DRL Midsem Exam

Questions and Detailed Solutions

Q1 Wristband Modes & Value Iteration

A mobile health-monitoring wristband must choose among three modes each hour, based on the current activity (resting, moderate, high):

- **Mode A:** low power, moderate accuracy
- **Mode B:** medium power, high accuracy
- **Mode C:** high power, very high accuracy

The immediate benefit (reward) from each mode changes based on the user's physical activity.

a) State why Reinforcement Learning (not supervised learning) fits this setting, and give a one-line distinction between immediate reward and long-term value. [1.5 Marks]

b) Using the same wristband scenario:

- **Case 1:** Initially, assume the user's activity pattern is independent of the modes chosen (choosing Mode A or B today does not affect future rewards).
- **Case 2:** Later, a firmware update introduces an adaptation rule: If the wristband samples in Mode C for too long, the battery temperature rises, reducing future rewards for power-hungry modes.

i) Classify Case 1 as a Multi-Armed Bandit (MAB) or a Finite MDP, with one justification. [1 Mark]
ii) Classify Case 2 again and justify does the added dependency change the problem class? [1 Mark]

c) For the MDP given below, calculate the Values of states = {resting, moderate activity, high activity} (in the same order), using synchronous Value Iteration.

- Actions = {Mode A, Mode B} are allowed from each state
- Discount factor: $\gamma = 0.1$
- Solve for 1 iteration only [4 Marks]

Solution

a) RL Suitability

Why RL fits: This is a sequential decision-making problem where the "correct" mode is unknown (no labels), and the agent must learn to maximize cumulative benefits (accuracy vs. power) through trial-and-error interactions.

Distinction: Immediate Reward (R_t) is the instant feedback after an action, whereas Long-term Value (G_t) is the expected sum of discounted future rewards.

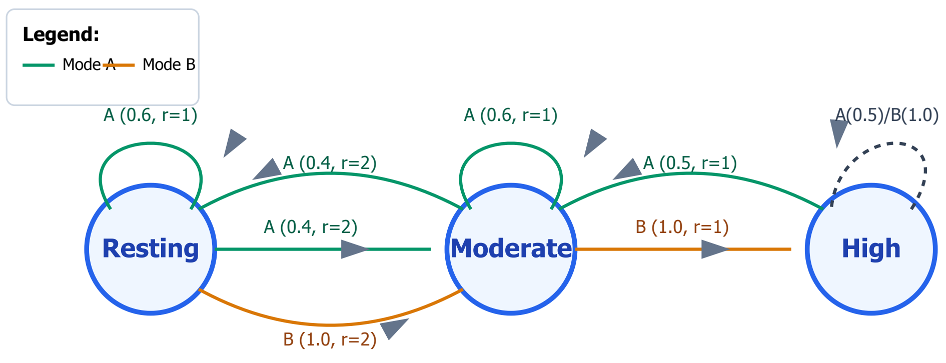
b) Classification

i) Case 1: Multi-Armed Bandit (MAB). Justification: The problem is stateless. The probability of rewards depends only on the current action, and actions do not influence future states.

ii) Case 2: Finite MDP. Justification: The firmware update creates a state dependency (battery temperature). The current action affects the *next state* and future rewards, satisfying the Markov property.

c) Value Iteration Calculation

Formula: $V_1(s) = \max_a \sum P(s'|s,a)R(s,a,s')$ (since $V_0 = 0$)



State	Calculations (Max Q)	V1 Value
Resting	$Q(A) = 0.6(1) + 0.4(2) = 1.4$ $Q(B) = 1.0(2) = 2.0$	2.0
Moderate	$Q(A) = 0.4(2) + 0.6(1) = 1.4$ $Q(B) = 1.0(1) = 1.0$	1.4

State	Calculations (Max Q)	V1 Value
High	$Q(A) = 0.5(1) + 0.5(1) = 1.0$ $Q(B) = 1.0(1) = 1.0$	1.0

Q2 ICU Ventilator MDP

In an ICU environment, clinicians must periodically adjust a patient's ventilator settings to maintain optimal blood oxygen saturation (SpO_2). The system can be modeled as a finite MDP where the patient's oxygenation levels are defined as $lowO_2$, $highO_2$, or $OptimalO_2$.

