CMPSCI 687 Homework 1

Due September 26, 2017, 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). We recommend that you use LaTeX. The assignment should be submitted as a single .pdf on Moodle. The automated system will not accept assignments after 11:55pm on September 26.

Part One: Written (30 Points Total)

1. (15 Points) Given an MDP $M = (\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$ and a fixed policy, π, the probability that the action at time $t = 0$ is $a \in \mathcal{A}$ is:

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

   Write similar expressions (using only the terms defined in M) 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.
   - 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 $a' \in \mathcal{A}$.

2. (2 Points) How many deterministic 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 $|\mathcal{A}|$).

3. (2 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 is $\theta = 0$ and $\dot{\theta} = 0$ 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?

4. (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?

5. (10 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 $(\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$ (your specifications of these terms may use English rather than math, but be precise).

Part Two: Programming (40 Points Total)

Download the C++ project from Moodle. This code is set up to run the Cross-Entropy Method (CEM) on the pendulum domain (both with deterministic initial state and with a stochastic initial state) and on the gridworld that we discussed in class. However, **1)** parts of the code for CEM are missing from the file `CrossEntropyMethod.cpp` (search that file for "@TODO:" to see where code is missing) and **2)** the hyperparameters for CEM have not been provided (they are entered manually when the program is executed). Fill in the missing code and find hyperparameters that make CEM work as well as possible on each domain (each domain will have its own set of hyperparameters, so you should be finding three sets of hyperparameters).

- (-10 points if missing) Print your code (just the parts that you filled in) and include it with your submission.
- (10 Points) What were the best parameters that you found for the gridworld? Include a plot of the results of using these parameters. The plot should have "episodes" on the horizontal axis and "expected return" on the vertical axis. It should include error bars showing standard error (not standard deviation), and should use at least 1000 trials (you may want to run fewer trials when testing out hyperparameters during your search for good hyperparameter values).
- (10 Points) The same as the previous question, but for the pendulum domain with deterministic starts. The plot need only show 100 trials.
- (10 Points) The same as the previous question, but for the pendulum domain with stochastic start states. The plot need only show 30 trials.
- (10) Write a brief statement describing your experience getting CEM working. Include responses to the following questions:
  - How long did it take (wall-time) to get working?
  - Roughly how many times would you estimate you ran an agent for a full lifetime (multiple episodes starting from an initial policy) total on each domain?
  - Do you think that, given the computational time spent adjusting hyperparameters, you could have solved each problem many times over with a brute force search (e.g., randomly trying policies)?

– Compare the number of episodes required by CEM to the numbers of episodes that you answered for questions 3 and 4 of the first part of this assignment. If they differ significantly, why do you think that this happened?

– Which domains were harder to find good hyperparameters for? Why do you think that is?

– Given your experience with CEM, what types of applications do you think BBO algorithms like CEM will work well for, and what types of applications will they not work well for?