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 :

6406533044690. ✔ YES

6406533044691. ✖ NO

Question Number : 148 Question Id : 640653904191 Question Type : MCQ Calculator : Yes
Correct Marks : 0

Question Label : Multiple Choice Question

Note:

For numerical answer type questions, enter your answer correct upto two decimal places without rounding up or off unless stated otherwise.

Options :

6406533044692. ✔ Instructions has been mentioned above.

6406533044693. ✖ This Instructions is just for a reference & not for an evaluation.

Sub-Section Number :	2
Sub-Section Id :	640653134121
Question Shuffling Allowed :	Yes

Question Number : 149 Question Id : 640653904192 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

Consider following assertion reason pair:

Assertion: Reinforcement learning is a type of unsupervised learning algorithm as both don't have correct labels.
Reason: In unsupervised learning, a reward like quantity is not maximized.

Options :

6406533044694. ✖ Assertion and Reason are both true and Reason is a correct explanation of Assertion.

6406533044695. ✖ Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

6406533044696. ✖ Assertion is true but Reason is false.

6406533044697. ✔ Assertion is false but Reason is true.

Question Number : 150 Question Id : 640653904193 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Which of these statements is true regarding the rewards obtained in an MDP?

Options :

6406533044698. ✖ The reward r_{t+1} obtained on choosing an action a_t depends only on a_t .

6406533044699. ✖ The reward r_{t+1} obtained on choosing an action a_t depends only on s_t and a_t .

6406533044700. ✔ The reward r_{t+1} obtained on choosing an action a_t depends only on s_t, a_t and s_{t+1} .

6406533044701. ✖ None of these

Question Number : 151 Question Id : 640653904196 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider a reinforcement learning agent trying to balance a pole in a continuous environment. The agent receives a reward of +1 for each time step the pole remains balanced and 0 otherwise. Which of the following statements accurately describes the differences between Monte Carlo and Temporal Difference (TD) learning in this scenario?

Options :

6406533044712. ✔ Monte Carlo methods update the value function based on complete episodes, while TD methods update after each step.

6406533044713. ✖ TD methods are guaranteed to converge to the optimal policy, while Monte Carlo methods may not converge.

6406533044714. ✖ Monte Carlo methods are less sensitive to the choice of the discount factor compared to TD methods.

6406533044715. ✖ TD methods are more effective in environments with high variance and stochasticity compared to Monte Carlo methods.

6406533044716. ✖ None of these.

Question Number : 152 Question Id : 640653904197 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider a reinforcement learning agent learning to control a robotic arm in a simulated environment. The agent receives a reward of +1 for successfully placing an object in target location and 0 otherwise. Which of the following statements accurately describes a difference between SARSA and Q-learning in this scenario?

Options :

6406533044717. ✔ SARSA updates its action-value function using the action taken in the next state, while Q-learning updates its action-value function using the maximum action-value across all possible actions in the next state.

6406533044718. ✖ SARSA is less sensitive to the choice of policy compared to Q-learning, making it more robust in environments with frequent changes.

6406533044719. ✖ SARSA converges more quickly than Q-learning in environments with high reward variance.

6406533044720. ✖ Q-learning is inherently more computationally efficient than SARSA due to fewer updates required per episode.

6406533044721. ✖ None of these.

Question Number : 153 Question Id : 640653904199 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

In Q-learning, what is the impact of maximization bias on the algorithm's performance, especially in environments with noisy or stochastic rewards?

Options :

6406533044723. ✔ Maximization bias can lead to an overestimation of action values, causing the algorithm to favor suboptimal actions and potentially delaying convergence to the optimal policy.

6406533044724. ✖ Maximization bias tends to enhance the algorithm's performance by consistently selecting actions with higher estimated values, leading to quicker convergence.

6406533044725. ✖ Maximization bias may result in excessive exploration, allowing the algorithm to discover better strategies in environments with highly variable rewards.

6406533044726. ✖ Maximization bias typically has a minor effect on the performance of Q-learning, as the exploration process naturally mitigates its influence over time.

6406533044727. ✖ None of these.

Question Number : 154 Question Id : 640653904201 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider the following scenario in Double Q-Learning:

- You have two Q-value functions, Q_A and Q_B.
- The agent is in state s and takes action a, receiving reward r and transitioning to state s'.
- The update step involves using both Q_A and Q_B.

Which of the following best describes the update rule for $Q_A(s,a)$ in Double Q-Learning?

