LECTURER: TAI LE QUY

INTRODUCTION TO REINFORCEMENT LEARNING

UNIT 3

DYNAMIC PROGRAMMING

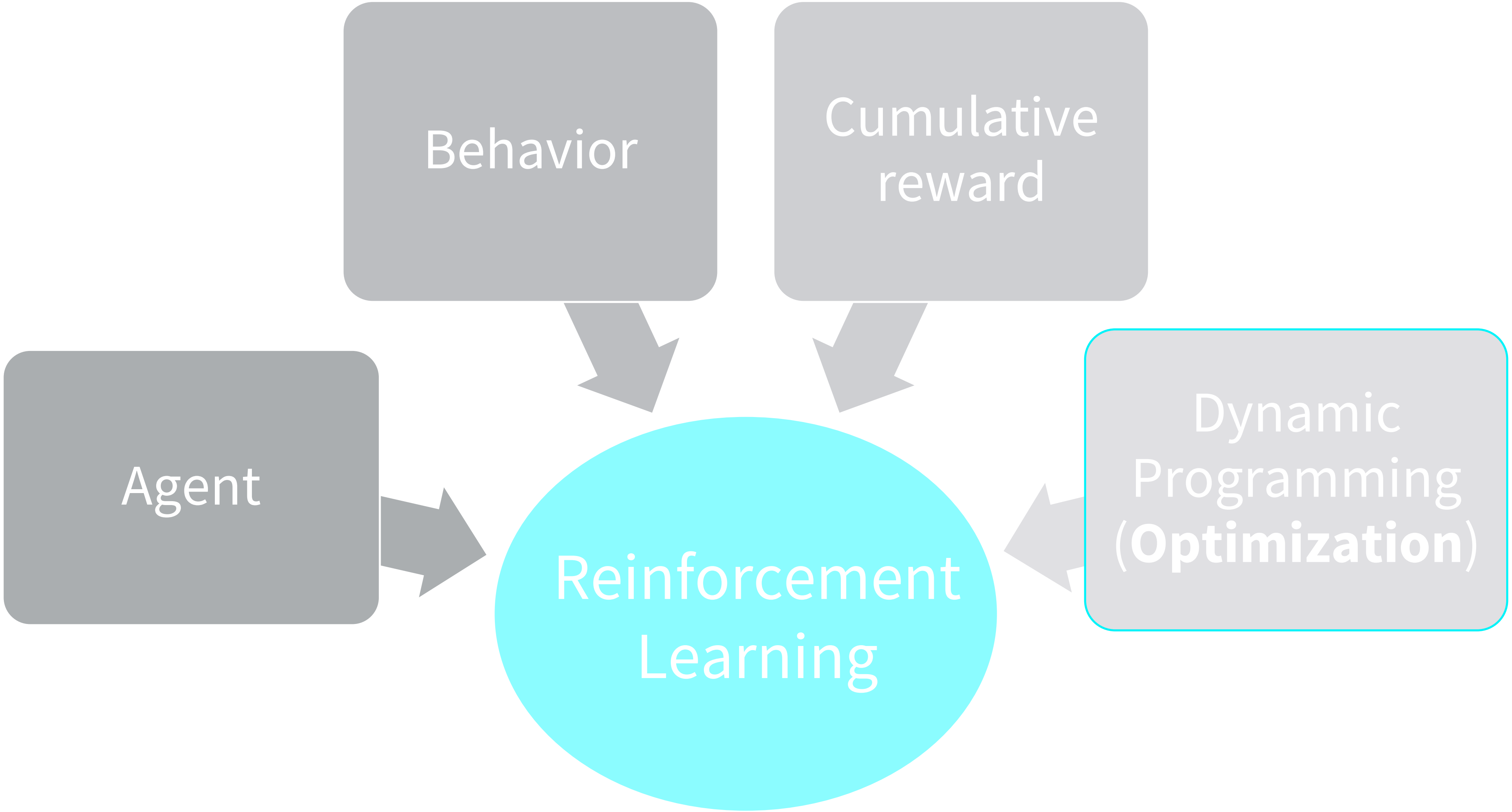
STUDY GOALS

- Explain the importance of policies and actions in Reinforcement Learning (RL)
- Evaluate and compare policies using value functions
- Describe how dynamic programming is applied to RL
- Utilize Bellman equations to optimize a RL problem and assess their effectiveness in finding an optimal policy

