

# 1 Introduction

What is Reinforcement Learning? RL is the computational approach to learning from interactions. RL problems involve learning "what to do":

- by mapping states to actions
- while maximizing total reward

## Three central Characteristics:

- Closed-loop problems: actions of agents influence its later observations
- No direct instructions: the agent is not told which action to take.
- No initial knowledge: the agent has to figure out the consequences of its actions.

## Differences to Supervised- and Unsupervised learning

- Difference to Supervised learning: 1. Agent must learn from own experience. 2. No supervisor tell what to do
- Difference to unsupervised learning: 1. No predefined data. 2. Agent tries to maximize reward.

## Reinforcement Maximize reward.

- **Given:** an environment and a reward signal
- **Find:** a behavior that maximizes total reward over time

The agent takes action in the environment and gets back a state and a reward. Challenge is finding actions by exploring, exploit knowledge about good actions to maximize reward.

## Challenges in RL

- How maximize reward? We need to exploit our knowledge about the past.
- How find highly rewarded actions? We need to explore the environment.

## Dilemma: We have a trade-off between **Explore** and **Exploit**:

- to obtain reward an agent must favor actions that have proven to be beneficial in the past (Exploitation).
- to discover such actions, an agent has to try actions that it has not selected before (Exploration).

## 1.1 Examples

- a animal learn to walk.
- playing chess
- trash collection robots: has to decide search more trash or go to recharge.
- Computer games
- OpenAI Gym is a python library for reinforcement learning

### Commonalities

- interaction between agent and environment
- agents seek to achieve a goal in their environment, despite uncertainty
- actions affect the future state of the environment
- has to take into account indirect, delayed consequences of actions
- consequences of actions can't be fully predicted
- agents use experience
- interaction with environment is essential

## 1.2 Elements of Reinforcement Learning

- **Environment:** for example: the universe, one street, coffee machine, chess board ...
- **States:** represents the environments current condition: a position, temperature, etc.
- **Actions:** for each state there is a set of actions. E.g. turning steering wheel, move in a particular direction, ...
- **Policy:** A policy completely determines its behavior. It defines which action an agent can take in a given state. E.g. if car leaves street, steer in the other direction, if pressure too high, decrease current to heating element, move to a position with high expected future reward. Is a property of the agent. An agent has a specific policy at a specific time.
- **Reward signal:** Is given out from the environment and encodes how good the agent is doing currently. Given from environment, not the agents knowledge. It is the primary basis for altering the policy. E.g.: winning or losing the game, car stays on the street, does not crash.
- **Value functions:** is compressed knowledge about the future, it encodes an agent's experience. E.g.: eat cake, feel good. Turn steering wheel to avoid a crash.

**Policy** The behavior of an agent is called a policy. The policy:

- maps a state to a probability distribution over actions
- samples from this distribution to select an action

We can say an RL agent follows a policy (behaves a certain way) and updates its policy (to try and maximize reward). The policy completely determines its behavior, deciding which action to take at a given state. The policy is that, what the agent should learn???

**Policy implementation** Is a mapping from states to actions. Could be

- a lookup table
- a simple function
- a search process
- a DNN
- a combination

**Reward Signal** . Defines the goal in a RL problem. The reward:

- is a single real number
- is perceived by the agent at each time step
- defines what is good and what is bad
- is immediate and defines features of the problem
- primary basis for changing the policy

**Value Function** Defines what is good in the long run.

- is the total amount of reward an agent can expect to accumulate in the future starting from that state
- usually an estimate
- often used to choose an action

**Reward vs. Value**

- rewards are immediate, part of environment
- values are predictions of future reward, part of the agent
- rewards are used, in order to estimate value
- the only purpose of estimation values is to get more reward in the long run
- most RL algos focus on efficient value estimation
- a state might yield a low immediate reward but still have a high value because it is usually followed by beneficial states.

