DEPARTMENT OF COMPUTER SCIENCE
ST. FRANCIS XAVIER UNIVERSITY

St. Francis Xavier University
Department of Computer Science

CSCI-531 - Reinforcement Learning
Practice Exercises: MDPs and Dynamic Programming

Part I: MDP Fundamentals

1. **MDP Components Identification**

   Consider a smart thermostat that learns to control room temperature efficiently. The thermostat can set the temperature to Low, Medium, or High, and the room temperature depends on outside weather and current setting.

   (a) Identify and define the four MDP components for this scenario:

   (b) Write the transition probability notation for: ”Given the room is Cold and we set thermostat to High, there's a 70% chance the room becomes Comfortable.”

   (c) Design a reward function that encourages energy efficiency while maintaining comfort.

2. **Markov Property Analysis**

   (a) For each scenario, determine if it satisfies the Markov Property. If not, suggest how to modify the state representation:

      1. A chess AI where the state is the current board position
      2. A stock trading bot where the state is only today's stock price
      3. A robot navigating where the state includes position, velocity, and battery level
      4. A recommendation system where the state is the user's last clicked item

   (b) An autonomous vehicle's state is defined as: `(current_speed, position, destination)`. Explain why this might not be Markovian and propose a better state representation.

3. (12 points) **Return Calculations**

   An agent receives the following reward sequence: $r_0 = 5, r_1 = -2, r_2 = 8, r_3 = 1, r_4 = 3$

(a) Calculate the undiscounted return G_0.

(b) Calculate the discounted return G_0 with $\gamma = 0.8$.

(c) Calculate G_2 with $\gamma = 0.9$.

4. **Simple MDP Analysis**

Consider a 2-state MDP with states $S = \{s_1, s_2\}$ and actions $A = \{a_1, a_2\}$:

Transition probabilities:

- $p(s_1|s_1, a_1) = 0.7, p(s_2|s_1, a_1) = 0.3$
- $p(s_1|s_1, a_2) = 0.2, p(s_2|s_1, a_2) = 0.8$
- $p(s_1|s_2, a_1) = 0.4, p(s_2|s_2, a_1) = 0.6$
- $p(s_1|s_2, a_2) = 0.1, p(s_2|s_2, a_2) = 0.9$

Rewards: $r(s_1, a_1) = 2, r(s_1, a_2) = 1, r(s_2, a_1) = 0, r(s_2, a_2) = 3$

Discount factor: $\gamma = 0.9$

(a) Consider the deterministic policy $\pi(s_1) = a_1, \pi(s_2) = a_2$. Write the system of Bellman equations for this policy.

(b) Solve the system to find $v_\pi(s_1)$ and $v_\pi(s_2)$.

(c) Write the Bellman optimality equations for both states.

Part II: Dynamic Programming

5. **Policy Evaluation Algorithm**

Consider a 3×3 gridworld where an agent starts at position (0,0) and wants to reach the goal at (2,2). The agent can move Up, Down, Left, Right, but there's a 10% chance of staying in place on each move.

Actions that would move outside the grid result in staying in the current position. Reward: -1 per step, +10 for reaching goal. Discount factor: $\gamma = 0.9$

(a) Define the policy evaluation update equation for this problem.

(b) Consider a simple policy: "always move right if possible, otherwise move up". Perform 2 iterations of policy evaluation starting with $V_0(s) = 0$ for all states. Show calculations for at least 3 states.

6. **Value Iteration vs Policy Iteration**

(a) Compare Value Iteration and Policy Iteration algorithms by filling in the table:

Aspect	**Policy Iteration**	**Value Iteration**
Update Equation		
Convergence		
Per Iteration Cost		
When to Use		

(b) For the 2-state MDP from Question 4, perform one iteration of Value Iteration starting with $V_0(s_1) = 0, V_0(s_2) = 0$. Show all calculations.

7. **Policy Improvement**

Given the value function $v_\pi(s)$ for some policy π: - $v_\pi(s_1) = 5.2$ - $v_\pi(s_2) = 8.7$ - $v_\pi(s_3) = 3.1$

And the MDP parameters from a 3-state system: - $\gamma = 0.8$ - Rewards: $r(s_1, a_1) = 2, r(s_1, a_2) = 1$ - Transitions from s_1: $p(s_2|s_1, a_1) = 0.6, p(s_3|s_1, a_1) = 0.4$ - Transitions from s_1: $p(s_1|s_1, a_2) = 0.3, p(s_2|s_1, a_2) = 0.7$

(a) Calculate $q_\pi(s_1, a_1)$ and $q_\pi(s_1, a_2)$.

(b) What action should the improved policy π' take in state s_1?

(c) (3 points) Is the current policy π optimal in state s_1? Justify your answer.