The AI can *increase*, *decrease*, or *maintain* pressure. With 0.4 probability, increasing/decreasing does not change health, whereas maintaining always retains the condition.

Goal: Stabilize SpO_2. "Once for **four consecutive recordings**, the SpO_2 is observed to be Optimal, then the patient stabilizes."

a) Formulate the MDP. State components clearly. Diagrammatically represent model dynamics. [2.5 Marks]

b) Analyze the impact of the following reward designs on patient safety. Which do you prefer and why? [2 Marks]

- **Design A:** Moderating $lowO_2 \rightarrow OptimalO_2$, Reward = +10.
- **Design B:** Moderating $OptimalO_2 \rightarrow HighO_2$, Reward = -70.

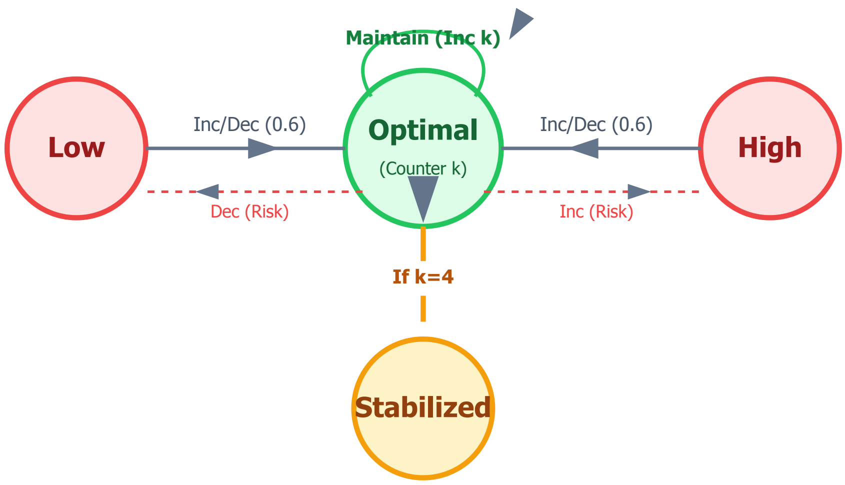
c) Is this episodic or continuous? Justify (< 30 words). [1.5 Marks]

d) If $\gamma = 0.01$, what is the impact on stability? Justify (< 30 words). [1.5 Marks]

Solution

a) MDP Formulation

- **States:** $\{Low, High, Optimal, Stabilized\}$. (Note: 'Optimal' implicitly tracks the consecutive counter).
- **Actions:** $\{Increase, Decrease, Maintain\}$.
- **Dynamics:** 'Maintain' in Optimal increments a counter; if count=4 $\rightarrow$ Stabilized. Inc/Dec have 0.6 prob to change state.



b) Reward Design

Preferred: Design B.

Why: In a clinical/ICU setting, *safety* is paramount. Design B's large penalty (-70) discourages risky exploration that causes over-ventilation, whereas Design A might encourage "cycling" (breaking health to fix it again for rewards).

c) Task Type

Episodic. The task terminates when the specific condition "patient stabilizes" is met, distinguishing it from continuous tasks.

d) Impact of $\gamma = 0.01$

Agent becomes myopic. It prioritizes immediate rewards and ignores the long-term goal (4 steps away), causing failure to stabilize.

Q3 Multi-Armed Bandit (Interventions)

Interventions: **S** (SMS), **T** (Teleconsult), **R** (Lab Reminder), **C** (Counseling).

