COMP4200 / COMP 5430 AI: Homework VI
REINFORCEMENT LEARNING

UMASS - LOWELL

Name: ______________________________

Student ID: ______________________________

Question	Points
1	20
2	13
3	15
4	20
Total	68

Instructions:

1. This examination contains 9 pages, including this page.
2. Write your answers in this booklet. If you must write on the back page, please indicate **very** clearly on the front of the page that you have written on the back of the page.
3. You **may** use any resources, including lecture notes, books, other students or other engineers, but you should provide a reference.
4. You may use a calculator. You may not share a calculator with anyone.

Question 1: RL: Q-Learning

[20 pts] Consider a system with two states and two actions.

(a) (16 points)Perform Q-learning for a system with two states and two actions, given the following training examples. The discount factor is $\gamma = 0.5$ and the learning rate is $\alpha = 0.5$. Assume that your Q-table is initialized to 0.0 for all values.

$a_1: S_1 \to S_2$

Start = S_1, *Action* = a_1, $R = 10$, *End* = S_2

$Q(a_1, S_1)$

	S_1	S_2
a_1	5.0	0
a_2	0	0

$0 + 0.5[10 + 0.5 \max[0,0] - 0]$

$= 5$

$a_2: S_2 \to S_1$

Start = S_2, *Action* = a_2, $R = -10$, *End* = S_1

$Q(a_2, S_2)$

	S_1	S_2
a_1	5.0	0.0
a_2	0.0	-3.75

$0 + 0.5[-10 + 0.5 \max[5.0, 0] - 0]$

$= -3.75$

$a_2: S_1 \to S_1$

Start = S_1, *Action* = a_2, $R = 10$, *End* = S_1

$Q(a_2, S_1)$

	S_1	S_2
a_1	5.0	0.0
a_2	6.25	-3.75

$0 + 0.5[10 + 0.5 \max[5, 0] - 0]$

$= 6.25$

$a_1: S_1 \to S_1$

Start = S_1, *Action* = a_1, $R = 10$, *End* = S_1

	S_1	S_2
a_1	9.0625	0
a_2	6.25	-3.75

$5 + 0.5[10 + 0.5 \max[5, 6.25] - 5]$

$= 9.0625$

(b) (4 points) What is the policy that Q-learning has learned?

$\pi(1) = a_1$, $\pi(2) = a_1$

Question 2: MDPs and RL: Wandering Merchant

[13 pts] There are N cities along a major highway numbered 1 through N. You are a merchant from city 1 (that's where you start). Each day, you can either travel to a neighboring city (actions *East* or *West*) or stay and do business in the current city (action *Stay*). If you choose to travel from city i, you successfully reach the next city with probability p_i, but there is probability $1 - p_i$ that you hit a storm, in which case you waste the day and do not go anywhere. If you stay to do business in city i, you get $r_i > 0$ in reward; a travel day has reward 0 regardless of whether or not you succeed in changing cities. The diagram below shows the actions and transitions from city i. Solid arrows are actions; dashed arrows are resulting transitions labeled with their probability and reward, in that order.

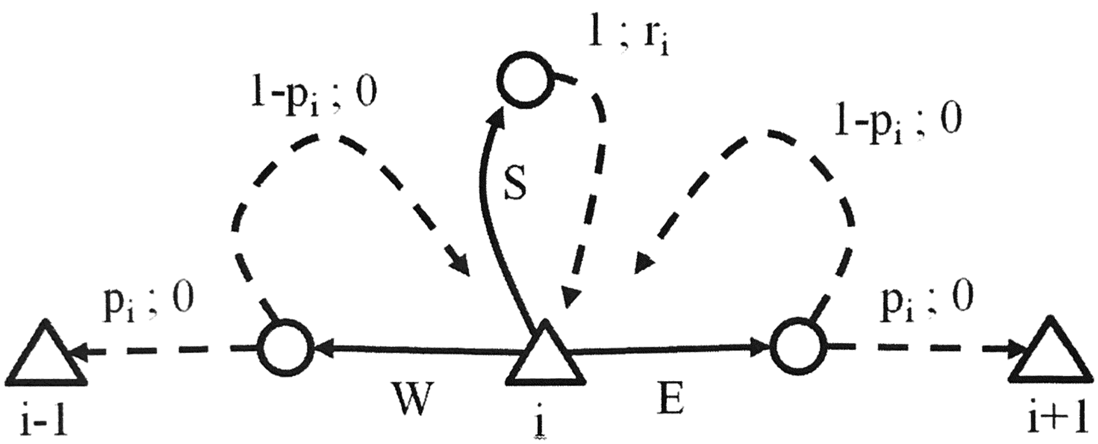

Figure 1: RL: Wandering Problem

(a) (2 points)If for $\forall i$, $r_i = 1$, $p_i = 1$, and there is a discount $\gamma = 0{:}5$, what is the value $V^{stay}(1)$ of being in city 1 under the policy that always chooses stay? Your answer should be a real number.

for all city (states) $i = 1, \dots, N$

$$V^{stay}(i) = r_i + \gamma V^{stay}(i)$$

$$V^{stay}(i) = 1 + 0.5\, V^{stay}(i)$$

$$\Rightarrow V^{stay}(i) = 2$$

$$\Rightarrow V^{stay}(1) = 2$$

(b) (2 points)If for $\forall i$, $r_i = 1$, $p_i = 1$, and there is a discount $\gamma = 0{:}5$, what is the optimal value $V^*(1)$ of being in city 1?

