Reinforcement Learning

Mid Semester Exam

22/10/2022

Sanjit K. Kaul (with inputs from Samaksh Gupta)

Instructions: You have an hour to work on the questions. Answers with no supporting steps will receive no credit. No resources, other than a pen/pencil, are allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 30 **marks** We have a single state 0 in which an agent may take two actions H and T. The action H results in a reward of 1 and that of T results in a reward of 0.

(a) (5 marks) Draw the MDP.

(b) (10 marks) Consider a policy that chooses actions with equal probability. Calculate the value of state 0. Assume a discount factor of $\gamma = 0.9$.

(c) (15 marks) Consider two episodes during which the agent visits 10 states. The episodes are summarized as $A_0, A_1, \ldots$. The first episode was generated using a policy that picks actions with equal probability and resulted in a sequence of actions $H, H, T, T, T, H, H, T, H, T$. The second was generated using a policy that chose action H with probability 0.8 and T with probability 0.2. $T, T, H, H, H, H, H, H, T, T$. Use the two episodes to calculate an estimate of value of state 0 when using a policy that chooses H with probability 0.2 and T with 0.8. Use every-visit Monte Carlo. Use the sample mean to calculate any averages. Assume a discount factor of $\gamma = 1$.

Question 2. 40 **marks** We have a single state 0 and a terminal state. An agent may take the actions H and T in state 0. The action H results in a reward of 1 with the agent continuing to stay in state 0. The action T results in a reward of 0 and a transition to the terminal state. Assume a discount factor of $\gamma = 1$.

(a) (5 marks) Draw the MDP.

(b) (35 marks) Calculate the value of state 0 assuming a behavior policy that chooses H and T with equal probability and a target policy that chooses H with probability 0.2 and T with probability 0.8.

Question 3. 30 **marks** Consider two episodes, where each episode is summarized as $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$. Episode 1 is $0, 1, 5, 1, 1, 4, 0, 2, 3, 0, 1, -3, 2, 1, 3$. Episode 2 is $1, 2, -3, 0, 1, -2, 0, 2, 2, 2, 2, 3$. Start with all q-values initialized to 0. Use EWMA with $\alpha = 0.2$. Let $\gamma = 1$. Use the episodes to update the action values using (a) Q-Learning and (b) SARSA.