Question Paper

Exam Date & Time: 23-Apr-2024 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
Second Semester Master of Engineering - ME
(Artificial Intelligence and Machine Learning) Degree Examination - April / May 2024

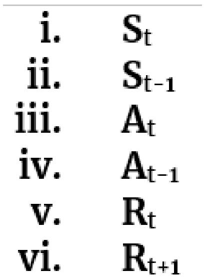
Reinforcement Learning [AML 5204]

Marks: 100 **Duration: 180 mins.**

Tuesday, April 23, 2024

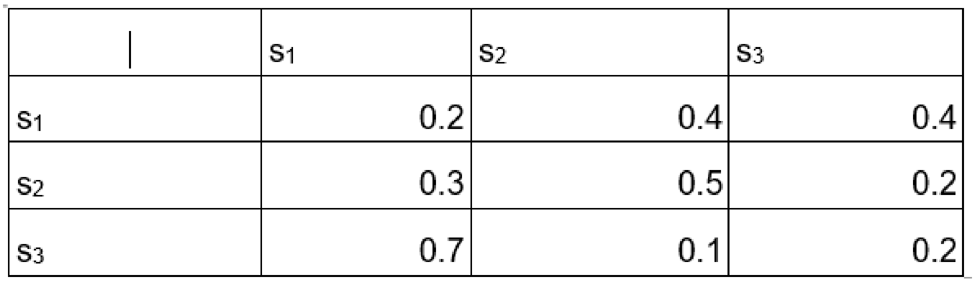
Answer all the questions.

1) [CO1, L5] Answer the following questions related to Markov property: (10)
In an MDP, on which of the following does the state $\mathbf{S_{t+1}}$ depend on? (Choose one or more that apply with a brief justification):



i. S_t
ii. S_{t-1}
iii. A_t
iv. A_{t-1}
v. R_t
vi. R_{t+1}

For a Markov process with the following probability transition matrix, rows represent the current state and columns represent the next state.



|	s1	s2	s3
s1	0.2	0.4	0.4
s2	0.3	0.5	0.2
s3	0.7	0.1	0.2

What are the values of the following transition probabilities: $\mathbf{P(S_{t+1}=s_1 \mid S_t=s_3)}$, $\mathbf{P(S_{t+1}=s_2 \mid S_t=s_3)}$ and $\mathbf{P(S_{t+1}=s_2 \mid S_t=s_1, S_{t-1}=s_1)}$?

2) [CO2, L3] Consider the state space S = { s_1, s_2 } and action space { a_1, a_2 }. Draw a 1 level backup diagram starting from state s_1 by clearly showing the branch probabilities. Use the backup diagram and write an expression for $v_\pi(s_1)$. (10)

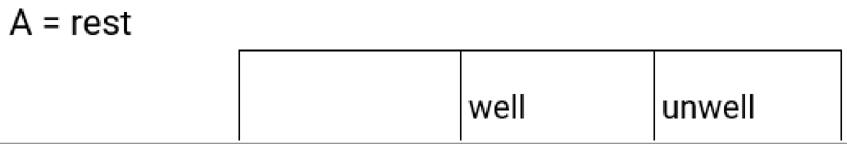
3) [CO2, L5] Answer the following questions using not more than two lines. (10)

- What is the difference between "one step reward" and " discounted long-term return"?
- Is long term return $\mathbf{G_t}$ a random variable? Justify why it is or not?
- What is meant by a policy (π)?

4) [CO1, L3] Consider a scenario where an individual can be either well or unwell on a particular day. The individual can either decide to rest or to work on the same day. (10)

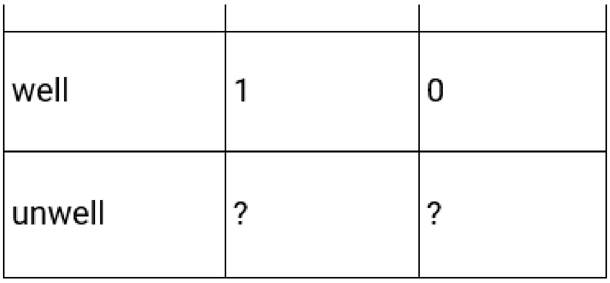
- When they are well and decide to work, there is a 1% chance that they become unwell the next day. However, if the person rests given that he/she is well, there is a 100% chance of being well the next day also.
- When the person is unwell and rests, there is a 90% chance of recovery, becoming well the next day. Otherwise, if they decided to work there is a 80% chance of still being unwell on the next day.