EXPLAIN SIMPLY

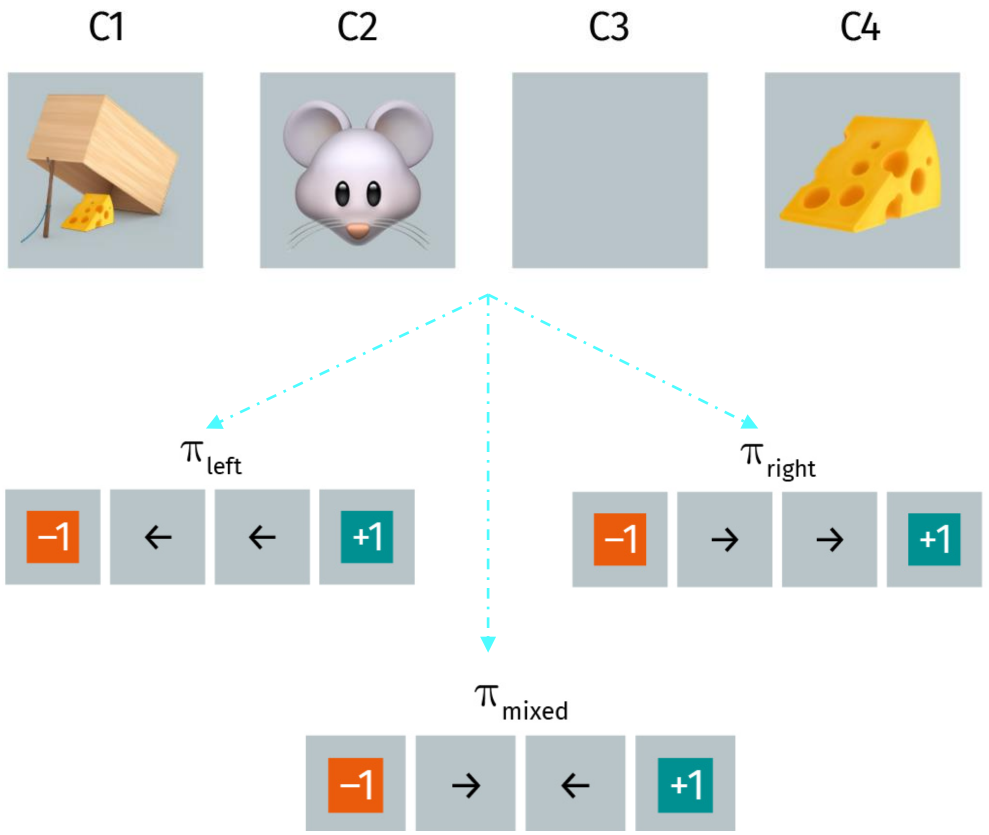
1. Explain the role of policies and actions in Reinforcement Learning
2. Describe how Bellman equations are used to compute state values and how they enable the iterative process of finding an optimal policy
3. Discuss the benefits of policy and value iteration in solving Reinforcement Learning problems

REINFORCEMENT LEARNING AND DYNAMIC PROGRAMMING



Policy is a mapping from states to actions

- Optimal policy = highest cumulative reward
- Defines agent's state actions
- Agent learns policies by experience



Source of the image: Nair, 2023.

POLICIES AND ACTIONS

Deterministic

- Map each state to one action
- Execute same action in a state
- Simpler, but prone to local optima

vs

Stochastic

- Map state to multiple actions
- Sample action from probability distribution
- Complex, but promote exploration and better solutions

Compute state or state-action pair value for achieving goal state

- Used to compare policies
- Assign value to states based on expected return from policy
- Return is sum of discounted rewards over trajectory

$$
\begin{aligned}
G_t &= r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \cdots \\
&= \Sigma_{\mathrm{t=0}}^{\infty} \; (\gamma^t \cdot r_t)
\end{aligned}
$$

