RL MOCK TEST

1. A Markov Decision Process is being used on a random variable of state xt, action at and reward rt. How is the reward conditional probability for the process defined?

A. P(rt | at+1, xt+1)

B. P(rt+1 | at, xt)

C. P(rt+1 | at+1, xt+1)

D. P(rt | at, xt)

2. You find out that the next state is terminal when calculating the target for a Deep Q network. What will the target for the particular state be in the given scenario?

A. r

B. r+γ

C. r + γ ∗ maxaQ(s′,a)

D. None of the above

3. You are asked to approximate the given functions using neural networks. Which of these options is the most valid with respect to the given context?

1. State-value function

2, Action-value function

3. Policy function

A. 1 and 3 cannot be approximated

B. Only 1 and 2 can be approximated

C. All 1, 2 and 3 can be approximated

D. 2 and 3 cannot be approximated

4. When performing GPI(Generalised Policy Iteration), you want to see if both the evaluation process and the improvement process are stabilized.

When can one be sure of the stabilization in the given context?

A. When the value function and policy are not optimal

B. When a policy has been found that is greedy with respect to its own evaluation function

C. When a evaluation function has been found that is greedy with respect to its own policy

D. None of these

5. You have a ϵ-greedy policy where the ϵ is a hyperparameter that controls the tradeoff between exploration and exploitation. What happens to the actions in the given scenario?
1. ϵ is too small
2. ϵ is too large

A. Actions are biased to be more greedy in both 1 and 2.
B. Actions explore more. / Actions are biased to be more greedy.
C. Actions are biased to be more greedy. / 2. Actions explore more.
D. Actions explore more in both 1 and 2.

6. An action space has 20 joints, each joint having its own id, value and direction. How many torques will the action space contain?

A. 20
B. 10
C. 40
D. 5

7. You have a probabilistic policy π(a|s) to help the agent decide the action that he/she should take in a given state. When does this policy become a deterministic policy?

A. When π(a|s) = 1
B. When π(a|s) ---> a
C. When π(s) ---> a
D. When π(s) = 1

8. You are using policy improvement to maximise the total rewards, by finding the best action for each state.
What should the value of π′(a|s) be such that the best action is chosen?

A. 0
B. 1
C. a
D. None of these

9. The discount rate in the infinite horizon model is used to keep the return finite. What happens when the discount rate approaches 1?

A. Only the immediate reward is counted
B. The rewards further in the future is counted more
C. The agent becomes more farsighted
D. Both 1 and 3
E. Both 2 and 3

10. Read the given statements carefully and choose the correct option.
S1: The target variable is Non-stationary or unstable.
S2: A Target Network is used where instead of using one network for learning, two networks are used.

A. S1 is an unavoidable problem which arises in Reinforcement learning.
B. S1 is a consequence of S2
C. S1 and S2 are not related
D. S2 is a solution for S1

11. You are given a Q table of size mxn to calculate the current state-action value for a problem. Assume that the given problem has 5 different actions and 10 different states.
In the given scenario, what will the size mxn of Q be?

A. 5 x 10
B. 10 x 5
C. 11 x 6
D. 6 x 11

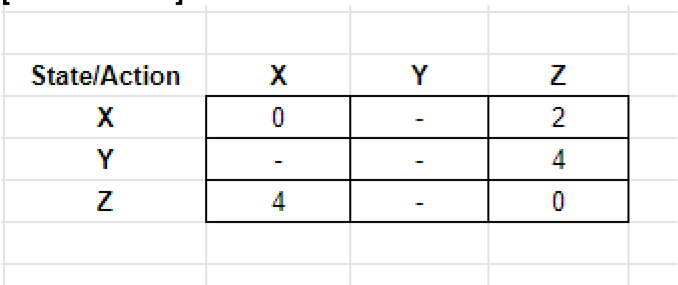
12. When training a model using Deep Q learning, you are using two networks: the actual network and the target network. What is the purpose of the target network?

A. The actual network weights are assigned to the target network for every N no of frames /iterations.
B. To update the actual network gradients for every frame/action
C. Make labels for training purposes
D. None of these

13. It is given that during a GPI(Generalised Policy Iteration), value and policy functions interact until they are optimal and thus consistent with others.
In which of the given cases can you say that the value function and policy is stabilised?
Case 1: When it is consistent with the current policy
Case 2: When it is greedy with respect to the current value function

A. Case 1: Policy is stabilised / Case 2: Value function is stabilised
B. Case 1: Value function is stabilised / Case 2: Policy is stabilised
C. Case 1: Value function and Policy is stabilised
D. Case 2: Value function and Policy is stabilised

14. You have a Q matrix which is a zero matrix and a reward matrix as given alongside. Assume that the discount rate is 0.5. What is the value of Q(X,Z) in the given context? [Assume α =1]



State/Action	X	Y	Z
X	0	-	2
Y	-	-	4
Z	4	-	0

A. 2
B. 4
C. 100
D. 10

15. An agent follows a policy π and for each encountered state(of the actual returns that have followed that state) it maintains an average of all the rewards.
What can be inferred from the given scenario?

A. The average will converge to the state’s value under the policy π, as the number of times that state is encountered approaches infinity
B. The averages will similarly converge to the action values.
C. Both 1 and 2
D. None of these

16. The condition (for value function) given alongside holds between the value of s and the value of its possible successor states for a policy π and any state s.