Given,
State space S = {well, unwell}
Action space A = {rest, work}
Draw a state transition diagram representing the states and transition probabilities for each action the individual takes.Write down the transition probabilities in the form of a transition matrix for the two possible actions:

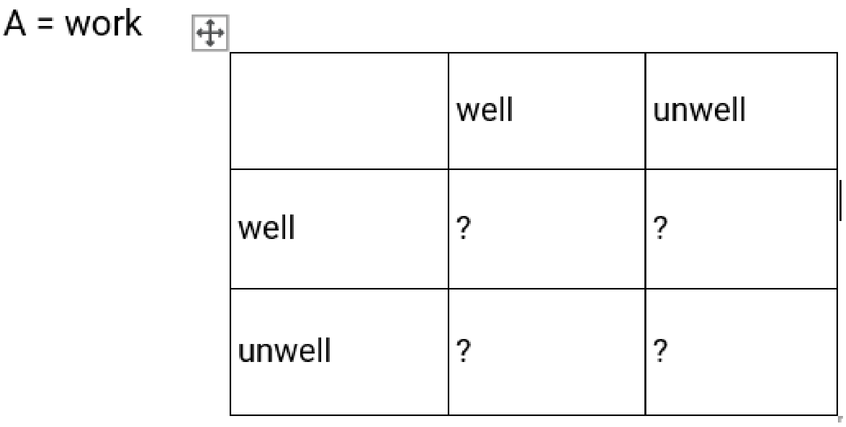


A = rest

	well	unwell



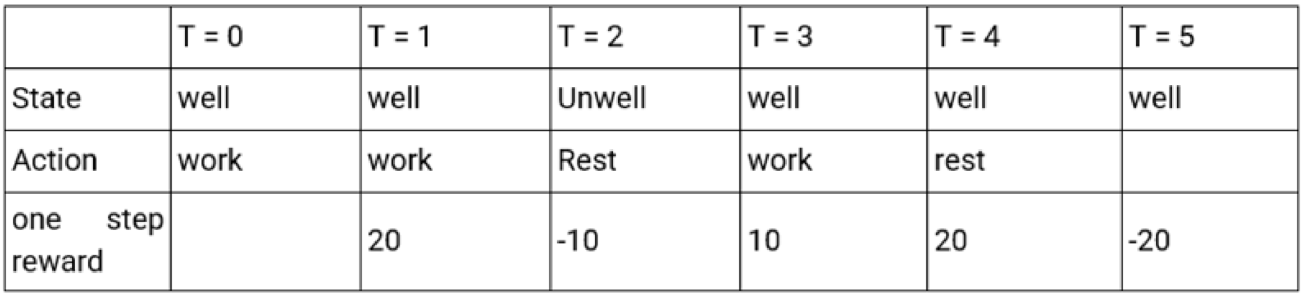
well	1	0
unwell	?	?



A = work

	well	unwell
well	?	?
unwell	?	?

5) [CO3, L3] A particular individual undergoes the following state transitions over 5 days because of the actions they took: (10)

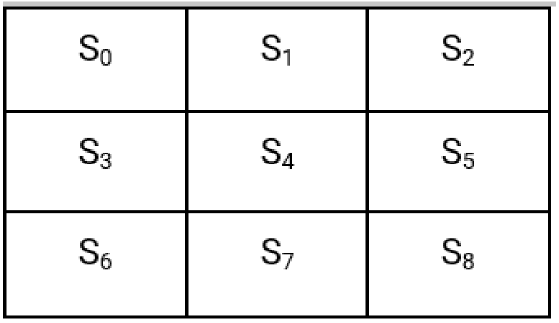


	T = 0	T = 1	T = 2	T = 3	T = 4	T = 5
State	well	well	Unwell	well	well	well
Action	work	work	Rest	work	rest	
one step reward		20	-10	10	20	-20

Discount factor gamma = 0.1

- Calculate G_0, G_1, G_2, G_3, G_4.
- For the above individual, what are the values of states well and unwell: v_π(well) and v_π(unwell)?

