Reinforcement learning (More)

# **Youtube (deepLizard)**

Reinforcement learning is an area of machine learning that focuses how an agent might act in an environment to maximise some given reward. Reinforcement learning aims on how a given agent might take actions in an environment to maximise reward.

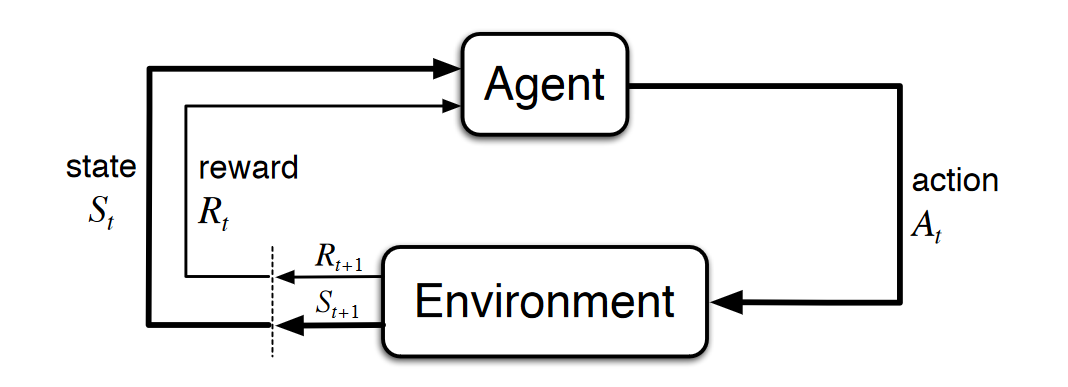
***Markov Decision Processes (MDPs)***

In an MDP we have a decision maker called the agent. The agent interacts with the environment its placed in. the interactions occur sequential over time. At each time step it will get some representation of the environment state. Given the representation the agents selects an action to take. The environment is then transitioned into some new state and the agent is given a reward as a consequence of its previous action.

*Components of an MDP*

* Environment
* Agent
* Possible states
* Possible actions
* Rewards

This process of selecting an action from a given state, transitioning to a new state and receiving a reward, happens sequentially, over and over again which creates a trajectory that shows the sequences of states, actions and rewards. Throughout the process, it’s the agents goal, to maximise the total amount of rewards that it receives from taking actions in given states of the environment. This means that the agent wants to maximise not just the immediate reward but the cumulative reward that it will receive over time. We can think of the process of receiving a reward as an arbitrary function that maps state-action pairs to rewards. At each time t, we have f**(St,At)=Rt+1.**

The trajectory representing the sequential process of selecting an action from a state, transitioning to a new state, and receiving a reward, can be represented as follows:

S0,A0,R1,S1,A1,R2,S2,A2,R3,⋯

This can be illustrated diagrammatically as follows

Since the sets S and R are finite, the [random variables](https://en.wikipedia.org/wiki/Random_variable) Rt and St have well defined probability distributions. In other words, all the possible values that can be assigned to Rt and St have some associated probability. These distributions depend on the *preceding* state and action that occurred in the previous time step t−1.

For example, suppose s′∈S and r∈R. Then there is *some* probability that St=s′ and Rt=r. This probability is determined by the particular values of the *preceding* state s∈S and action a∈A(s). Note that A(s) is the set of actions that can be taken from state s.