Assignment 9

Reinforcement Learning
Prof. B. Ravindran

1. State True or False for the following statements:
   Statement 1: DQN is an **on-policy** technique.
   Statement 2: Actor-Critic is a **policy gradient** method.

   (a) Both the statements are True.
   (b) Statement 1 is True and Statement 2 is False.
   (c) Statement 1 is False and Statement 2 is True.
   (d) Both the statements are False.

   Sol. (c)
   DQN uses Q learning, which is **off-policy**.
   Actor-critic is based on policy gradient theorem(used in the actor training).

2. What are the reasons behind using an experience replay buffer in DQN?

   (a) Random sampling from experience replay buffer breaks correlations among transitions.
   (b) It leads to efficient usage of real-world samples.
   (c) It guarantees convergence to the optimal policy.
   (d) None of the above

   Sol. (a), (b)
   Sampling randomly from replay buffer breaks the strong correlations between the transitions, leading to better updates in the neural network. Moreover, a sample can be used for more than one update.

3. **Statement:** DQN is implemented with current and target network.
   Reason: Using target network helps in avoiding chasing a non-stationary target.

   (a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.
   (b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.
   (c) Assertion is true, Reason is false
   (d) Both Assertion and Reason are false

   Sol. (a)
   The target network provides a stationary target for certain number of steps stabilizing the updates in DQN

4. Policy gradient methods can be used for continuous action spaces.

   (a) True
   (b) False

   Sol. (a)
   As policy gradient directly approximate the policy itself, they can be used for continuous action spaces.

5. **Assertion:** Actor-critic updates have lesser variance than REINFORCE updates.
Reason: Actor-critic methods use TD target instead of G_t.

   (a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

   (b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

   (c) Assertion is true, Reason is false

   (d) Both Assertion and Reason are false

   Sol. (a)
   Using G_t in REINFORCE involves returns from entire trajectory which causes higher variance in updates than TD-target in Actor-Critic.

6. Choose the correct statement for Policy Gradient Theorem for average reward formulation:

   (a) $\frac{\partial\rho(\pi)}{\partial\theta} = \sum_s d^\pi(s) \sum_a \frac{\partial\pi(s,a)}{\partial\theta}$

   (b) $\frac{\partial\rho(\pi)}{\partial\theta} = \sum_s v^\pi(s) \sum_a \frac{\partial\pi(s,a)}{\partial\theta} q^\pi(s,a)$

   (c) $\frac{\partial\rho(\pi)}{\partial\theta} = \sum_s d^\pi(s) \sum_a \frac{\partial\pi(s,a)}{\partial\theta} q^\pi(s,a)$

   (d) None of the above

   Sol. (c)
   As proved in the video of REINFORCE, (c) represents the correct statement of Policy Gradient Theorem

7. Suppose we are using a policy gradient method to solve a reinforcement learning problem. Assuming that the policy returned by the method is not optimal, which among the following are plausible reasons for such an outcome?

   (a) The search procedure converged to a locally optimal policy.

   (b) The search procedure was terminated before it could reach an optimal policy.

   (c) An optimal policy could not be represented by the parameterisation used to represent the policy.

   (d) None of these

   Sol. (a), (b) and (c)
   (a), (b) and (c) are all plausible reasons.

8. State True or False:
Monte Carlo policy gradient methods typically converge faster than the actor-critic methods, given that we use similar parameterisations and that the approximation to the Q^π used in the actor-critic method satisfies the compatibility criteria.

   (a) True

   (b) False

   Sol. (b)
   MC policy gradient algorithms generally suffer from large variance due to long episode lengths which can slow down convergence. Actor-critic methods, by relying on value function estimates lead to reduced variance, and hence, faster convergence.

9. When using policy gradient methods, if we make use of the average reward formulation rather than the discounted reward formulation, then is it necessary to assign a designated start state, s_0?

   (a) Yes
   (b) No
   (c) Can't say

   Sol. (b)
   We use the concept of a designated start state to allow a single value that can be assigned to a policy for evaluation. This is true for the average reward formulation even without a designated start state, by using the long term expected reward per step, $\rho(\pi)$, where:

$$\rho(\pi) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[r_1 + r_2 + ... + r_N|\pi]$$

10. State True or False:
    Exploration techniques like softmax (or other equivalent techniques) are not needed for DQN as the randomisation provided by experience replay provides sufficient exploration.

    (a) True
    (b) False

    Sol. (b)
    Some technique to ensure exploration is still required. As with the original Q-learning algorithm, if we only consider greedy transitions with respect to the action-value function, a large part of the state space will likely remain unexplored.