Observations: `(1, S, 7), (2, T, 5), (3, R, 6), (4, C, 4), (5, S, 8), (6, T, 6), (7, R, 7), (8, C, 5)`

a) Estimate best intervention using Exponential Recency-Weighted Average ($\alpha = 0.5$). Initial values = 0. [3 Marks]

b) Significance of α? What if $\alpha = 1$? [1.5 Marks]

c) Significance of confidence level in UCB? [1 Mark]

d) Analyze steps 1-5 for ϵ-case occurrences (Random vs Greedy). [2 Marks]

Solution

a) Estimation Table ($\alpha = 0.5$)

Step	Action	Reward	Update: $Q_n = Q_o + 0.5(R - Q_o)$	S	T	R	C
1	S	7	$0 + 0.5(7 - 0) = 3.5$	**3.5**	0	0	0
2	T	5	$0 + 0.5(5 - 0) = 2.5$	3.5	**2.5**	0	0
3	R	6	$0 + 0.5(6 - 0) = 3.0$	3.5	2.5	**3.0**	0
4	C	4	$0 + 0.5(4 - 0) = 2.0$	3.5	2.5	3.0	**2.0**
5	S	8	$3.5 + 0.5(8 - 3.5) = 5.75$	**5.75**	2.5	3.0	2.0
6	T	6	$2.5 + 0.5(6 - 2.5) = 4.25$	5.75	**4.25**	3.0	2.0
7	R	7	$3.0 + 0.5(7 - 3.0) = 5.0$	5.75	4.25	**5.0**	2.0
8	C	5	$2.0 + 0.5(5 - 2.0) = 3.5$	5.75	4.25	5.0	**3.5**

Best Intervention: S (Value: 5.75)

b) Significance of α

Controls memory/forgetting factor. If $\alpha = 1$, the agent is memoryless and only uses the most recent reward as the value estimate.

c) UCB Confidence Level

Controls the **exploration bonus**. Higher confidence compels the agent to visit uncertain (less sampled) actions.

d) ϵ-case Analysis (Steps 1-5)

- **Definitely Occurred (Steps 2, 3, 4):** The agent chose T, R, C (Values=0) when S had a higher value (3.5). Choosing a suboptimal action implies exploration (random selection).
- **Possibly Occurred (Step 5):** The agent picked S (Max Value). While this looks greedy, ϵ-greedy strategies can still pick the optimal action randomly by chance.

Q4 Model-Free Control & MC

a) In a model-free MDP, explain why estimating only $v_\pi(s)$ is insufficient for selecting actions during control. [2 Marks]

b) Why might first-visit MC fail with a deterministic policy? Suggest a fix. [2 Marks]

c) Chatbot Episode Update.

States: s_0 (engaged), s_1 (disengaged). Actions: a_1, a_2, a_3.

Episode:

$$(s_0, a_1, 2, s_0) \rightarrow (s_0, a_3, 0, s_1) \rightarrow (s_1, a_2, 3, s_1) \rightarrow (s_1, a_2, -1, Term).$$

$\gamma = 0.8$.
Compute returns and provide the Revised Q-table and Updated Policy table.
[3.5 Marks]

Solution

a) Insufficiency of V(s)

In Model-Free RL (unknown dynamics P), knowing only state values $V(s)$ does not tell you which action leads to the best next state. You need **Action-Values** $Q(s, a)$ to directly compare the expected returns of different actions available in the current state.

b) Deterministic Policy Failure

A deterministic policy only ever visits one action per state, leaving others unsampled. MC cannot estimate values for unvisited actions. **Fix:** Use **Exploring Starts** or an **ϵ-soft policy** to ensure all actions have non-zero probability.

c) Chatbot Episode Calculation

1. Calculate Returns (G_t)

(s_0, a_1) **@ t=0:**
$G_0 = 2 + 0.8(0) + 0.8^2(3) + 0.8^3(-1)$
$= 2 + 1.92 - 0.512 = \mathbf{3.408}$

(s_0, a_3) **@ t=1:**
$G_1 = 0 + 0.8(3) + 0.8^2(-1)$
$= 2.4 - 0.64 = \mathbf{1.76}$

(s_1, a_2) **@ t=2:**
$G_2 = 3 + 0.8(-1) = \mathbf{2.2}$
(Ignore t=3 visit for First-Visit MC)

2. Revised Q-Table

State	a1	a2	a3
s0	**3.408**	0.5	**1.76**
s1	0	**2.2**	1

3. Updated Policy (Greedy)

State	a1	a2	a3
s0	**1.0**	0	0
s1	0	**1.0**	0