Chapter 1 - Introduction

1.1 Introduction

Reinforcement learning is a field that focuses on teaching agents how to map situations to actions to maximize a numerical reward signal. Unlike supervised learning where the learner is provided with labeled examples, the learner in reinforcement learning must discover the most rewarding actions through trial and error. It stands as a unique paradigm in machine learning, distinct from supervised and unsupervised learning.

2 most important distinguishing features of reinforcement learning:

- **Trial-and-Error Search**: The learner must experiment to find the most rewarding actions.
- **Delayed Reward**: Actions may affect not just immediate rewards but also future states and subsequent rewards.

Definitions

- **Problem**: The task that needs to be solved.
- **Solution Methods**: Algorithms or techniques that solve the problem.
- **Field**: The area of study that explores the problem and solution methods.

It is crucial to distinguish between these three concepts to avoid confusion.

Formalization

Reinforcement learning is formalized using dynamical systems theory, specifically as the optimal control of incompletely-known Markov Decision Processes (MDPs). An MDP captures essential aspects such as:

- Sensation: The agent's ability to sense its environment.
- Action: The actions the agent can take to affect the environment.
- Goal: The objectives the agent aims to achieve.

Comparison with Other Learning Paradigms

- **Supervised Learning**: Involves learning from labeled examples provided by an external supervisor. It is not ideal for interactive problems where labeled examples are hard to obtain.
- **Unsupervised Learning**: Focuses on discovering hidden structures in unlabeled data. Unlike reinforcement learning, it does not aim to maximize a reward signal.

Reinforcement learning is thus considered a third machine learning paradigm, alongside supervised and unsupervised learning.

Challenges and Characteristics of Reinforcement Learning

a. **Exploration vs Exploitation Dilemma:**
One of the unique challenges in reinforcement learning is the trade-off between **exploration** and **exploitation**. To maximize rewards:
   - **Exploitation**: The agent must use actions that have proven to be rewarding in the past.
   - **Exploration**: The agent must try new actions to improve its understanding of the environment.

   Both cannot be pursued exclusively without compromising the learning process. This dilemma is unique to reinforcement learning and doesn't appear in pure forms of supervised or unsupervised learning.

   On stochastic tasks, actions must be tried multiple times to get a reliable estimate of expected rewards. The exploration-exploitation dilemma has been a subject of intensive study yet remains unresolved.

b. **Holistic Approach to Goal-Directed Agents:**
Reinforcement learning focuses on the entire problem of a goal-directed agent interacting with an uncertain environment. This stands in contrast to other fields that may focus on subproblems without considering the bigger picture.

   Key Features
   - **Explicit Goals**: The agent has well-defined objectives.
   - **Sensory Capabilities**: The agent can perceive its environment.
   - **Action Selection**: The agent can perform actions to influence the environment.
   - **Uncertainty**: The agent operates despite incomplete knowledge of its environment.

c. **Planning and Real-Time Decision Making**
When reinforcement learning involves planning, it also considers how planning integrates with real-time decision-making and how environmental models are developed and refined.

d. **Interdisciplinary Connections**
Reinforcement learning is closely connected with various other fields:

   - **Engineering**: Addresses challenges like the "curse of dimensionality" in operations research and control theory.
   - **Psychology and Neuroscience**: Offers insights into human and animal learning.

e. **Trend Towards Simplicity in AI**
Reinforcement learning is also part of a larger trend in AI focusing on simple, general principles. This counters a previous belief that intelligence was a result of numerous special-purpose tricks and procedures.

   Weak vs Strong Methods

Weak Methods: Based on general principles like search or learning.
Strong Methods: Based on specific knowledge.

Modern AI research, including reinforcement learning, is leaning more towards weak methods, searching for general principles of learning, search, and decision-making

1.2 Examples and Applications of Reinforcement Learning

Chess Player:

- Interaction: The player interacts with the opponent and the chessboard.
- Goal: To win the game.
- Learning: Uses both planning and intuition to improve performance over time.

Petroleum Refinery Controller

- Interaction: The controller interacts with the refinery systems.
- Goal: To optimize yield, cost, and quality.
- Learning: Adjusts parameters based on real-time data, improving efficiency.

Gazelle Calf

- Interaction: The calf interacts with its natural environment.
- Goal: To stand and then to run.
- Learning: Quickly adapts to its environment to achieve the ability to run at high speeds.

Mobile Robot

- Interaction: The robot interacts with the room and its charging station.
- Goal: To collect trash and recharge its battery.
- Learning: Uses past experiences to decide when to collect trash and when to recharge.

Phil's Breakfast Routine

- Interaction: Phil interacts with his kitchen environment.
- Goal: To prepare and enjoy breakfast.
- Learning: Streamlines the process over time, based on needs and preferences.

Common Features

- Interaction: All examples involve interaction between an agent and its environment.
- Goal-Oriented: The agents have explicit goals they strive to achieve.
- Uncertainty: The outcomes of actions are not fully predictable, requiring continuous monitoring.
- Learning: Over Time The agents improve their performance based on their experiences.
- Planning and Real-Time Adaptation: Agents need to consider both immediate and future consequences of their actions.

