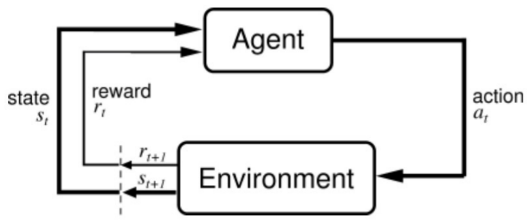
Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Summary of Lecture 2: Introduction to RL

Summarized By: Amirhossein Asadi

- **RL** is more important than ever. In AI, there are two main directions: one focuses on reaching **human-level intelligence**, and the other aims to surpass it, achieving **superhuman intelligence**. The first goal has already been reached in many areas, but true breakthroughs happen when AI exceeds human limits. **RL** is one of the main ways to achieve this. A great example is **AlphaGo**, which not only matched human performance but also went far beyond it. Now, the challenge is to keep pushing forward and unlock even greater possibilities.
- **Data-driven AI vs. RL:** Foundation models are heavily **data-driven**, focusing on cleaning and optimizing datasets. In contrast, **RL** aims for creative behaviors, requiring optimization during inference time to adapt and go beyond static data patterns.
- **RL in Recent Advancements: RL** has played a crucial role in recent AI breakthroughs, particularly in **RLHF** (Reinforcement Learning from Human Feedback), where a **reward model** helps refine LLMs. In simpler cases with a single state and action, this reduces to **bandits**. Moreover, in large scale reasoning-oriented RL, rewards from **rule-based verifiers** prevent reward hacking, leading to more reliable learning and significant progress in AI.
- **History:** Before 2013, **RL** had not fully flourished because representing policies relied on **ML** methods. Without the power of **DL**, these methods struggled to model complex functions, limiting **RL** to simpler problems where policies were straightforward, such as **PID** controllers.
- A **Markov Decision Processes** is a simple mathematical model for defining a task. It assumes an environment that provides a **state**, which the agent processes and maps to an **action**. This action is then executed in the environment, leading to a **new state** and a **reward**.



- **Goal:** A **MDP** is defined as a mathematical framework for modeling decision-making. Given the **MDP** components, our **goal** is to design a parameterized **policy** π_θ that maps states to actions:

$$\pi_\theta : S \to A$$

We aim to **optimize** this **policy** to maximize this expected cumulative **reward**:

$$\max_\pi \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) \mid \pi\right]$$

Horizon represents the number of time steps in the decision-making process, which can be finite or infinite.

By optimizing π_θ, we aim to learn a **policy** that maximizes long-term rewards.

- Sometimes, the **policy** is stochastic, in which case we have two sources of randomness and two nested expected value calculations: one for the environment's stochasticity and another for the policy itself.