Clear Response :	
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	64065389024
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 216 Question Id : 640653614938 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 :

6406532052806. ✓ YES

6406532052807. ✗ NO

Sub-Section Number :	2
Sub-Section Id :	64065389025
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 217 Question Id : 640653614939 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

What is the value of the following expression?

$$\sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') \cdot q_\pi(s', a')\right]$$

Options :

6406532052808. ✖ $v_\pi(s)$

6406532052809. ✔ $q_\pi(s, a)$

6406532052810. ✖ $v_*(s)$

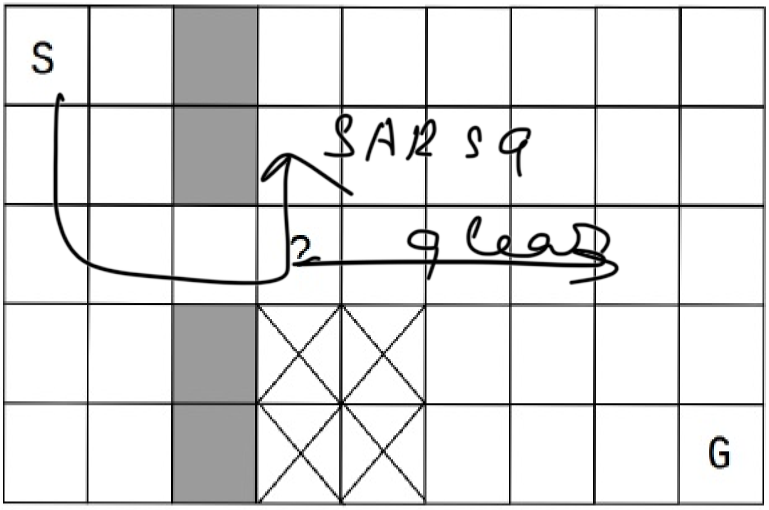
6406532052811. ✖ $q_*(s, a)$

Question Number : 218 Question Id : 640653614940 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 the following grid world where S is the start state and G is the goal. The gray cells are obstructions and the cells marked with the symbol X are bad states. Stepping into one of these bad states results in a reward of -100. Reaching the goal state results in a reward of 10. Every other transition results in a reward of -1. Note that unlike obstructions, the agent can step into bad states.



An agent is trained using SARSA. During training, the action selection strategy is ϵ-greedy with respect to the value function at every time step with a fixed value of $\epsilon = 0.1$. If an agent follows a policy that is greedy with respect to the value function learnt by SARSA, what action would it take in the state marked with the question mark? Choose the most appropriate answer.

Options :

6406532052812. ✔ UP

6406532052813. ✖ DOWN

6406532052814. ✖ LEFT

6406532052815. ✖ RIGHT

Question Number : 219 Question Id : 640653614941 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

Which of the following corresponds to an update for the actor in the case of one-step actor-critic method?

Options :

6406532052816. ✓ $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha\delta_t \frac{\nabla\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$

6406532052817. ✗ $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \frac{\nabla\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$

6406532052818. ✗ $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$

6406532052819. ✗ $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha\delta_t \nabla\pi(A_t \mid S_t, \boldsymbol{\theta}_t)$

Question Number : 220 Question Id : 640653614943 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

Which of the following is a correct Bellman equation for an SMDP? Note: $R(s, a, s') \implies$ reward is a function of only s, a and s'.

Options :

6406532052824. ✗ $V^*(s) = max_{a\in A(S)}[R(s, a, \tau, s') + \gamma^\tau P(s'|s, a)V^*(s')]$

6406532052825. ✗ $V^*(s) = max_{a\in A(S)}[\Sigma_{s',\tau} P(s'|s, a, \tau)(R(s, a, \tau, s') + \gamma V^*(s'))]$

6406532052826. ✓ $V^*(s) = max_{a\in A(S)}[\Sigma_{s',\tau} P(s', \tau|s, a)(R(s, a, \tau, s') + \gamma^\tau V^*(s'))]$

6406532052827. ✖ $V^*(s) = max_{a\in A(S)}[\Sigma_{s',\tau}P(s',\tau|s,a)(R(s,a,s')+\gamma V^*(s'))]$

Sub-Section Number : 3

Sub-Section Id : 64065389026

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 221 Question Id : 640653614944 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

Which of the following is the TD error when we use a 2-step return:

