CS 229 Machine Learning, spring 2020

Homework 5:

Reinforcement Learning

Due Saturday May 9, 11:59pm

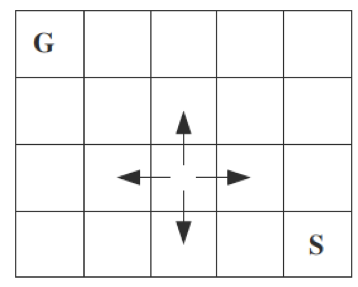
Submit by the **blackboard system**

**Question1: (3 pts) Difference between TD and MC**

Read the attached TD-MC chapter, and the example 6.1 Driving to home. Can you imagine a scenario in which a TD update would be better on average than an Monte Carlo update? Give an example scenario—a description of past experience and a current state—in which you would expect the TD update to be better.

**Question2: (7 pts) Off policy Q-learning and Sarsa algorithm**

Given the grid world in Figure 1, there are 4 deterministic actions: *up*, *down*, *left* and *right*. The goal is to reach the G, starting at S.



F

Figure 1 Grid world

The reward on reaching on the goal (G) is 10.

The reward on actions that would take the agent off the grid is -1 (agent stays still in this case).

The reward on entering F is -5

The reward on other actions is 0.

The discount factor γ = 0.9.

If you need, take a look at Sarsa-OffPolicy-Q-learning.pdf in the attached file.

Use off policy **Q learning** to learn the optimal values of Q\* (s, a). Please submit your own code on calculating the Q\*(s,a).

--1 (1pts) What is the Q\*(s,a) for each pair of s and a?

--2 (1pts) What is the V\*(s) for each s?

--3 (1pts) What are the actions of optimal policy?

Use **Sarsa** algorithm to learn the optimal values of Q\* (s, a). Please submit your own code on calculating the Q\*(s,a).

--1 (1pts) What is the Q\*(s,a) for each pair of s and a?

--2 (1pts) What is the V\*(s) for each s?

--3 (1pts) What are the actions of optimal policy?

Please discuss your observed difference between Sarsa algorithm and off-policy Q-learning (1pts) in your game.