6) [CO2, L6] Given a 3 X 3 grid world with 9 states, (10)



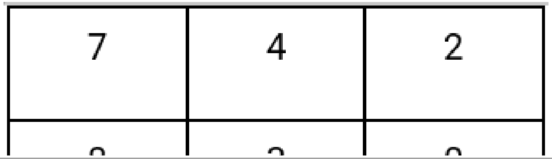
S_0	S_1	S_2
S_3	S_4	S_5
S_6	S_7	S_8

Action space consists of 4 actions to move: up, down, left, and right. The Agent cannot move outside the grid (It remains in the same state if the action intended to take the agent outside the gridworld). The transitions are deterministic, there is a 100% chance of the agent moving in the direction the action was chosen. For example, if the agent starts from S_4 and takes an action to move right, it moves to state S_5 with a probability 1.

It is also given that,

- S_8 and S_5 are the terminal states. Once the agent reaches these states, they cannot come out. The episode terminates once the agent reaches one of the terminal states.
- Transition to a terminal state gives a one-step-reward of +10, and all other transitions get a reward of -1.

Assume that the estimated optimal state values (V ~ v*(s)) are as follows:



7	4	2
[illegible]	[illegible]	[illegible]

8	3	0
7	1	0

Construct a deterministic policy π ~ π* using the above optimal state values. Display the policy using "arrow marks" on the gridworld. Assume the discount factor, gamma=1.
Hint: Choose the actions in a one-step greedy fashion using the bellman's optimality equation.

$$v_{\pi^*}(s) = \max_a q_{\pi^*}(a, s) = \max_a \left[T(s, a, s') \left(R(s, a, s') + \gamma \sum_{s' \in S} v_{\pi^*}(s') \right) \right]$$

Compute the best action to take for each of the following states shown with a question mark. Do not skip any of the calculations, show why the action you choose is the best action.

?	?	↓
?	→	X
↑	→	X

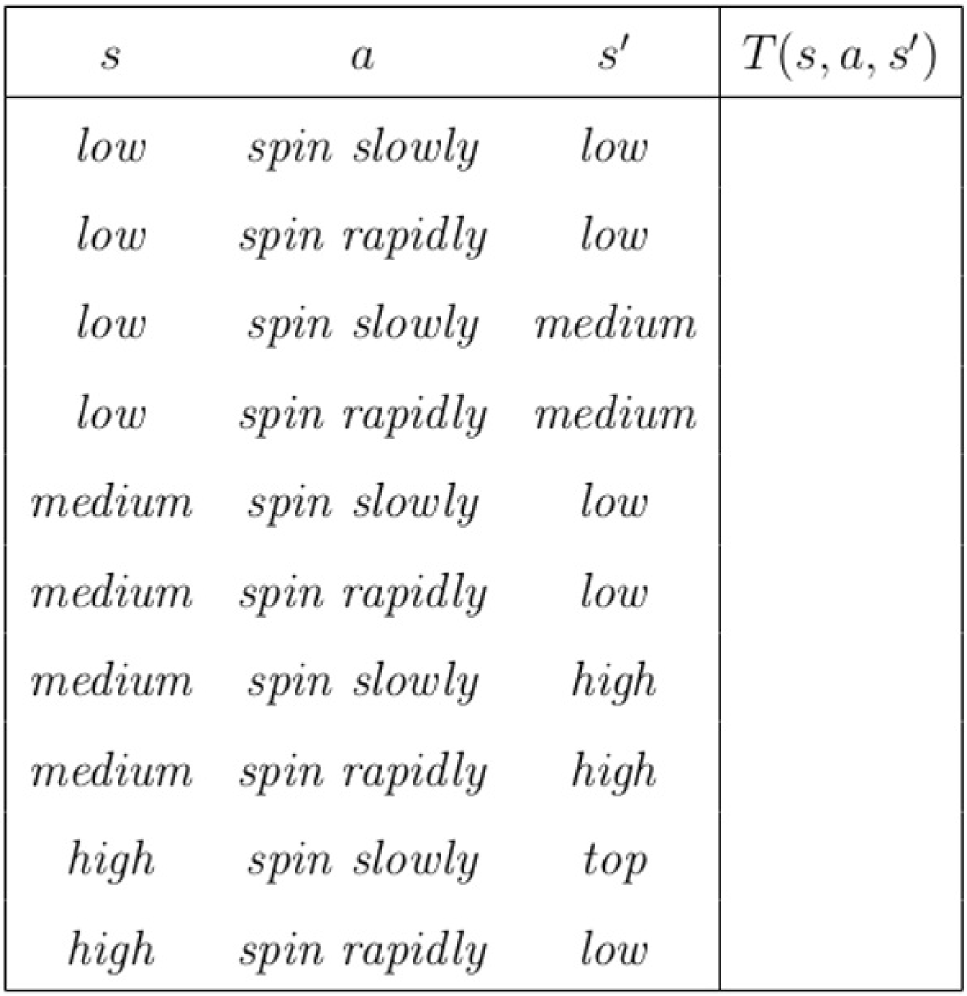
7) [CO3, L3] Consider a self-powered rover that operates on a slope. The rover can be in one of the following four states: low, medium, high, and top. (10)

