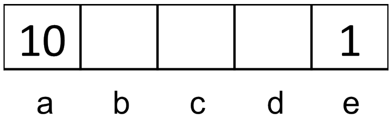
1. Question (5 Points) Value Iteration



Consider the gridworld MDP above for which *Left* and *Right* actions are 100% successful. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. Let the discount factor be $\gamma = 1$. Fill in the following quantities.

	$V_0(d)$	$V_1(d)$	$V_2(d)$	$V_3(d)$	$V_4(d)$	$V_5(d)$
Value						

2. Question (5 Points) Value Iteration



Consider the gridworld where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. Let the discount factor be $\gamma = 0.2$. Fill in the following quantities with $V^*(\cdot) = V_\infty(\cdot)$

	$V_\infty(a)$	$V_\infty(b)$	$V_\infty(c)$	$V_\infty(d)$	$V_\infty(e)$
Value					

3. Question (10 Points) Policy Search/Execution

10				1
a	b	c	d	e

Consider the gridworld above where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. The discount factor is given by $\gamma = 1$.

3.1) Consider the policy π_1 shown below, and evaluate the following quantities for this policy.

→	→	→	→	exit
a	b	c	d	e

	$V^{\pi_1}(a)$	$V^{\pi_1}(b)$	$V^{\pi_1}(c)$	$V^{\pi_1}(d)$	$V^{\pi_1}(e)$
Value					

3.2) Consider the policy π_2 shown below, and evaluate the following quantities for this policy.

exit	←	←	→	exit
a	b	c	d	e

	$V^{\pi_2}(a)$	$V^{\pi_2}(b)$	$V^{\pi_2}(c)$	$V^{\pi_2}(d)$	$V^{\pi_2}(e)$
Value					

4. Question (10 Points) Policy Search

10				1
a	b	c	d	e

Consider the gridworld above where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. Let the discount facto be given by $\gamma = 0.9$. Execute in the following tasks one round of policy iteration.

4.1) **(Policy Evaluation)** Consider the policy π_i shown below, and evaluate the following quantities for this policy.

exit	←	→	←	exit
a	b	c	d	e

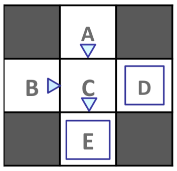
	$V^{\pi_i}(a)$	$V^{\pi_i}(b)$	$V^{\pi_i}(c)$	$V^{\pi_i}(d)$	$V^{\pi_i}(e)$
Value					

4.2) **(Policy Improvement)** Perform a policy improvement step. The current policy's values are the ones from Question **4.1**) (so make sure you first correctly answer Question **4.1**) before moving on). Make a cross for the correct answer:

$\pi_{i+1}(a) =$ ○ Exit ○ Right
$\pi_{i+1}(b) =$ ○ Left ○ Right
$\pi_{i+1}(c) =$ ○ Left ○ Right
$\pi_{i+1}(d) =$ ○ Left ○ Right
$\pi_{i+1}(e) =$ ○ Left ○ Exit

5. Question (5 Points) Model-Based Reinforcement Learning

Input Policy π



Assume: $\gamma = 1$

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

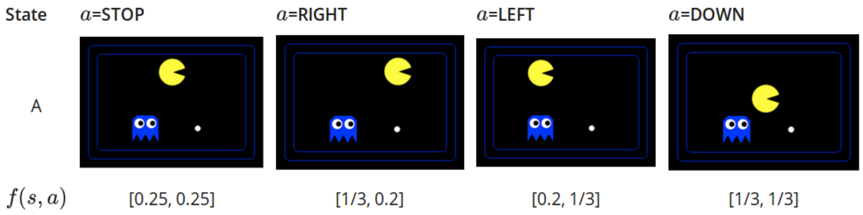
5.1) What model would be learned from the above observed episodes?

	T(A, south,C)	T(B, east, C)	T(C, south, E)	T(C, south, D)
Probability				

5.2) What are the estimates for the following quantities as obtained by direct evaluation:

	$\hat{V}^{\pi}(A)$	$\hat{V}^{\pi}(B)$	$\hat{V}^{\pi}(C)$	$\hat{V}^{\pi}(D)$	$\hat{V}^{\pi}(E)$
Value					

6. Question (5 Points, Bonus) Feature-Based Representation



Consider the two Pacman board states presented in two rows above. In each row, the agent considers possible actions to take; these are represented by the images. The agent is using feature-based representation to estimate the

$Q(s, a)$ value of taking an action in a state, and the features the agent uses are:

$$f_0 = \frac{1}{\text{Manhattan distance to closest food} + 1} \tag{1}$$

$$f_1 = \frac{1}{\text{Manhattan distance to closest ghost} + 1} \tag{2}$$

A possible feature representation is $f(s = A, a = \text{STOP}) = [0.25, 0.25]$, for example. The agent picks the action according to $\arg\max_a Q(s, a) = \mathbf{w}^T f(s, a) = w_0 f_0(s, a) + w_1 f_1(s, a)$, where the features $f_i(s, a)$ are defined as above and $\mathbf{w}$ is a weight vector.

6.1) Using the weight vector $\mathbf{w} = [0.2, 0.5]$, which action, of the ones shown above, would the agent take from state A?

○ STOP
○ RIGHT
○ LEFT
○ DOWN

6.2) Using the weight vector $\mathbf{w} = [0.2, -1]$, which action, of the ones shown above, would the agent take from state A?

○ STOP
○ RIGHT
○ LEFT
○ DOWN

7. Question (5 Points, Bonus) Exploration and Exploitation

For each of the following action-selection methods, indicate which option describes it best.

7.1) With probability p, select $\arg\max_a Q(s,a)$. With probability $1-p$, select a random action. It holds $p = 0.99$.

○ Mostly exploration
○ Mostly exploitation
○ Mix of both

7.2) Select action a with probability

$$p(a|s) = \frac{e^{Q(s,a)}/\tau}{\sum_{a\prime} e^{Q(s,a\prime)}/\tau}, \tag{3}$$

where τ is a temperature parameter that is decreasing over time.

○ Mostly exploration
○ Mostly exploitation
○ Mix of both

7.3) Always select a random action.

○ Mostly exploration
○ Mostly exploitation
○ Mix of both

7.4) Keep track of a count $K_{s,a}$ for each state-action tuple (s,a) of the number of times that tuple has been seen and select $\arg\max_a [Q(s,a) - K_{s,a}]$.

○ Mostly exploration
○ Mostly exploitation
○ Mix of both

7.5) Which method(s) would be advisable to use when doing Q-Learning?

○ **7.1**)
○ **7.2**)
○ **7.3**)
○ **7.4**)

8. Question (5 Points) Bellman Equation

On the slides the Bellman Equation was given by

$$V(s) = \max_a \left[\sum_{s\prime} p(s\prime|s,a) \left(R(s) + \gamma V(s\prime) \right) \right], \tag{4}$$

where $p(s\prime|s,a)$ defines the probability to end up in state $s\prime$ after being in state s and executing action a. Another form of the Bellman equation, used in the book *Artificial Intelligence - A Modern Approach*, is

$$V(s) = R(s) + \gamma \max_a \sum_{s\prime} p(s\prime|s,a) V(s\prime). \tag{5}$$

Show that both formulas are equivalent to each other by transforming Eq. (3) step by step to Eq. (4). Comment each step in order to receive full points: