Page No.
Date

Assignment No:- 5

1] What is Reinforcement learning? Draw suitable diagram and describe elements of Reinforcement learning.

Ans] i] Reinforcement learning is an autonomous self-teaching system that essentially learns by trial and error.

ii] It uses algorithms that learn from outcomes and decide which action to take next.

iii] It performs actions with the aim of maximizing rewards, or to achieve the best outcomes.

iv] After each action the algorithm recieves feedback that helps it determine whether the choice it made was correct, neutral or incorrect.

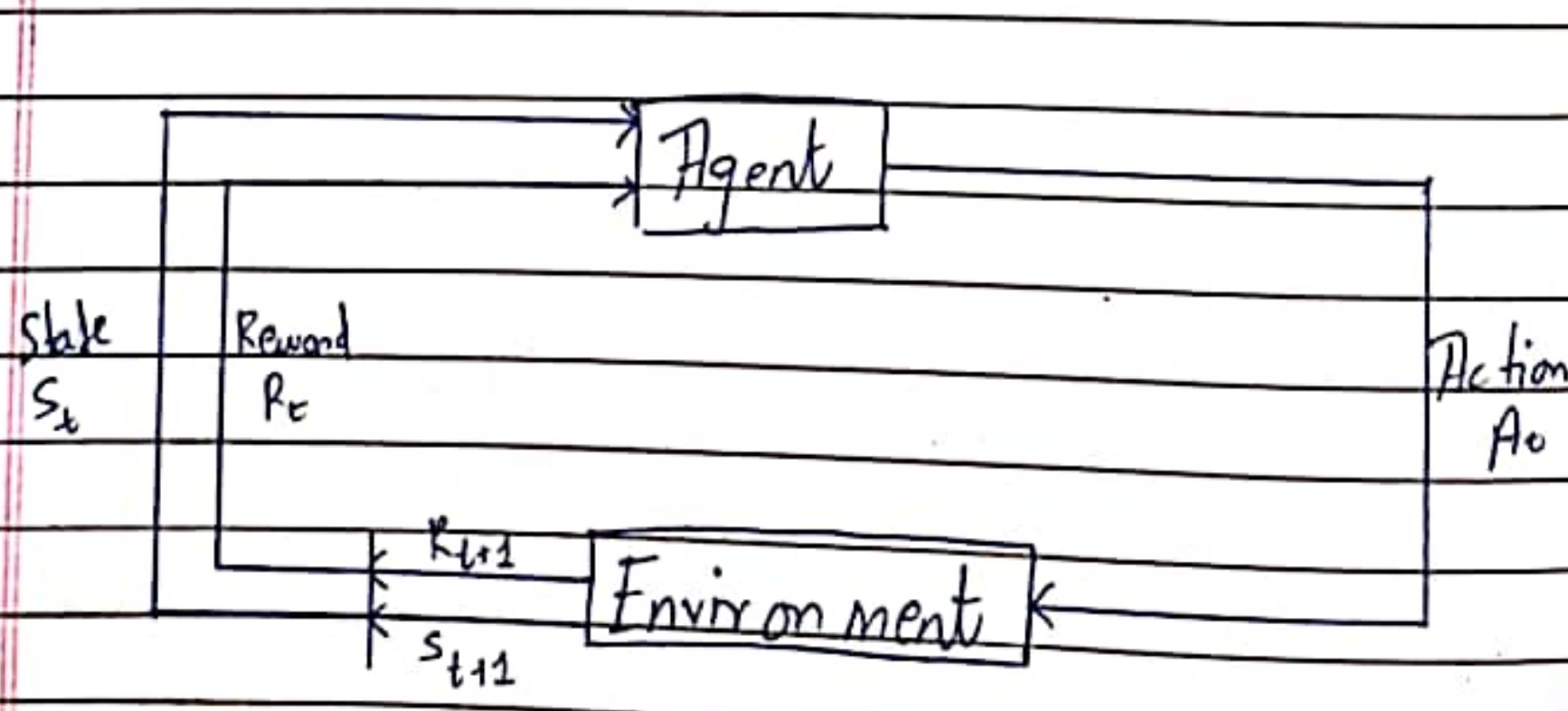

Fig:1 Block diagram of Reinforcement learning

There are five key elements of reinforcement learning:

i] The agent or the learner

ii] The environment the agent interacts with.

iii] The reward signal that the agent observes upon taking actions.

iv] Actions: The moves are chosen and performed by the agent to gain rewards.

v] The policy that the agent follows to take actions.

