CMPSCI 687 Homework 1

Due September 20, 2018, 11pm Eastern Time

Instructions: This homework assignment consists of a written portion and a programming portion. Collaboration is not allowed on any part of this assignment. Submissions must be typed (hand written and scanned submissions will not be accepted). You must use LaTeX. The assignment should be submitted on Moodle as a .zip (.gz, .tar.gz, etc.) file containing your answers in a .pdf file and a folder with your source code. Include with your source code instructions for how to run your code. You may not use any reinforcement learning or machine learning specific libraries in your code (you may use libraries like C++ Eigen and numpy though). If you are unsure whether you can use a library, ask on Piazza. If you submit by September 25, you will not lose any credit. The automated system will not accept assignments after 11:55pm on September 25.

Part One: Written (65 Points Total)

1. (Your grade will be a zero on this assignment if this question is not answered correctly) Read the class syllabus carefully, including the academic honesty policy. To affirm that you have read the syllabus, type your name as the answer to this problem.

2. (15 Points) Given an MDP $M = (\mathcal{S}, \mathcal{A}, P, d_R, d_0, \gamma)$ and a fixed policy, π, the probability that the action at time $t = 0$ is $a \in \mathcal{A}$ is:

$$\Pr(A_0 = a) = \sum_{s \in \mathcal{S}} d_0(s)\pi(s, a). \tag{1}$$

   Write similar expressions (using only $\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma$, and π) for the following:

   - The probability that the state at time $t = 3$ is either $s \in \mathcal{S}$ or $s' \in \mathcal{S}$.
   - The probability that the action at time $t = 16$ is $a' \in \mathcal{A}$ given that the action at time $t = 15$ is $a \in \mathcal{A}$ and the state at time $t = 14$ is $s \in \mathcal{S}$.
   - The expected reward at time $t = 6$ given that the action at time $t = 3$ is $a \in \mathcal{A}$, and the state at time $t = 5$ is $s \in \mathcal{S}$.
   - The probability that the initial state was $s \in \mathcal{S}$ given that the state at time $t = 1$ is $s' \in \mathcal{S}$.
   - The probability that the action at time $t = 5$ is $a \in \mathcal{A}$ given that the initial state is $s \in \mathcal{S}$, the state at time $t = 5$ is $s' \in \mathcal{S}$, and the action at time $t = 6$ is $a' \in \mathcal{A}$.

3. (3 Points) In 687-Gridworld, if we changed how rewards are generated so that hitting a wall (i.e., when the agent would enter an obstacle state, and is placed back where it started) results in a reward of -1, then what is $\mathbf{E}[R_t | S_t = 17, A_t = \text{AL}, S_{t+1} = 17]$?

4. (2 Points) How many stochastic policies are there for an MDP with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$? (You may write your answer in terms of $|\mathcal{S}|$ and $|\mathcal{A}|$).

5. (5 Points) Create an MDP (which may not have finite state or action sets) that does *not* have an optimal policy. The rewards for your MDP must be bounded.

6. (3 Points) Read about the Pendulum domain, described in Section 5.1 of this paper (Reinforcement Learning in Continuous Time and Space by Kenji Doya). Consider a variant where the initial state has the pendulum hanging down with zero angular velocity always (a deterministic initial state where the pendulum is hanging straight down with no velocity) and a variant where the initial angle is chosen uniformly randomly in $[-\pi, \pi]$ and the initial velocity is zero. Which variant do you expect an agent to require more episodes to solve? Why?

7. (1 Point) How many episodes do you expect an agent should need in order to find near-optimal policies for the gridworld and pendulum domains?

8. (5 Points) Select a problem that we have not talked about in class, where the agent does not fully observe the state. Describe how this problem can be formulated as an MDP by specifying $(\mathcal{S}, \mathcal{A}, P, [d_r \text{ or } R], d_0, \gamma)$ (your specifications of these terms may use English rather than math, but be precise).

9. (5 Points) Create an MDP for which there exist at least two optimal policies that have different variance of their returns. Describe the two optimal policies and derive the expected value and variance of their returns.

10. (2 Points) Create an MDP that always terminates, but which has no terminal states.

11. (2 Points) The sequence of states that results from running a fixed policy is a *Markov chain*. A state in a *Markov chain* has *period* k if every return to the state must occur in multiples of k time steps. More formally,

$$k = \gcd\{t > 0 : \Pr(S_t = s | S_0 = s) > 0\}.$$

Create an MDP and a policy that result in a state having a period of 3.

12. (5 Points) A Markov chain is *irreducible* if it is possible to get to any state from any state. An MDP is *irreducible* if the Markov chain associated with every deterministic policy is irredicible. A Markov chain is *aperiodic* if the period of every state is $k = 1$. The state of a Markov chain is *positive recurrent* if the expected time until the state recurs is finite. A Markov chain is *positive recurrent* if all states are positive recurrent. A Markov chain is *ergodic* if it is *aperiodic* and *positive recurrent*. An MDP is *ergodic* if the Markov chain associated with every deterministic policy is ergodic. Create an MDP that is ergodic, but *not* irreducible.

13. (3 Points) Create an MDP that is not ergodic.

14. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where R_t is *not* a deterministic function of S_t, A_t, and S_{t+1}.

15. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where R_t is a deterministic function of S_t.

16. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where R_t is a deterministic function of S_{t+1}.

17. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where the reward function, R, would be known.

18. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where the reward function, R, would *not* be known.

19. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where the transition function, P, would be known.

20. (2 Points) Describe a real-world problem and how it can be reasonably modeled as an MDP where the transition function, P, would *not* be known.

Part Two: Programming (25 Points Total)

Implement the 687-Gridworld domain described in class and in the class notes. Have the agent select actions uniformly randomly.

- (5 Points) Have the agent uniformly randomly select actions. Run 10,000 episodes. Report the mean, standard deviation, maximum, and minimum of the observed discounted returns.

- (5 Points) Find an optimal policy (you may do this any way you choose, including by reasoning through the problem yourself). Report the optimal policy here. Comment on whether it is unique.

- (10 Points) Run the optimal policy that you found in the previous question for 10,000 episodes. Report the mean, standard deviation, maximum, and minimum of the observed discounted returns.

- (5 Points) Using simulations, empirically estimate the probability that $S_{19} = 21$ (the state with water) given that $S_8 = 18$ (the state above the goal) when running the uniform random policy. Describe how you estimated this quantity (there is *not* a typo in this problem, nor an oversight).