reasoning steps of MDP formulation for RL

🧭 Thinking Process for Formulating an MDP from a New Problem

0. What is the for

When facing a new problem, we often don’t know in advance what the agent, environment, or rewards are.

This guide will help you systematically think through the process of identifying the components of a **Markov Decision Process (MDP)** step by step.

---

1. Who is the decision-maker? (Agent)

Ask:

> *Who or what makes decisions in this system?*

- The **agent** is the entity that chooses actions.
- It has some control or freedom to decide what to do at each step.
- If nothing in the system makes active decisions, the problem may not be suitable for reinforcement learning.

Examples:

- The robot’s controller in a navigation task.
- A trading algorithm adjusting portfolio weights.
- A recommendation system choosing which item to show a user.

---

2. What can the agent choose? (Action / Action Space)

Ask:

> *What are the agent's options? What can it control or change?*

- The **action space** defines all possible decisions the agent can make.
- Actions can be discrete (move left, right) or continuous (apply a certain torque).
- The action may be the decision variable $\mathbf{x}$ in your optimization formulation:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} J(\cdot)$$

- Sometimes, the action space might not be intuitive such as pointers in task allocation problem (if you don't know, skip this sentence for now).

Examples:

- Move up, down, left, or right on a grid.
- Change the throttle, roll, or pitch of a drone.
- Allocate 40% of funds to stock A, 60% to stock B.

---

3. What changes as a result of the action? (Environment)

Ask:

> *When the agent takes an action, what part of the system reacts or changes?*

- The **environment** is the world outside the agent that responds to the agent's actions (any other things other than the decision-maker's brain).
- It produces the next situation and feedback (reward)––in reality you need to design tho.

Examples:

- The drone's motion after a control input.
- The user's behavior after seeing a recommended video.
- The next market state after an investment decision.

---

4. How can we represent the changing situation? (State / State Space)

Ask:

> *How can we describe the situation at any moment in a numerical way?*

- The **state** is a compact mathematical representation (often a vector) of the current situation the agent is in.
- It should include enough information to predict what might happen next if an action is taken.
- It can be any mathematical form but typically they are either tensors, graphs, or a dictionary (i.e., key-value pairs).

Examples:

- For a mobile robot: position, velocity, and orientation.
- For a trading system: portfolio value, recent returns, and volatility.
- For a game: positions of all pieces on the board.

---

5. How does the world change from one state to another? (Transition Function)

Ask:

> *If the agent takes an action, how does the next state depend on the current state and action?*

- The **transition function** $P(s'|s,a)$ describes how the system evolves.
- It can be deterministic (fixed rules) or stochastic (probabilistic outcomes).

- Even if we don't know it explicitly, reinforcement learning methods can still learn from experience, if you use deep neural networks (or any other approximation approaches).
- If you chose *unconventional* action/state space, it might be counter-intuitive such as pointers as the action space with task status as the state space. If you don't know of it, ignore it for now.

Examples:

- Physics equations determine how a robot's position changes.
- User behavior models determine the probability of clicking a video.
- A weather model determines how conditions change after each action.

---

6. How good or bad is the result? (Reward Function)

Ask:

> *What immediate feedback should the agent receive after each action?*

- The **reward function** $R(s, a, s')$ defines the goal of the problem.
- It assigns numerical feedback based on how desirable the new state is.
- Designing a good reward often requires creativity and understanding of the task's true objective.

Examples:

- +100 for reaching the goal, -100 for crashing.
- +1 for a successful click, 0 otherwise.
- Profit increase minus transaction cost.

---

7. (Optional) How long does the decision process last? (Time Horizon and Discounting)

Ask:

> *Is this a one-step or multi-step decision problem? What is a time step here? Should future rewards matter?*

- If the task unfolds over multiple time steps, we define a **time horizon** and a **discount factor** \gamma.
- The discount factor controls how much future rewards influence the agent's decisions.
- \gamma \approx 0: Focus on immediate rewards––often not desired in most RL.
- \gamma \approx 1: Consider long-term outcomes.

