CCE Mid-Term Exam II : Reinforcement Learning

Answer all the Questions; Total Marks: 40

1. Basic Probability Questions:

   (a) Consider repeated independent tosses of a coin whose probability of heads is p, $0 < p < 1$. What is the probability that the second time a head appears is on the fifth toss ? (2)

   (b) A fair dice is rolled repeatedly till we get at least one 5 and one 6. What is the probability that we need five rolls? (3)

2. Consider a multi-armed bandit with five arms with the ith arm associated with a parameter p_i that decides the reward which one gets by pulling arm i. Specifically assume that the reward obtained from pulling arm i is obtained as follows: When arm i is pulled at time t, a coin for which heads occurs with probability p_i is tossed once. The reward R^i_{t+1} obtained by pulling arm i with $i = 1, \ldots, 5$, is obtained according to

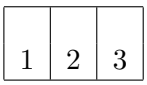
$$R^i_{t+1} = \begin{cases} 1 - p_i & \text{if coin toss results in heads} \\ 0 & \text{otherwise} \end{cases}$$

   Let $p_1 = 1/4$, $p_2 = 1/3$, $p_3 = 1/2$, $p_4 = 2/3$ and $p_5 = 3/4$ denote the values of p_i for the five arms respectively.

   (a) Suppose the goal of the decision maker is to pull the arm that gives the maximum expected reward. Assume the decision maker knows the reward distribution as described above. Then find the arm that the decision maker should pull? In other words, find the arm that gives the maximum expected reward. (2)

   (b) For a given arm i, suppose we can decide the parameter to be any value $p_i \in [0, 1]$, not necessarily restricted to one of the above mentioned values. Using basic calculus, find the parameter p_i that maximizes the expected reward $E[R^i_{t+1}]$ for this arm? (3)

3. Consider a gridworld with three cells as described below.



   In each of the three states, two actions are allowed. These are "move left" or "move right". These actions are picked according to a random policy π which picks both of these actions with equal probability of 0.5 in any state. When in state 1, if left action is picked, the state remains 1 and a reward of -1 is obtained. On the other hand, when in state 1, the right action is picked, the next state is 2 and reward of $+1$ is obtained.

   When in state 2, if left action is picked, the next state is 1 and a reward of 0 is obtained. On the other hand, when in state 2, the right action is picked, the next state is 3 and a reward of $+1$ is obtained.

   Finally, when in state 3, the left action results in a move to state 1 with a reward of 0 while the right action results in no change in state but with a reward of -1. Let the discount factor $\gamma = 0.9$.

(a) Write down all the current states (s), actions (a), rewards (r) and next states (s') in the form of a table? (1.5)

(b) Similar to the recycling robot example done in the class, draw a state transition diagram to compactly represent the information in part (a) above by drawing states, actions as well as weighted arcs indicating weights of both transition probabilities as well as single-stage rewards? (1.5)

(c) Find the values $v_\pi(1)$, $v_\pi(2)$ and $v_\pi(3)$ of the three states by writing down the corresponding Bellman equation for the policy π? (3)

4. In an infinite horizon discounted reward problem under a given policy π and with discount factor $\gamma \in (0, 1)$, we add a constant $C > 0$ to each single-stage reward obtained at each instant.

(a) What change in Q-value function do you observe as a result of the above change? Explain through a calculation? (2)

(b) Upon computation of the above Q-value function under policy π suppose we apply an ϵ-greedy scheme to update the policy. Argue whether or not the action you obtain when the constant single-stage reward is added (as above) is going to the same as what you would obtain without the additional reward? (2)

(c) What happens if tasks are episodic (with discounting) instead of continuing? (2)

5. Consider the 4×4 grid world with 14 non-terminal states and 2 terminal states T_1 and T_2 as below (this is the same setting as of Example 4.1 of the Text Book):

T_1	1	2	3
4	5	6	7
8	9	10	11
12	13	14	T_2

The actions that an agent can choose in any state are (i) go up, (ii) go down, (iii) go right or (iv) go left, respectively. Upon taking an action, the state moves one step in the chosen direction if it is feasible. If the chosen direction takes the state out of the grid (e.g., the action "move left" in state 4), the state remains unchanged. Assume the reward $R_t = -1$ on all transitions and discount factor $\gamma = 1$. Thus, this is an undiscounted, episodic task.

The value function for the equiprobable random policy (that picks all actions with equal probability of 1/4) is found to be:

0	-14	-20	-22
-14	-18	-20	-20
-20	-20	-18	-14
-22	-20	-14	0

Suppose now a new state 15 is added to the gridworld just below state 14, and its actions, left, up, right, and down, take the agent to states 13, 14, T_2, and 15, respectively.

(a) Assume that transitions from the original states are unchanged. Find $v_\pi(15)$ for the equiprobable random policy? (2)

(b) Assume now that the dynamics of state 14 is also changed so that the 'down' action in state 14 results in the next state being 15. Find $v_\pi(15)$ for the equiprobable random policy in this case? (3)

6. Consider an MDP with a single non-terminal state (NT) and two terminal states T_1 and T_2, respectively. When in state NT, the next state is NT itself with probability 0.6, T_1 with probability 0.2 or T_2 with probability 0.2, respectively. If the transition from NT is to either NT itself or to T_1, the single-stage reward is 1. On the other hand, if the transition from NT is to T_2, the single-stage reward is 0. We observe two episodes of this MDP both starting in state NT. The first episode terminates in T_1 and gives a total reward of 10. The second episode terminates in T_2 and gives a total reward of 4.

   (a) Write down both the episodes completely by writing the sequence of states visited and the single-stage rewards obtained? (2)

   (b) Write down the first visit and every visit estimates of the value of NT? (2)

   (c) How many estimates are obtained for the first visit and the every visit estimators? (1)

   (d) Find the value of state NT obtained from (i) the first visit estimators and (ii) the every visit estimators? (2)

7. Consider a Monte-Carlo estimator of average returns where the sequence of returns G_1, G_2, ..., G_n all start in the same state. Further, the return G_i has a corresponding weight e^{W_i} (the exponential of W_i), $i = 1, \ldots, n$. Suppose we form the estimate:

$$V_n \stackrel{\triangle}{=} \frac{\sum_{k=1}^{n} e^{W_k} G_k}{\sum_{k=1}^{n} e^{W_k}},\ n \geq 1.$$

   (a) Obtain an incremental update rule for V_n? The update rule should express V_{n+1} in terms of V_n. (3)

   (b) Write down an off-policy Monte-Carlo control algorithm in this case? (2)

   (c) For part (b) above, identify the connection of W_i with the importance sampling ratio $\rho_{t_i:T(t_i)-1}$? (1)