Mahindra University Hyderabad
École Centrale School of Engineering
End Semester Examination (Regular)

Program: B. Tech. Branch: AI/CSE Year: IV Semester: I
Subject: Reinforcement Learning and Autonomous Systems (AI 4102)

Date: 22/12/2023 **Start Time: 10:00 AM**
Time Duration: 3 hours **Max. Marks: 100**

Instructions:

1) The submitted answer sheet should only contain your final answer. Use separate sheets for rough work.
2) You must show your solution procedure in the submitted answer sheet.

Q1: Contextual Bandits **(20 marks)**

(a) Consider the probabilistic graphical model (PGM) shown in Fig. 1.

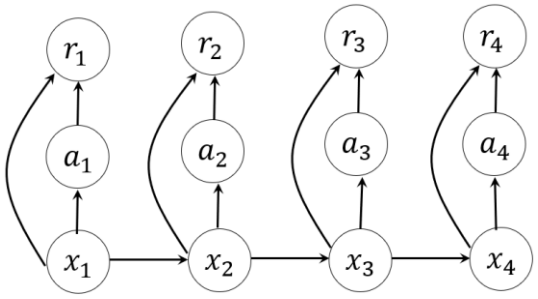


Fig. 1

Is the above PGM for multi-armed bandit, contextual bandit, or reinforcement learning? Explain you answer. Your answer shouldn't be more than **four lines**. **(4 marks)**

(b) There is a algorithm for bandits called "explore then commit", abbreviated as ETC, that was briefly discussed during lecture. In this question, you have to write a psuedocode for the same after reading the following description of ETC.

Consider that there are K actions in the action space and N contexts in the context space. The **exploration phase** starts first. It continues till for each of the N contexts all the K actions has been taken atleast M times. After the exploration phase ends, **commit phase** begins. In commit phase, for each context, the agent chooses the action with the highest context-action value. Write a psuedocode for ETC. **(8 marks)**
IMPORTANT: The final psuedocode should be **neat** and not exceed **a page** of you answer script.

[P.T.O]

(c) What is the most glaring flaw of ETC from the perspective of "learning"? Explain it with a numerical example. **(4 marks)**
IMPORTANT: There is only one GLARING flaw that has been repeated during lecture time and again. No partial points for writing anything else.

(d) There is one more flaw of ETC that happens when there are a few contexts that are **very rare**. Point it out and suggest how to modify the psuedocode to account for this flaw. You DO NOT have to write the entire psuedocode again. **(4 marks)**
HINT: Commit phase for different contexts will start at different times.

Q2: Markov Decision Process **(20 marks)**

Consider a smart residential area that draws energy from a transmission grid to satisfy its electricity demand. It also has a battery unit to store power for electricity demand of the residential area. The battery is also associated with "health".

We consider a time slotted model. Let the electricity demand of the residential area at time t be d_t units. d_t is an iid process and the probability of demand, d, is θ_d. The battery level at time t is $b_t \in \{0,1,\dots,N_b\}$ where N_b is the battery capacity. The health of the battery at time t is $h_t \in \{0,1,\dots,H\}$ where 0 is the lowest and H is the highest battery health. In a given time slot, if the battery level is greater than 80% of the battery capacity, then the health of the battery will go down by one level with a probability ϕ (unless it is at its lower level) and stay in the current health level with probability $1-\phi$. If the battery level is less than or equal to 80%, there is no change in battery health.

In every time slot, the residential area has to decide the amount of energy to draw from the transmission grid, u_t. There are two cases:

1. If $u_t \leq d_t$, then it means that energy drawn from the grid is less than the demand and hence the remaining energy to satisfy the demand has to be drawn from the battery.
2. If $u_t > d_t$, then the energy drawn from the grid is greater than the demand. In this case, d_t units of energy is used to serve the demand and the remaining $u_t - d_t$ units is used for charging the battery. However, if the battery health is h_t, then $ceil\left(\alpha_{h_t} \cdot (u_t - d_t)\right)$ units of energy, where $\alpha_{h_t} \in [0,1]$ and $ceil(\cdot)$ is the ceiling function, is used to charge the battery and remaining is lost as heat. α_{h_t} is monotonic decreasing in h_t.

