1) imagine an investor trying to optimize your trading strategy for ~~your~~ four different stocks labeled A, B, C, D. Each stock has its own unique potential to profit, which is unknown for you. To maximize your return over a series of 100 trades, you decide to implement an ε-greedy strategy with ε being 0.1. the actual return from each stock follow

Stock A : 70% chance of +1 return,
30% chance of 0.

stock B : 50% chance of +2 return
50% chance of 0.

Stock C : 10% chance of +5 return
90% chance of 0.

Stock D. : guaranteed return of + 0.5

Ans (a) show how do you model this as an Reinforcement Learning problem

Ans) b) An investor intends to buy 100 times (each time buying one share of one stock)

The strategies the investor chose

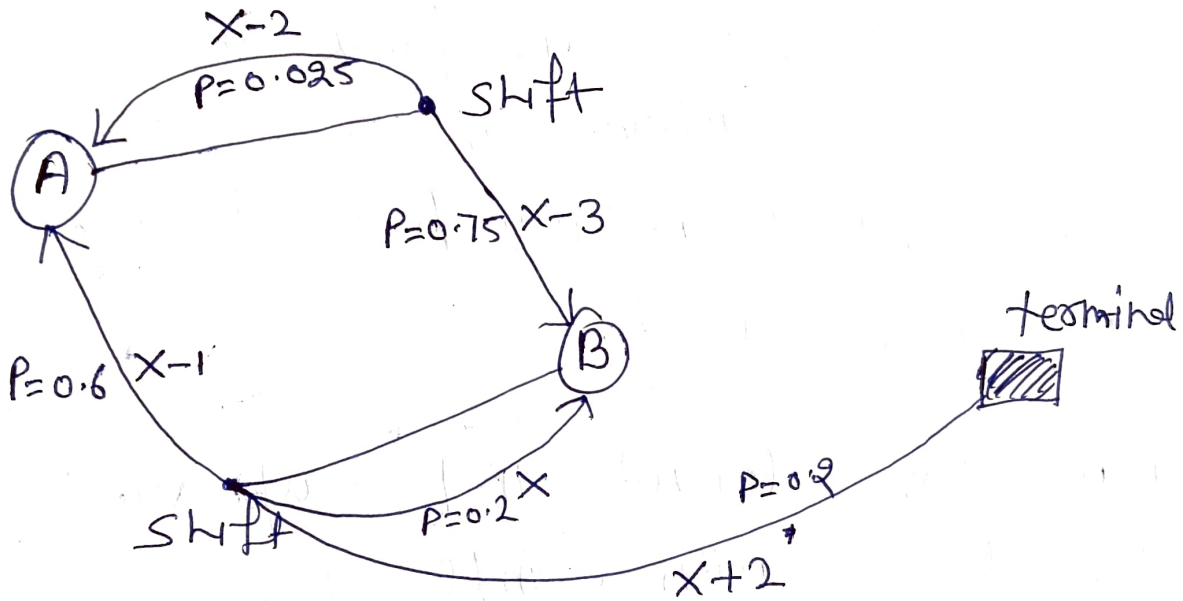
(i) follow ε-greedy for initial 25 trading and only exploit the information for next 75 purchase

(ii) follow ε greedy for initial 75 trading only exploid 25 next purchases

(iii) follow ε greedy for all 100 purchases support the investor with your analysis. show all step, tabulate ans to all the option and write conclusion

(c) What MDP, POMDP and CMDP? What are they? Suggest one RL technique that is used to solve problem stated with them It is adequate if you write just one line for each

Q-2: Consider the MDP given below containing 2 states A & B with action shift that may result in A, B or terminal state. The Reward obtained are indicated along the edges in the figure (X-2, X-3, X-1, X+2). Treat X to be 6. The transition probability are also given along the edges. discount factor = 0.4.



(a) Evaluate the given deterministic policy where the shift always executes the higher probability action. Improve it upto 1 iterations, use dynamic programming solution to MDP.

(b) With value iteration of dynamic programming determine the value of states A and B. Let the value of A & B initialized to 1, show 1 iteration

Q=3 (a) What are the most important issue when you have to learn the value function using a first visit Monte Carlo using for a deterministic policy. Explain also provide possible solution

Q=3(b) Explain any 3 most significant action selection strategy used in RL and mention how each selection method balances exploration and exploitation. provide your ans in table

Q=3(c) If we utilize policy gradient method to address a Reinforcement learning model and find that policy if provide not optimal, what could be the possible exploitation for HIS. State methods selected 3 learn

Q.4

(a) Why Alphago use a ~~Separate~~ Separate policy network and a seprate value network

(b) How does the MCTS ensure an action with the highest value is found in real time? If the best action can be selected only by MCTS, why is any prior learning of $Q(S,a)$ required.

4(c) We have learned that supervised learning that learns with samples from a given distribution does not capture the online ~~value~~ nature of interactions as required for reinforcement learning quite well.

(i) Why does Alphago use supervised learning to learn the initial policy (and even further)

ii) In what way the shortcoming of supervised learning are mitigated in Alphago?

4)(d) How does DQN handle the challenge referred to in the part (c) of this question

5 a) Consider the following ways of organizing Reinforcement learning techniques.

(i) Model based vs Model free

(ii) value based vs policy based

(iii) On policy vs off policy.

Write a statement or two on each of the points (for both category.) explaining the kind of problem these RL techniques are suited to. Provide your response in neatly organized table.

5(b) Consider the learning scenario. A human expert is presented with two trajectories taken by two drivers in the highway stretch. The human expert must rank which of the trajectory better. the agent learns this expertise (to locate a better trajectory by giving two unseen trajectory) observing the expert decision from many such example. Explain how you precisely

model this as an approach R2 problem.
show all the element of your model.
Making necessary assumption