Unit 5 - Week 3

Assignment 3

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

Due on 2020-02-19, 23:59 IST.

1) Consider the following policy-search algorithm for a multi-armed binary bandit: *1 point*

$\forall a, \quad \pi_{t+1}(a) = \pi_t(a)(1-\alpha) + \alpha(1_{a=a_t} R_t + (1 - 1_{a=a_t})(1 - R_t))$
where $1_{a_t=a}$ is 1 if $a = a_t$ and 0 otherwise. Which of the following is true for the above algorithm?

- It is L_{R-I} algorithm
- It is $L_{R-\epsilon P}$ algorithm
- It would work well if the best arm had probability of 0.9 of resulting in +1 reward and the next best arm had probability of 0.5 of resulting in +1 reward
- It would work well if the best arm had probability of 0.3 of resulting in +1 reward and the worst arm had probability of 0.25 of resulting in +1 reward

No, the answer is incorrect.
Score: 0
Accepted Answers:
It would work well if the best arm had probability of 0.9 of resulting in +1 reward and the next best arm had probability of 0.5 of resulting in +1 reward

2) Assertion: Contextual bandits can be modeled as a full reinforcement learning problem. *1 point*
Reason: We can define an MDP with n states where n is the number of bandits and each of them having k actions corresponding to the arms in each bandit, with
single action leading to termination of episode, giving a reward which corresponds to the state and action(bandit and arm)

- 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 and Reason is false
- Both Assertion and Reason are false

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

3) In a full RL setting is it possible for an agent to land up in the same state after performing an action *1 point*

- Yes
- No

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

4) Lets assume for some full RL problem we are acting according to a policy π. At some time t we are in a state S where we took action A_1. After few time steps at time t' the same state S was reached where we performed an action $A_2 (\neq A_1)$. What can you say about the policy π. *1 point*

- It is not a Stationary policy
- It may be Stationary policy
- It is a Non-Stationary policy
- It may be a Non-Stationary policy

No, the answer is incorrect.
Score: 0
Accepted Answers:
It may be Stationary policy
It may be a Non-Stationary policy

5) Which of the following statements are true. In RL problems *1 point*

- We assume that the agent determines the reward based on the current state and action
- Our main aim is to get a net positive reward
- At one time step we can perform only one action
- Zero rewards may be possible

No, the answer is incorrect.
Score: 0
Accepted Answers:
At one time step we can perform only one action
Zero rewards may be possible

6) Stochastic gradient ascent/descent update occurs in the right direction at every step *1 point*

- True
- False

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

7) The baseline in the REINFORCE update can depend on which of the following(without voiding any of the steps in the proof of REINFORCE)? *1 point*

- R_{n-1}
- Action taken(a_n)
- R_n
- None of the above

No, the answer is incorrect.
Score: 0
Accepted Answers:
R_{n-1}
R_n

8) Which of the following is true for an MDP? *1 point*

- $Pr(S_{t+1}, R_{t+1}|S_t, A_t) = Pr(S_{t+1}, R_{t+1})$
- $Pr(S_{t+1}, R_{t+1}|S_t, A_t, S_{t-1}, A_{t-1}, S_{t-2}, A_{t-2}, \ldots, S_0, A_0) = Pr(S_{t+1}, R_{t+1}|S_t, A_t)$
- $Pr(S_{t+1}, R_{t+1}|S_t, A_t) = Pr(S_{t+1}, R_{t+1}|S_0, A_0)$
- $Pr(S_{t+1}, R_{t+1}|S_t, A_t) = Pr(S_t, R_t|S_{t-1}, A_{t-1})$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$Pr(S_{t+1}, R_{t+1}|S_t, A_t, S_{t-1}, A_{t-1}, S_{t-2}, A_{t-2}, \ldots, S_0, A_0) = Pr(S_{t+1}, R_{t+1}|S_t, A_t)$

9) Remember for discounted returns, *1 point*
$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots$
Where γ is a discount factor. Which of the following best explains what happens when $\gamma > 1$, (say $\gamma = 5$)?

- Nothing, $\gamma > 1$ is common for many RL problems
- Theoretically nothing can go wrong, but this case does not represent any real world problems
- The agent will learn that delayed rewards will always be beneficial and so will not learn properly
- None of the above is true

No, the answer is incorrect.
Score: 0
Accepted Answers:
The agent will learn that delayed rewards will always be beneficial and so will not learn properly

10) Which of the following expressions is a possible update made when using REINFORCE for the following policy? $\pi(a_1; \theta) = \sin^2(\theta)$, consider only two *1 point*
actions, a_1 and a_2 are possible. Consider α can be taken to be any constant(Any constant obtained in the expression can be absorbed into α)

- $\alpha(R - b)\cot(\theta)$
- $-\alpha(R - b)\tan(\theta)$
- Neither (a) or (b)
- Both (a) and (b)

No, the answer is incorrect.
Score: 0
Accepted Answers:
Both (a) and (b)