Sessional 2

Reinforcement Learning
Even Sem - 2024

Manipal School of Information Science, MAHE

---

1. [10 points] Choose the most appropriate technique for each of the following scenarios:
   Techniques:
   - Policy Evaluation by Dynamic Programming. 1
   - Policy Improvement by Dynamic Programming. 2
   - Monte Carlo Prediction (Policy Evaluation) 3
   - Monte Carlo Control (Policy Improvement) 4

   Use Cases:
   1. Train a robot to navigate through a frozen lake. Transition probabilities and the one step rewards are known for every possible state transition. 2
   2. 3 robots are trained to perform a floor cleaning task. You are provided with a new unknown factory where you can run the robots. The task is to choose the best robot. 3
   3. Train an agent that would decide to buy, sell or keep a particular stock(Ex, reliance stock) each day, given the stock price data of the last 5 years. Agent needs to learn an optimal policy by looking at the past data. 4
   4. Compare two agents which are trained to navigate through a grid world to reach the terminal state, you are provided with a model of the environment and you do not know the dynamics (transition probabilities or one step rewards) of the environment. 3

2. [10 points] Consider the state space $S = \{s_1, s_2, s_3\}$ 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 $v_\pi(s_1)$.

3. [10 points] An individual can be classified as either **normal, heavy or obese** depending on their weight. Assume the weights are measured once per month.
   - If the person **exercises**, there is a 10% chance of losing weight and transitioning to the lower weight category*[obese -> heavy or heavy -> normal]* and 90% chance of remaining in the same category in the following month. If the person is in normal state they remain in normal state with 100% chance by exercising.
   - If the person does **not exercise,** there is a 20% chance for the person to move into the higher weight category and 80% chance of remaining in the same category in the following month. Once they are obese, they will remain obese with 100% chance if they don't exercise.

   The person gets a reward of -1 for moving from a lower weight category to a higher weight category and gets a reward of -2 for staying obese[obese -> obese transition].

Answer the following questions:

1. What is the state space and action space?
2. Write down the different transition probabilities, P(s' | s, a).
   Example, P(normal | no exercise, obese) = ?, write for all possible transitions.
3. What are the one step rewards R(s, a, s') for all transitions.

4. [10 points] Given the following policy:
   - *π(exercise | normal) = 0.3*
   - *π(exercise | heavy) = 0.4*
   - *π(exercise | obese) = 0.7*

   Draw a 2 level backup diagram starting from the state "normal". The levels of the backup diagram should represent the start state, action and the end state with the appropriate policy and transition probabilities written over the branches. Using the backup diagram, write the expression for $v_\pi(normal)$. Simplify the equation by assigning the values of transition probabilities and the action probabilities. Assume the discount factor, gamma=0.

5. [10 points] Given a 3 X 3 grid world with 9 states,

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. 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 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
8	3	0
7	1	0

Come up with 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.

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

Solutions

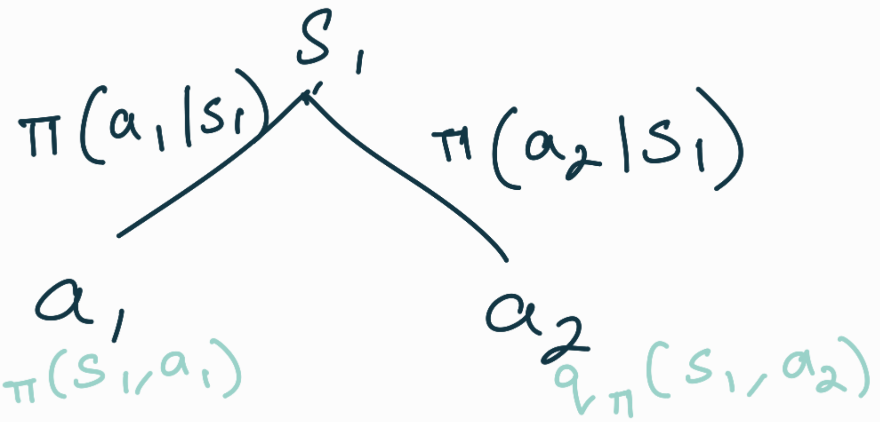
Q1) See above

Q2) Given, State space

$$S = \{ s_1, s_2, s_3 \}$$

$$A = \{ a_1, a_2 \}$$

One level backup diagram starting from state s_1



$$v_\pi(s_1) = \pi(a_1|s_1)\, q_\pi(s_1, a_1) + \pi(a_2|s_1)\, q_\pi(s_1, a_2)$$

$$= \sum_{a \in A} \pi(a|s_1)\, q_\pi(s_1, a)$$

$$\boxed{v_\pi(s_1) = \sum_{a \in A} \pi(a|s_1)\, q_\pi(s_1, a)}$$

Q 3) $S = \{ \text{normal}, \text{heavy}, \text{Obese} \}$

$A = \{ \text{exercise}, \text{no exercise} \}$

S	A	S'	P(s' | s, a)	r(s, a, s')
normal	ex	normal	1	0
normal	no ex	heavy	0.2	-1
normal	no ex	normal	0.8	0
heavy	ex	normal	0.1	0
heavy	ex	heavy	0.9	0
heavy	no ex	heavy	0.8	0
heavy	no ex	obese	0.2	-1
obese	ex	heavy	0.1	0
obese	ex	obese	0.9	-2
obese	no ex	obese	1	-2

