CMPSCI 687 Homework 2

Due October 17, 2017, 11pm Eastern Time

Instructions: This homework assignment consists of only a written portion. You may discuss concepts related to the written portion with other students, but should not discuss how to solve the specific questions with other students. 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 the due date specified above.

Written Portion (65 Points Total)

1. (10 Points) Apply value iteration to the gridworld used in class (with stochastic state transitions and zero reward for hitting obstacles, as it was originally presented). Remember that R_t is -10 if S_{t+1} is the state with water, and R_t is $+10$ if S_{t+1} is the bottom-right state. Use $\gamma = 1.0$. Begin with the value of every state being zero. Draw the value function as a 5×5 grid with two cells missing (the obstacles), with numbers in each cell of the grid correspond to the current estimate of the value of that state. After computing v_{k+1}, round all values to three decimal places before continuing (your answer should include three decimal places, and future computations should use the rounded values). Show the first ten iterations of value iteration. Below is the initial value function:

v_0:

0	0	0	0	0
0	0	0	0	0
0	0	N/A	0	0
0	0	N/A	0	0
0	0	0	0	0

2. (10 Points) Prove that multiplying all rewards (of a finite MDP with bounded rewards) by a positive scalar does not change which policies are optimal, using either of the definitions of optimal policies that we covered in class (that is, show it for at least one of the definitions that we covered in class).

3. (5 Points) Prove that adding a positive constant to all rewards (of a finite MDP with bounded rewards) can change which policies are optimal, using either of the definitions of optimal policies that we covered in class.

4. (5 Points) Your boss asked you to estimate the state-value function associated with a known policy, π, for a specific MDP. You misheard and instead estimated the action-value function. This estimation was very expensive, and so you do not want to do it again. Explain how you could easily retrieve the value of any state given what you have already computed.

5. (10 Points) Consider a finite MDP with bounded rewards, where all rewards are negative. That is, $R_t < 0$ always. Let $\gamma = 1$. The MDP is finite horizon, with horizon L, and also has a deterministic transition function and initial state distribution (rewards may be stochastic). Let $H = (S_0, A_0, R_0, S_1, A_2, R_1, \ldots, S_{L-1}, A_{L-1}, R_{L-1})$ be any history that can be generated by a deterministic policy, π. Prove that the sequence $v^\pi(S_0), v^\pi(S_1), \ldots, v^\pi(S_{L-1})$ is strictly increasing.

6. (15 Points) The Bellman operator for q-functions is:

$$\mathcal{T} : \mathcal{Q} \to \mathcal{Q},$$

where $\mathcal{Q}$ is the set of all functions, $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and

$$Tq(s,a) \coloneqq \sum_{s'} P(s,a,s') \left(R(s,a,s') + \gamma \max_{a'} q(s',a') \right).$$

Prove that the Bellman operator for q-functions is a contraction mapping.

7. (10 Points) A researcher proposes an estimator, $\hat{J}$, of J. The estimator uses data to estimate the performance of a policy. That is, $\hat{J}(\pi, H)$ corresponds to the estimator's estimate of $J(\pi)$, where H is a history produced by running π for one episode. Specifically:

$$\hat{J}(\pi, H) = \sum_{t=0}^{\infty} \gamma^t \left(R_t - R(S_t, A_t, S_{t+1}) \right) + \sum_{t=0}^{\infty} \gamma^t \sum_{s'} P(S_t, A_t, s') R(S_t, A_t, s').$$

Now consider the case where we have a data set, D_n, that includes $n \in \mathbb{N}_{>0}$ i.i.d. histories, i.e., $D_n = (H_1, \ldots, H_n)$, each produced by running the policy π. We construct a new estimator, $\hat{J}_n(\pi, D_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{J}(\pi, H_i)$. Prove that $\hat{J}_n(\pi, D_n)$ converges in probability to $J(\pi)$. That is, for all ϵ,

$$\lim_{n \to \infty} \Pr \left(|\hat{J}_n(\pi, D_n) - J(\pi)| > \epsilon \right) = 0.$$