Assignment 7

Reinforcement Learning
Prof. B. Ravindran

1. Which of the following is the corrected n-step truncated return?

   (a) $R_{t+n} + \gamma^n V_t(s_{t+n})$

   (b) $R_{t+1} + \gamma R_{t+2} + \gamma^2(R_{t+3}) + ... + \gamma^{n-1} R_{t+n} + \gamma^n V_t(s_{t+n})$

   (c) $\gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3(R_{t+3}) + ... + \gamma^n R_{t+n} + V_t(s_{t+n})$

   (d) None of the above.

   Sol. (b)
   From the definition of $G_t^{(n)}$ given in "Eligibility Traces" video.

2. Which of the following is correct way to implement replacing traces?

   (a) $E_t(s) = \begin{cases} 1 & ;\text{if } s_t \neq s \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   (b) $E_t(s) = \begin{cases} (\gamma\lambda)E_{t-1}(s) + 1 & ;\text{if } s_t = s \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   (c) $E_t(s) = \begin{cases} (\gamma\lambda)E_{t-1}(s) + 1 & ;\text{if } s_t \neq s \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   (d) $E_t(s) = \begin{cases} 1 & ;\text{if } s_t = s \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   Sol. (d)
   (d) correctly defines replacing traces

3. **Assertion:** TD(1), a way of implementing Monte Carlo with eligibility traces, extends Monte Carlo algorithm for continuing tasks.
   Reason: TD(1), implemented with eligibility traces, is an offline 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 and Reason is false

   (d) Both Assertion and Reason are false

   Sol. (c)
   TD(λ) with $\lambda = 1$ gives the Monte Carlo return as its target and hence TD(1) extends Monte Carlo for online implementation for continuing tasks. TD(1) however is an online algorithm with eligibility traces.

4. **Assertion:** It is NOT necessary for behavior policy of an off-policy method to have non-zero probability of selecting all actions.
   Reason: If probability of certain actions in estimation policy is zero, then probability of selecting corresponding actions in behavior policy can be zero, and it would not cause trouble in learning procedure.

(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)
actions for which probability in both policies is zero can simply be ignored and thus, there can be actions having zero probability in a policy. So both statements are true and reason is correct explanation for assertion.

5. Considering episodic tasks and for $\lambda \in (0, 1)$, one-step return always gets assigned the maximum weight in the λ-return. True or false.

   (a) False

   (b) True

   Sol. (a)
   It depends on the length of an episode (as well as the value of λ). For example, consider an episode of length 3 and a value of $\lambda = 0.7$.

6. **Assertion:** It is possible to use the forward view of eligibility traces based target for updates before the end of an episode.
   Reason: G_t^λ is the weighed average of terms of the form G_t^i, where each G_t^i is computable after a finite number of steps i, making G_t^λ also computable after a finite number of steps in all cases.

   (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 assertion is wrong, the forward view of eligibility traces based target cannot be computed till the end of the episode, making updates before the end of an episode impossible. The reason is wrong as well, each G_t^i is computable after i steps, however i has no upper bound, making G_t^λ impossible to calculate in a finite amount of time in some cases.

7. Which of the following is a way to use replacing traces in SARSA(λ) such that we only update the estimate for the action, a, taken in the state, s, in a trajectory at timestep t?

   (a) $E_t(s,a) = \begin{cases} \gamma\lambda E_{t-1}(s,a) + 1 & ;\text{if } s_t = s, a_t = a \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   (b) $E_t(s,a) = \begin{cases} 1 & ;\text{if } s_t = s, a_t = a \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

   (c) $E_t(s,a) = \begin{cases} \gamma\lambda E_{t-1}(s,a) + 1 & ;\text{if } s_t = s, a_t = a \\ 0 & ;\text{if } s_t = s, a_t \neq a \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

(d) $E_t(s,a) = \begin{cases} 1 & ;\text{if } s_t = s, a_t = a \\ 0 & ;\text{if } s_t = s, a_t \neq a \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

Sol. (d)
(d) is an aggressive form of replacing traces update for SARSA(λ) where we update estimate only for estimate for action taken in a state by making traces for other actions zero.

8. **Assertion:** When using an ϵ-greedy exploration strategy, and Watkins $Q(\lambda)$, the ϵ value must be kept low.
Reason: Traces will become too short if a high value of ϵ is used, negating many of the advantages of using eligibility traces.

(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)
Explained in the video "Eligibility Trace Control" in minutes 21:00 to 24:30. Every time an exploratory action is taken, the trace ends, causing traces to become too short.

9. Consider the current state is s and the action recommended by the policy, a_1, is taken. The possible reason(s) behind setting $\forall a \neq a_1,\ E_t(s,a) = 0$, is/are:

(i) Rewards obtained by taking a_1 in s should not be attributed to actions other than a_1 taken when in state s previously.
(ii) It assumed that the time steps between reaching s are large enough to decay the eligibility trace to 0.

Which of the above is/are the correct reason(s)?

(a) Only (i)
(b) Only (ii)
(c) Both (i) and (ii)
(d) None of the above

Sol. (a)
Explained in the video "Eligibility Trace Control".

10. In off-policy TD(λ), which of the following is a correct way to update eligibility trace, if π is the policy whose Q-values we want to estimate, μ is the behaviour policy, and $\rho_t = \frac{\pi(s_t,a_t)}{\mu(s_t,a_t)}$?

(a) $E_t(s) = \begin{cases} \rho_t\gamma\lambda E_{t-1}(s) + 1 & ;\text{if } s_t = s \\ \rho_t\gamma\lambda E_{t-1}(s) & ;\text{otherwise} \end{cases}$

(b) $E_t(s) = \begin{cases} \rho_t(\gamma\lambda E_{t-1}(s) + 1) & ;\text{if } s_t = s \\ \rho_t\gamma\lambda E_{t-1}(s) & ;\text{otherwise} \end{cases}$

(c) $E_t(s) = \begin{cases} \gamma\lambda E_{t-1}(s) + 1 & ;\text{if } s_t = s \\ \rho_t \gamma\lambda E_{t-1}(s) & ;\text{otherwise} \end{cases}$

(d) $E_t(s) = \begin{cases} \rho_t(\gamma\lambda E_{t-1}(s) + 1) & ;\text{if } s_t = s \\ \gamma\lambda E_{t-1}(s) & ;\text{otherwise} \end{cases}$

Sol. (b)
The correct equation for this update is (b) as it takes care of importance sampling ratio for both cases, $s_t = s$ and otherwise, by multiplying ρ_t to both terms.