Number of Questions :	29
Number of Questions to be attempted :	29
Section Marks :	50
Display Number Panel :	Yes
Section Negative Marks :	0
Group All Questions :	No
Enable Mark as Answered Mark for Review and Clear Response :	Yes
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	640653103511
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 124 Question Id : 640653698544 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 0

Question Label : Multiple Choice Question

THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL :REINFORCEMENT LEARNING (COMPUTER BASED EXAM)"

ARE YOU SURE YOU HAVE TO WRITE EXAM FOR THIS SUBJECT?
CROSS CHECK YOUR HALL TICKET TO CONFIRM THE SUBJECTS TO BE WRITTEN.

(IF IT IS NOT THE CORRECT SUBJECT, PLS CHECK THE SECTION AT THE TOP FOR THE SUBJECTS REGISTERED BY YOU)

Options :

6406532332918. ✔ YES

6406532332919. ✖ NO

Question Number : 125 Question Id : 640653698545 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 0

Question Label : Multiple Choice Question

Note:

1. Always enter your answer correct upto 3 decimal places without rounding off for numerical questions.

Options :

6406532332920. ✔ Useful Data has been mentioned above.

6406532332921. ✖ This data attachment is just for a reference & not for an evaluation.

Sub-Section Number :	2
Sub-Section Id :	640653103512
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 126 Question Id : 640653698546 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

In the context of a multi arm bandit (MAB) problem and stationary reward distribution, consider following:

Assertion: UCB minimizes the regret better than ϵ-greedy approach.

Reason: ε-greedy approach keeps the probability of choosing a suboptimal arm constant.

Options :

6406532332922. ✔ Assertion and Reason are both true and Reason is a correct explanation of Assertion.

6406532332923. ✖ Assertion and Reason are both true and Reason is not a correct explanation of

Assertion.

6406532332924. ✖ Assertion is true and Reason is false

6406532332925. ✖ Both Assertion and Reason are false.

Question Number : 127 Question Id : 640653698549 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

In the policy improvement step of policy iteration for a finite MDP, if the tie among actions, which have the same maximum value, is broken randomly, what would happen to the convergence of the algorithm?

Options :

6406532332933. ✖ The algorithm's convergence is independent of how ties are broken

6406532332934. ✔ The algorithm may oscillate between multiple optimal policies and may never converge or convergence may be delayed

6406532332935. ✖ The algorithm will certainly not converge

6406532332936. ✖ None of these

Question Number : 128 Question Id : 640653698550 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

Real-time dynamic programming, or RTDP, is an on-policy trajectory-sampling version of which of the following algorithms?

Options :

6406532332937. ✔ Value iteration.

6406532332938. ✖ Q-learning.

6406532332939. ✖ SARSA.

6406532332940. ✖ TD.

6406532332941. ✖ None of these.

Question Number : 129 Question Id : 640653698558 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

Which of the following is the correct 5-step truncated corrected return starting from time *t* ?

Options :

6406532332955. ✖ $G_t^{(5)} = r_{t+1} + \gamma V_t(s_{t+1})$

6406532332956. ✖ $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma^5 V_t(s_{t+1})$

6406532332957. ✔ $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma^5 V_t(s_{t+5})$

6406532332958. ✖ $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma V_t(s_{t+5})$

6406532332959. ✖ $G_t^{(5)} = \sum_{i=t+1}^{T} [\gamma^{i-1} r_i]$, where T is the terminal state index.

Question Number : 130 Question Id : 640653698565 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

Which of the following is the correct way to represent policy for policy search methods?

