Question Paper

Exam Date & Time: 01-Jul-2022 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

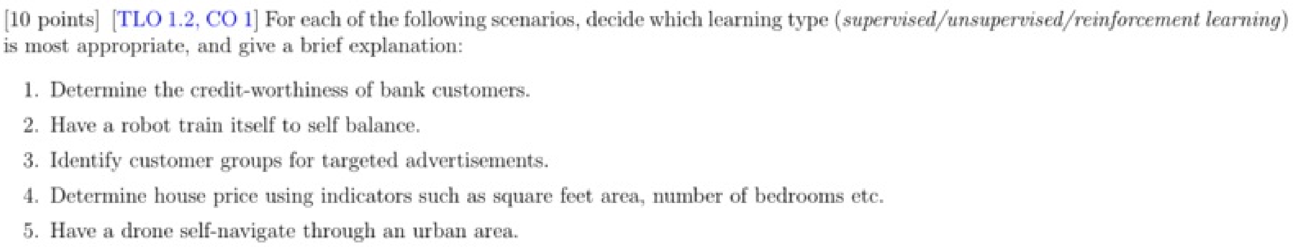
MANIPAL SCHOOL OF INFORMATION SCIENCES, MANIPAL
SECOND SEMESTER MASTER OF ENGINEERING- ME (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Reinforcement Learning [AML 5204]

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

Friday, July 1, 2022

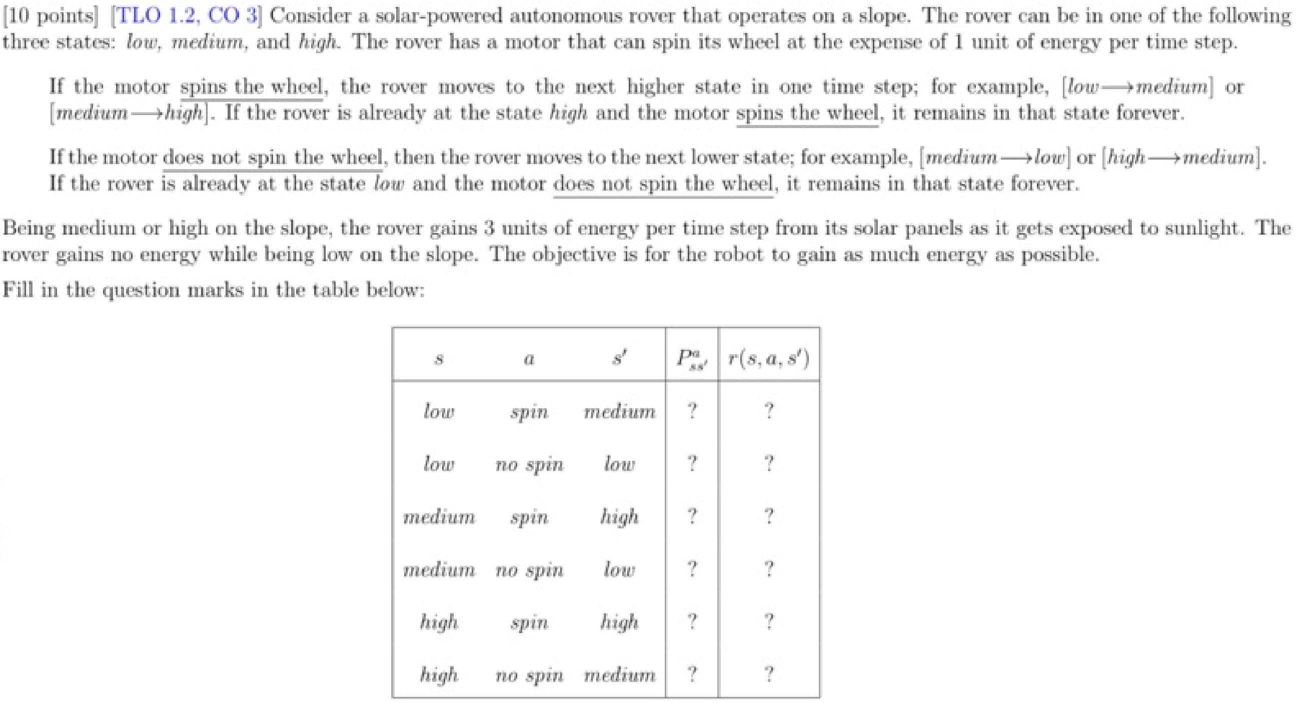
Answer all the questions.

