Introduction to Reinforcement Learning

Foundations - What is RL? Agent-Environment Interaction

Sarwan Ali

Department of Computer Science
Georgia State University

 Understanding Reinforcement Learning 

Today's Learning Journey

1. What is Reinforcement Learning?
2. Agent-Environment Interaction
3. Key Concepts and Examples
4. Rewards and Goals

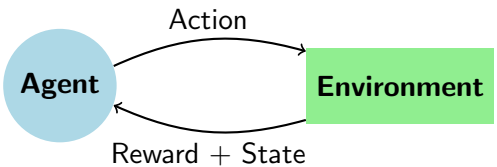
What is Reinforcement Learning?

Definition

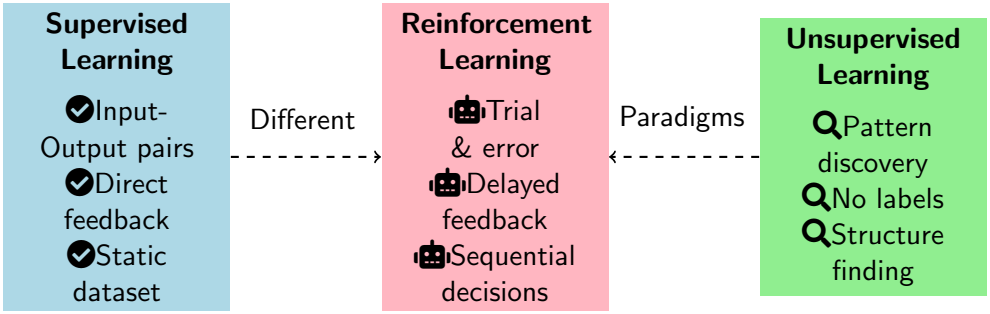
Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment to maximize cumulative reward.

Key Characteristics:

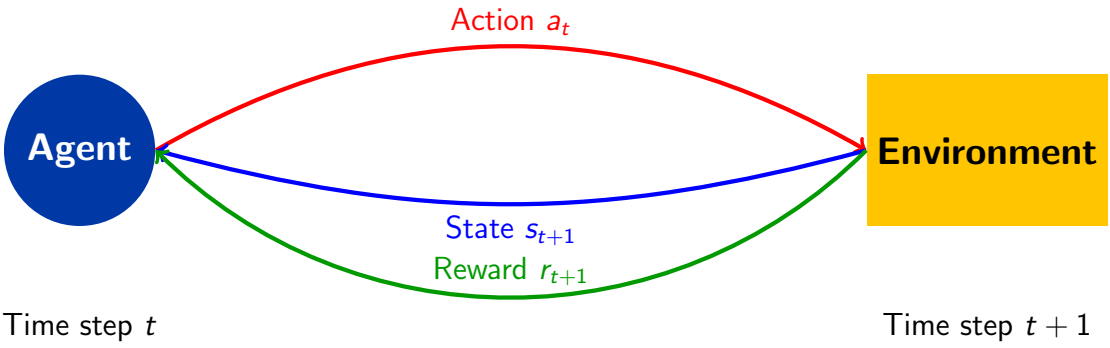
- Learning through trial and error
- No explicit supervision
- Delayed rewards
- Sequential decision making
- Goal: maximize long-term reward



RL vs Other Machine Learning Paradigms



The RL Framework: Agent-Environment Interaction



The RL Loop

At each time step t: Agent observes state s_t, takes action a_t, receives reward r_{t+1} and new state s_{t+1}

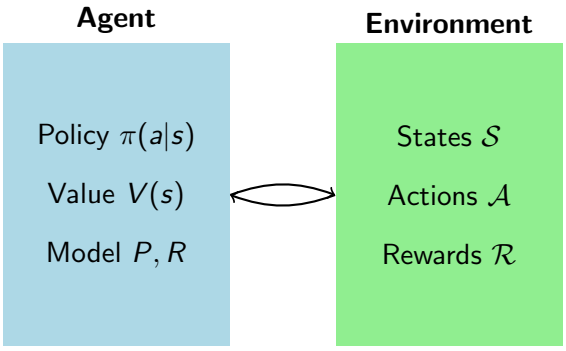
Components of RL System

Agent Components

- **Policy** π: Action selection strategy
- **Value Function** V: Expected future rewards
- **Model** (optional): Environment representation