Options :

6406532332987. ✔ $\pi(a|s) = \rho_a, \sum_i \rho_i = 1 \text{ and } 1 \geq \rho_i \geq 0 \forall i$

6406532332988. ✖ $\pi(a|s) = \rho_a, \sum_i \rho_i \geq 1 \text{ and } 1 \geq \rho_i \geq 0 \forall i$

6406532332989. ✖ $\pi(a|s) = \rho_a, \sum_i \rho_i \leq 1 \text{ and } 1 \geq \rho_i \geq 0 \forall i$

6406532332990. ✖ $\pi(a|s) = \rho_a, \sum_i \rho_i = 1 \text{ and } 2 \geq \rho_i \geq 0 \forall i$

Sub-Section Number :	3
Sub-Section Id :	640653103513
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 131 Question Id : 640653698547 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1 Max. Selectable Options : 0

Question Label : Multiple Select Question

Consider a grid world with obstacles and boundary walls. An agent can take only 4 actions, i.e. north, south, east or west. In a state, an action that might lead to an obstacle or out of boundary, does not change the state. The agent is trained sufficiently. Which of the following settings will lead to the agent finding the shortest (i.e. least number of steps) to the goal? Note that a trajectory ends if the agent reaches the goal.

Options :

6406532332926. ✔ +10 reward for reaching the goal, -1 reward for taking each step.

6406532332927. ✔ 0 reward for reaching the goal, -1 reward for taking each step.

6406532332928. ✖ 10 reward for reaching the goal, 0 reward for taking each step.

6406532332929. ✔ +1 reward for reaching the goal, discounting factor (γ) is set to 0.9.

6406532332930. ✔ +1 reward for reaching the goal, discounting factor (γ) is set to 0.1.

6406532332931. ✖ +1 reward for reaching the goal, discounting factor (γ) is set to 0.

Question Number : 132 Question Id : 640653698564 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1 Max. Selectable Options : 0

Question Label : Multiple Select Question

What are the advantages of policy search methods over other approaches?

Options :

6406532332982. ✔ They can lead to simpler solution description.

6406532332983. ✔ They offer better convergence as compared to function approximation based methods.

6406532332984. ✔ In continuous action setting, they work better than value function based approaches.

6406532332985. ✔ They are robust to partial observability.

6406532332986. ✖ None of these.

Question Number : 133 Question Id : 640653698568 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1 Max. Selectable Options : 0

Question Label : Multiple Select Question

Consider following update rule regarding an actor-critic algorithm and choose the correct options:

$$\theta_{t+1} \leftarrow \theta_t + \alpha(G_{t:t+1} - \hat{v}(S_t, w))\frac{\nabla\pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$$

Options :

6406532333000. ✔ θ represents the policy parameters.

6406532333001. ✔ $\hat{v}(S_t, w)$ is the critic.

6406532333002. ✖ θ is the critic.

6406532333003. ✖ $\hat{v}(S_t, w)$ actor.

6406532333004. ✖ None of these

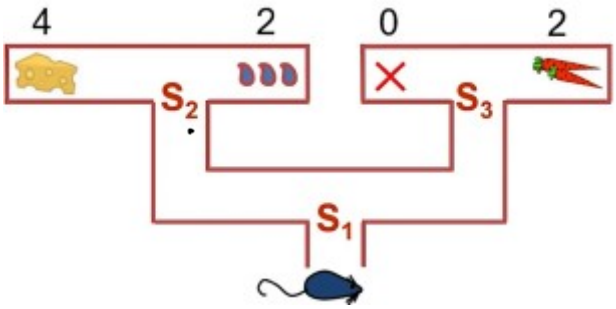
Sub-Section Number :	4
Sub-Section Id :	640653103514
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 134 Question Id : 640653698548 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Short Answer Question

Consider following diagram: There are only two actions from each state, the $\pi(s_i, left) = 0.4$ and $\pi(s_i, right) = 0.6$ for $i \in \{1, 2, 3\}$. The scalar rewards are mentioned at the terminating states. All transitions are deterministic and discounting factor($\gamma = 1$). What is $v_\pi(s_1)$?



2.8
x 4
11 2

0.6 x (0.4x0 + 0.6x2)
0.6 x 1.2
0.72

$v_{\pi}(s_1) = 0.4 \times (0.4 \times 4 + 0.6 \times 2)$
$= 0.4 \times (1.6 + 1.2)$
$= 0.4 \times 2.8$
$= 1.12 +$
0.72
1.84

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

1.80 to 1.85

Sub-Section Number :	5
Sub-Section Id :	640653103515
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 135 Question Id : 640653698551 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

In every visit Monte Carlo methods, multiple samples for one state are obtained from a single trajectory. Which of the following is true?