1) 

[10 points] [TLO 1.2, CO 1] For each of the following scenarios, decide which learning type (*supervised/unsupervised/reinforcement learning*) is most appropriate, and give a brief explanation:

1. Determine the credit-worthiness of bank customers.
2. Have a robot train itself to self balance.
3. Identify customer groups for targeted advertisements.
4. Determine house price using indicators such as square feet area, number of bedrooms etc.
5. Have a drone self-navigate through an urban area.

(10)

2) 

[10 points] [TLO 1.2, CO 3] Consider a solar-powered autonomous rover that operates on a slope. The rover can be in one of the following three states: *low*, *medium*, and *high*. The rover has a motor that can spin its wheel at the expense of 1 unit of energy per time step.

If the motor spins the wheel, the rover moves to the next higher state in one time step; for example, [$low \longrightarrow medium$] or [$medium \longrightarrow high$]. If the rover is already at the state *high* and the motor spins the wheel, it remains in that state forever.

If the motor does not spin the wheel, then the rover moves to the next lower state; for example, [$medium \longrightarrow low$] or [$high \longrightarrow medium$]. If the rover is already at the state *low* and the motor does not spin the wheel, it remains in that state forever.

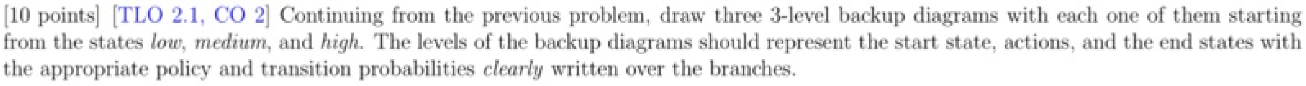
Being medium or high on the slope, the rover gains 3 units of energy per time step from its solar panels as it gets exposed to sunlight. The rover gains no energy while being low on the slope. The objective is for the robot to gain as much energy as possible.

Fill in the question marks in the table below:

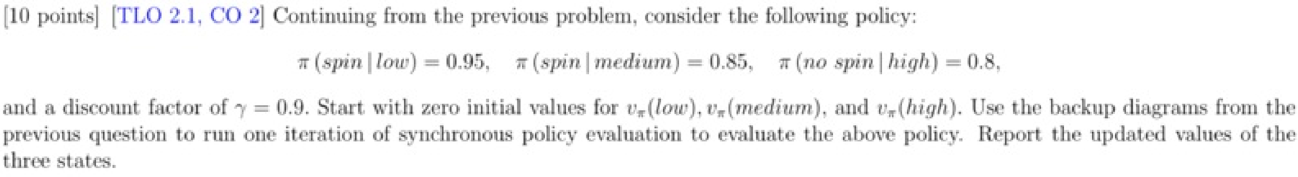
s	a	s'	$P^a_{ss'}$	$r(s,a,s')$
low	*spin*	*medium*	?	?
low	*no spin*	*low*	?	?
medium	*spin*	*high*	?	?
medium	*no spin*	*low*	?	?
high	*spin*	*high*	?	?
high	*no spin*	*medium*	?	?

(10)

3) (10)



[10 points] [TLO 2.1, CO 2] Continuing from the previous problem, draw three 3-level backup diagrams with each one of them starting from the states *low*, *medium*, and *high*. The levels of the backup diagrams should represent the start state, actions, and the end states with the appropriate policy and transition probabilities *clearly* written over the branches.