**Model environment** Model-based methods vs model-free methods.

**Boundaries** Boundary between Agent and Environment? **Fill summary of Example with the fish** this is temporal-difference-learning.

### 1.3 Example: Gridworld Environment

- **States:** we have 11 states: where is the fish?
- **Actions:** In all states we have the same actions. The fish can move. When he moves left in the left bottom the state doesn't change.
- **Loop:** the agent takes an action and receives a new state and a reward.
- **How get food and survive?** We set up a table with the states and the values for the value function.

**Lookup gridworld learning**

## 2 Multi-armed Bandits

Here we are in a so called *nonassociative* setting, this is we only have one situation or state.

### 2.1 k-armed bandits

Is inspired by the "slot machines" in casinos. The problem is the following: we have  $k$  different options or choices for an action and receive a reward from a probability distribution. We try to maximize the return over time. Each of the action we can take have an expected reward:

$$q_*(a) = \mathbb{E}[R_t | A_t = a]$$

If we would know this, we always would choose the action with the highest reward. But we don't know this, we only have the estimates  $Q_t(a)$ , which is the estimated value of action  $a$  at time step  $t$ . We want  $Q_t(a)$  close to  $q_*(a)$ .

At every time step, there is one action  $a$  with the highest value, this is the greedy action. If we choose this action, we are *exploiting* our knowledge about the values of the actions, but we don't gain new knowledge. If we choose a non-greedy action, we are exploring and gain knowledge about the values of action. There is always a trade-off between *explore* and *exploit*. The need to balance exploration and exploitation is a distinctive challenge that arises in reinforcement learning.

### 2.2 Action-value Methods

We are looking for methods to estimate the value of actions. The true value of an action is the mean reward when selected. We can estimate this with:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i 1_{A_i=a}}{\sum_{i=1}^{t-1} 1_{A_i=a}}$$

By the law of large numbers,  $Q_t(a)$  converges to  $q_*(a)$ . The greedy action selection method is

$$A_t = \arg \max_a Q_t(a)$$

This is called the *sample-average* method. This is maybe not the best method. Instead of this you could randomly select another method to explore new knowledge and therefore maybe get higher rewards in the future. This is called the  *$\epsilon$ -greedy* method: With probability  $(1 - \epsilon)$  the agent take the greedy action and with probability  $\epsilon$  it select an action randomly. This method also ensures then  $Q_t(a) \rightarrow q_*(a)$  for all  $a$  and the probability of selecting the optimal solution converges to greater than  $1 - \epsilon$ .

**Implementation**  $Q_{t+1}$  can be computed from  $Q_t$ :

$$\begin{aligned} Q_{t+1} &= \frac{1}{t} \sum_{i=1}^t R_i \\ &= \frac{1}{t} (R_t + (t-1) \frac{1}{(t-1)} \sum_{i=1}^{t-1} R_i) \\ &= Q_t + \frac{1}{t} (R_t - Q_t) \end{aligned}$$

This update rule occurs often in RL. The general form is

$$newEstimate = oldEstimate + stepSize[target - old]$$

### 2.3 Optimistic initial values

All methods we discussed so far, depends on the initial values to some extend. If we have prior knowledge we can use that to get better results. If we don't have prior knowledge but set high (optimistic) values (higher than the highest possible reward), then the algorithm starts exploring. Because when an action is taken, the reward is lower than the current value, so the value of the action is reduced and has lower values then the other actions.

### 2.4 Upper-Confident bound action selection

$\epsilon$ -greedy selection selects a non-greedy action but randomly with no preference. It would be better to select among the non-greedy actions according to their potential of being optimal.

$$A_t = \arg \max_a \left[ Q_t(a) + \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

The square root term represents something like a variance or uncertainty in the estimates of  $a$ 's values. If  $a$  is chosen often, this term gets reduced. The  $\ln$  ensures, that the increases get smaller over time, if  $a$  is not selected, but is still unbounded.