Options :

6406532052828. ✖ $V(S_{t+1}) - V(S_t)$

6406532052829. ✖ $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

6406532052830. ✖ $R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+1}) - V(S_t)$

6406532052831. ✔ $R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(S_t)$

Sub-Section Number : 4

Sub-Section Id : 64065389027

Question Shuffling Allowed : Yes

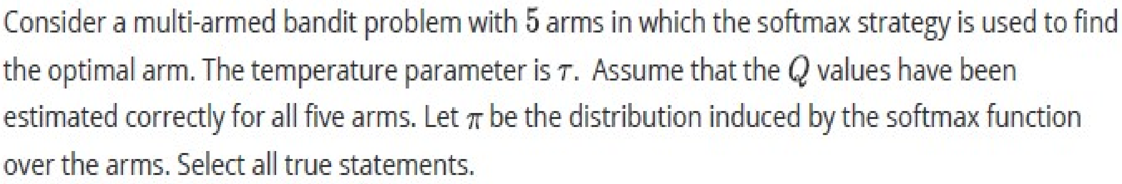
Is Section Default? : null

Question Number : 222 Question Id : 640653614945 Question Type : MSQ Is Question

Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

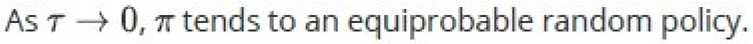
Correct Marks : 4 Max. Selectable Options : 0

Question Label : Multiple Select Question

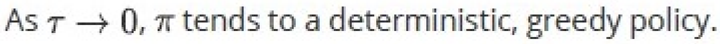


Consider a multi-armed bandit problem with 5 arms in which the softmax strategy is used to find the optimal arm. The temperature parameter is τ. Assume that the Q values have been estimated correctly for all five arms. Let π be the distribution induced by the softmax function over the arms. Select all true statements.

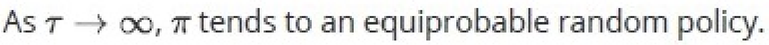
Options :

6406532052832. ✖ 

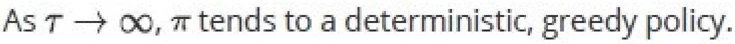
As $\tau \rightarrow 0$, π tends to an equiprobable random policy.

6406532052833. ✓ 

As $\tau \rightarrow 0$, π tends to a deterministic, greedy policy.

6406532052834. ✓ 

As $\tau \rightarrow \infty$, π tends to an equiprobable random policy.

6406532052835. ✖ 

As $\tau \rightarrow \infty$, π tends to a deterministic, greedy policy.

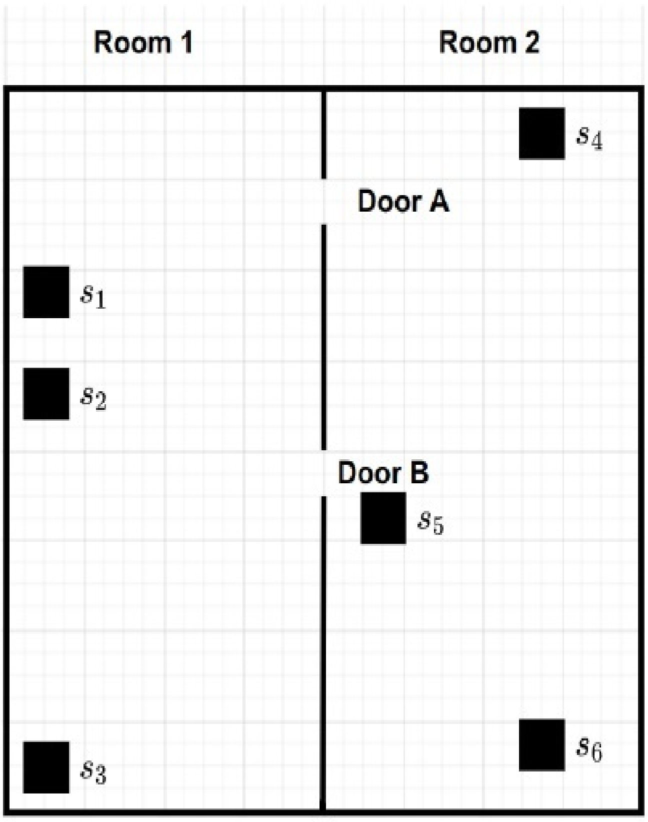
Sub-Section Number :	5
Sub-Section Id :	64065389028
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 223 Question Id : 640653614942 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