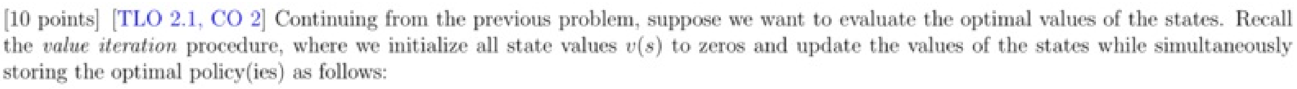
4) 

[10 points] [TLO 2.1, CO 2] Continuing from the previous problem, consider the following policy:

$$\pi\,(spin\,|\,low) = 0.95, \quad \pi\,(spin\,|\,medium) = 0.85, \quad \pi\,(no\ spin\,|\,high) = 0.8,$$

and a discount factor of $\gamma = 0.9$. Start with zero initial values for $v_\pi(low)$, $v_\pi(medium)$, and $v_\pi(high)$. Use the backup diagrams from the previous question to run one iteration of synchronous policy evaluation to evaluate the above policy. Report the updated values of the three states.

(10)

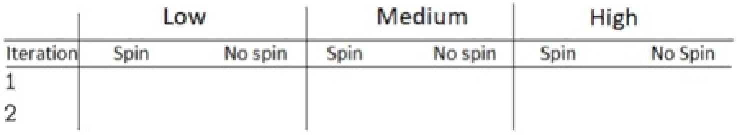
5) 

[10 points] [TLO 2.1, CO 2] Continuing from the previous problem, suppose we want to evaluate the optimal values of the states. Recall the *value iteration* procedure, where we initialize all state values $v(s)$ to zeros and update the values of the states while simultaneously storing the optimal policy(ies) as follows:

(10)

$$v(s) = \max_a \left[P^a_{ss'} \left(r(s,a,s') + \gamma \sum_{s' \in S} v(s') \right) \right].$$

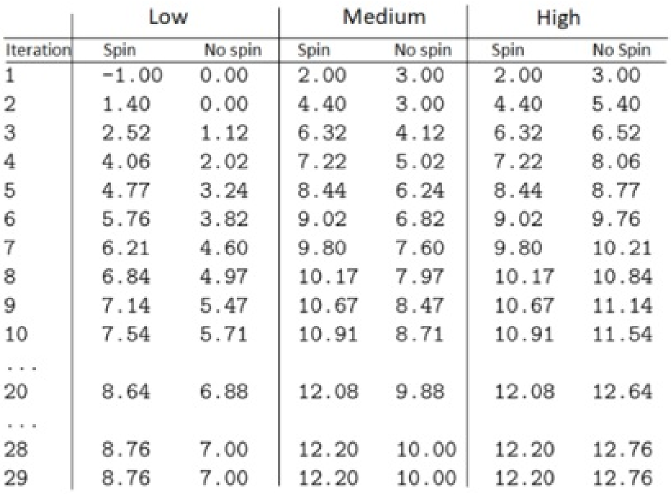
Run two iterations of the *synchronous* version of the value iteration procedure and fill in the entries in the table below:



	Low		Medium		High	
Iteration	Spin	No spin	Spin	No spin	Spin	No Spin
1						
2						

6) [10 points] [TLO 2.1, CO 2] Repeat the previous problem but now run two iterations of the *asynchronous* version of the value iteration procedure. (10)

7) [10 points] [TLO 2.1, CO 2] For the solar-powered rover from Question-1, suppose running a synchronous version of the value iteration procedure results in the following table: (10)



	Low		Medium		High	
Iteration	Spin	No spin	Spin	No spin	Spin	No Spin
1	-1.00	0.00	2.00	3.00	2.00	3.00
2	1.40	0.00	4.40	3.00	4.40	5.40
3	2.52	1.12	6.32	4.12	6.32	6.52
4	4.06	2.02	7.22	5.02	7.22	8.06
5	4.77	3.24	8.44	6.24	8.44	8.77
6	5.76	3.82	9.02	6.82	9.02	9.76
7	6.21	4.60	9.80	7.60	9.80	10.21
8	6.84	4.97	10.17	7.97	10.17	10.84
9	7.14	5.47	10.67	8.47	10.67	11.14
10	7.54	5.71	10.91	8.71	10.91	11.54
...						
20	8.64	6.88	12.08	9.88	12.08	12.64
...						
28	8.76	7.00	12.20	10.00	12.20	12.76
29	8.76	7.00	12.20	10.00	12.20	12.76

In iteration 8, what are the optimal values of the states *low*, *medium*, and *high*? What are the corresponding optimal policies? Describe the optimal policy after convergence of the value iteration procedure in plain English in one sentence.