The rover has a motor that can spin its wheel

- slowly at the expense of 20 unit of energy per time step;
- or rapidly at the expense of 40 units of energy per time step.

If the motor spins the wheel slowly, with probability 0.5 it moves to the next higher state in one time step, and with probability 0.5, it slides all the way down the slope to the low state.
On the other hand, if the motor spins the wheel rapidly, with probability 0.8 it moves to the next higher state in one time step, and with probability 0.2, it slides all the way down the slope to the low state.

The rover's motion terminates once it reaches the top state. The rover is low on the slope and aims to reach the top with minimum energy consumption. Fill the entries in the table below:



s	a	s'	$T(s,a,s')$
low	*spin slowly*	*low*	
low	*spin rapidly*	*low*	
low	*spin slowly*	*medium*	
low	*spin rapidly*	*medium*	
medium	*spin slowly*	*low*	
medium	*spin rapidly*	*low*	
medium	*spin slowly*	*high*	
medium	*spin rapidly*	*high*	
high	*spin slowly*	*top*	
high	*spin rapidly*	*low*	

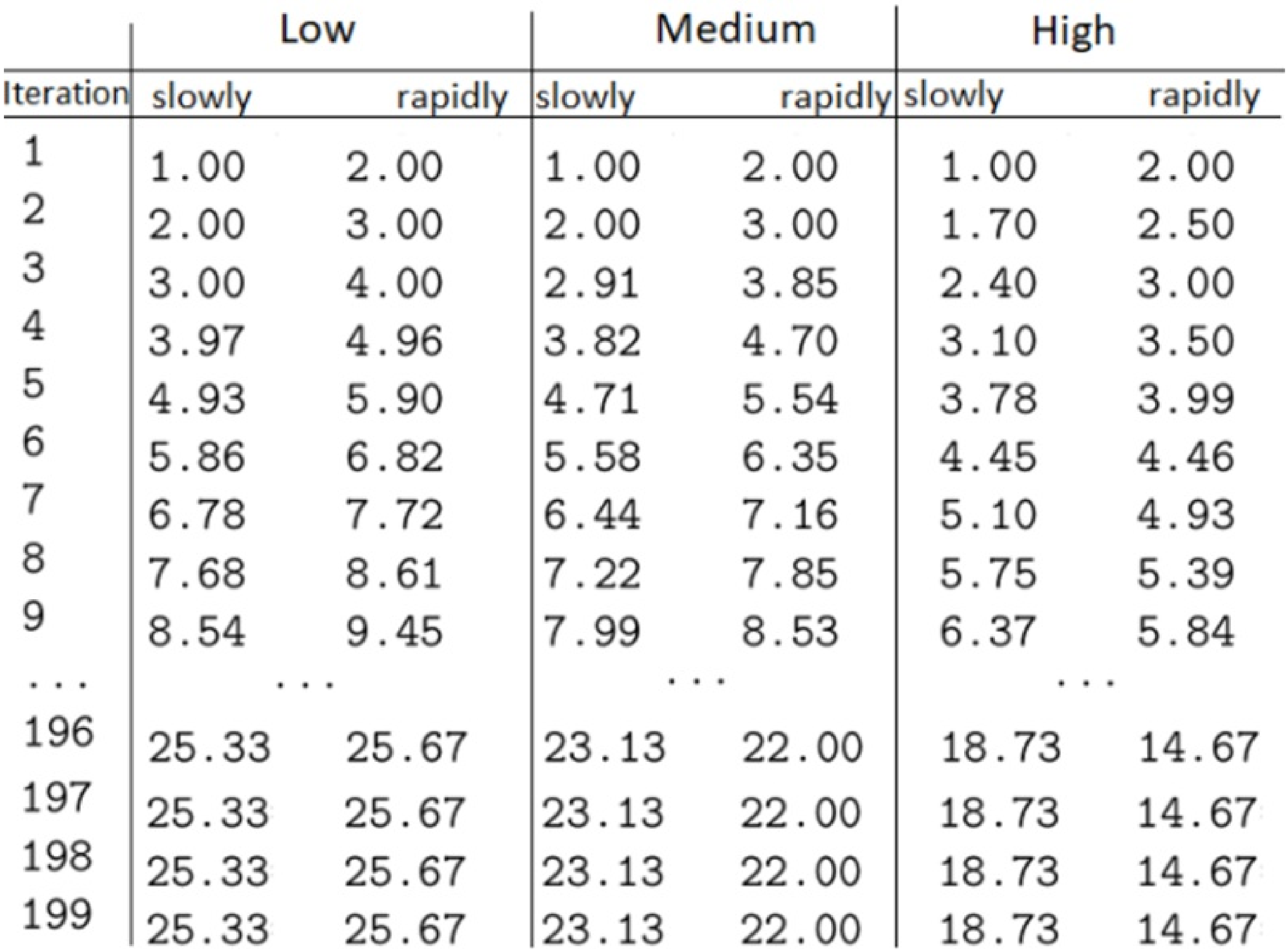
8) [CO3, L3] Continuing from the previous problem, suppose the rover gains 4 units of energy per time step from its solar panels any time it transitions upward from the medium position as it gets exposed to sunlight. The rover gains no energy while being low on the slope. Consider the following policy: (10)

- π (spin rapidly | low) = 0.7,
- π (spin rapidly | medium) = 0.8,
- π (spin slowly | high) = 0.2.

