CS 594: Modern Reinforcement Learning
Homework 1
Due Sunday of Week 2 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 iteration:** Do Exercise 4.5 on page 82 of the text. This asks you to adapt the pseudocode of the policy iteration algorithm to learn q-values rather than values. While doing so, make sure to fix the bug pointed out in Exercise 4.4.

2. **Monte Carlo estimates:** Do Exercise 5.5 on page 105 of the text, which asks you to compare the Monte Carlo estimate you get from just using your first visit to averaging across all visits on the given example.

3. **Importance sampling:** Consider the MDP, behavior, and target policies from Example 5.5 on page 106. What is the (first-visit ordinary) importance sampling estimate of the value of s on a trajectory where left is played 10 times in a row and then the terminal is reached?

4. **Sarsa:** In class we discussed how Q-Learning can be viewed as a special case of Expected Sarsa. Is Sarsa a special case of Expected Sarsa? Explain.

5. **Convergence:** In class we discussed how there is flexibility in setting the α_t parameters. One choice sometimes used in practice is to set α_t at some initially high value α and then decay it over time until it eventually reaches a lower value α'. Does this satisfy the Robbins-Monro conditions for convergence? Explain.