Q-2] Explain Exploration vs Exploitation dilemma.

Ans] Exploitation: Exploitation is referred as greedy approach in which agents try to get more rewards by using estimated value but not the actual value. So, in this technique, agents make the best decision based on current information.

Exploration: Unlike exploitation, in exploration techniques, agents primarily focus on improving their knowledge about each action instead of getting more rewards so that they can get long term benefits. So, in this technique, agents works on gathering more

information to make the best overall decision.

Let's understand the dilemma of exploration vs exploitation.

In Reinforcement learning whenever agents get a situation in which it has to make a difficult choice between whether to continue the same work or explore something new at a specific time.

This situation results in exploration-exploitation dilemma because the knowledge of an agent about the state, actions, rewards and resulting states is always partial.

3] Describe Epsilon Greedy Algorithm.

Ans:i] In Q-learning we select an action based on its reward. The agent always chooses the optimal action. Hence, it generates the maximum reward possible for the given state.

ii] In Epsilon Greedy action selection, the agent uses both exploitations to take advantage of prior knowledge and exploration to look for new problems.

iii) Epsilon Greedy policy is defined as a technique to maintain a balance between exploration and exploitation.

iv) To choose between exploration and exploitation, a very simple method is to select randomly.

v) This can be done by choosing exploitation most of the time with a little exploration.

vi) To find if the agent will select exploration or exploitation at each step, a random number is generated between 0 and 1 and compare it to the position.

vii) If this random number is greater than ε, then the next action would be decided by exploitation method.

viii) In the case of exploitation method, the agent will take the action with the highest Q-value for the current state.

Ex: random_number = random()

if random_number > epsilon:
    // choose next action via exploitation
else:
    // choose next action via exploration

Page No.
Date

4) Draw suitable diagram and describe Markov Decision Process (MDP)

~~Ans)~~ Q.

Ans) i) In Mathematics, a Markov Decision Process (MDP) is a discrete-time stochastic control process.

ii) It describes a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker

iii) MDP gives us a way to formalize sequential decision making.

iv) This formalization is the basis for structuring problems that are solved with reinforcement learning.

v) Components of an MDP are:

- Agent
- Environment
- State
- Action
- Reward

vi) In an MDP, we have a decision maker called an agent.

vii) Agent interacts with the environment it's placed in.

viii) These interactions occur sequentially over time.

ix) At each time step, the agent will get some representation of the environment state.

x) Given this representation, the agent selects an action to take.

x] The environment is then transitioned into a new stak.

xi] The agent is given a reward as a consequence of the previous action.

xii] The process of selecting an action from a given state, transitioning to a new state, a recieving a reward happens sequentially over and over again.

xiii] Which creates something called a trajectory that shows the sequence of stakes, action, and reward.

xiii] Throughout this process, it's the agents goal to maximize the total amount of rewards that it recives from taking actions in given states.

xiv] This means that the agent wants to maximize not just the immediate reward, but the cummulative reward it recieves overtime.

