CCE Final Exam : Reinforcement Learning

Answer all the Questions; Total Marks: 50

1. Consider a multi-armed bandit problem with a single arm. Suppose the sequence of rewards obtained when pulling this arm on the first five pulls is $R_i = i$, $i = 1, \ldots, 5$. On subsequent pulls, this sequence gets repeated, i.e., $R_6 = 1$, $R_7 = 2$, etc. Let Q_n denote the action-value estimate after the arm has been pulled n times.

   (a) Write down the expression for Q_n (only) in terms of the rewards obtained? (1)

   (b) Obtain $Q_1, \ldots, Q_{15}$, i.e., Q-value estimates for the first 15 instants? (2)

   (c) Based on these 15 values, do you expect the Q_n sequence to converge? Give a proper justification for your belief? (2)

   (d) If you expect the Q_n sequence to converge, identify the point where it will converge. Again give a proper justification for your answer? (1)

   (e) Can one incrementally obtain Q_{n+1} (the Q-value at instant $n+1$) in terms of Q_n (the Q-value at instant n)? If your answer is yes, provide the algorithm for doing so? (1)

   (f) Compute $Q_1, \ldots, Q_5$ (the first five Q-values) using the incremental update algorithm and verify that the values obtained are the same as those of $Q_1, \ldots, Q_5$ obtained in part (b) above. (1)

   (g) In the incremental update algorithm (part (e) above), if the step-size $1/(n+1)^{0.8}$ is used in place of the one used in the algorithm, then clearly the sequence of Q-values obtained at every update changes. Can one find the point of convergence in this case? Do you expect it to be different from the point of convergence of $\{Q_n\}$ in part (d) above? Give some arguments to substantiate your claim? (2)

2. Consider the infinite horizon discounted reward setting with a discount factor $\gamma \in (0, 1)$. Let π be a given stochastic policy according to which actions are chosen in every state. Let $v_\pi(s)$ denote the value function under policy π and $q_\pi(s, a)$ denote the action value function under policy π. Assume that the transition probabilities are specified by the four-argument probability function $p(s', r \mid s, a)$. Then

   (a) Give an equation for q_π in terms of v_π and the four-argument p? (2)

   (b) Give an equation for v_π in terms of q_π and the policy π? (2)

   (c) Write down the Bellman equation for v_π in terms of the four-argument function p, the policy π and v_π itself? (2)

   (d) Give arguments to show that the Bellman equation has a solution and, if so, identify the solution? (2)

   (e) Answer whether or not one can write down the Q-Bellman equation as well, i.e., an equation for q_π in terms of the four-argument function p, the policy π and q_π itself? (2)

3. Consider a house with four rooms that are identified with the states of a Markov decision process and numbered as states 1,2, DG and UG, respectively, see Figure 1. The states DG and UG are goal states. If the process enters into either of these states, it just stays there and does not come out of them. Here DG is the desirable goal state and UG is the undesirable goal state. The transition dynamics is as follows: There are two actions A_1 and A_2 that are

feasible in state 1 and only an action A feasible in state 2. When action A_1 is chosen in state 1, the process moves to state 2 with probability 0.7 giving a reward of 0 and moves to state DG with probability 0.3 and gives a reward of 1. When action A_2 is chosen in state 1, the process remains in state 1 with probability 0.5 giving a reward of 0 and to DG with probability 0.3 giving a reward of 1 and to state UG with probability 0.2 giving a reward of 0. When in state 2, upon selecting action A, the process moves to state 1 with probability 0.3 and a reward of 0.5 and to state UG with probability 0.7 and a reward of 0. Assume no discounting, i.e., $\gamma = 1$. Further, since DG and UG are goal states, the process does not come out of these states once it gets inside any of them and moreover, $V^*(\text{DG}) = V^*(\text{UG}) = 0$.

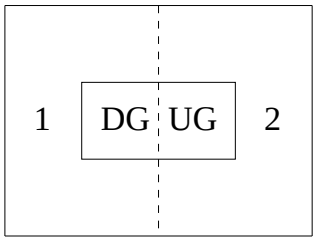

Figure 1: Room Transition Model

(a) Write down the Bellman equations in the state-value function for this problem? (2)

(b) Write down the value iteration scheme for numerically obtaining the solution to this Bellman equation? (1)

(c) Starting with initial value estimates of 0 for both the states, find the value estimates at $n = 5$ using the value iteration procedure, i.e., find $V_5(1)$ and $V_5(2)$, respectively. (2)

(d) Suppose we consider the following deterministic policy (π): pick action A_1 in state 1 and action A in state 2. For this policy π, (i) write down the Bellman equation and (ii) find the state values (i.e., the value function) $V_\pi(1)$ and $V_\pi(2)$ under this policy by solving this Bellman equation? (2)

(e) Under the above policy π, when starting in state 1, suppose we encounter the following episode of states: 1, 2, 1, 2, 1, 2, 1, 2, 1, DG. Write down the Monte-Carlo value function estimates for $V_\pi(1)$ and $V_\pi(2)$ using (i) the first visit method and (ii) the every visit method, respectively. (3)

4. Consider a random walk with nonterminal states A, B, C and D and with two terminal states E and F. Assume that the states are arranged along a straight line with (from left to right) A following E, B following A, C following B, D following C, and F following D. Thus the nonterminal state closest to E is A while the nonterminal state closest to F is D. From each nonterminal state, a transition happens either to one state prior to that state or the state just next to it with probability 1/2 each. However, when a terminal state is hit, the process terminates in that state. A transition from state D to F fetches a reward of 1. Rewards on all other transitions are 0. Assume the discount factor $\gamma = 1$. Suppose that the following episode is encountered:

$$A \xrightarrow{0} B \xrightarrow{0} C \xrightarrow{0} D \xrightarrow{1} F,$$

where the number (0 or 1) above the arrow indicates the reward earned on that transition. Assume that the initial estimate of the values $V(A)$, $V(B)$, $V(C)$ and $V(D)$ is 0.6 for all states $A, \ldots, D$. Let the step-size $\alpha = 0.1$.

(a) Apply three iterations of temporal difference learning (or n-step TD with $n = 1$) on all the four nonterminal states, i.e., find $V(A), \ldots, V(D)$ after three iterations of regular TD on each? (2)

(b) Apply three iterations of 2-step TD (i.e., n-step TD with $n = 2$) and again find $V(A), \ldots, V(D)$ after three iterations of 2-step TD? (3)

(c) Apply three iterations of 3-step TD (i.e., n-step TD with $n = 3$) and again find $V(A), \ldots, V(D)$ after three iterations of 3-step TD? (3)

(d) Which of the three algorithms (from amongst 1-step TD, 2-step TD and 3-step TD) do you expect to converge faster to the TD fix-point and why? (2)

5. Consider the case of value function approximation using an approximation architecture.

(a) Write down and briefly explain the update rule for semi-gradient TD(0) algorithm to estimate the value function under a given policy? (1)

(b) Write down the update rule for TD(0) under a linear approximation architecture? (2)

(c) Write down the expected SARSA update rule when linear function approximation is used to approximate the Q-values? (2)

(d) Write down the Q-learning algorithm with linear function approximation, i.e., when the Q-values are approximated using a linear function approximator? (2)

(e) Suppose we use linear function approximation for both sets of Q-values in the double Q-learning algorithm. Write down the pseudo-code for this (double Q-learning) algorithm with linear function approximation? (3)