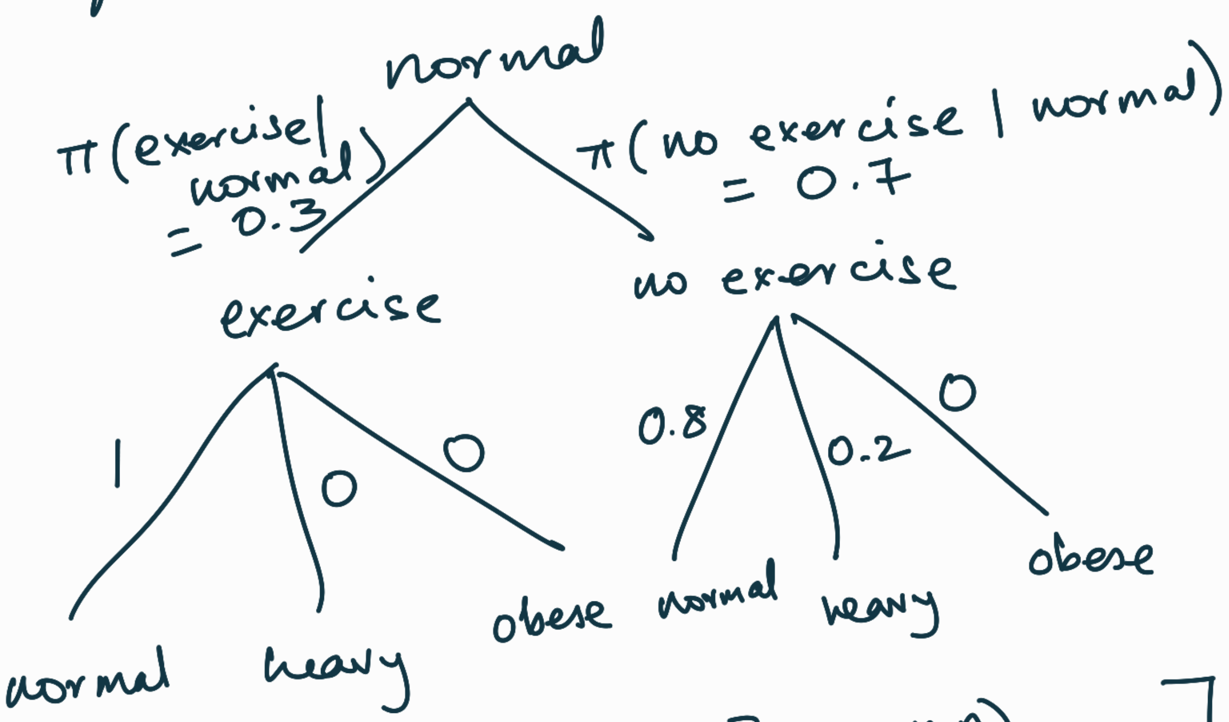
Q 4) Given,

$\pi(\text{exercise} \mid \text{normal}) = 0.3$

$\pi(\text{exercise} \mid \text{heavy}) = 0.4$

$\pi(\text{exercise} \mid \text{obese}) = 0.7$

2 level backup diagram starting from normal.



$$v_\pi(\text{normal}) = \pi(\text{exercise}|\text{normal})\left[r(n, ex, n) + \gamma v_\pi(\text{normal})\right]$$

$$+ \pi(\text{no exercise}|\text{normal})\left[0.8\left(r(n, noex, n) + \gamma v_\pi(\text{normal})\right) + 0.2\left(r(n, noex, h) + \gamma v_\pi(\text{heavy})\right)\right]$$

$$v_\pi(\text{normal}) = 0.3\left[0 + 0 \cdot v_\pi(\text{normal})\right]$$

$$+ 0.7\left[0.8\left(0 + 0 \cdot v_\pi(\text{normal})\right) + 0.2\left(-1 + 0 \cdot v_\pi(\text{heavy})\right)\right]$$

$$= 0.7(-0.2) = \underline{\underline{-1.4}}$$

Q5)

[10 points] Given a 3 X 3 grid world with 9 states,

S_0	S_1	S_2
S_3	S_4	S_5
S_6	S_7	S_8

Assume that the estimated optimal state values ($V \sim v^*(s)$) are as follows:

7	4	2
8	3	0
7	1	0

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

for starting from state S_0,

action,

up, $r(S_0, up, S_0) + \gamma\, v_\pi(S_0)$

$= -1 + 1 \times 7 = 6$

down, $r(S_0, down, S_3) + \gamma\, v_\pi(S_3)$

$= -1 + 1 \times 8 = 7$

left, $r(S_0, left, S_0) + \gamma\, v_\pi(S_0)$

$= -1 + 1 \times 7 = 6$

right, $r(s_0, \text{right}, s_1) + \gamma V_\pi(s_1)$

$= -1 + 1 \times 4 = -3$

greedy policy given start state is $s_0 = \max(\text{up}, \text{down}, \text{left}, \text{right})$

$= \max(6, 7, 6, -3)$

$= 7$

which is for moving down.

↓		

Repeat the above steps for all states and fill the arrow marks of all the states.

↓	←	↓
←	→	
↑	→	