Reinforcement Learning

Mid Sem Exam Retake, 29/11/2024

Sanjit K. Kaul

Question 1. 30 **marks** You would like to estimate the value function using every-visit Monte Carlo for a set $\{1, 2, 3, 4, 5\}$ of non-terminal states and the terminal state 0. The set of actions is $\{-2, -1, 1, 2\}$. You start with an initial estimate of the value function. Suppose you generate two episodes using a policy that chooses actions with equal probability. The first episode results in the state sequence $4, 3, 5, 1, 0$. The corresponding action sequence is $-2, 2, 2, -1$ and the reward sequence is $-10, -5, 5, 10$. The corresponding sequences from the second episode are the state sequence $2, 3, 3, 4, 0$, the action sequence $1, -1, -2, 1$, and the rewards $-6, 8, 4, -4$. Derive your estimates of the value function at the end of each episode. Assume $\gamma = 1$ and use the sample mean to estimate expected values.

Derive your estimates of the value function at the end of each episode for an ϵ-greedy policy, with $\epsilon = 0.1$.

Question 2. 50 **marks** You are given a stochastic policy π, which picks any action from the set of all actions $A(s)$ with probability *greater* than $\epsilon/|A(s)|$. For the policy, write down $v_\pi(s)$ in terms of $q_\pi(s, a)$ and $\pi(a|s)$. Now consider a policy μ that is an ϵ-greedy policy, which in any state s chooses the greedy action to be the action that maximizes $q_\pi(s, a)$. Show that $q_\pi(s, \mu(s))$, which is the action value function $q_\pi(s, a)$ with a chosen as per policy μ and the reward-to-go as per policy π, is at least as large as $v_\pi(s)$, $\forall s$. [*Hint: To do so, you will want to write $q_\pi(s, \mu(s))$ in terms of $q_\pi(s, a)$ and the ϵ-greedy policy μ. Also, in the expression for $v_\pi(s)$ rewrite the probabilities as a sum of two terms, one of them being $\epsilon/|A(s)|$.*]

Suppose $v_*(s)$ is the optimal state value function and $q_*(s, a)$ is the corresponding action value function. Assume that the optimal policy is an ϵ-greedy policy. Write down $v_*(s)$, for any s, in terms of $q_*(s, a)$ and the corresponding optimal ϵ-greedy policy and expand the resulting expression in terms of $v_*(s)$ and the MDP PMF(s) $p(s', r|s, a)$. We will call the resulting system of equations as the Bellman equations for $v_*(s)$.

Further suppose $v_\mu(s) = v_\pi(s)$, $\forall s$. Show that $v_\pi(s)$ satisfies the above derived Bellman equations for $v_*(s)$. To do so, first write down $v_\mu(s)$ in terms of $q_\mu(s, a)$ and the policy μ, wherein the ϵ-greedy action selection as per policy μ must be made explicit. Next rewrite the obtained expression in terms of the MDP, the ϵ-greedy action selection as per μ, and $v_\mu(s)$. Finally, rewrite the last obtained expression in terms of $v_\pi(s)$ and the MDP. Argue that the final expression results in a system of equations that is obtained by replacing $v_*(s)$ by $v_\pi(s)$ in Bellman equations for $v_*(s)$ that we derived earlier.

Question 3. 20 **marks** Given a Markov Decision process, are rewards R_t and R_{t+1} independent? Prove your claim.