RL

Section Id :	64065364225
Section Number :	7
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	24
Number of Questions to be attempted :	24
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 :	640653134770
Question Shuffling Allowed :	No

Question Number : 161 Question Id : 640653906888 Question Type : MCQ Calculator : Yes
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 :

6406533052571. ✔ YES

6406533052572. ✖ NO

Question Number : 162 Question Id : 640653906889 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 :

6406533052573. ✔ Instructions has been mentioned above.

6406533052574. ✖ This Instructions is just for a reference & not for an evaluation.

Sub-Section Number : 2
Sub-Section Id : 640653134771
Question Shuffling Allowed : Yes

Question Number : 163 Question Id : 640653906890 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider the 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 :

6406533052575. ✖ Assertion and Reason are both true and Reason is a correct explanation of Assertion.

6406533052576. ✖ Assertion and Reason are both true and Reason is not a correct explanation of Assertion.

6406533052577. ✖ Assertion is true but Reason is false.

6406533052578. ✔ Assertion is false but Reason is true.

Question Number : 164 Question Id : 640653906891 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

If $q_\pi(s, a) = 1.5$ for some state s in an MDP, which of the following statements is always true?

Options :

6406533052579. ✖ The expected return starting from state s and following policy π is equal to 1.5.

6406533052580. ✔ The expected return starting from state s, taking action a and then following policy π is equal to 1.5.

6406533052581. ✖ The return starting from state s is equal to 1.5 in some episode.

6406533052582. ✖ The return starting from state s, taking action a is equal to 1.5 in some episode.

6406533052583. ✖ None of these

Question Number : 165 Question Id : 640653906894 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider a reinforcement learning agent navigating a grid world environment. The agent receives rewards of +1 for reaching the goal state and 0 otherwise. Which of the following statements accurately describes the differences between Monte Carlo and TD learning in this scenario?

Options :

6406533052593. ✔ Monte Carlo updates are unbiased estimators of the true value function, while TD updates may introduce bias.

6406533052594. ✖ TD updates are guaranteed to converge to the optimal value function, while Monte Carlo updates may not converge.

6406533052595. ✖ Monte Carlo updates require less memory and computational resources compared to TD updates.

6406533052596. ✖ TD updates are more robust to noise and stochasticity in the environment compared to Monte Carlo updates.

6406533052597. ✖ None of these

Question Number : 166 Question Id : 640653906895 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Consider a reinforcement learning agent learning to navigate a grid world environment. The agent receives rewards of +1 for reaching the goal state and 0 otherwise. Which of the following statements accurately describes a difference between SARSA and Q-learning in this scenario?

Options :

6406533052598. ✔ SARSA updates its action-value function based on the action actually taken in the next state, while Q-learning updates its action-value function based on the maximum action-value in the next state.

6406533052599. ✖ SARSA is guaranteed to converge to the optimal policy under certain conditions, while Q-learning may diverge or oscillate without additional modifications.

6406533052600. ✖ SARSA is more computationally efficient than Q-learning, requiring fewer updates to converge to the optimal policy.

6406533052601. ✖ SARSA and Q-learning exhibit similar performance in terms of convergence speed and solution quality in this scenario.

6406533052602. ✖ None of these

Question Number : 167 Question Id : 640653906897 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

In Q-learning, how does maximization bias affect the performance of the algorithm in complex environments?

Options :

6406533052604. ✔ Maximization bias can lead to overestimation of action values, resulting in suboptimal policies and slower convergence to the optimal policy.

6406533052605. ✖ Maximization bias helps to accelerate learning by prioritizing actions with higher estimated values, leading to faster convergence to the optimal policy.

6406533052606. ✖ Maximization bias reduces the exploration-exploitation trade-off, resulting in more exploratory behavior and improved generalization to unseen states.

6406533052607. ✖ Maximization bias has minimal impact on the performance of Q-learning in complex environments, as it tends to balance out over time through exploration.

6406533052608. ✖ None of these

Question Number : 168 Question Id : 640653906899 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 :

6406533052610. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma \max_{a'} Q_A(s',a') - Q_A(s,a)\right]$

6406533052611. ✔ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma Q_A(s', \arg\max_{a'} Q_B(s',a')) - Q_A(s,a)\right]$

6406533052612. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma Q_B(s', \arg\max_{a'} Q_A(s',a')) - Q_A(s,a)\right]$

6406533052613. ✖ $Q_A(s,a) \leftarrow Q_A(s,a) + \alpha\left[r + \gamma \max_{a'} Q_B(s',a') - Q_A(s,a)\right]$

6406533052614. ✖ None of these

Question Number : 169 Question Id : 640653906901 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 :

6406533052616. ✖ $Q(s,a) = V(s) + A(s,a)$

6406533052617. ✖ $Q(s,a) = V(s) + A(s,a) - \max A(s,a')$

6406533052618. ✔

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

6406533052619. ✖ $Q(s,a) = V(s) + A(s,a) - \min A(s,a')$

6406533052620. ✖ None of these

Question Number : 170 Question Id : 640653906902 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What problem does the dueling architecture aim to address that is commonly faced by standard DQN?

