6406531933041. ✔ I have written answers on the answer sheets

6406531933042. ✖ Not applicable

RL

Section Id :	64065339131
Section Number :	7
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	14
Number of Questions to be attempted :	14
Section Marks :	50
Display Number Panel :	Yes
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 :	64065382957
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 136 Question Id : 640653578969 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 :

6406531933043. ✔ YES

6406531933044. ✖ NO

Sub-Section Number :	2
Sub-Section Id :	64065382958
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 137 Question Id : 640653578970 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3

Question Label : Multiple Choice Question

Consider two policies π_1 and π_2 for a finite MDP that has 4 states. The value functions for these two policies are given below:

$$v_{\pi_1}(.) = [1.5 \quad 10 \quad -3 \quad -1]$$

$$v_{\pi_2}(.) = [1.2 \quad 9.8 \quad -3.1 \quad -1.5]$$

Which of the following statements is true?

Options :

6406531933045. ✔ $\pi_1 > \pi_2$

6406531933046. ✖ $\pi_1 < \pi_2$

6406531933047. ✖ $\pi_1 = \pi_2$

6406531933048. ✖ π_1 and π_2 cannot be compared

Question Number : 138 Question Id : 640653578973 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3

Question Label : Multiple Choice Question

Suppose an (ϵ, δ) PAC algorithm returns an arm a in a multi-armed bandit setting. What can be said about arm a?

Options :

6406531933055. ✖ The probability that the expected reward of arm a is ϵ close to the expected reward of the optimal arm is at least δ.

6406531933056. ✖ The probability that the expected reward of arm a is ϵ close to the expected reward of the optimal arm is $(1-\delta)$

6406531933057. ✔ The probability that the expected reward of arm a is ϵ close to the expected reward of the optimal arm is at least $(1-\delta)$

↳ this term makes it closer to PAC Guarantee

6406531933058. ✖ The probability that the expected reward of arm a is ϵ close to the expected reward of the optimal arm is δ

Question Number : 139 Question Id : 640653578974 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3

Question Label : Multiple Choice Question

In the context of a multi-armed bandit problem with stationary reward distributions, consider the following:

Assertion: UCB minimizes the regret better than the softmax approach.

Reason: Softmax approach assigns a low probability of picking a sub-optimal arm that has a very low expected reward.

Options :

6406531933059. ✖ Assertion and Reason are both true and Reason is a correct explanation of the Assertion.

6406531933060. ✔ Assertion and Reason are both true and Reason is not a correct explanation of the Assertion.

6406531933061. ✖ Assertion is true and Reason is false

6406531933062. ✖ Both Assertion and Reason are false.

Sub-Section Number : 3

Sub-Section Id : 64065382959

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 140 Question Id : 640653578971 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

Is the following statement true or false?

In the synchronous version of policy evaluation, two copies of the value function are used, one for step k and another for step $k+1$.

Options :

6406531933049. ✔ TRUE

6406531933050. ✖ FALSE

Sub-Section Number : 4

Sub-Section Id : 64065382960

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 141 Question Id : 640653578972 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 4

Question Label : Multiple Choice Question

Consider these statements with respect to a finite MDP:

(1) Expected return if the agent starts from state s, takes action a and then follows policy π

(2) Expected return if the agent starts from state s and behaves according to policy π

Match the statement with the corresponding value function.

Options :

6406531933051. ✖ (1): $v_\pi(s)$ (2): $q_\pi(s,a)$

6406531933052. ✔ (1): $q_\pi(s,a)$ (2): $v_\pi(s)$

6406531933053. ✖ (1): $q_\pi(s,a)$ (2): $q_\pi(s,a)$

6406531933054. ✖ (1): $v_\pi(s)$ (2): $v_\pi(s)$

Sub-Section Number : 5

Sub-Section Id : 64065382961

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 142 Question Id : 640653578975 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 4 Max. Selectable Options : 0

Question Label : Multiple Select Question

Consider a bandit problem with 4 actions labeled as $1, 2, 3, 4$. We use an ϵ-greedy strategy for action selection and sample averages for estimating the action values. The initial estimates of the action values for all actions are zero. Suppose the initial sequence of actions and rewards is:

$$\begin{array}{ll} A_1 = 1 & R_1 = -1 \\ A_2 = 2 & R_2 = 1 \\ A_3 = 2 & R_3 = -2 \\ A_4 = 2 & R_4 = 2 \\ A_5 = 3 & R_5 = 0 \end{array}$$

Here, A_t denotes the action at time step t and R_t is the corresponding reward. On some of these time steps, a non-greedy, random action could have been chosen. On which time steps did this certainly happen?

