Assignment 6

Reinforcement Learning
Prof. B. Ravindran

1. 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 $TD(0)$.

2. 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 $TD(0)$.

3. 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

4. Which of the following statements are true for SARSA?

   (a) It is a TD method.

   (b) It is an off-policy algorithm.

   (c) It uses bootstrapping to approximate full return.

   (d) It always selects the greedy action choice.

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

5. **Assertion:** In Expected-SARSA, we may select actions off-policy.
   Reason: In the update rule for Expected-SARSA, we use the estimated expected value of the next state under the policy π rather than directly using the estimated value of the next state that is sampled on-policy.

   (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 assertion and reason are both true. In expected SARSA, we use the estimated expected value of the next state under the policy π rather than the estimated value of the next state that is sampled on-policy and this allows us to select actions off-policy. Note: We assume that we have access to the policy π (or some estimation of the policy probabilities) in order to compute the expectation.

6. **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.

7. 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.

8. **Assertion:** Rollout algorithms take advantage of the policy improvement property.
   Reason: Rollout algorithms selects action with the highest estimated values.

   (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)
   Refer lecture on UC-Trees.

9. 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):

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

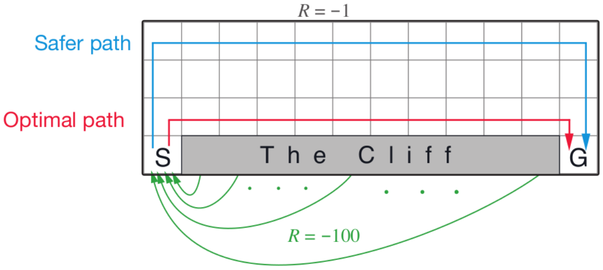

Figure 1: CliffWorld with two possible policies

(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.

10. Which of the following are **True** for TD(0) ? (Assume that the environment is truly Markov)

(a) It uses the full return to update the value of states.

(b) Both TD(0) and Monte-Carlo policy evaluation converge to the same value function, given a finite number of samples.

(c) Both TD(0) and Monte-Carlo policy evaluation converge to the same value function, given an infinite number of samples.

(d) TD error is given by "$\delta = v_{new}(s_t, a_t) - v_{old}(s_t, a_t)$".

Sol. (c)