Reinforcement Learning

Tom Wang

2025

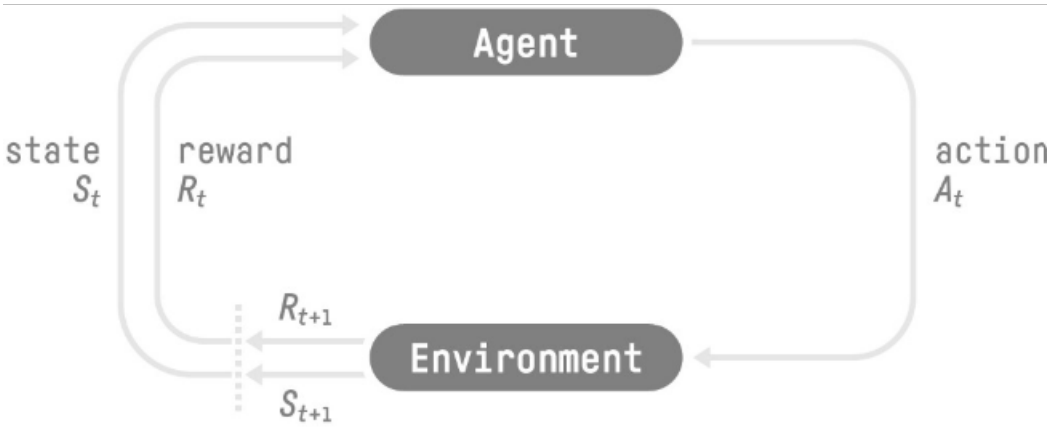
1 Introduction

The idea behind Reinforcement Learning is that an agent (an AI) will learn from the environment by interacting with it (through trial and error) and receiving rewards (negative or positive) as feedback for performing actions.

Formal definition:

Definition. *Reinforcement learning is a framework for solving control tasks (also called decision problems) by building agents that learn from the environment by interacting with it through trial and error and receiving rewards (positive or negative) as unique feedback.*

1.1 Framework



1. Our Agent receives state S_0 from the environment.
2. Based on that state, the agent selects an action A_0.
3. The environment transitions to a new state S_1 and provides a reward R_1.

This RL loop outputs a sequence of state-action-reward tuples: (S_0, A_0, R_1, S_1). The agent's goal is to maximize the total reward it receives over time.

1.1.1 Reward Hypothesis

Definition. *The reward hypothesis states that all goals can be described by the maximization of the expected cumulative reward.*

1.2 Markov Property

RL process might be called **Markov Decision Process (MDP)** if it satisfies the Markov property. The Markov property states that the future state depends only on the current state and action, not on the past states and actions.

Definition. *Observations/States are the **information our agent gets from the environment**.*

It will be good that we are always given a complete description of the state of the world. However, in practice, we might not have access to the complete state of the world. In such cases, we can use the **observation (partial description of the state)** to represent the state of the world.

1.2.1 Action Space

Definition. *The action space is the set of all possible actions that the agent can take.*

- Discrete action space: The agent can take a finite number of actions.
- Continuous action space: The agent can take an infinite number of actions.

1.2.2 Reward and the discounting

The reward is the only feedback to the agent to evaluate its actions.

The cumulative reward is the sum of all rewards the agent receives over time.

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t \tag{1}$$

where τ is the trajectory, r_t is the reward at time t, and γ is the discount factor.

Definition. *The discount factor γ is a value between 0 and 1 that determines the importance of future rewards.*

We have it since the reward in the future is less valuable than the reward now. Typically, $\gamma \in (0.95, 0.99)$.

1.3 Types of tasks

Definition. *A task is an instance of a Reinforcement Learning problem.*

- ***Episodic tasks**: Tasks that have a well-defined starting and ending point.*
- ***Continuing tasks**: Tasks that continue indefinitely.*

1.3.1 Episodic tasks

An episodic task creates an episode, which is a sequence of states, actions, and rewards that ends in a terminal state.

For instance, think about Super Mario Bros: an episode begins at the launch of a new Mario Level and ends when you're killed or you reached the end of the level.

1.3.2 Continuing tasks

Continuing tasks continue indefinitely. In this case, the agent must learn how to choose the best actions and simultaneously interact with the environment.

For example, think about a robot that needs to learn how to walk. The robot will keep walking until the battery runs out.

1.4 Exploration and Exploitation

Definition. *Exploration is the process of trying out new actions to discover the best possible actions.*

Definition. *Exploitation is the process of selecting the best-known actions to maximize the reward.*

The general trap is that if we exploit too much, we might miss out on better actions. On the other hand, if we explore too much, we might not get enough reward.