Options :

6406533044729. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma \max_{a'} Q_A(s',a') - Q_A(s,a)\right]$

6406533044730. ✔ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma Q_A(s', \arg\max_{a'} Q_B(s',a')) - Q_A(s,a)\right]$

6406533044731. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma Q_B(s', \arg\max_{a'} Q_A(s',a')) - Q_A(s,a)\right]$

6406533044732. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma \max_{a'} Q_B(s',a') - Q_A(s,a)\right]$

6406533044733. ✖ None of these

Question Number : 155 Question Id : 640653904203 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

How is the Q-value $Q(s,a)$ computed in the dueling architecture?

Options :

6406533044735. ✖ $Q(s,a) = V(s) + A(s,a)$

6406533044736. ✖ $Q(s,a) = V(s) + A(s,a) - \max A(s,a')$

6406533044737. ✔ $Q(s,a) = V(s) + A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a')$

6406533044738. ✖ $Q(s,a) = V(s) + A(s,a) - \min A(s,a')$

6406533044739. ✖ None of these

Question Number : 156 Question Id : 640653904204 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What key issue in standard DQN does the dueling architecture seek to address?

Options :

6406533044740. ✖ The problem of ensuring faster convergence in environments with high-dimensional state spaces.

6406533044741. ✖ The challenge of distinguishing between actions in environments with a large number of potential moves.

6406533044742. ✔ The inefficiency in accurately evaluating actions that do not significantly affect the overall value of the state.

6406533044743. ✖ The difficulty in learning effective value functions when rewards are sparse and infrequent.

6406533044744. ✖ None of these.

Question Number : 157 Question Id : 640653904207 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following best captures the key difference between policy gradient methods and Q-learning in reinforcement learning?

Options :

6406533044747. ✔ Policy gradient methods directly optimize the policy by computing gradients of the expected reward, while Q-learning indirectly improves the policy by learning a value function that estimates the expected rewards for actions.

6406533044748. ✖ Q-learning focuses on optimizing the policy parameters directly, while policy gradient methods estimate the action-value function to guide policy improvement.

6406533044749. ✖ Policy gradient methods require a model of the environment's dynamics to compute gradients, while Q-learning does not rely on any such model.

6406533044750. ✖ Q-learning is used primarily for continuous action spaces, while policy gradient methods are better suited for discrete action spaces.

6406533044751. ✖ None of these.

Question Number : 158 Question Id : 640653904208 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider following assertion reason pair:

Assertion: A3C can provide better performance compared to A2C, provided the updates are small enough.
Reason: In A2C, if a thread runs for a long time, other threads have to wait for it to finish.

Options :

6406533044752. ✔ Assertion and Reason are both true and Reason is a correct explanation of Assertion.

6406533044753. ✖ Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

6406533044754. ✖ Assertion is true but Reason is false.

6406533044755. ✖ Assertion is false but Reason is true.

Question Number : 159 Question Id : 640653904209 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What role do "meta-policies" play in Hierarchical Reinforcement Learning?

Options :

6406533044756. ✖ They define the specific actions to take in each state.

6406533044757. ✔ They manage the high-level decisions and selection of subtasks.

6406533044758. ✖ They directly compute the rewards for each subtask.

6406533044759. ✖ They optimize the low-level action policies.

6406533044760. ✖ None of these.

Question Number : 160 Question Id : 640653904210 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

In HRL, what is a ”subtask” typically used for?

Options :

6406533044761. ✖ To evaluate the performance of the overall policy

6406533044762. ✖ To execute the final decision made by the high-level policy

6406533044763. ✔ To break down complex tasks into more manageable components

6406533044764. ✖ To directly interact with the environment and collect rewards

6406533044765. ✖ None of these.

Question Number : 161 Question Id : 640653904211 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

How does HRL handle long-term dependencies in tasks?

Options :

6406533044766. ✖ By using recurrent neural networks (RNNs)

6406533044767. ✖ By focusing on immediate rewards only

6406533044768. ✔ By leveraging hierarchical structures to manage dependencies

6406533044769. ✖ By reducing the state space through dimensionality reduction

6406533044770. ✖ None of these.

Question Number : 162 Question Id : 640653904212 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What is a common challenge when implementing HRL in practice?