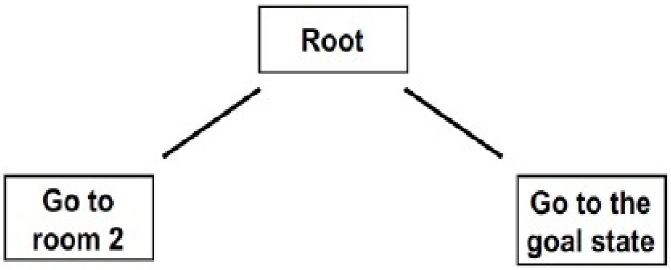
Question Label : Multiple Select Question

Consider the following grid world:



The grid world has two doors A and B and six specific states/positions s_1 to s_6. π_1 represents the optimal policy to reach from *anywhere* in Room 1 to Room 2. At any state in the grid world, an agent can take only 4 actions (up, down, left, right), the actions that take the agent out of the grid world or on obstacles (i.e. walls) don't change the state.

Following is the hierarchy of tasks:



Consider following trajectory:

$$s_1 \to \ldots \to \text{Door A} \to \ldots \to s_5$$

Above trajectory **CANNOT** be consistent with:

Options :

6406532052820. ✔ Hierarchical optimal policy

6406532052821. ✖ Recursively optimal policy

6406532052822. ✔ Flat optimal policy

6406532052823. ✖ None of these

Question Number : 224 Question Id : 640653614946 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

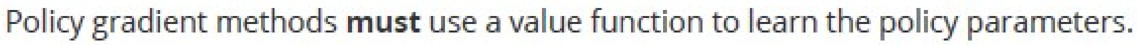
Question Label : Multiple Select Question

Select all true statements.

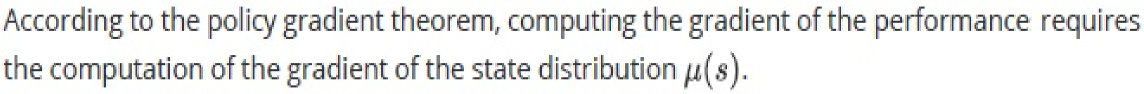
Options :

6406532052836. ✓

Policy gradient methods use a parameterized policy that can select actions without consulting a value function.

6406532052837. ✗ 

Policy gradient methods **must** use a value function to learn the policy parameters.

6406532052838. ✗ 

According to the policy gradient theorem, computing the gradient of the performance requires the computation of the gradient of the state distribution $\mu(s)$.

6406532052839. ✓

In REINFORCE with baseline, the baseline **cannot** be a function of the actions.

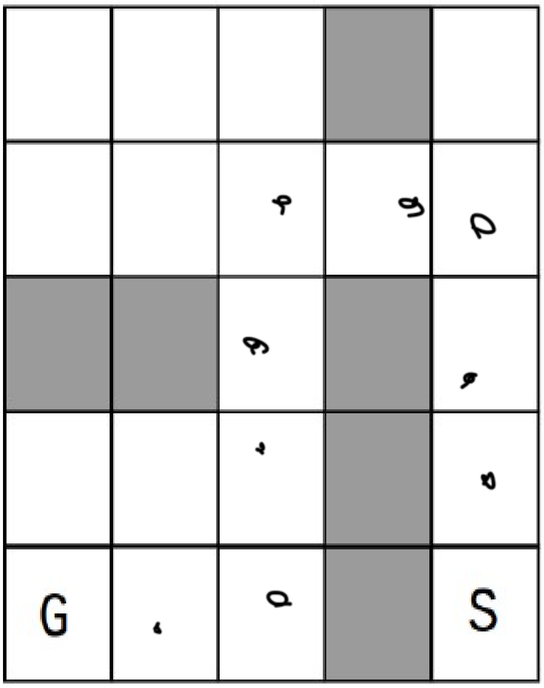
Sub-Section Number :	6
Sub-Section Id :	64065389029
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 225 Question Id : 640653614947 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3

Question Label : Short Answer Question

Consider the following grid-world in which all transitions are deterministic. S is the start state, G is the goal and the gray cells are obstructions.