Options :

6406533052621. ✖ The inability to handle large action spaces

6406533052622. ✖ The difficulty of learning value functions when rewards are sparse

6406533052623. ✖ The slow convergence rate of policy optimization

6406533052624. ✔ The inefficiency in estimating Q-values for actions that have little impact on the overall state value

6406533052625. ✖ None of these

Question Number : 171 Question Id : 640653906905 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following statements best describes the primary difference between policy gradient methods and value function methods?

Options :

6406533052628. ✖ Policy gradient methods directly optimize the value function, while value function methods optimize the policy parameters.

6406533052629. ✔ Value function methods approximate the policy by learning the value function and deriving the policy from it, while policy gradient methods directly optimize the policy parameters by computing gradients of expected rewards.

6406533052630. ✖ Policy gradient methods approximate the value function and use it to optimize the policy, while value function methods directly optimize the policy using gradients of the log-probability of actions.

6406533052631. ✖ Value function methods directly optimize the policy parameters, while policy gradient methods focus on approximating the value function.

6406533052632. ✖ None of these

Question Number : 172 Question Id : 640653906907 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What is a major challenge that Hierarchical Reinforcement Learning (HRL) aims to address?

Options :

6406533052638. ✖ Overfitting to specific environments

6406533052639. ✔ The curse of dimensionality in large state spaces

6406533052640. ✖ Lack of exploration in early stages

6406533052641. ✖ Handling non-stationary environments

6406533052642. ✖ None of these

→ but create sub-sub problems and solving them optimally

Question Number : 173 Question Id : 640653906908 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

In HRL, what is an "option" typically composed of?

$O(\pi_o, T_o, I_o)$

Options :

6406533052643. ✖ A single action

6406533052644. ✖ A sequence of random actions

6406533052645. ✔ A policy, a termination condition, and an initiation set

6406533052646. ✖ A fixed sequence of primitive actions

6406533052647. ✖ None of these

Question Number : 174 Question Id : 640653906909 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

How does HRL typically improve learning efficiency?

Options :

6406533052648. ✖ By increasing the learning rate

6406533052649. ✖ By reducing the action space

6406533052650. ✔ By decomposing a complex task into simpler subtasks

6406533052651. ✖ By using a single policy for all tasks

6406533052652. ✖ None of these

Question Number : 175 Question Id : 640653906910 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

What distinguishes Hierarchical Reinforcement Learning from traditional reinforcement learning?

Options :

6406533052653. ✖ The use of Q-learning instead of policy gradients

6406533052654. ✔ The explicit decomposition of tasks into a hierarchy of subtasks

6406533052655. ✖ The use of continuous action spaces

6406533052656. ✖ The ability to learn in real-time environments

6406533052657. ✖ None of these

Question Number : 176 Question Id : 640653906911 Question Type : MCQ Calculator : Yes Correct Marks : 2

Question Label : Multiple Choice Question

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

Options :

6406533052658. ✖ Expansion

6406533052659. ✔ Selection

6406533052660. ✖ Backup

6406533052661. ✖ Simulation

6406533052662. ✖ None of these

UCT = UCB + MCTS

Sub-Section Number : 3

Sub-Section Id : 640653134772

Question Shuffling Allowed : Yes

Question Number : 177 Question Id : 640653906892 Question Type : MSQ Calculator : Yes Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Which of the following are valid equations for $q_\pi(s,a)$?

Options :

6406533052584. ✖ $q_\pi(s,a) = \pi(a|s) \cdot v_\pi(s)$

6406533052585. ✔ $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \cdot [r + \gamma \cdot v_\pi(s')]$

6406533052586. ✖ $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \cdot [r + \gamma \cdot q_\pi(s',a)]$

6406533052587. ✔ $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \cdot [r + \gamma \cdot \sum_{a'} \pi(a'|s') \cdot q_\pi(s',a')]$

6406533052588. ✖ None of these

$v_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s,a)$

Question Number : 178 Question Id : 640653906893 Question Type : MSQ Calculator : Yes Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Which of the following statements are true with regards to Monte Carlo value approximation methods?

Options :

6406533052589. ✔ To evaluate a policy using these methods, a subset of trajectories in which all states are encountered at least once are enough to update all state-values.