Value function measures how **good** state or action is

- **State-value** function: measure state value under policy

$$V^*(s_t) = \max_{\pi} E\,[G_t \mid s_0 = s_t]$$

- **Action-value** function: measure value of state-action pair

$$Q^*(s_t, a_t) = \max_{\pi} E\,[G_t \mid s_0 = s_t, a_0 = a_t]$$

Bellman equations are the cornerstone of Reinforcement Learning

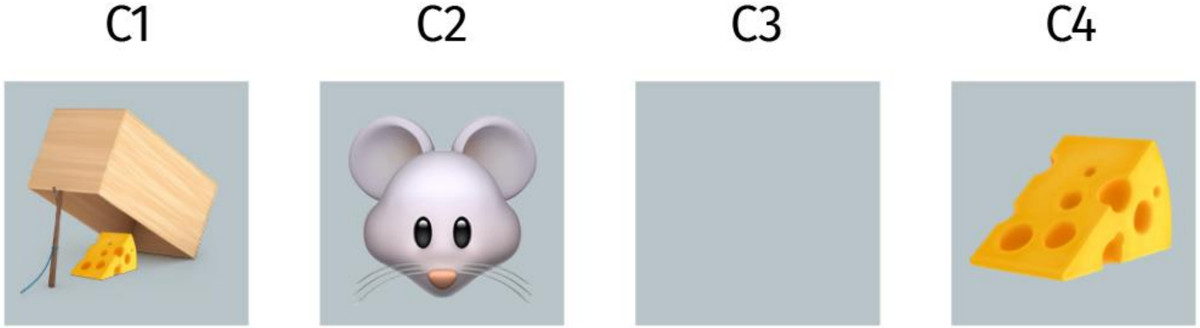
- Recursive value function

$$V^{\pi}(s_t) = \mathrm{E}\left[\mathrm{r_t} + \gamma \cdot V^{\pi}(s_{t+1}) \mid s_0 = s_t\right]$$

- Estimate value based on future rewards

$$Q^{\pi}(s_t, a_t) = E\left[r_t + Q^{\pi}(s_{t+1}, a_{t+1}) \mid s_0 = s_t, a_0 = a_t\right]$$

- Evaluate policies



MDP <S, A, R, P>

State space $S = \{C1, C2, C3, C4\}$

Action space $A = \{left, right\}$

Reward: R

Transition: $P = p(s_{t+1} \mid s_t, a_t)$

Source of the image: Nair, 2023.

Evaluate and improve policies for optimal rewards

Policy evaluation

- Initialize value of state V(S)
- Update V(s) using Bellman equation
- Repeat until convergence

Policy improvement

- Get best estimate for each state
- Look ahead to find best policy
- Pick action with highest reward

Value iteration is a *greedy* variant of policy iteration

- Combine policy improvement and truncated policy evaluation
- Select actions using the value from a pass of policy evaluation

Initialize V(s)

Repeat until convergence

    for each state s, update V(s) using Bellman optimality

    Update policy to be greedy with respect to V(s)

DYNAMIC PROGRAMMING

POLICIES AND ACTIONS

Define agent behavior

VALUE FUNCTIONS

Estimate long-term reward potential

BELLMAN EQUATION

Estimate optimal solution

POLICY AND VALUE ITERATION

Optimize policies

REVIEW STUDY GOALS

- Explain the importance of policies and actions in Reinforcement Learning (RL)
- Evaluate and compare policies using value functions
- Describe how dynamic programming is applied to RL
- Utilize Bellman equations to optimize a RL problem and assess their effectiveness in finding an optimal policy

SESSION 3

TRANSFER TASK

Case study

A company is designing a robot to navigate in a maze-like environment. The robot must make decisions at each intersection to reach the destination

Task

Using dynamic programming concepts, design an optimal policy for the robot to navigate the maze. Discuss how value functions Bellman equations can be used to estimate rewards, and policy and value iteration to find the optimal policy

Please present your results.

The results will be discussed in plenary.

1. Dynamic programming is a _________ for solving complex problems
   a) Programming language
   b) Genetic algorithm
   c) Optimization method
   d) Machine learning method

2. Policy iteration consists of two parts: policy evaluation and _________
   a) Policy improvement
   b) Policy update
   c) Value improvement
   d) Value update

3. The Bellman equations for v and q are ________ relationships
   a) Optimal
   b) Recursive
   c) Numerical
   d) Inequality

LIST OF SOURCES

Images

Nair, 2023.