**Homework 6: Reinforcement Learning**

**Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Question 1: Model-Based RL: Grid**

0.0/4.0 points



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) =



**Question 2: Model-Based RL: Cycle**

0.0/26.0 points

*We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.*

***This section will reset, so send me a screenshot of your correct answer.***Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given samples of what an agent experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, we will first estimate the model (the transition function and the reward function), and then use the estimated model to find the optimal actions.

To find the optimal actions, model-based RL proceeds by computing the optimal V or Q value function with respect to the estimated T and R. This could be done with any of value iteration, policy iteration, or Q-value iteration. Last week you already solved some exercises that involved value iteration and policy iteration, so we will go with Q value iteration in this exercise.

Consider the following samples that the agent encountered. (Note that nan stands for not-a-number and indicates that this entry cannot be estimated from the samples.)



|  |  |  |
| --- | --- | --- |
|  |  |  |

**Part 1**

We start by estimating the transition function, T(s,a,s') and reward function R(s,a,s') for this MDP. Fill in the missing values in the following table for T(s,a,s') and R(s,a,s').

|  |
| --- |
| Discount Factor, γ = 0.5 |
|

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| s  | a  | s'  | T(s,a,s')  | R(s,a,s')  |
| A  | Clockwise  | B  | M  | N  |
| A  | Clockwise  | C  | O  | P  |
| A  | Counterclockwise  | B  | 0.400  | 0.000  |
| A  | Counterclockwise  | C  | 0.600  | -1.000  |
| B  | Clockwise  | A  | 0.200  | 8.000  |
| B  | Clockwise  | C  | 0.800  | 0.000  |
| B  | Counterclockwise  | A  | 0.600  | 0.000  |
| B  | Counterclockwise  | C  | 0.400  | 0.000  |
| C  | Clockwise  | A  | 0.800  | 0.000  |
| C  | Clockwise  | B  | 0.200  | -9.000  |
| C  | Counterclockwise  | A  | 0.200  | 0.000  |
| C  | Counterclockwise  | B  | 0.800  | -5.000  |

 |

|  |  |
| --- | --- |
| M |   |
| N |   |
| O |   |
| P |   |

 |

**Part 2**

Now we will run Q-iteration using the estimated T and R functions. The values of Qk(s,a), are given in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A  | B  | C  |
| Clockwise  | -1.48  | 0.82  | -1.88  |
| Counterclockwise  | -0.82  | -0.54  | -3.42  |

Fill in the values for Qk+1(s,a).

|  |  |  |  |
| --- | --- | --- | --- |
|  | A  | B  | C  |
| Clockwise  |   |   |   |
| Counterclockwise  |   |   |   |

**Part 3**

Suppose Q-iteration converges to the following Q\* function, Q\*(s,a).

|  |  |  |  |
| --- | --- | --- | --- |
|  | A  | B  | C  |
| Clockwise  | -1.856  | 0.609  | -2.194  |
| Counterclockwise  | -1.136  | -0.78  | -3.87  |

What is the optimal action, either Clockwise or Counterclockwise, for each of the states?

|  |  |  |
| --- | --- | --- |
| A  | B  | C  |
|   |   |   |

**Question 3: Direct Evaluation**

0.0/10.0 points



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





















**Question 4: Temporal Difference Learning**

0.0/10.0 points

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function Vπ for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming γ =1, α = 1/2, what are the value estimates after the TD learning update? (note: the value will change for one of the states only)























**Question 5: Model-Free RL: Cycle**

0.0/12.0 points

*We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.*

***This section will reset, so send me a screenshot of your correct answer.***Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning.

Assume, the discount factor, γ is 0.5 and the step size for Q-learning, α is 0.5.

Our current Q function, Q(s,a), is as follows.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A  | B  | C  |
| Clockwise  | 0.372  | 0.836  | 2.097  |
| Counterclockwise  | 6.999  | -1.686  | 3.443  |

The agent encounters the following samples.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |
| --- | --- | --- | --- |
| s  | a  | s'  | r  |
| A  | Clockwise  | C  | -7.0  |
| C  | Counterclockwise  | B  | 4.0  |

 |

Process the samples given above. Below fill in the Q-values after both samples have been accounted for.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A  | B  | C  |
| Clockwise  |   |   |   |
| Counterclockwise  |   |   |   |

**Question 6: Q-Learning Properties**

0.0/4.0 points

In general, for Q-Learning to converge to the optimal Q-values...

It is necessary that every state-action pair is visited infinitely often.

It is necessary that the learning rate α (weight given to new samples) is decreased to 0 over time.

It is necessary that the discount γ is less than 0.5.

It is necessary that actions get chosen according to argmaxaQ(s,a).

**Question 7: Exploration and Exploitation**

0.0/10.0 points

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

A: With probability p, select argmaxaQ(s,a). With probability (1-p), select a random action. p – 0.99

Mostly exploration

Mostly exploitation

Mix of both

B: Select action a with probability



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

Mostly exploration

Mostly exploitation

Mix of both

C: Always select a random action.

Mostly exploration

Mostly exploitation

Mix of both

D: Keep track of a count, Ks,a, for each state-action tuple, (s,a), of the number of times that tuple has been seen and select argmaxa[Q(s,a) – Ks,a].

Mostly exploration

Mostly exploitation

Mix of both

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

A

B

C

D

### Question 8: Feature-Based Representation: Actions

0.0/6.0 points

Consider the two Pacman board states presented in two rows below. 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:

* f0 = 1/(Manhattan distance to closest food + 1)
* f1 = 1/(Manhattan distance to closest ghost + 1)

For example, the feature representation f(s = A, a = STOP) = [1/4, ¼].



The agent picks the action according to argmaxaQ(s,a) = wTf(s,a)=w0f0(s,a) + w1f1(s,a), where the features fi(s,a)are as defined above, and w is a weight vector. Using the weight vector w = [0.2, 0.5], which action, of the ones shown above, would the agent take from state A?

STOP

RIGHT

LEFT

DOWN

Using the weight vector w = [0.2, -1], which action, of the ones shown above, would the agent take from state A?

STOP

RIGHT

LEFT

DOWN

**Question 9: Feature-Based Representation: Update**

0.0/18.0 points

Consider the following feature based representation of the Q-function:



with:



**Part 1**

Assume w1 = 1, w2 = 10. For the state s shown below, find the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.











Based on this approximate Q-function, which action would be chosen:

West

South

**Part 2**

Assume Pac-Man moves West. This results in the state s’ shown below.



The reward for this transition is r = +10 -1 = 9 (+10: for food pellet eating, -1 for time passed). Fill in the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.









What is the sample value (assuming γ = 1)?




**Part 3**

Now let's compute the update to the weights. Let α = 0.5.