1.3 Elements of Reinforcement Learning

Understanding the various elements that make up a reinforcement learning system is essential. A typical system comprises four main sub-elements: a policy, a reward signal, a value function, and optionally, a model of the environment.

Policy

- Definition: A policy is a mapping from perceived states of the environment to actions that the agent takes. In psychological terms, it's akin to a set of stimulus-response rules.
- Importance: It is the core of a reinforcement learning agent and determines its behavior.
- Nature: Policies can be deterministic or stochastic, specifying probabilities for each action.

Reward Signal

- Definition: The reward signal defines the goal of the reinforcement learning problem. The environment provides the agent with a numerical reward at each time step.
- Importance: It shapes the agent's understanding of good and bad outcomes.
- Nature: Can be stochastic, depending on both the state and actions.

Value Function

- Definition: The value function represents the long-term desirability of states, accounting for the amount of reward an agent can expect over the future.
- Importance: Unlike the reward, which is immediate, the value function aids in long-term planning.
- Estimation: Estimating value functions is often the most critical part of reinforcement learning algorithms.

Environment Model (Optional)

- Definition: A model mimics the environment's behavior, predicting the next state and reward for a given state-action pair.
- Usage: Employed for planning, allowing the agent to think ahead before taking actions.
- Types of Methods: Methods using models are termed "model-based," while those that don't are called "model-free."

Interaction among Elements

- Policy and Reward: The reward signal is the primary basis for altering the policy.
- Reward and Value: While rewards are primary and immediate, values are secondary and long-term. Values are essentially predictions of future rewards.

- Model and Planning: In model-based methods, the model is used to simulate future scenarios for better planning.

Complexity

Reinforcement learning spans from simple trial-and-error methods to complex systems that involve high-level planning and model learning.

1.4 Limitations and Scope of Reinforcement Learning

Focus on State

- Importance of State: Reinforcement learning heavily relies on the concept of "state" as input for the policy, value function, and model. Informally, the state can be considered as a snapshot that tells the agent "how the environment is" at a particular moment.
- Scope: The book focuses on decision-making aspects and does not delve into the issues of constructing or learning the state representation.

Estimation of Value Functions

- Value Function Centric: Most reinforcement learning methods discussed focus on estimating value functions.
- Alternatives: There are methods like genetic algorithms, simulated annealing, and other optimization techniques that don't estimate value functions but can still solve reinforcement learning problems.

Evolutionary Methods

- Characteristics: Use multiple static policies, each interacting over time with separate instances of the environment. Policies that gain the most reward are carried over to the next generation, and the process repeats.
- Limitations: Do not learn while interacting with the environment. Ignore much of the problem's useful structure, such as the state-action mapping.

Why Not Evolutionary Methods?

While evolutionary methods can be effective in some cases, they are not particularly well-suited for problems where the agent can learn from individual interactions.

They lack the ability to take advantage of the nuances of individual state-action experiences, making them generally less efficient than learning methods for many reinforcement learning problems.

1.5 An Extended Example: Tic-Tac-Toe

The Game and the Challenge

- The game of Tic-Tac-Toe involves two players taking turns to mark a 3x3 grid.
- The objective is to construct a player that learns to maximize its chances of winning.

Classical Solutions and Their Limitations

- Classical "minimax" solutions from game theory are not applicable here as they assume perfect play from the opponent.

- Classical optimization methods like dynamic programming require a complete specification of the opponent's behavior.

Reinforcement Learning (RL) Approach

Value Function Approach

- **Initialization**: A table of numbers is initialized for each possible state of the game. Each number represents the estimated probability of winning from that state.
- **Policy**: Most of the time, the agent makes a "greedy" move to the state with the highest estimated value. Occasionally, it explores other moves.
- **Learning**: After each greedy move, the agent updates the value of the state based on the value of the resulting state. This is done using Temporal-Difference (TD) learning methods.

Convergence

If managed properly, the RL approach can converge to an optimal policy against any fixed, imperfect opponent.

Comparing Evolutionary Methods and Value Function Methods

- **Evolutionary Methods**: Evaluate a policy by playing many games but ignore the details of individual state-action experiences.
- **Value Function Methods**: Allow the evaluation of individual states, making the search more efficient by leveraging intermediate experiences during play.

Key Features and Flexibility of RL

- **Learning while Interacting**: RL emphasizes learning from interactions with the environment.
- **Delayed Rewards**: RL takes into account the delayed effects of actions.
- **Applicability**: RL is not just for games or episodic tasks but can also be applied to continuous tasks, large state spaces, and when part of the state is hidden.
- **Models**: RL can work both with and without a model of the environment.
- **Levels**: RL can operate at various levels within a hierarchical system, from low-level actions to high-level strategies.

Conclusion

The Tic-Tac-Toe example illustrates the versatility and key features of Reinforcement Learning. It shows how RL can be applied to a variety of problems and how it can adapt and learn optimal policies from interactions with the environment. It also highlights the differences between RL and other traditional methods, showcasing why RL is often more suitable for problems involving interaction and delayed rewards.