The agent gets a reward of 10 when it reaches the goal state. For all other transitions, the reward is 0. Actions that take the agent out of the grid or into the obstructions leave the state unchanged. If $\gamma = 0.8$, find the maximum possible return for the agent starting from state S. Enter your answer correct to three decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

$(0.8)^9$

Answers Type : Range

Text Areas : PlainText

0.134

Possible Answers :

0.12 to 0.15

Question Number : 226 Question Id : 640653614948 Question Type : SA Calculator : None

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

Correct Marks : 3

Question Label : Short Answer Question

Compute $v_\pi(s)$ if it is given that $q_\pi(s,.) = [1, \quad 4, \quad -1]$ and $\pi(. \mid s) = [0.2 \quad 0.5 \quad 0.3]$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

$0.2 + 2 - 0.3$

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

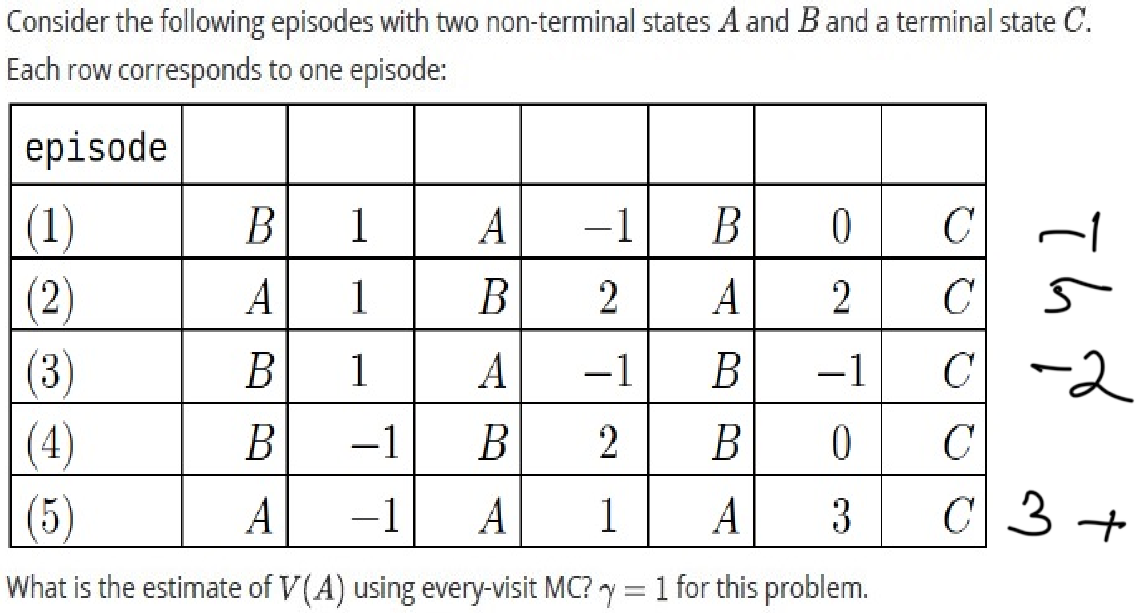
1.9

Question Number : 227 Question Id : 640653614949 Question Type : SA Calculator : None

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

Correct Marks : 3

Question Label : Short Answer Question



Consider the following episodes with two non-terminal states A and B and a terminal state C. Each row corresponds to one episode:

episode							
(1)	B	1	A	−1	B	0	C
(2)	A	1	B	2	A	2	C
(3)	B	1	A	−1	B	−1	C
(4)	B	−1	B	2	B	0	C
(5)	A	−1	A	1	A	3	C

-1

5

-2

$3 + 4 + 3$

What is the estimate of $V(A)$ using every-visit MC? $\gamma = 1$ for this problem.