Options :

6406531933063. ✖ 1

6406531933064. ✖ 2

6406531933065. ✖ 3

6406531933066. ✔ 4

6406531933067. ✔ 5

Question Number : 143 Question Id : 640653578977 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 4 Max. Selectable Options : 0

Question Label : Multiple Select Question

Which of the following expressions evaluate to $q_*(s, a)$?

Options :

6406531933073. ✔ $\max_{\pi} \quad q_\pi(s, a)$

6406531933074. ✖ $\max_{a'} \sum_{s', r} p(s', r \mid s, a') \cdot [r + \gamma v_*(s')]$

6406531933075. ✔ $\sum_{s',r} p(s', r \mid s, a) \cdot [r + \gamma v_*(s')]$

6406531933076. ✖ $\max_a \quad q_\pi(s, a)$

Sub-Section Number :	6
Sub-Section Id :	64065382962
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 144 Question Id : 640653578976 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3 Max. Selectable Options : 0

Question Label : Multiple Select Question

Consider the following expression for a deterministic policy π:

$$\sum_{s',r} p(s', r \mid s, \pi(s)) \cdot [r + \gamma v_\pi(s')]$$

What is this equal to? Select all appropriate options.

Options :

$v_\pi(s) = q_\pi(s \mid a) = q_\pi(s \mid \pi(s))$

6406531933068. ✔ $v_\pi(s)$

6406531933069. ✖ $q_\pi(s, a)$, where a could be any arbitrary action

6406531933070. ✖ $v_*(s)$

6406531933071. ✔ $q_\pi(s, \pi(s))$

6406531933072. ✖ $q_*(s, a)$, where a could be any arbitrary action

Sub-Section Number : 7

Sub-Section Id : 64065382963

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 145 Question Id : 640653578978 Question Type : SA Calculator : None

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

Correct Marks : 4

Question Label : Short Answer Question

Consider the following statements, all of which are regarding policy iteration run on a finite MDP:

(1) Policy iteration can be used to find a deterministic optimal policy.

(2) Policy iteration is an algorithm that is exclusively used to evaluate the value function for a given policy.

(3) We can use the optimal value function output by policy iteration to find out all possible optimal policies, both deterministic and stochastic.

How many of these statements are true?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

Question Number : 146 Question Id : 640653578979 Question Type : SA Calculator : None

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

Correct Marks : 4

Question Label : Short Answer Question

The sequence of rewards for a continuing task with $\gamma = 0.9$ is given below:

$$R_1 = 1$$
$$R_2 = 2$$
$$R_3 = 3$$
$$R_t = 1, \quad t \geq 4$$

Find the return G_0. Your answer should have exactly two places after the decimal point.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

12.51 to 12.53

Question Number : 147 Question Id : 640653578980 Question Type : SA Calculator : None

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

Correct Marks : 4

Question Label : Short Answer Question

Consider a multi-armed bandit setting with 5 arms. An ϵ-greedy strategy with $\epsilon = 0.1$ is used to find the optimal arm. What is the probability of picking the optimal arm as the number of time steps tends to infinity? Your answer should have exactly two places after the decimal point.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.91 to 0.93

Sub-Section Number : 8

Sub-Section Id : 64065382964

Question Shuffling Allowed : No

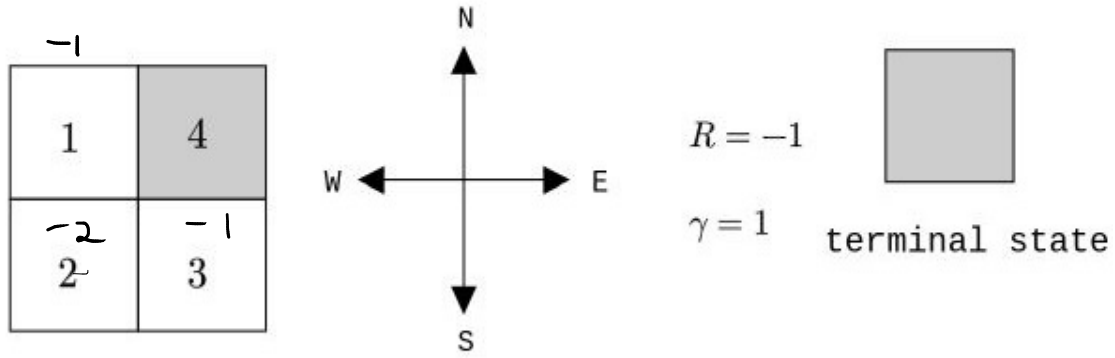
Is Section Default? : null

Question Id : 640653578981 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 : (148 to 153)

Question Label : Comprehension

Consider a 2×2 grid world with deterministic transitions. The reward is -1 on each time step. This is an episodic task with $\gamma = 1$. States are numbered as $1, 2, 3, 4$ (refer to the figure). The state 4 is a terminal state. All four actions are permitted at each non-terminal state. Actions that take an agent out of the grid-world leave the state unchanged. For instance, the action west from state 1 keeps the agent in the same state.