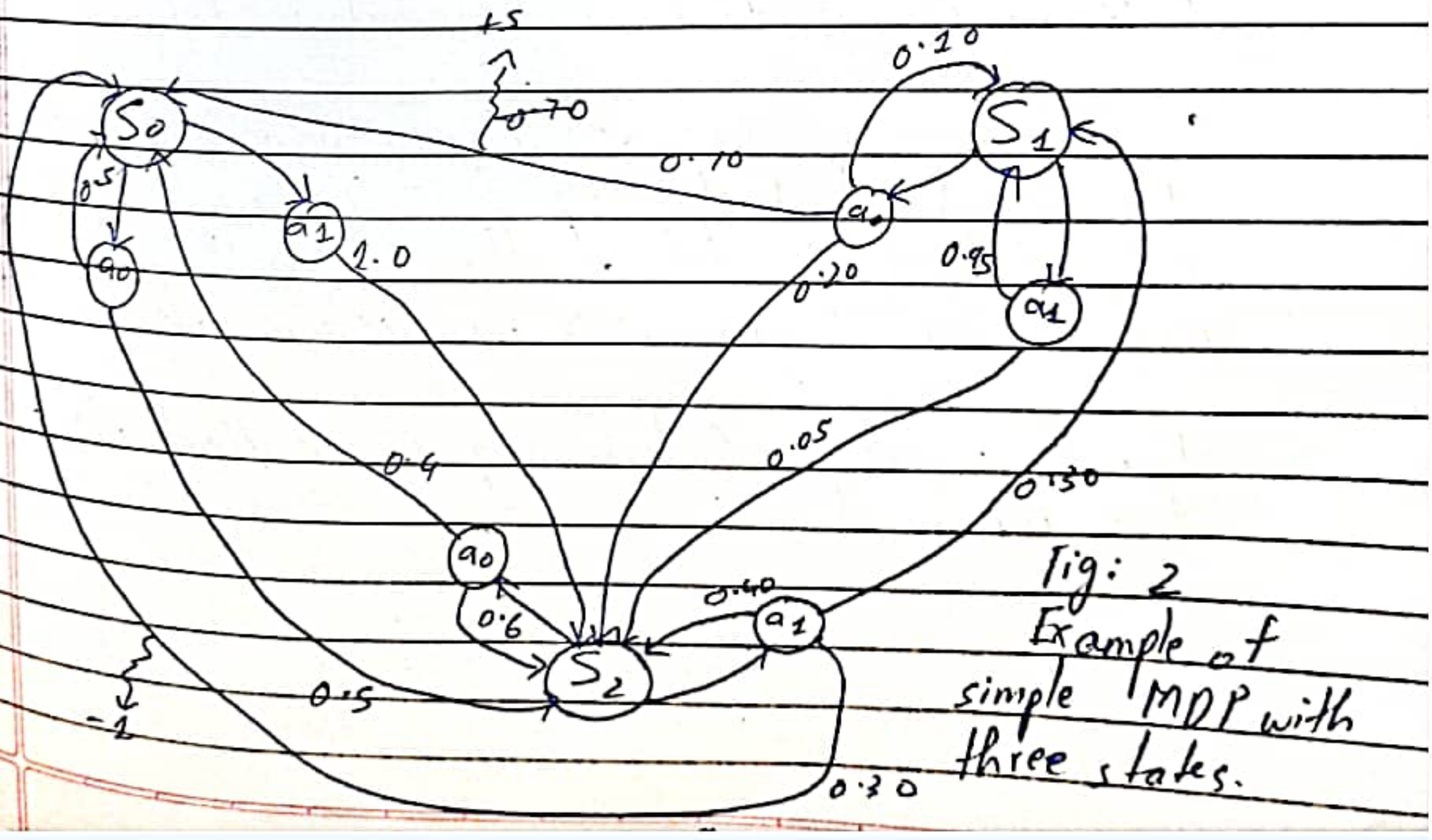

Fig: 2 Example of simple MDP with three states.

5) Derive necessary equation and brief Q learning algorithm

Ans:- i) Q-learning is a model free reinforcement learning algorithm to learn the value of an action in a particular state.

The value of doing action a in state 's' can be denoted by -

$$V(s) = \max_a Q(s, a)$$

Constraint equation that must be hold at equilibrium when Q-values are correct.

$$Q(s, a) = R(s) + \gamma \Sigma_{s'} P(s'|s, a) \max_{a'} Q(s', a')$$

We can use this equation as an update equation for an iteration process that calculates exact Q-values given an estimated model.

Q-function possesses a very important property, the agent that learns a Q-function does not need a model of the form $P(s'|s, a)$ either for learning or for action selection.

For this reason, Q-learning is called model free method.

Date

For no model state transition, update equation is -

$$Q(s,a) \leftarrow Q(s,a) + \alpha (R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$