Response Type : Numeric

Evaluation Required For SA : Yes

$\frac{\cancel{12}}{\cancel{6}} \simeq 2$

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

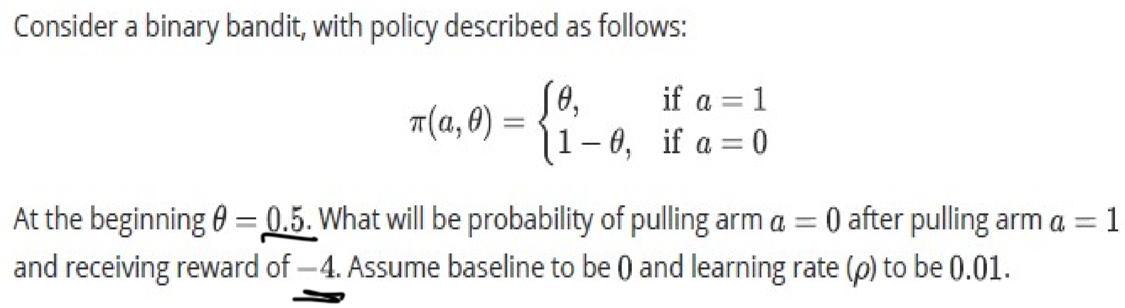
2

Question Number : 228 Question Id : 640653614950 Question Type : SA Calculator : None

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

Correct Marks : 3

Question Label : Short Answer Question



Consider a binary bandit, with policy described as follows:

$$\pi(a,\theta) = \begin{cases} \theta, & \text{if } a = 1 \\ 1-\theta, & \text{if } a = 0 \end{cases}$$

At the beginning $\theta = 0.5$. What will be probability of pulling arm $a = 0$ after pulling arm $a = 1$ and receiving reward of -4. Assume baseline to be 0 and learning rate (ρ) to be 0.01.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

0..48

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

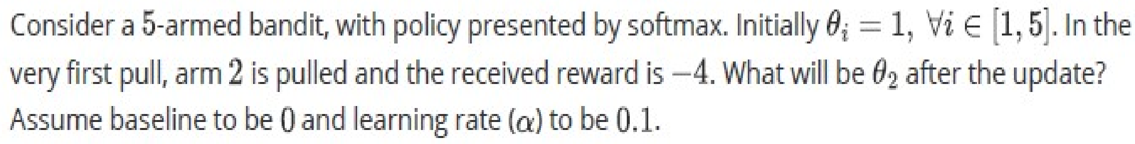
0.48 ✓

Question Number : 229 Question Id : 640653614951 Question Type : SA Calculator : None

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

Correct Marks : 3

Question Label : Short Answer Question



Consider a 5-armed bandit, with policy presented by softmax. Initially $\theta_i = 1,\ \forall i \in [1,5]$. In the very first pull, arm 2 is pulled and the received reward is -4. What will be θ_2 after the update? Assume baseline to be 0 and learning rate (α) to be 0.1.

Response Type : Numeric

$\theta_2 =$

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.68

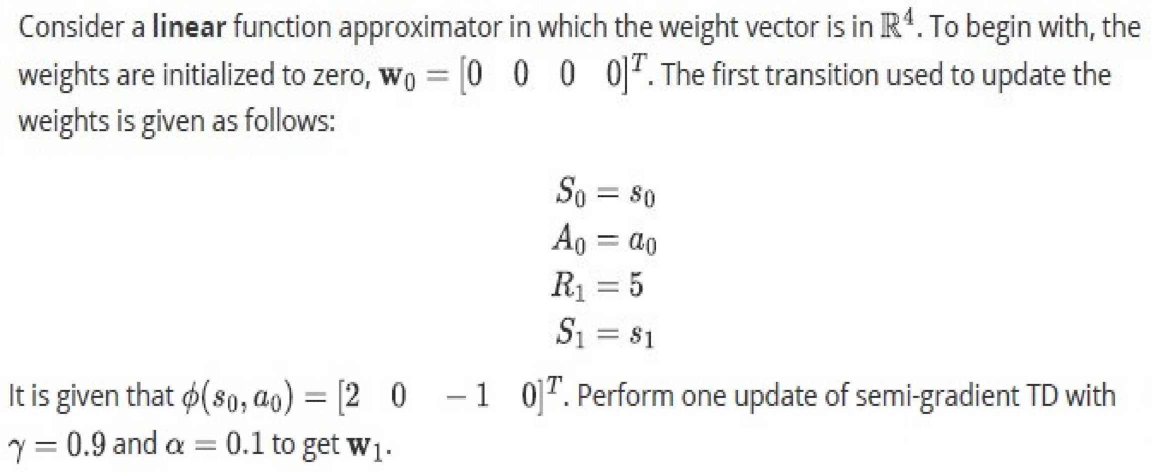
Sub-Section Number :	7
Sub-Section Id :	64065389030
Question Shuffling Allowed :	No

