ECE4010 Homework 5

Q1. [15 pts] The following is the dataset that we have.

	x_1	x_2	y
1	4	1	2
2	2	8	-14
3	1	0	1
4	3	2	-1

In this dataset, we have 4 observations, where x_1 and x_2 are independent variables and y is the response variable. To answer the question "how are both x_1 and x_2 related to y", we have figured out our model to be

$$\hat{y} = \boldsymbol{\omega} \cdot \boldsymbol{x}$$

where $\boldsymbol{\omega} \in \mathbb{R}^3$. Let learning rate $\alpha = 0.05$ and the current estimate is $\boldsymbol{\omega} = (0, -0.017, -0.048)$.

What is the next estimate of $\boldsymbol{\omega} \in \mathbb{R}^3$ if we use stochastic gradient descent? You need to show how you get this value.

Q2. [45 pts] In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your utility is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you Draw, you receive no utility. There is no discount ($\gamma = 1$).

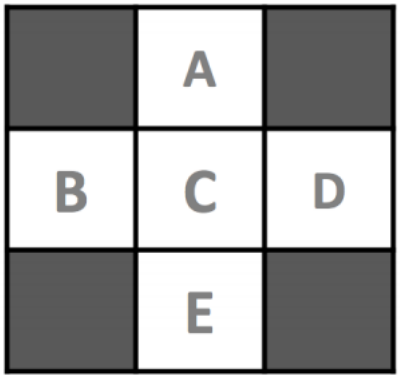
(a) [15 pts] What are the states and the actions for this MDP?

(b) [15 pts] What is the transition function and the reward function for this MDP?

(c) [15 pts] Use value iteration method to find the optimal policy for this MDP (you need to list the value iteration steps based on the states and transition functions. The example list is shown below. In this example list, there are n states and we need 2 rounds to make value iteration converge. You should explicitly write down each state's name (or value in this question). You may also need to add/delete rows based on the number of iterations you calculate.)

$U(s)$	s_1	s_2	…	s_n
$U_0(s)$				
$U_1(s)$				
$U_2(s)$				

Q3. [40 pts] Consider the Gridworld example that we looked at in lecture. We would like to use TD learning to find the values of these states.



Suppose we observe the following $(\mathrm{s}, \mathrm{a}, s', R(\mathrm{s}, \mathrm{a}, s'))$ transitions and rewards*:

(B, East, C, 2), (C, South, E, 4), (C, East, A, 6), (B, East, C, 2)

*Note that the $\mathrm{R}(\mathrm{s}, \mathrm{a}, s')$ in this notation refers to observed reward, not a reward value computed from a reward function.

The initial value of each state is 0. Let $\gamma = 1$ and $\alpha = 0.5$

(a) [20 pts] What are the learned values for each state from TD learning after all four observations?

(b) [20 pts] What are the learned Q-values from Q-learning after all four observations? Use the same $\alpha = 0.5$, $\gamma = 1$ as before.