Options :

6406533044771. ✔ Finding suitable reward functions for all levels

6406533044772. ✖ Scaling the approach to very large state spaces

6406533044773. ✖ Ensuring the subtasks are independent of each other

6406533044774. ✖ Integrating HRL with existing non-hierarchical methods

6406533044775. ✖ None of these.

Question Number : 163 Question Id : 640653904213 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Choose the correct formula for UCT:

Options :

6406533044776. ✖ $\text{argmax}_{a \in A(s)} \left[Q(s,a) - 2c\sqrt{\frac{2lnN(s)}{N(s,a)}} \right]$

6406533044777. ✔ $\text{argmax}_{a \in A(s)} \left[Q(s,a) + 2c\sqrt{\frac{2lnN(s)}{N(s,a)}} \right]$

6406533044778. ✖ $\text{argmax}_{a \in A(s)} \left[Q(s,a) + 2c\sqrt{\frac{2lnN(s,a)}{N(s)}} \right]$

6406533044779. ✖ $\max_{s'} \left[v(s') + 2c\sqrt{\frac{2lnN(s')}{N(s',a)}} \right]$

6406533044780. ✖ None of these.

Sub-Section Number : 3

Sub-Section Id : 640653134122

Question Shuffling Allowed : Yes

Question Number : 164 Question Id : 640653904194 Question Type : MSQ Calculator : Yes

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

How many of the following are equal to $v^*(s)$? Here, s is some arbitrary state in the set S.

Options :

6406533044702. ✔ $\max_\pi v_\pi(s)$

6406533044703. ✔ $\max_a q^*(s,a)$

6406533044704. ✔ $\max_a \sum_{a,s'} p(s',r|s,a) \cdot [r + \gamma v^*(s')]$

6406533044705. ✔ $\max_a \sum_{a,s'} p(s',r|s,a) \cdot [r + \gamma \cdot \max_{a'} q^*(s',a')]$

6406533044706. ✖ None of these.

Question Number : 165 Question Id : 640653904195 Question Type : MSQ Calculator : Yes Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Select the correct statements about Generalized Policy Iteration (GPI).

Options :

6406533044707. ✔ GPI lets policy evaluation and policy improvement interact with each other regardless of the details of the two processes.

6406533044708. ✔ At the end of evaluation, the policy is not greedy with respect to the value function computed

6406533044709. ✔ GPI converges only when a policy has been found which is greedy with respect to its own value function.

6406533044710. ✔ The policy and value function found by GPI at convergence will both be optimal.

6406533044711. ✖ None of these.

Sub-Section Number : 4

Sub-Section Id : 640653134123

Question Shuffling Allowed : Yes

Question Number : 166 Question Id : 640653904198 Question Type : SA Calculator : None Correct Marks : 2

Question Label : Short Answer Question

Consider a Q-learning algorithm with a learning rate (α) of 0.2 and a discount factor (γ) of 0.85. If the current Q-value $Q(s, a)$ is 8, the reward R received is 6, and the maximum Q-value for the next state $\max_{a'} Q(s', a')$ is 14, what is the updated Q-value after taking the action a in state s?

$Q(s,a) = 8 + 0.2[6 + 0.85 \times 14 - 8]$

$= 8 + 0.2[6 + 11.9 - 8]$

$= 8 + 0.2[9.9]$

$= 8 + 1.98 = 9.98$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

9.93 to 10.3

Question Number : 167 Question Id : 640653904206 Question Type : SA Calculator : None Correct Marks : 2

Question Label : Short Answer Question

Consider a binary bandit problem where the action a can be either 0 or 1.
The policy is parameterized by θ and is represented using a Bernoulli distribution with probability $p = \pi_\theta(a = 1)$. Suppose the reward $R(a)$ for action a is defined as follows:

- $R(a = 1) = 1$
- $R(a = 0) = 10$

If the current policy parameter θ results in $p = 0.7$, and the reward obtained for action $a = 1$ is 5, what is the policy gradient for this parameter setting?

