Assignment 6
Reinforcement Learning
Prof. B. Ravindran

1. **Assertion:** In order to use importance sampling for off-policy Monte-Carlo policy evaluation, we require knowledge of the state transition probabilities of the MDP.
   Reason: We require knowledge of the probability of each trajectory x_i according to the estimation policy and according to the behaviour policy. Both of these values are dependent on the state transition probabilities of the MDP.

   (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. (d)
   The importance factor only depends on the ratio between the probability of each trajectory x_i according to the estimation policy and according to the behaviour policy. The state transition probabilities cancel out in this ratio. Consequently, we do not require knowledge of the state-transition probabilities for off-policy Monte-Carlo policy evaluation.

2. Which of the following are true?

   (a) Dynamic programming methods use full backups and bootstrapping.

   (b) Temporal-Difference methods use sample backups and bootstrapping.

   (c) Monte-Carlo methods use sample backups and bootstrapping.

   (d) Monte-Carlo methods use full backups and no bootstrapping.

   Sol. (a),(b)
   Refer to the lecture on $TD(0)$.

3. Which of the following statement(s) is/are true for the UC-Trees algorithm?

   (a) We typically require a simulation model for the environment.

   (b) It uses ϵ-greedy exploration for action selection.

   (c) It computes an upper confidence bound on the value of (state, action) pairs.

   (d) It is a variation of the Monte-Carlo tree search algorithm.

   Sol. (a),(c),(d)
   Refer to the lecture on UC-Trees.

4. Consider the following statements:

   (i) $TD(0)$ methods uses unbiased sample of the return.

   (ii) $TD(0)$ methods uses a sample of the reward from the distribution of rewards.

   (iii) $TD(0)$ methods uses the current estimate of value function.

Which of the above statements is/are true?

(a) (i), (ii)

(b) (i),(iii)

(c) (ii), (iii)

(d) (i), (ii), (iii)

Sol. (c)
Refer the lecture on $TD(0)$.

5. **Assertion:** Q-learning can use asynchronous samples from different policies to update Q values.
Reason: Q-learning is an off-policy learning algorithm.

(a) Assertion and Reason are both true and Reason is a correct explanation of Assertion.

(b) Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

(c) Assertion is true but Reason is false.

(d) Assertion is false but Reason is true.

Sol. (a)
The learned action-value function directly approximates optimal the optimal action-value function independent of the policy being followed.

6. Suppose, for a 2 player game that we have modeled as an MDP, instead of learning a policy over the MDP directly, we separate the deterministic and stochastic result of playing an action to create 'after-states' (as discussed in the lectures). Consider the following statements:

(i) The set of states that make up 'after-states' may be different from the original set of states for the MDP.
(ii) The set of 'after-states' could be smaller than the original set of states for the MDP.

Which of the above statements is/are True?

(a) Only (i)

(b) Only (ii)

(c) Both (i) and (ii)

(d) Neither (i) nor (ii).

Sol. (c)
In tic-tac-toe for example, if the agent plays first, then the original set of states will all have an even number of marks (knot/cross) on the board. All after-states on the other hand, will all have an odd number of marks on the board. Hence (i) is True.
As discussed in the lectures, several of the original states of the MDP could map to the same after-state (once an action is played), so (ii) is True.

7. Consider the environment given below (CliffWorld discussed in lecture):
Suppose we use ϵ-greedy policy for exploration with a value of $\epsilon = 0.1$. Select the correct option(s):

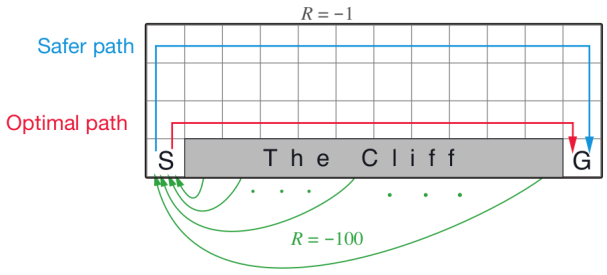

Figure 1: CliffWorld with two possible policies

(a) Q-Learning finds the optimal (red) path.

(b) Q-Learning finds the safer (blue) path.

(c) SARSA finds the optimal (red) path.

(d) SARSA finds the safer (blue) path.

Sol. (a), (d)
Q-Learning being off-policy finds the estimates for the optimal policy whereas SARSA being on-policy, finds a safer path due to -100 reward of falling in the cliff during exploration.

8. **Assertion:** Having a simulator/model is an advantage when using rollouts based methods.
Reason: Multiple trajectories can be sampled from the model from any given state.

(a) Assertion and Reason are both true and Reason is a correct explanation of Assertion.

(b) Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

(c) Assertion is true but Reason is false.

(d) Assertion and Reason are both false.

Sol. (a)
Having a model is necessary when using rollouts based methods. If a simulator is not present, the state from which multiple trajectories are being created needs to be reached multiple times, multiple trajectories will actually need to be "played". In comparison, simulating trajectories in a model that is present, without actually having to "play" them in the environment is significantly better.

9. Consider an MDP with two states A and B. Given the single trajectory shown below (in the pattern of state, reward, next state...), use on-policy TD(0) updates to make estimates for the values of the 2 states.

A, 3, B, 2, A, 5, B, 2, A, 4, END

Assume a discount factor $\gamma = 1$, a learning rate $\alpha = 1$ and initial state-values of zero. What are the estimated values for the 2 states at the end of the sampled trajectory? (Note: You are not asked to compute the true values for the two states.)

(a) $V(A) = 2, V(B) = 10$

(b) $V(A) = 8, V(B) = 7$

(c) $V(A) = 4, V(B) = 12$

(d) $V(A) = 12, V(B) = 7$

Sol. (c)
The TD(0) update rule is: $V_{new}(s_t) = V_{old}(s_t) + \alpha[R_{t+1} + \gamma V_{old}(s_{t+1}) - V_{old}(s_t)]$

Given the parameters $\gamma = 1$, and $\alpha = 1$, this rule simply becomes:

$V_{new}(s_t) = R_{t+1} + V_{old}(s_{t+1})$

Starting with state-values of zero and making updates along the sampled trajectory, we have the following updates:

V(A)= 3 + V(B) = 3
V(B)= 2 + V(A) = 5
V(A)= 5 + V(B) = 10
V(B)= 2 + V(A) = 12
V(A)= 4

So, at the end of the trajectory, we have estimates: V(A)=4, V(B)=12

10. Which one of the following statements are **True** for SARSA?

    (a) It uses bootstrapping to approximate full return.

    (b) It is an on-policy algorithm.

    (c) It is a TD method.

    (d) It always selects the greedy action choice.

    Sol. (a),(b),(c)
    (d) is false. SARSA requires adequate exploration of the state space to converge to an optimal policy.