Environment Components

- **State Space** $\mathcal{S}$: All possible states
- **Action Space** $\mathcal{A}$: All possible actions
- **Reward Function** R: Feedback mechanism
- **Transition Function** P: State dynamics



Formal Mathematical Framework

Markov Decision Process (MDP)

An RL problem is formalized as an MDP: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$

Components:

$\mathcal{S}$: State space (1)

$\mathcal{A}$: Action space (2)

P : Transition probability (3)

R : Reward function (4)

γ : Discount factor (5)

Key Equations:

$$P(s'|s,a) = \Pr[S_{t+1} = s' | S_t = s, A_t = a] \quad (6)$$

$$R(s,a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a] \quad (7)$$

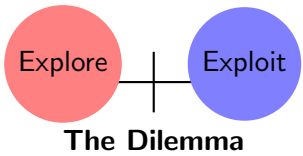
$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (8)$$

where G_t is the return (cumulative discounted reward)

The Learning Process: Exploration vs Exploitation

Exploration

- Try new actions to discover better strategies
- Risk short-term loss for long-term gain
- Essential for learning



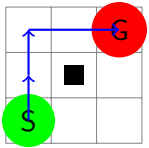
Exploitation

- Use current knowledge to maximize reward
- Play it safe with known good actions
- Optimize current performance

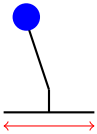
Restaurant Example

Exploitation: Always go to your favorite restaurant
Exploration: Try new restaurants to find potentially better ones

Classic RL Examples



Grid World
Navigate to goal while avoiding obstacles



Cart-Pole
Balance pole by moving cart left/right

Game Playing
Learn optimal strategies through self-play

Common Characteristics

Sequential decision making, delayed rewards, learning from interaction, goal-oriented behavior

Types of RL Environments

Deterministic
Same action → Same outcome

Stochastic
Same action → Random outcomes

Fully Observable
Agent sees complete state

Partially Observable
Agent has limited information

Episodic
Clear start and end points

Continuing
No natural ending

The Reward Hypothesis

Reward Hypothesis

All goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (reward).

Reward Design Principles:

- ✔ Clear objective signal
- ✔ Immediate when possible
- ✔ Scaled appropriately
- ✘ Avoid reward hacking
- ✘ Don't over-engineer

Examples:

- Chess: +1 win, -1 loss, 0 draw
- Robot navigation: -1 per step, +100 at goal
- Stock trading: Portfolio value change
- Game playing: Score difference

Policy and Value Functions

Policy π

Defines agent's behavior
$\pi(a|s) = \Pr[A_t = a|S_t = s]$

Types:

- Deterministic: $a = \pi(s)$
- Stochastic: $\pi(a|s)$

Value Functions

Estimate expected future rewards

State Value: $V^\pi(s) = \mathbb{E}[G_t|S_t = s]$
Action Value:
$Q^\pi(s, a) = \mathbb{E}[G_t|S_t = s, A_t = a]$

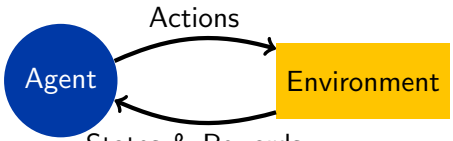
Relationship

$V^\pi(s) = \sum_a \pi(a|s)Q^\pi(s, a)$

Summary: RL Foundations

Key Takeaways

- **RL is learning through interaction** - no explicit teacher, just rewards
- **Agent-environment loop** is the core framework
- **MDP formalization** provides mathematical foundation
- **Exploration vs exploitation** is the fundamental tradeoff
- **Reward design** is crucial for success



The RL Paradigm

Next Topics

Comparison with supervised and unsupervised learning

Questions?

❓ Discussion ❓

Think about: What RL problems do you encounter in daily life?