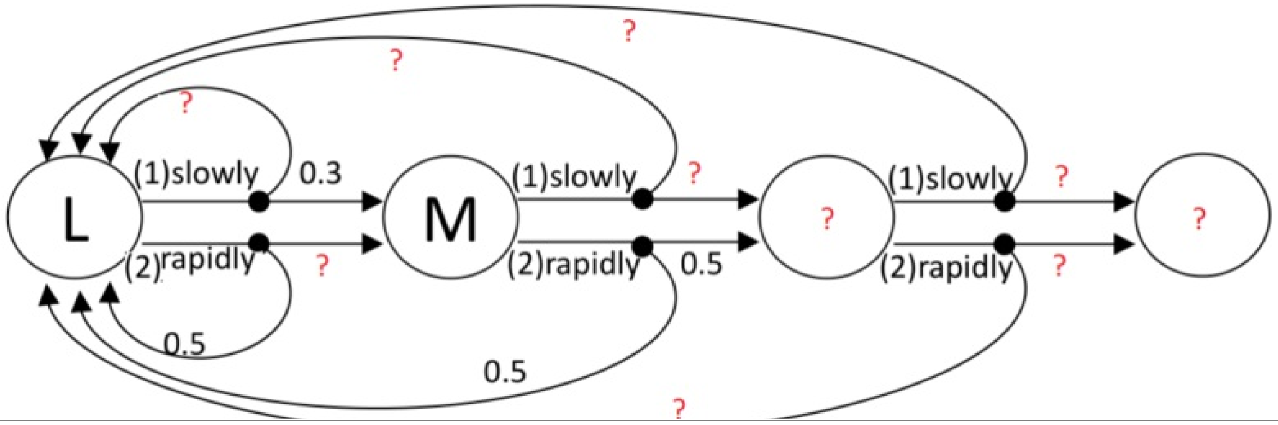
8) (10)

[10 points] [TLO 1.2, CO 3] 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*. The rover has a motor that can spin its wheel

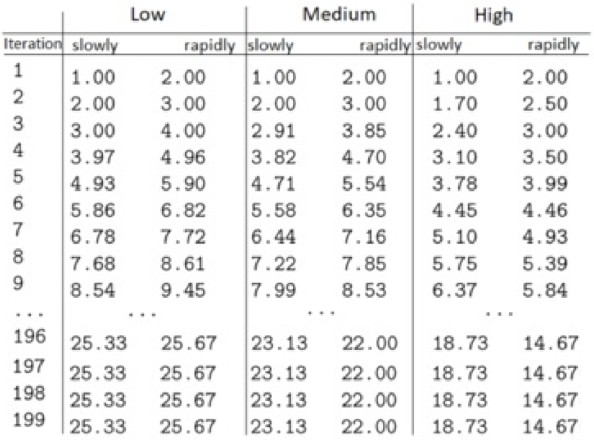
- *slowly* at the expense of 1 unit of energy per time step;
- or *rapidly* at the expense of 2 units of energy per time step.

If the motor spins the wheel slowly, with probability 0.3 it moves to the next higher state in one time step. With probability 0.7, 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 0.5 it moves to the next higher state in one time step. With probability 0.5, 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 in the questions marks in the graphical representation below showing the MDP corresponding to the rover movement:



9) [10 points] [TLO 2.1, CO 2] Continuing from the previous problem, suppose running a synchronous version of the value iteration 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.

10) (10)

[10 points] [TLO 1.2, CO 3] Suppose you have the following deal for a contract: you will be paid 1 lakh Rupees each year for the next 20 years; if the interest rate is 5 percent, how much money do you need to get right away to be indifferent between a one-time payment and a yearly payment as explained above for the next 20 years? I'm looking at an intuitive as well as a mathematical answer (both should be brief).

-----End-----