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes
Answers Type : Range
Text Areas : PlainText
Possible Answers :
1.35 to 1.45

$9 \times 0.7 \times 0.3$
9×0.21
$= 1.89$

Sub-Section Number : 5
Sub-Section Id : 640653134124
Question Shuffling Allowed : Yes

Question Number : 168 Question Id : 640653904200 Question Type : SA Calculator : None Correct Marks : 4
Question Label : Short Answer Question

In a Q-learning algorithm, you are using an ϵ-greedy policy to choose actions.
You observe the following Q-values for a state s at a given time:

$Q(s, a_1) = 12$
$Q(s, a_2) = 14$
$Q(s, a_3) = 17$

$\frac{0.2}{3} = 0.06$
$\frac{0.2}{3} = 0.06$
$0.2/3 + 0.8 = 0.866$

Suppose that $\epsilon = 0.2$, meaning that with 20% probability, a random action is selected and with 80% probability, the action with the highest Q-value is chosen. Due to the maximization bias, if the estimated Q-values are consistently overestimated, how much greater is the expected value of the action selection due to the bias, given that the true Q-values are known to be 10, 12, and 14 for a_1, a_2, and a_3 respectively?

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes
Answers Type : Range
Text Areas : PlainText
Possible Answers :
2.85 to 2.95

$E[R] = 0.792 + 0.924 + 14.722$
$= 16.438$
$E[R] = 0.66 + 0.792 + 12.124$
$= 13.576$
$= 2.862$

Question Number : 169 Question Id : 640653904202 Question Type : SA Calculator : None
Correct Marks : 4

Question Label : Short Answer Question

Consider an agent using linear function approximation to estimate the value function in a reinforcement learning problem. The value function $V(s)$ is approximated as $\hat{V}(s;\mathrm{w}) = \mathrm{w}^\top \mathrm{x}(s)$, where w is the weight vector and $\mathrm{x}(s)$ is the feature vector for state s.

Given the following parameters and observations:

- Current weight vector: $\mathrm{w} = [0.5, -0.2, 0.3]$
- Feature vector for state s: $\mathrm{x}(s) = [1, 2, 3]$
- Observed reward: $r = 4$
- Discount factor: $\gamma = 0.9$
- Learning rate: $\alpha = 0.2$
- Next state feature vector: $\mathrm{x}(s') = [2, 0, 1]$
- Value estimate for next state: $\hat{V}(s'; \mathrm{w}) = \mathrm{w}^\top \mathrm{x}(s')$

$V(s, w) = 0.5 - 0.4 + 0.9$
$= 1.0$

$V(s', w) = 1 - 0 + 0.3 = 1.3$

$TD = r + \gamma V(s', w) - V(s, w)$
$= 4 + 0.9 \times 1.3 - 1$
$= 4.17$

$W_{new} = W + \alpha \delta_T x(s)$

Calculate the updated weight vector w after one step of gradient descent using the TD(0) update rule and enter its L1 norm.

$= [0.5, -0.2, 0.3]$
$+ 0.2 \times 4.17 [1, 2, 3]$
$0.834 [1, 2, 3]$

$= 0.5, -0.2, 0.3$
$+ \quad 0.834, 1.668, 2.502$
$1.334, 1.468, 2.802$

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes
Answers Type : Range
Text Areas : PlainText
Possible Answers :
4.3 to 4.4

5.604

Question Number : 170 Question Id : 640653904205 Question Type : SA Calculator : None
Correct Marks : 4

Question Label : Short Answer Question

Consider a simple bandit problem where the action a is continuous and the policy is parameterized by θ. The policy $\pi_\theta(a)$ is given by a Gaussian distribution with mean μ_θ and fixed standard deviation σ. Assume the reward R for an action a is $R(a) = -(a-3)^2 + 5$. If the current policy parameter θ results in $\mu_\theta = 2$ and $\sigma = 1$, what is the policy gradient for this parameter setting given the action $a = 4$ and the corresponding reward $R = 4$?

$R(a) = -(a-3)^2 + 5$

$\nabla_\theta J(\theta) = R \times (a - 2)$
$= 4 \times (4 - 2)$
$= 4 \times 2 = 8$

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes
Answers Type : Range

Text Areas : PlainText

Possible Answers :

7.95 to 8.05

Intro to Big Data

Section Id :	64065364136
Section Number :	8
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	22
Number of Questions to be attempted :	22
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 :	No
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	640653134125
Question Shuffling Allowed :	No

Question Number : 171 Question Id : 640653904214 Question Type : MCQ Calculator : Yes

Correct Marks : 0

Question Label : Multiple Choice Question

THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL : INTRODUCTION TO BIG DATA (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 :

6406533044781. ✔ YES

6406533044782. ✖ NO

Sub-Section Number :	2
Sub-Section Id :	640653134126
Question Shuffling Allowed :	Yes

Question Number : 172 Question Id : 640653904215 Question Type : MSQ Calculator : Yes