Options :

6406532332942. ✔ There is an increase in bias of the estimates.

6406532332943. ✖ There is an increase in variance of the estimates.

6406532332944. ✖ It does not affect the bias or variance of estimates.

6406532332945. ✖ Both bias and variance of the estimates increase.

6406532332946. ✖ None of these.

Question Number : 136 Question Id : 640653698559 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is the correct formula for computing G_t^λ?

Options :

6406532332960. ✔ $G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^n$

6406532332961. ✖ $G_t^\lambda = (-1+\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^n$

6406532332962. ✖ $G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n} G_t^n$

6406532332963. ✖ $G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n+1} G_t^n$

6406532332964. ✖ None of these.

Question Number : 137 Question Id : 640653698560 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is the correct coefficient of G_t^{T-t-1} for computing G_t^{λ}?

Options :

6406532332965. ✖ $(1-\lambda)\lambda^{T-t-1}$

6406532332966. ✖ $(1-\lambda)\lambda^{T-t}$

6406532332967. ✖ λ^{T-t-1}

6406532332968. ✖ $(1-\lambda)\lambda^{T-1}$

6406532332969. ✔ $(1-\lambda)\lambda^{T-t-2}$

6406532332970. ✖ None of these.

Question Number : 138 Question Id : 640653698563 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

The output layer is a fully connected linear layer in DQNs. Can we use a sigmoid or ReLU activation at the output layer instead?

Options :

6406532332978. ✔ We can't use either ReLU or sigmoid because it becomes difficult to backpropagate the errors if the last layer has a non-linear activation function.

6406532332979. ✖ We can use ReLU but not sigmoid. This is because all the Q values need not lie between 0 and 1.

6406532332980. ✖ We cannot use either ReLU or sigmoid. This is because the Q value could be any real number.

6406532332981. ✖ None of these

Question Number : 139 Question Id : 640653698566 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is the correct update rule for policy parameter θ if the policy is represented in soft-max fashion for a multi arm bandit?

Options :

6406532332991. ✖ $\Delta\theta_i = \alpha(r - \bar{r})(1 - \pi(a_i, \theta))$

6406532332992. ✖ $\Delta\theta_i = \alpha(r - \bar{r})(-\pi(a_i, \theta))$

6406532332993. ✔ $$\Delta\theta_i = \begin{cases} \alpha(r - \bar{r})(1 - \pi(a_i, \theta)), & \text{if action } a_i \text{ is chosen.} \\ \alpha(r - \bar{r})(-\pi(a_i, \theta)), & \text{otherwise.} \end{cases}$$

6406532332994. ✖ None of these.

Question Number : 140 Question Id : 640653698567 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is the correct update rule for MC with policy gradient?

Options :

6406532332995. ✔ $\theta_{t+1} \leftarrow \theta_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$

6406532332996. ✖ $\theta_{t+1} \leftarrow \theta_t - \alpha G_t \frac{\pi(A_t|S_t, \theta_t)}{\nabla \pi(A_t|S_t, \theta_t)}$

6406532332997. ✖ $\theta_{t+1} \leftarrow \theta_t + \alpha G_t \frac{\nabla \pi(S_t|A_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$

6406532332998. ✖ $\theta_{t+1} \leftarrow \theta_t + \alpha R_t \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$

6406532332999. ✖ None of these

Question Number : 141 Question Id : 640653698570 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Consider a grid world. The agent is in state s and it can take either of the 4 actions $up, left, right, down$ according to policy π. Consider following table and choose the best action:

Action	$q_\pi(s,a)$	$A_\pi(s,a)$
left	0.25	0.11
right	0.68	0.89
up	0.33	0.88
down	0.89	2.8

Options :

6406532333006. ✖ Left

6406532333007. ✖ Right

6406532333008. ✖ Up

6406532333009. ✖ Down

6406532333010. ✔ Insufficient information

6406532333011. ✖ None of these.