Draw three 2-level backup diagrams with each one of them starting from the states low, medium, and high, respectively. The levels of the backup diagrams should represent the start state, actions, and the end states with the appropriate policy and transition probabilities written over the branches.

9) [CO3, L3] Continuing from the previous question, start with zero initial values for v_{π}(low), v_{π}(medium), and v_{π}(high), and a discount factor γ= 0.9. Use the backup diagrams from the previous question to run two iterations of policy evaluation to evaluate the above policy. Report the updated values of the three states. (10)

10) [CO4, L4] For the solar-powered rover from Question-7, suppose running the value iteration (policy improvement) procedure with zero initial values results in the following table: (10)



	Low		Medium		High	
Iteration	slowly	rapidly	slowly	rapidly	slowly	rapidly
1	1.00	2.00	1.00	2.00	1.00	2.00
2	2.00	3.00	2.00	3.00	1.70	2.50
3	3.00	4.00	2.91	3.85	2.40	3.00
4	3.97	4.96	3.82	4.70	3.10	3.50
5	4.93	5.90	4.71	5.54	3.78	3.99
6	5.86	6.82	5.58	6.35	4.45	4.46
7	6.78	7.72	6.44	7.16	5.10	4.93
8	7.68	8.61	7.22	7.85	5.75	5.39
9	8.54	9.45	7.99	8.53	6.37	5.84
...	...		...		...	
196	25.33	25.67	23.13	22.00	18.73	14.67
197	25.33	25.67	23.13	22.00	18.73	14.67
198	25.33	25.67	23.13	22.00	18.73	14.67
199	25.33	25.67	23.13	22.00	18.73	14.67

Clearly show the steps as to how the value of the state **low** is computed in iteration-1. After convergence of the value iteration procedure, what are the optimal values of the states? What are the corresponding optimal policies? Describe the optimal policy after convergence of the value iteration procedure in plain English in one sentence.

-----End-----