Is Section Default? : null

Question Id : 640653614952 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 : (230 to 231)

Question Label : Comprehension



Consider a **linear** function approximator in which the weight vector is in $\mathbb{R}^4$. To begin with, the weights are initialized to zero, $\mathbf{w}_0 = [0 \quad 0 \quad 0 \quad 0]^T$. The first transition used to update the weights is given as follows:

$$\begin{aligned} S_0 &= s_0 \\ A_0 &= a_0 \\ R_1 &= 5 \\ S_1 &= s_1 \end{aligned}$$

It is given that $\phi(s_0, a_0) = [2 \quad 0 \quad -1 \quad 0]^T$. Perform one update of semi-gradient TD with $\gamma = 0.9$ and $\alpha = 0.1$ to get $\mathbf{w}_1$.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 230 Question Id : 640653614953 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2.5

Question Label : Short Answer Question TD = 5

Find the TD error for this transition.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

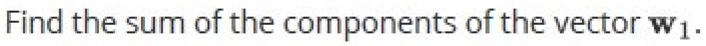
5 ✓

Question Number : 231 Question Id : 640653614954 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



Find the sum of the components of the vector $\mathbf{w}_1$.

$$w_0 + \alpha \delta \phi$$

$$0 \quad 0 \quad 0 \quad 0 + 0.5[2 \quad 0 \quad -1 \quad 0]$$

$$\begin{array}{rcccc} = & 0 & 0 & 0 & 0 \\ & 1 & 0 & -0.5 & 0 \\ \hline & 1 & 0 & -0.5 & 0 \end{array}$$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.5 ✓

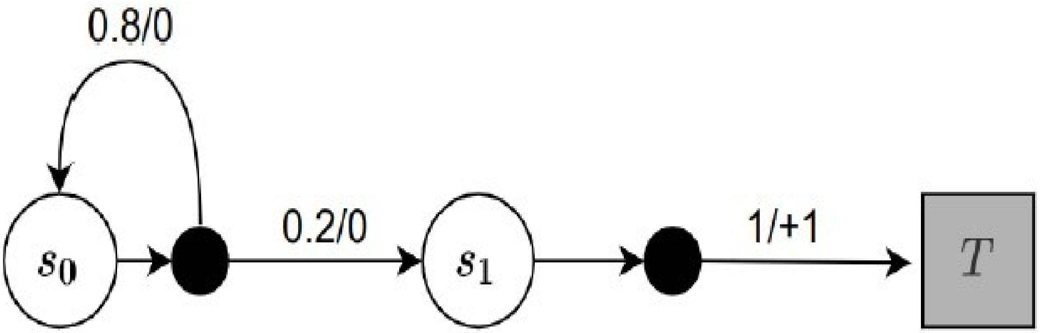
Sub-Section Number :	8
Sub-Section Id :	64065389031
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653614955 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 : (232 to 233)

Question Label : Comprehension

Consider following MDP, assume $\lambda = 1, \gamma = 0.8$:



The edges have the value p/r, where p denotes the transition probability and r is the immediate expected reward. s_0 and s_1 are non-terminal states while T is a terminal state.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 232 Question Id : 640653614956 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 3.5

Question Label : Short Answer Question

What is the minimum length of the trajectory sampled from given MDP, for which value of the accumulating eligibility trace for state s_0 is greater than 2?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

5

✓

Question Number : 233 Question Id : 640653614957 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction

Time : 0

Correct Marks : 1.5

Question Label : Multiple Choice Question

If the eligibility traces are replacing in nature, which state would have highest eligibility trace at the end of a trajectory, and what will be the value of the eligibility trace of the corresponding state?

Options :

6406532052848. ✖ State with highest eligibility trace = s_0 and $e(s_0)$ = 1

6406532052849. ✔ State with highest eligibility trace = s_1 and $e(s_1)$ = 1

6406532052850. ✖ State with highest eligibility trace = s_0 and $e(s_0)$ = 0.8

6406532052851. ✖ State with highest eligibility trace = s_1 and $e(s_1)$ = 0.8

Statistical Computing

Section Id :	64065341449
Section Number :	11
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	6
Number of Questions to be attempted :	6
Section Marks :	50
Display Number Panel :	Yes
Section Negative Marks :	0
Group All Questions :	No
Enable Mark as Answered Mark for Review and	Yes