SYS 6582: Reinforcement Learning **Spring 2018**

Homework 1

Handed Out: Feb 5, 2018 *Due: Feb* 12, 2018

Instructions: Solve each of the following exercises. Give quantitative answers where required, and always explain your reasoning. The textbook exercise numbers refer to the second of version of Sutton & Button (Complete Draft). Finally, pledge your solutions.

1. (Exercise 3.8, S & B) Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodesthe successive runs through the mazeso you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

2. (Exercise 3.12, S & B) The Bellman equation (3.14) must hold for each state for the value function v_pi shown in Figure 3.3 (right) of Example 3.6. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, 0.4, and +0.7. (These numbers are accurate only to one decimal place.)

3. (Exercise 3.13, S & B) What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the stateaction pair (s, a). Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.

4. (Exercise 3.14, S & B) In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

5. (Exercise 3.15, S &B) Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

6. (Exercise 3.20, S & B) Give the Bellman Optimality equations for q_* for the recycling robot.

7 Write a function in Python that can generate an episode for the student MRP example. Each episode should include states and rewards. Use your function to generate three episodes, and append the results as comment following your code. Submit your code on Collab.

8 Write a function in Python that can generate an episode for the student MDP example under a given policy. Each episode should include states, actions, and rewards. Then use your function to generate three episodes under the random policy. Paste results as comment in your code. Submit your code on Collab.
Note: for simplicity, we will assume that there are only two possible actions: *Study* and *Relax* (i.e., Facebook, Pub, or Sleep).

9 Solve the oil well leasing problem. (You can write a program in python, create an excel spreadsheet, or solve it by hand.)