Final Exam: Reinforcement Learning

Answer all the Questions; Total Marks: 30

1. Recall the policy gradient theorem done in the class for the episodic setting. Let $J(\theta)$ denote the total reward when policy is π_θ and is given by

$$J(\theta) = V_{\pi_\theta}(s_0),$$

for some state s_0. Then the theorem says that

$$\nabla J(\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla \pi_\theta(s,a) Q^{\pi_\theta}(s,a). \tag{0.1}$$

(a) Give precise definitions of $d^{\pi_\theta}(s)$ and $Q^{\pi_\theta}(s,a)$ in (0.1)? (2)

(b) Let $|A(s)|$ denote the number of feasible actions in state s. Under the above assumptions, argue whether or not $\nabla J(\theta)$ also equals

i.
$$\nabla J(\theta) = \sum_s d^{\pi}(s) \sum_a \nabla \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - \max_a Q^{\pi_\theta}(s,a)).$$
(2)

ii.
$$\nabla J(\theta) = \sum_s d^{\pi}(s) \sum_a \nabla \pi_\theta(s,a) \left(Q^{\pi_\theta}(s,a) - \sum_b \frac{Q^{\pi_\theta}(s,b)}{|A(s)|} \right).$$
(2)

2. Consider the MDP model as shown in Figure 1.

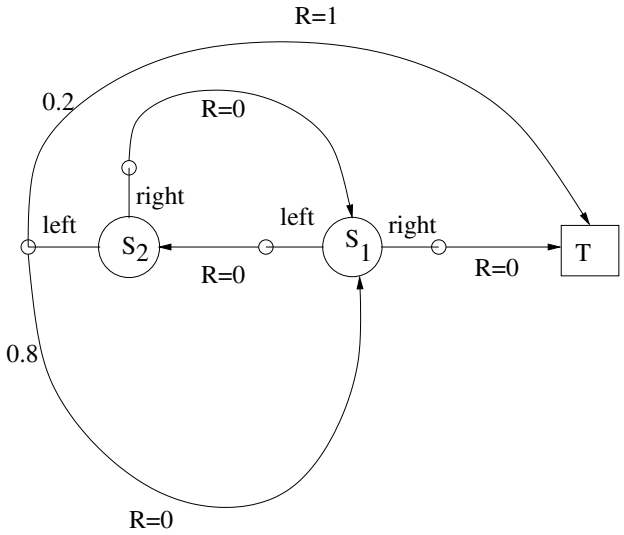

Figure 1: The MDP Model

Here S_1 and S_2 are two non-terminal states and T is the terminal state. There are two actions - left and right available in both the states S_1 and S_2, respectively. In state S_1, the left action

results in the next state being S_2 while the right action results in the next state being T. However, both actions result in a single-stage reward of 0. In state S_2, the right action results in the next state being S_1 but with an immediate reward of 0. However, when the left action is chosen in state S_2, with probability 0.8, the next state is S_1 with an immediate reward of 0, while with probability 0.2, the next state is T, with an immediate reward of 1. Let the discount factor $\gamma = 1$ since this is an episodic task.

Let S_1 be the starting state in each episode. Assume that the target policy (π) always selects the left action in both the states S_1 and S_2. However, the behavior policy b selects both the left and the right actions with probability of 0.5 in any of the states S_1 and S_2. Let G_0 denote the return starting from time 0. Answer the following questions:

(a) Compute the expected return $E_b\left[\prod_{t=0}^{T-1}\frac{\pi(A(t)\mid S(t))}{b(A(t)\mid S(t))}G_0\right]$? (2)

(b) Estimate the variance of the return by computing the following second moment of the return: $E_b\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A(t)\mid S(t))}{b(A(t)\mid S(t))}G_0\right)^2\right]$? (2)

(c) For this example, argue by giving precise arguments as to which of the two algorithms: (a) ordinary importance sampling and (b) weighted importance sampling you believe will perform better? (2)

3. (a) What is the difference between Monte-Carlo technique and TD(0) technique. When do you prefer one method over the other ? Justify your answer. (3)

(b) Consider a MDP with two states A and B. Say you follow a policy π and observe following episodes (in the same order):

1. B 1
2. B 1
3. B 1
4. $A, 0, B, 0$

This means that the first three episodes started in state B and terminated immediately with a reward of 1. And that the last episode started in state A, transitioned to B with a reward of 0, and then terminated from B with a reward of 0.

Compute the value function corresponding to states A and B using Monte Carlo technique and TD(0) technique respectively. Assume initial estimates of both the states to be 0, the step-size sequence for MC to be $\frac{1}{n+1}$ and a constant step-size of 0.5 for TD(0). (3)

4. Consider the infinite horizon discounted cost problem i.e., in the non-episodic or continuing case. Answer the following questions:

(a) Write the SARSA, expected SARSA, Q-learning update rule for the look-up table case? (3)

(b) From a practitioner's perspective, discuss merits and demerits of these algorithms. (3)

5. Consider the TD(0) algorithm with linear function approximation for estimating the value of a policy π in the discounted reward setting with discount factor $\gamma = 0.5$. Let the underlying Markov chain under the given policy π have just two states 0 and 1 with transition probabilities $p_{0,0} = 0.7$, $p_{0,1} = 0.3$, $p_{1,0} = 1$ and $p_{1,1} = 0$, respectively. Let the features associated with the two states 0 and 1 be $X(0) = [1, 0]^T$ and $X(1) = [1, 1]^T$ respectively.

   (a) Write now the general TD recursion with linear function approximation for the discounted reward setting assuming $X(s)$ is the d-dimensional feature vector for state s. (1)

   (b) Let the single-stage reward when in state S_t upon transition to state S_{t+1} be R_{t+1}. In particular, when $S_t = 0$, then $R_{t+1} = 5$ when $S_{t+1} = 0$. Further, when $S_t = 0$, then $R_{t+1} = 3$ when $S_{t+1} = 1$. On the other hand, when $S_t = 1$, then $R_{t+1} = 3$. Identify the point where the above algorithm is likely to converge? (5)