An equiprobable random policy is one where every action has an equal chance of being picked from each state.

NOTE: The answers to all the six sub-questions should be integers.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 148 Question Id : 640653578982 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1.5

Question Label : Short Answer Question

Let π be an equiprobable random policy. What is $v_\pi(1)$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

-6

Question Number : 149 Question Id : 640653578983 Question Type : SA Calculator : None

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

Correct Marks : 1.5

Question Label : Short Answer Question

Let π be an equiprobable random policy.

What is $v_\pi(2)$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

-8

Question Number : 150 Question Id : 640653578984 Question Type : SA Calculator : None

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

Correct Marks : 1.5

Question Label : Short Answer Question

Let π be an equiprobable random policy.

What is $v_\pi(3)$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

-6

Question Number : 151 Question Id : 640653578985 Question Type : SA Calculator : None

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

Correct Marks : 1.5

Question Label : Short Answer Question

Let π be an equiprobable random policy.

What is $v_\pi(4)$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

Question Number : 152 Question Id : 640653578986 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

If v_* is the optimal value function, compute

$v_*(1) + v_*(2) + v_*(3) + v_*(4)$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

-4

Question Number : 153 Question Id : 640653578987 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

How many deterministic optimal policies does this MDP have?

Response Type : Numeric

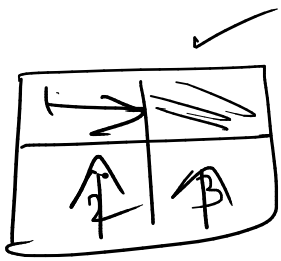
Evaluation Required For SA : Yes

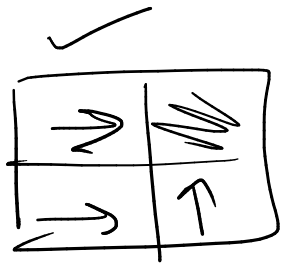
Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :





2

Sub-Section Number :	9
Sub-Section Id :	64065382965
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653578988 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 : (154 to 155)

Question Label : Comprehension

Consider a MDP for which the state set is $\mathcal{S} = \{s_1, s_2, s_3\}$ and $A = \{a_1, a_2\}$ with s_3 being a terminal state. The set of actions that can be taken are the same in both non-terminal states. The following are three trajectories experienced by an agent:

T_1: $s_2,\ a_1,\ -1,\ s_2,\ a_2,\ 2,\ s_3$

T_2: $s_1,\ a_2,\ 2,\ s_2,\ a_2,\ 3,\ s_1,\ a_2,\ 1,\ s_3$

T_3: $s_2,\ a_2,\ 4,\ s_3$

This is an episodic task with $\gamma = 1$.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 154 Question Id : 640653578989 Question Type : SA Calculator : None

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

Correct Marks : 2

$T_1 \rightarrow$ feist step s_2

$TR = -1 + 2 = 1$

Question Label : Short Answer Question

Find the estimate $V(s_2)$ for first-visit MC.

$T_2 \rightarrow 2^{nd}$ step s_2

$TR = 3 + 1 = 4$

Response Type : Numeric

Evaluation Required For SA : Yes

$T_3 \rightarrow TR = 4$

Show Word Count : Yes

Answers Type : Equal

$$\text{Feist Visit MC} = \frac{\text{Returns}}{\text{total visits}} = \frac{4}{2} = 3$$

Text Areas : PlainText

Possible Answers :

3

Question Number : 155 Question Id : 640653578990 Question Type : SA Calculator : None

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

Correct Marks : 2

$T_1 \rightarrow -1 + 2 = 1$

$2 \quad + \Rightarrow 3$

Question Label : Short Answer Question

$T_2 \Rightarrow 4$

$T_3 \Rightarrow 4$

Find the estimate $V(s_2)$ for every-visit MC. $= \frac{\text{total returns}}{\text{total no. of } S_2} = \frac{11}{4} = 2.75$

Your answer should have exactly two places after the decimal point.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

2.74 to 2.76

Statistical Computing

Section Id :	64065339132
Section Number :	8
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	10
Number of Questions to be attempted :	10
Section Marks :	35
Display Number Panel :	Yes
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 :	64065382966
Question Shuffling Allowed :	No
Is Section Default? :	null