Since all city offer same reward ($r_i = 1$), there is no incentive to move to another city to do business, so the optimal policy is always stay $\Rightarrow V^*(1) = 2$

(c) (2 points) If the r_i's and p_i's are known positive numbers and there is almost no discount, i.e. $\gamma = 1$, describe the optimal policy. You may define it formally or in words, e.g. "always go east", but your answer should precisely define how an agent should act in any given state. Hint: You should not need to do any computation to answer this question.

The optimal policy is alway move towards the city with highest reward. Once there, stay there and do business forever.

(d) (2 points) If the optimal value of being in city 1 is positive, i.e. $V^* > 0$, what is the largest k for which $V_k(1)$ could still be zero? Be careful of off-by-one errors.

Assuming $r_i > 0$, then the largest k is 0.
because $V_1(s) = \max\{r_1 + 0, \dots\} > 0$

(e) (2 points) If all of the r_i and p_i are positive, what is the largest k for which $V_k(s)$ could still be zero for some state s? Be careful of off-by-one errors.

Since $r_i > 0$, the largest k is 0, because

$V_1(s) = \max\{ r_1 + 0, \dots \} > 0$

(f) (3 points)Suppose we experience the following sequence of states, actions, and rewards: (s=1, a=stay, r=4), (s=1, a=east, r=0), (s=2, a=stay, r=6), (s=2, a=west, r=0), (s=1, a=stay, r=4, s=1). What are the resulting Q(s, a) values if the learning rate is 0.5, the discount is 1, and we start with all Q(s, a) = 0? Fill in the table below; each row should hold the q-values after the transition specified in its first column. You may leave unchanged values blank.

$< s, a, r, s' >$	Q(1,S)	Q(1, E)	Q(2,W)	Q(2,S)
Initial	0	0	0	0
$< 1, S, 4, 1 >$	2			
$< 1, E, 0, 2 >$		0		
$< 2, S, 6, 2 >$				3
$< 2, W, 0, 2 >$			1	
$< 1, S, 4, 1 >$	4			

$(1,S,4,1) \Rightarrow Q(1,S) \leftarrow 0.5[4 + 1 \times 0] + 0.5(0) = 2$

$(1,E,0,2) \rightarrow Q(1,E) \leftarrow 0.5[0 + 1 \times 0] + 0.5(0) = 0$

$(2,S,6,2) \Rightarrow Q(2,S) \leftarrow 0.5[6 + 1 \times 0] + 0.5(0) = 3$

$(2,W,0,1) \Rightarrow Q(2,W) \leftarrow 0.5[0 + 1 \times 2] + 0.5(0) = 1$

$(1,S,4,1) \rightarrow Q(1,S) \leftarrow 0.5[4 + 1 \times 2] + 0.5(2) = 4$

Question 3: MDPs and RL: Flippers Folly

[15 pts] In Flipper's Folly, a player tries to predict the total number of heads in two coin flips. The game proceeds as follows (also show by Figure)

(a) From the state state (XX, choose the special action *begin* (only possible action)

(b) Flip a coin and observe the result, arriving i the state HX or TX

(c) Guess what the total number of heads will be : $a \in 0, 1, 2$

(d) Flip a coin and observe the result, arriving in one of the states HH, HT, TH, TT

(e) Count the total numbers of heads in the two flips: $c \in 0, 1, 2$

(f) Receive reward

$$R(s, a, s') = \begin{cases} 2.a^2 - c^2 & ifc \geq a \\ -3 & ifc < a \end{cases}$$

where c is the total number of heads in s'

Note that the rewards depend only on the action and the landing state, and that all rewards for leaving the start state are zero. The MDP for this game has the following structure, where all legal transitions have a 0.5 probability. Assuyme a **discount** rate of 1.