6406533052590. ✖ Monte-Carlo value function approximation methods need knowledge of the full model.

6406533052591. ✔ Monte-Carlo methods update state-value estimates only at the end of an episode.

6406533052592. ✖ None of these

Question Number : 179 Question Id : 640653906906 Question Type : MSQ Calculator : Yes Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

In the context of actor-critic methods, what is the effect of replacing the return G_t with the TD target?

Options :

6406533052633. ✖ It increases the variance in the estimate of the gradient of the performance.

6406533052634. ✔ It decreases the variance in the estimate of the gradient of the performance.

6406533052635. ✔ It introduces a bias in the estimate of the gradient of the performance.

6406533052636. ✖ It doesn't introduce any bias in the estimate of the gradient of the performance.

6406533052637. ✖ None of these

Sub-Section Number :	4
Sub-Section Id :	640653134773
Question Shuffling Allowed :	Yes

Question Number : 180 Question Id : 640653906896 Question Type : SA Calculator : None Correct Marks : 2

Question Label : Short Answer Question

Consider a Q-learning algorithm with a learning rate (α) of 0.1 and a discount factor (γ) of 0.9. If the current Q-value $Q(s, a)$ is 10, the reward R received is 5, and the maximum Q-value for the next state $\max_{a'} Q(s', a')$ is 12, what is the updated Q-value after taking the action a in state s?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

10.57 to 10.59

Question Number : 181 Question Id : 640653906904 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:

- $R(a = 1) = 5$
- $R(a = 0) = 1$

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 :

7.09 to 7.19

Sub-Section Number :	5
Sub-Section Id :	640653134774
Question Shuffling Allowed :	Yes

Question Number : 182 Question Id : 640653906898 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:

- $Q(s, a_1) = 10$
- $Q(s, a_2) = 12$
- $Q(s, a_3) = 15$

Suppose that $\epsilon = 0.1$, meaning that with 10% probability, a random action is selected and with 90% 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 8, 10, and 12 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

= 2.931

Question Number : 183 Question Id : 640653906900 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: $\alpha = 0.1$
- Next state feature vector: $\mathrm{x}(s') = [2, 0, 1]$
- Value estimate for next state: $\hat{V}(s';\mathrm{w}) = \mathrm{w}^\top \mathrm{x}(s')$

$V(s,w) = w \cdot x(s)$

$= [0.5, -0.2, 0.3] \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

$= 0.5 - 0.4 + 0.9$

$= 1.0$

$V(s',w) = [0.5, -0.2, 0.3] \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$

$= 1 - 0 + 0.3 = 1.3$

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

$TD = r + \gamma V(s',w) - V(s,w)$

$= 4 + 0.9 \times 1.3 - 1$

$= 3 + 1.17 = 4.17$

$w_{new} = w + \alpha \cdot TD \cdot x(s)$

$= [0.5, -0.2, 0.3] + 0.1 \times 4.17$

$= [0.917, 0.634, 1.551]$ $[1, 2, 3]$

$Norm = 3.102$

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

3.05 to 3.15

0.917
0.634
1.551
3.102

Question Number : 184 Question Id : 640653906903 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 $R = 2$?

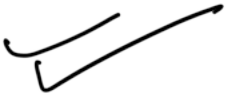
$R(a) = -(a-3)^2 + 5$

$\mu_\theta = 2, \sigma = 1$

$a = 4, R = 2$

Response Type : Numeric
Evaluation Required For SA : Yes

Show Word Count : Yes
Answers Type : Range
Text Areas : PlainText
Possible Answers :
3.95 to 4.05



$\mu_\theta(a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(a-\mu_\theta)^2}{2\sigma^2}}$ $\mu_\theta = 2$ $\sigma = 1$

$= \frac{1}{\sqrt{2\pi}} e^{\frac{-(a-2)^2}{2}}$ $\Rightarrow (a-2)$

log $\nabla\theta J_\theta = R(a) * \nabla\theta \log \pi(\theta)(a)$

$= 2 \times (4-2)$

$= 2 \times 2 = 4$

Algorithmic Thinking

Section Id :	64065364226
Section Number :	8
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	15
Number of Questions to be attempted :	15
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 :	640653134775
Question Shuffling Allowed :	No

Question Number : 185 Question Id : 640653906912 Question Type : MCQ Calculator : Yes
Correct Marks : 0
Question Label : Multiple Choice Question
THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL : ALGORITHMIC THINKING IN BIOINFORMATICS (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 :
6406533052663. ✔ YES
6406533052664. ✖ NO

Sub-Section Number :	2
Sub-Section Id :	640653134776
Question Shuffling Allowed :	Yes