In the given scenario, which of the following are implicit in case of an episodic problem?

$$\sum_{a} \pi(a|s) \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_{\pi}(s') \Big]$$

A. The actions, a, are taken from the set A(s), the next states, s', are taken from the set S.

B. The rewards, r, are taken from the set R

C. The actions, a, are taken from the set A(s), the next states, s', are taken from the set S+.

D. Both 1 and 2

E. Both 2 and 3

17. You notice that a traditional Deep Q Network leads to unstable training and low quality policy problems when learning.

What is a valid solution to this problem?

1. Using two separate Q-value estimators, each of which is used to update the other.

2. Use multiple independent estimators to update the opposite estimator

A. Only 1

B. Only 2

C. Either 1 or 2

D. Neither 1 nor 2

18. Assume that you take a policy and roll out an episode(game). For every state, an action is sampled, performed in the environment and the reward and the next state is observed. The decisions given below are taken when the rewards are observed.

Now, suppose you lost the game, which results in a negative reward and the probabilities are pushed down. A negative reward is obtained even though some actions were really good.

What can you do to deal with this problem?

A. Blame all the actions equally

B. Blame all the actions equally and slightly modify the actual reward

C. Discount the rewards at each time step/state

D. Blame all the actions equally and discount the rewards at each time step/state

19. Which of the following statements about Monte Carlo methods is true?

S1: Monte Carlo methods can only be applied to episodic MDPs

S2: Monte Carlo methods learn directly from episodes of experience

A. Only S1

B. Only S2

[✓ C. Both S1 and S2]

D. Neither S1 nor S2

20. An agent wants to learn the optimal policy while it is continuing to explore. To do so, the learning policy π from the episodes is generated using another policy α.

What can be said about the method used to learn the policy in the given scenario?

[✓ A. Agent uses the Off-policy method to learn the policy]

B. Agent used the On-policy method to learn the policy

C. Agent uses either On-Policy or Off-policy method to learn the policy

D. None of these

21. You want to determine the time required to train a network using Deep Q Learning. In the given context, which of the following factors does the time required to train a network depend on?

1. Layers in the neural network

2. Markov Decision Process

[✓ A. Only 1]

B. Only 2

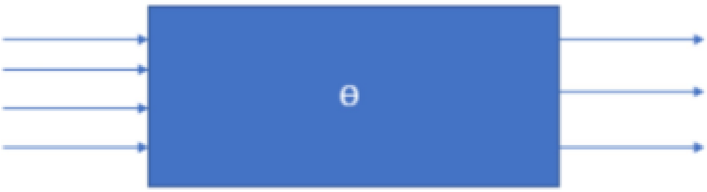
C. Both 1 and 2

D. Neither 1 nor 2

22. Assume that you have 5 states (S1, S2,..., S5) in the environment and 5 actions (A1, A2,....., A5) that can be taken from each of these states. You are using Monte-Carlo prediction to find an unbiased estimate of qπ(s,a) for each state-action pair.
What will the state-action values for qπ(S1,A1) and qπ(S3,A5) be if the episode is tracked as given alongside? [Assume γ=1 and initialised q(s,a)=0]
Episode 1: (S1, A1, 6) , (S2, A2, 4) , (S3, A5, 10) , (S4, A3, 6) , (S5, A8, 8)

A. qπ(S1,A1) = 34 / qπ(S3,A5) = 24
B. qπ(S1,A1) = 34 / qπ(S3,A5) = 20
C. qπ(S1,A1) = 30 / qπ(S3,A5) = 20
D. qπ(S1,A1) = 28 / qπ(3,A5) = 14

23. You are using the Deep Q network architecture given alongside to find the optimal action in the given state. You are performing one feedforward operation before taking an action.
What is provided as an input to the network and why?



A. A state is provided as an input because Q value for all possible action is obtained by performing one feedforward operation
B. A state is provided as an input because max Q value is obtained by performing one feedforward operation
C. A state is provided as an input because max Q value for all possible actions is obtained by performing one feedforward operation
D. A state is provided as an input because max Q value for one action is obtained by performing one feedforward operation

24. When training a Double Deep Q network the parameters of the Q-network are updated until an episode ends. You notice that the network has destabilized by falling into feedback loops between the predicted and target Q-values.
What might have caused this issue?

A. Not updating the parameters of the target network
B. Constantly shifting the predicted and target Q-values to update the network
C. By fixing the target value for the entire episode
D. Both 1 and 3
[✓ E. Both 1 and 2]

25. You have an undiscounted Markov Process with two states A and B. The transition matrix and reward matrix are unknown. Two sample episodes are observed as given alongside with their respective rewards.
A+2 ---> A+3 ---> B-5 ---> A+2 ---> B-3 ---> terminate
B-2 ---> A+3 ---> B-3 ---> terminate
You are using the Monte Carlo model for policy evaluation. In the given context, what will the value of the state-value functions V(A) and V(B) be during the first visit?

A. V(A) = 1 , V(B) = -8
B. V(A) = 1/2 , V(B) = -5/2
[✓ C. V(A) = 1 , V(B) = -4]
D. V(A) = -1/2 , V(B) = -4