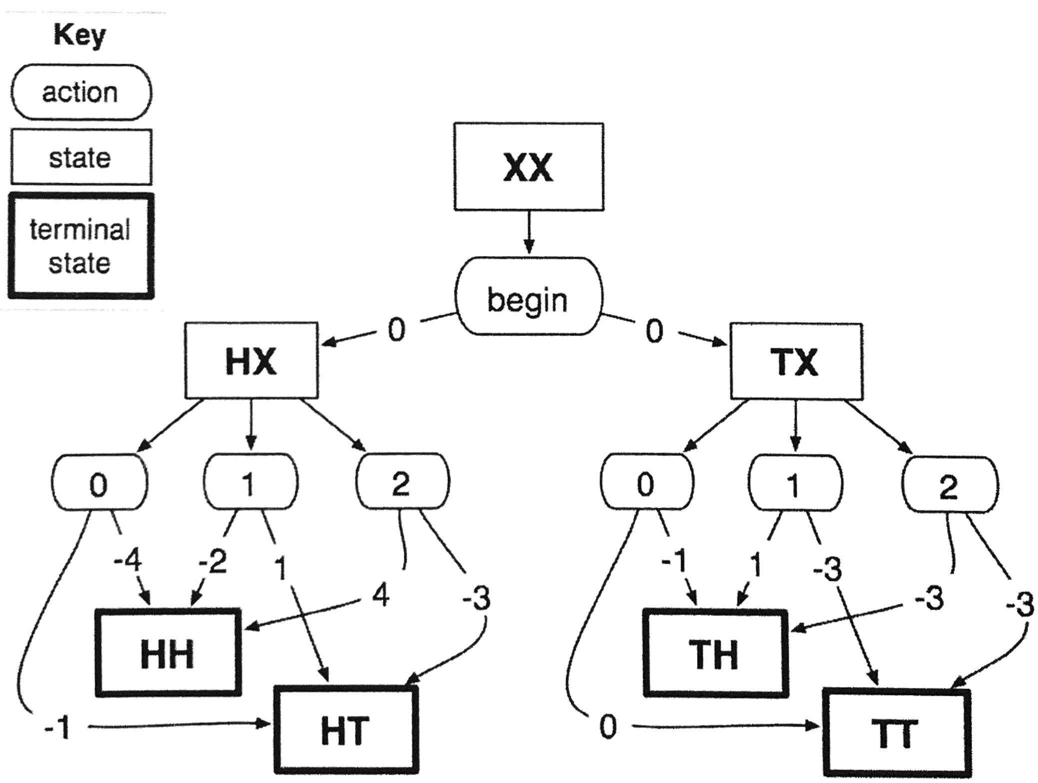

Figure 2: RL: Flippers Folly

(a) (3 points) What is the value of the start state under the policy of always guessing $a = 2$?

$$\frac{1}{2}\left[\frac{1}{2}(4-3)+\frac{1}{2}(-3-3)\right] = -\frac{5}{4}$$

(b) (5 points) Run value iteration on this MDP until convergence. *Hint:* values and the q-values of terminal state are always zero.

	V		
k	XX	HX	TX
0	0	0	0
1	0	0.5	−0.5
2			
3			
4			
5			

value iteration converges after 1 iteration

(c) (2 points) What is the optimal policy for this MDP?

$\pi^*(XX) = \text{begin}$

$\pi^*(HX) = 2$

$\pi^*(TX) = 0$

(d) (5 points) Run q-learning in this MDP with the following $< s, a, r, s' >$ observations. Use a learning rate of 0.5. Leave zero entries blank.

Observations				Q(s,a)						
s	a	r	s'	(XX, begin)	(HX, 0)	(HX, 1)	(HX, 2)	(TX, 0)	(TX, 1)	(TX, 2)
				0	0	0	0	0	0	0
XX	begin	0	HX							
HX	0	-1	HT		-0.5					
XX	begin	0	HX		-0.5					
HX	2	4	HH		-0.5		2			
XX	begin	0	HX	1	-0.5		2			

Question 4: RL: Q-Learning

[20 pts] Consider a system with two states and two actions. You perform actions and observe the rewards and transitions listed below. Each step lists the current state, reward, action and resulting transition as S_i:R = r; a_k : $S_i \to S_j$.

(a) (16 points) Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-table entries are initialized to zero.

S_1;R = -10; a_1 : $S_1 \to S_1$

Q	S_1	S_2
a_1	-5	0
a_2	0	0

$Q(a,s) \leftarrow Q(a,s) + \alpha\,(R(s) + \gamma \max[Q(a',s')] - Q(a,s))$

$= 0 + 0.5\,(-10 + 0.5 \cdot \max[0,0] - 0)$

$= -5$

S_1;R = -10; a_2 : $S_1 \to S_2$

	S_1	S_2
a_1	-5	0
a_2	-5	0

$Q(a_2,S_1) \leftarrow 0 + 0.5\,(-10 + 0.5 \max[0,0] - 0)$

$= -5$

S_2;R = +20; a_1 : $S_2 \to S_1$

	S_1	S_2
a_1	-5	8.75
a_2	-5	0

$Q(a_1,S_2) = 0 + 0.5\,(-10 +$

$0.5 \max[-5,-5] - 0)$

$= 8.75$

S_1;R = -10; a_2 : $S_1 \to S_2$

	S_1	S_2
a_1	-5	8.75
a_2	-5.312	0

$Q(a_2,S_1)$

$= -5 + 0.5\,[-10 +$

$0.5 \max[8.75, 0] - (-5)]$

$= -5.3125$

(b) (4 points) What is the policy that Q-learning has learned?

$\pi(S_1) = a_1$

$\pi(S_2) = a_1$