Reinforcement Learning

Final Exam

11/12/2020

Sanjit K. Kaul

Instructions: Exam is open book and notes. No other resources, human or otherwise are allowed. Violation of the policy will be treated as use of unfair means and plagiarism. You will receive the F grade in this course and all other additional academic penalties, if any, will apply.

Every question requires a summary answer in the provided paragraph in the Google form for the quiz. The Google form will stop accepting responses at $16:29:59$. At $16:25$, an assignment will be posted on Classroom for you to upload a scanned PDF form of your detailed work. This PDF submission will be allowed till $17:00$ hours. Note that any work in the PDF that seems in addition of the provided paragraph answers will not be considered for evaluation.

I. Model for our Enviroment

Our environment has states A, B, T_1, and T_2, where T_1 and T_2 are terminal states. The agent can take either the action *left* or the action *right* in the states A and B. When in state A, choosing right has the agent move to T_1 and choosing left has the agent transition to state B. When in state B, choosing right has the agent move to T_1 and choosing left has the agent transition to T_2.

If the agent chooses *right* in state A, it gets a reward of 1 with probability 0.5 and that of 0 otherwise. If the agent chooses *left* in state A, it gets a random reward identical to that it gets on choosing *right*.

If the agent chooses *right* in state B, it gets a reward of -1 with probability 0.5 and that of 0 otherwise. If the agent chooses *left* in state B, it gets a random reward governed by the Gaussian distribution with mean -1 and standard deviation 5.

Question 1. 20 **marks** Draw the MDP. Derive the optimal policy for the agent and the given environment using Policy Iteration. Start with an initial policy that chooses the actions left and right with equal probability in states A and B. In the Google form, write down the optimal values of A and B and the optimal policy.

Question 2. 30 **marks** Consider the Q-learning algorithm. Assume an ϵ-greedy behavior policy with $\epsilon = 0.01$. Further assume that all Q values are initialized to zero at the start of the algorithm. You are given five episodes that were generated while learning using the algorithm. The sequence shown for each episode must be read as $S_0, A_0, R_1, S_1, A_1, \ldots, S_T$. We have

- Episode 1: A, *left*, 1, B, *left*, 3, T_2.
- Episode 2: A, *right*, 0, T_1.
- Episode 3: A, *left*, 0, B, *left*, -1.5, T_2.
- Episode 4: A, *right*, 1, T_1.
- Episode 5: A, *left*, 1, B, *left*, -0.5, T_2.

For each episode, derive the a priori probability with which each action was picked. For every time step in each episode derive the updated Q values. State the greedy policy that is obtained using the Q values you obtain at the end of the five episodes.

In the Google form, state the Q values obtained at the end of each episode. Also, mention the greedy policy at the very end.

Question 3. 30 **marks** Consider the episodes shown in the question above. Assume that the learning algorithm used is SARSA. As in the question above, assume $\epsilon = 0.01$. Also assume that all Q values are initialized to zero at the beginning.

For each episode, derive the a priori probability with which each action was picked. For every time step in each episode derive the updated Q values. State the *greedy* policy that is obtained using the Q values you obtain at the end of the five episodes.

In the Google form, state the Q values obtained at the end of each episode. Also, mention the greedy policy at the very end.

Question 4. 20 **marks** Let the policy $\pi(a|s,\theta)$ be given by the softmax function

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}},$$

where the function $h(s,b,\theta)$ provides a valuation of a certain state-action pair. Let $h(s,a,\theta) = \theta^T x(s,a)$. Do the following. **(a)** Derive $\nabla_\theta \sum_a \pi(a|s,\theta)$. **(b)** Derive $\nabla_\theta \frac{\pi(a|s,\theta)}{\sum_a \pi(a|s,\theta)}$.

In the Google form mention your answer for (a) and explain why you think it is correct.