Question Number : 142 Question Id : 640653698571 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is correct Q-learning update rule with options ? Make suitable assumptions, the symbols have usual meaning. *s* is the state encountered at timestep *t*.

Options :

6406532333012. ✖ $Q(s,o) \leftarrow Q(s,o) - \alpha[\bar{r} + \gamma^{t+\tau} max_a Q(s',a) - Q(s,o)]$

6406532333013. ✖ $Q(s,o) \leftarrow Q(s,o) - \alpha[\bar{r} + \gamma^{\tau} max_a Q(s',a) - Q(s,o)]$

6406532333014. ✔ $Q(s,o) \leftarrow Q(s,o) - \alpha[\bar{r}_{t+\tau} + \gamma max_a Q(s',a) - Q(s,o)]$

6406532333015. ✖ $Q(s,o) \leftarrow Q(s,o) - \alpha[\bar{r}_{t+\tau} + \gamma^{\tau} max_a Q(s',a) - Q(s,o)]$

6406532333016. ✖ $Q(s,o) \leftarrow Q(s,o) - \alpha[\bar{r} + \gamma max_a Q(s',a) - Q(s,o)]$

6406532333017. ✖ None of these

Question Number : 143 Question Id : 640653698572 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following is a markov option? Assume the agent is traversing a grid world.

Options :

6406532333018. ✔ In state s_1 move 1 step right, in $s_2 m$ move down.

6406532333019. ✖ In state s_1 move 1 step right, take 10 steps left.

6406532333020. ✖ In state s_1 take 10 steps left, move 1 step right.

6406532333021. ✖ In state s_1 take 10 steps left, move 1 step right, in s_2 move 1 step up.

6406532333022. ✖ None of these.

Question Number : 144 Question Id : 640653698573 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

In SMDP, consider the case when τ is fixed for all state, action pairs. Will we always get the same policy for conventional Q-learning and SMDP Q learning then? Provide answer for the three cases when $\tau = 3$, $\tau = 2$, $\tau = 1$.

Options :

6406532333023. ✖ yes, yes, no

6406532333024. ✖ no, no, no

6406532333025. ✖ yes, yes, yes

6406532333026. ✔ no, no, yes

Question Number : 145 Question Id : 640653698574 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Consider following steps involved in Monte Carlo Tree Search:

A Simulation

B Backup

C Expansion

D Selection

D C A B

Organize the above steps in their chronological order:

Options :

6406532333027. ✖ ABCD

6406532333028. ✔ DCAB

6406532333029. ✖ BCAD

6406532333030. ✖ CBAD

6406532333031. ✖ None of these

Question Number : 146 Question Id : 640653698575 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

Consider following assertion and reason pair in context of MCTS:
Assertion: The goal of a rollout algorithm is not to estimate a complete optimal action-value function, q^*, or a complete action-value function, q_π, for a given policy π.
Reason: Rollout algorithm produces Monte Carlo estimates of action values for each current state and for a given policy usually called the rollout policy.

Options :

6406532333032. ✔ Assertion and Reason are both true and Reason is a correct explanation of Assertion.

6406532333033. ✖ Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

6406532333034. ✖ Assertion is true and Reason is false

6406532333035. ✖ Both Assertion and Reason are false.

Question Number : 147 Question Id : 640653698576 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Multiple Choice Question

UCT is different from the original version of MCTS in which of the following steps?

Options :

6406532333036. ✖ Expansion

6406532333037. ✔ Selection

6406532333038. ✖ Backup

6406532333039. ✖ Simulation

Sub-Section Number :	6
Sub-Section Id :	640653103516
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 148 Question Id : 640653698552 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Which of the following statements are **FALSE** about solving MDPs using dynamic programming?

Options :

6406532332947. ✖ If the state space is large or computation power is limited, it is preferred to update only some states through random sampling or selecting states seen in trajectories.

6406532332948. ✔ Knowledge of transition probabilities is not necessary for solving MDPs using dynamic programming.

6406532332949. ✔ Methods that update only a subset of states at a time guarantee performance equal to or better than classic DP.

