CS 594: Modern Reinforcement Learning
Homework 2
Due Sunday of Week 4 11:59 PM

You may discuss the assignment with other students, but if you do you must note on your submission who you discussed it with. The actual submission must be entirely your own work. It must be submitted via gradescope. Please make sure to tag which page(s) each problem is answered on at the appropriate step in the submission process.

1. **Policy Gradients:** Do Exercise 13.3 on page 329, which asks you to derive the given form for $\nabla \ln \pi$ with the softmax parameterization.

2. **Occupancy Measures:** Consider the simple MDP illustrated in Exercise 3.22 on Page 66. Suppose we use the policy that randomizes equally likely between left and right, a discount factor of $\gamma = 0.5$, and that we start in the top state. What is the discounted occupancy measure if we terminate the episode at time 1 (i.e. including the start state and the state after the first action)? At time 3? At time 9? What is the discounted occupancy measure if we let the process continue forever?

3. **Baselines and Variance:** Consider the same MDP and policy as in the previous problem, but now with $\gamma = 1$ and episodes that last for two actions. Suppose you start in the top state. What is variance of a Monte Carlo estimate with no baseline. What is the variance if $v_\pi(s)$ is used as a baseline? Now suppose you are equally likely to start in any of the three states. What is the variance now in each case?

4. **MARL:** Give an example of a situation where Independent Q-learning will fail to converge even if the Robbins Munro conditions are satisfied. A high-level description is fine; you do not need to fully work it out.

5. **Deadly Triad:** For each pair of features of the deadly triad, give an example of an algorithm which has those two features but not the third. (You should have three examples in total.)