Examples:

- A single advertisement choice → short horizon.
- Robot navigation to a far goal → long horizon.

---

8. (Optional) Can the agent fully observe the environment? (Observation / POMDP)

Ask:

> *Does the agent see the whole state, or only partial information?*

- If the agent cannot observe the full state, the problem becomes a **Partially Observable MDP (POMDP)**.
- Then, we define observations o_t and an observation function $O(o_t|s_t)$.
- If one observation can describe more than one state, then it's PO.

Examples:

- A drone's camera sees only part of the world and the other parts are required for the task.
- A stock trader sees only public market data, not hidden signals.

---

9. How does the agent decide what to do? (Policy)

Ask:

> *Given the current state, how does the agent choose its next action?* That thinking/reasoning is the policy.

> Choosing the best action given the current state: does it align with your optimization problem?

- The **policy** $\pi(a|s)$ is the mapping from states to actions.
- It represents the agent's strategy.
- The goal of reinforcement learning is to find the optimal policy π^* that maximizes expected cumulative reward.

---

10. Final Check: Can all components form a consistent MDP?

Ask:

> *Do all the defined elements fit together coherently?*

- Before proceeding, verify that everything fits together: $\text{MDP} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$
- Ask yourself:
  - Are the state and action definitions clear and measurable?
  - Does each action produce a meaningful state transition?
  - Does the reward truly reflect the task objective?
- If all answers are yes, the problem is well-formulated as an MDP.

⚙️ Example Applications of the Thinking Process

Example 1: Autonomous Drone Landing

Step	Thinking	Definition
1	The decision-maker?	Drone control system (agent).
2	What can it choose?	Control inputs: throttle, roll, pitch, yaw.
3	What changes?	Drone's position, velocity, and attitude (environment).
4	How to represent it?	State = [x, y, z, vx, vy, vz, roll, pitch, yaw].
5	How does it evolve?	Physics + wind disturbance (stochastic transition).
6	What's good or bad?	Reward = -distance from landing pad; crash = -100.
7	Time horizon?	Multi-step, $\Delta t = 0.025s$ discount factor ≈ 0.99.
8	Partial observation?	Yes, sensors have noise.
9	Policy?	Neural network mapping states → control actions.

Example 2: Robot in a Grid World

Step	Thinking	Definition
1	Agent?	Robot navigating the grid.
2	Actions?	Move up/down/left/right.
3	Environment?	The grid and walls.
4	State representation?	(x, y) position.
5	Transitions?	Deterministic or 10% chance to slip.
6	Reward?	+10 at goal, -1 per step.
7	Horizon?	Finite (until goal reached). Δt : one movement
8	Observation?	Fully observable.
9	Policy?	Table or NN mapping position → move.

Example 3: Investment Portfolio Management

Step	Thinking	Definition
1	Agent?	Portfolio management algorithm.
2	Actions?	Asset allocation ratios.
3	Environment?	Financial market dynamics.
4	State?	Current prices, holdings, volatility.
5	Transition?	Price changes with uncertainty.
6	Reward?	Profit minus risk or transaction cost.
7	Horizon?	Long-term; discount factor near 1.

Step	Thinking	Definition
8	Partial observation?	Yes, market uncertainty.
9	Policy?	Strategy mapping market state → allocation.

Example 4: Energy-efficient Building Control

Step	Thinking	Definition
1	Agent?	HVAC controller.
2	Actions?	Heating/cooling power level.
3	Environment?	Building thermal dynamics.
4	State?	Indoor temp, outdoor temp, humidity.
5	Transition?	Thermodynamic equations with delay.
6	Reward?	Comfort score - energy cost.
7	Horizon?	Long-term operation.
8	Observation?	Sensors only (partial).
9	Policy?	Control rule adjusting power level.