6406532332950. ✖ DP methods bootstrap but do not sample.

Question Number : 149 Question Id : 640653698562 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Which of the following are on policy algorithms?

Options :

6406532332972. ✔ SARSA

6406532332973. ✔ Expeted SARSA

6406532332974. ✖ Q-Learning

6406532332975. ✖ DQN

6406532332976. ✖ DDQN

6406532332977. ✖ None of these

Sub-Section Number : 7

Sub-Section Id : 640653103517

Question Shuffling Allowed : No

Is Section Default? : null

Question Id : 640653698553 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Question Numbers : (150 to 153)

Question Label : Comprehension

You are training an agent with TD(λ) algorithm. There are total 10 states s_i for $i \in [0, 8]$ and a terminal state s_T. Following is the first trajectory the agent observes:

$$state = s_0 \rightarrow action = a_0 \rightarrow reward = 0 \rightarrow$$

$$state = s_1 \rightarrow action = a_1 \rightarrow reward = 0 \rightarrow$$

$$state = s_1 \rightarrow action = a_2 \rightarrow reward = 0 \rightarrow$$

$$state = s_2 \rightarrow action = a_3 \rightarrow reward = +10 \rightarrow state = s_T$$

Assuming following:

- discount factor($\gamma = 1.0$) .
- Lambda ($\lambda = 0.8$).
- Learning rate ($\alpha = 1.0$).
- $V(s)$ is initialized to 0 $\forall s$.
- s_T is terminal state.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 150 Question Id : 640653698554 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Short Answer Question

What will be the eligibility trace of state s_5 i.e. $e(s_5)$ once the episode concludes but before the next episode begins?

$S_5 \rightarrow 0$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

Question Number : 151 Question Id : 640653698555 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Short Answer Question

What will be the eligibility trace of state s_0 i.e. $e(s_0)$ once the episode concludes but before the next episode begins?

$S_0 \rightarrow 0 + (0.8)^3$

$= 0.512$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.512

Question Number : 152 Question Id : 640653698556 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Short Answer Question

What will be the eligibility trace of state s_1 i.e. $e(s_1)$(if eligibility trace is accumulating), once the episode concludes but before the next episode begins?

$S_1 \rightarrow 0.8 + 1 = 1.8$
$\times 0.8$
$= 1.44$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1.44

Question Number : 153 Question Id : 640653698557 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Short Answer Question

What will be the eligibility trace of state s_1 i.e. $e(s_1)$ (if eligibility trace is replaced for a state encountered in an episode), once the episode concludes but before the next episode begins?

$0.8 + 1 = 1.8$
0.8

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.8

Sub-Section Number : 8

Sub-Section Id : 640653103518

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 154 Question Id : 640653698561 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Short Answer Question

If you have to train a DQN for an Atari-like game with only 10 actions, how many neurons would the final layer have?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

10

Question Number : 155 Question Id : 640653698569 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2

Question Label : Short Answer Question

Consider following steps for one iteration of a basic actor-critic algorithm:

1. Update the parameter w using data $(s, r + \gamma\hat{v}(s', w))$
2. $\theta \leftarrow \theta + \alpha\hat{\delta}(s, a) \cdot \nabla_\theta log\pi_\theta(a|s)$
3. Take action $a \sim \pi_\theta(a|s)$ and receive (s, a, s', r')
4. Compute $\hat{\delta}(s, a) = r + \gamma\hat{v}(s', w)\hat{v}(s, w)$

What is the correct order of the steps? If you think the correct order is step 1, step 2, step 3 and step 4, then enter "1234" (without quotes and spaces).

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

3142

Advanced Algorithms

Section Id :	64065349296
Section Number :	7
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	11
Number of Questions to be attempted :	11
Section Marks :	50
Display Number Panel :	Yes
Section Negative Marks :	0
Group All Questions :	No
Enable Mark as Answered Mark for Review and Clear Response :	Yes
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	640653103519
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 156 Question Id : 640653698577 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 0

Question Label : Multiple Choice Question

THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL : ADVANCED ALGORITHMS (COMPUTER BASED EXAM)"