Unit 9 - Week 7

Assignment 7

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2020-03-18, 23:59 IST.

1) Which of the following is equal to $G_t^{(2)}$ *1 point*

- $R_{t+1} + \gamma V(s_{t+1})$
- $R_{t+1} + \gamma R_{t+2} + \gamma^2 V(s_{t+2})$
- $R_{t+1} + \gamma R_{t+2} + \gamma^2 V(s_{t+3})$
- None of the above

No, the answer is incorrect.
Score: 0
Accepted Answers:
$R_{t+1} + \gamma R_{t+2} + \gamma^2 V(s_{t+2})$

2) **Assertion:** It is possible to use the forward view of eligibility traces based target for updates before the end of an episode. *1 point*
Reason: G_t^i 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^i also computable after a finite number of steps in all cases

- Both Assertion and Reason are true, and Reason is correct explanation for Assertion
- Both Assertion and Reason are true, but Reason is not correct explanation for assertion
- Assertion is true, Reason is false
- Both Assertion and Reason are false

No, the answer is incorrect.
Score: 0
Accepted Answers:
Both Assertion and Reason are false

3) Consider the current state is s and the action recommended by the policy, a_1, is executed. Which of the following is the correct reasoning behind setting $\forall a \neq a_1,\ E_t(s,a) = 0$? *1 point*

- Rewards obtained by playing a_1 in s should not be attributed to actions other than a_1 played when in state s previously
- It assumed that the time steps between reaching s are large enough to decay the eligibility trace to 0
- Both (a) and (b)
- None of the above

No, the answer is incorrect.
Score: 0
Accepted Answers:
Rewards obtained by playing a_1 in s should not be attributed to actions other than a_1 played when in state s previously

4) **Assertion:**When using an ϵ-greedy exploration strategy, and Watkins $Q(\lambda)$, the $epsilon$ value must be kept low *1 point*
Reason: Traces will become too short if a high value of ϵ is used, negating many of the advantages of using eligibility traces.

- Both Assertion and Reason are true, and Reason is correct explanation for Assertion
- Both Assertion and Reason are true, but Reason is not correct explanation for assertion
- Assertion is true, Reason is fals
- Both Assertion and Reason are false

No, the answer is incorrect.
Score: 0
Accepted Answers:
Both Assertion and Reason are true, and Reason is correct explanation for Assertion

5) Recollect estimation policy(π) and behavior policy(μ) in importance sampling. *1 point*
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 estiamation policy is zero, then probability of selecting corresponding actions in behavior policy can be zero, and it would not cause trouble in learning procedure

- Assertion and Reason are both true and Reason is a correct explanation of Assertion
- Assertion and Reason are both true and Reason is not a correct explanation of Assertion
- Assertion is true but Reason is false
- Assertion and Reason are both false

No, the answer is incorrect.
Score: 0
Accepted Answers:
Assertion and Reason are both true and Reason is a correct explanation of Assertion

6) 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? *1 point*

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

No, the answer is incorrect.
Score: 0
Accepted Answers:
$E_t(s,a) = \begin{cases} 1 & ;\text{if } s_t = s, a_t = a \\ 0 & ;\text{if } s_t = s, a_t \neq a \\ (\gamma\lambda)E_{t-1}(s) & ;\text{otherwise} \end{cases}$

7) 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 $\rho_t = \frac{\pi(s_t,a_t)}{\mu(s_t,a_t)}$? *1 point*

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

No, the answer is incorrect.
Score: 0
Accepted Answers:
$E_t(s) = \begin{cases} \rho_t(\gamma\lambda E_{t-1}(s) + 1) & ;\text{if } s_t = s \\ \rho_t\gamma\lambda E_{t-1}(s) & ;\text{otherwise} \end{cases}$

8) If $\lambda = 1$ then we add only G_t at the end of the episode. In MC we do the same thing *1 point*

- Yes
- No

No, the answer is incorrect.
Score: 0
Accepted Answers:
Yes

9) Consider the following trajectory $s_3, s_2, s_1, s_2, s_2, s_4, s_5, s_6$ *1 point*
What would be the eligibility value $E_8(s_2)$, for state s_2 after the $8th$ time step if we use accumulating trace. Discount factor = γ , trace decay parameter = λ, initial value of eligibility is zero for all states

- $\gamma^3\lambda^3(\gamma^2\lambda^2 + 2)$
- $\gamma^8\lambda^8$
- $\gamma^3\lambda^3(\gamma^3\lambda^3 + \gamma\lambda + 1)$
- $\gamma\lambda$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\gamma^3\lambda^3(\gamma^3\lambda^3 + \gamma\lambda + 1)$

10) Which of the following statements are true? *1 point*

- TD(1) with replacing trace is identical to first visit Monte Carlo algorithm
- TD(1) with replacing trace is identical to every visit Monte Carlo algorithm
- TD(1) with replacing trace is identical to last visit Monte Carlo algorithm
- TD(1) with accumulating trace is identical to every visit Monte Carlo algorithm

No, the answer is incorrect.
Score: 0
Accepted Answers:
TD(1) with replacing trace is identical to first visit Monte Carlo algorithm
TD(1) with accumulating trace is identical to every visit Monte Carlo algorithm