The cost of drawing u units of energy from the transmission grid is u^2 per time slot. The objective of the smart residential area is to minimize the β-discounted cost of drawing energy from the transmission grid. Answer the following questions: **[P.T.O]**

(a) What is the state and state space for this problem? **(3 marks)**

(b) What is the action and action space for this problem? **(2 marks)**

(c) Define the cost for all state-action pair for this problem? **(3 marks)**

(d) Write an equation for b_{t+1} in terms of b_t, d_t, h_t, and u_t, and α_{h_t}. **(4 marks)**

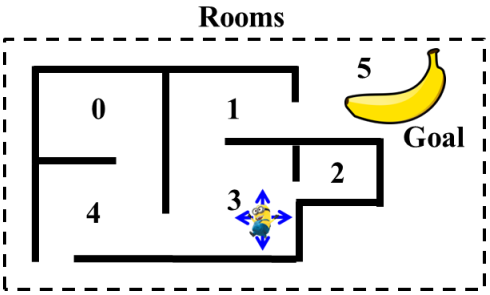
(e) What is the Bellman optimality equation for this problem? **(8 marks)**

Q3: Reinforcement Learning **(20 marks)**

Answer the following questions:

(a) Briefly explain Discount factor (γ), Learning Rate (α) and $\epsilon-$ greedy? Explain the significance of these parameters to build a model? **(10 Marks)**

(b) Explain briefly what is Q-learning with a suitable example? Explain with Reward and Q-Matrices to solve the following environment? **(10 Marks)**



Q4: Deep Reinforcement Learning **(20 marks)**

(a) Any reinforcement learning algorithm must ensure that every state-action pair is sampled infinitely often. How does policy gradient reinforcement learning algorithms ensures that this criterion is met? **(2 marks)**

(b) In general, ε-greedy strategy is used for exploration in Deep Q Learning. However, the drawback of ε-greedy strategy is that **during exploration**, those actions that are less likely to be optimal has the same probability of getting chosen as compared to those actions that are more likely to be optimal. Write the **pseudocode** of a function whose **inputs** are the DQN

model, the current state, and the ε value and whose **output** is the chosen action. The function that you write should satisfy the following conditions: **(5 marks)**

i. NO state-action pair should have **zero** probability of getting chosen.

ii. For a given state, the action that is more likely to be optimal should have higher probability of getting chosen during the exploration phase.

HINT: Think about the final layer of policy gradient.
IMPORTANT: The final psuedocode that you write should be **neat** and not exceed a **half a page** of you answer script.

The remaining questions of Q4 if motivated by the idea of extending Deep Q Learning to continuous actions. As you may know that Deep Q-Learning that was taught during lecture can't be directly extended to continuous action space. One way to extended Deep Q-Learning to continuous action space is to approximate Q-functions with specific forms that can be easily maximized with respect to the actions. One such form of Q-function is:

$$\hat{Q}(x, a; w) = V(x; w) - \big(a - \mu(x; w)\big)^T P(x; w)\big(a - \mu(x; w)\big) \quad (1)$$

where $a \in \mathbb{R}^N$, N being the dimensionality of the action space. In equation (1), $V(x; w)$, $\mu(x; w)$, and $P(x; w)$ are scalar, vector, and a matrix respectively of appropriate dimensions that are obtained as an output of **a neural network** with input, x, and weights, w (weights means the parameters of the neural network). If you are thinking how the matrix $P(x; w)$ can be an output of a neural network, simply think of it as a vector that is reshaped before plugging into equation (1). Answer the following questions:

(c) What is the number of outputs in the output layer of the neural network? **(2 marks)**

(d) Is training this neural network a regression or a classification problem? **(2 marks)**

(e) Why Deep Q-Learning taught during lecture can't be directly extended to continuous action space? **(2 marks)**

(f) Derive a closed-form expression for the optimal action by maximizing $\hat{Q}(x, a; w)$ in equation (1) with respect to action a. Assume that $P(x; w)$ is positive definite. **(4 marks)**

(g) How would you include exploration in continuous action space? Your answer should not exceed **five sentences**. **(3 marks)**

[P.T.O]

Q5: State Estimation **(20 marks)**

(a) Derive an expression for the recursive relation of Bayes filter. All the notations used during the derivation should be defined clearly. You DO NOT have to explain what state estimation is (you will lose points if you do). **(13 marks)**

(b) Consider Q2. In many situations, the agent may not know to the health of the battery; the agent have to estimate it. In every time slot, the agent knows the electricity demand, the battery level, and the energy drawn from the grid. Derive an expression for the belief of battery health, $bel(h_t)$. **(7 marks)**

HINT: Off course you can use Bayes filter to derive $bel(h_t)$. But you don't really need it. Your answer to Q2-(d) has everything that you need to answer this question.