Reading-Writing Assignment: Chapter 3

Brian (Boyuan) Lu, blu1

Finite Markvo decision processes, known as MDPs, is a particular way to implement the Agent-environment interfaces. The MDPs contains Agent which is the learner and decision maker, Environment that interact with agents and comprises everything out of agent. Normally, agents and environments would interact with each other continuously, agent will estimate and pass the most possible actions for the next time step to the environment, environments will evaluate the picked action and response back with typical reward and environment state to the agent. And then the agent will estimate a new decision based on the reward and state sent from environment, and provide the most possible action to nest step. MDP framework is an efficient method for goal-directed learning, which means the environment has specific goals and is able to provide reasonable rewards and states to the agent based on the goals it holds.

Generally, the goal of the agents in reinforcement learning is to maximize the reward it can receive from the environment. Thus, one particular feature the Reinforcement learning has is to the reward signal to formalize the idea of goal. Episode is the concept we use to break the agent-environment interaction into subsequences. In this way, each start of an episode is a fresh start that is not influenced by previous episodes. Also, the end of every episode would have different results or states.

Another distinctive feature of reinforcement learning is value functions. Value functions is used in agent algorithm to estimate if an action is good or not. It follows policies which are defined by particular ways of acting. In this way, the agent was able to evaluate through the passing reward and environment state, with the help of value functions, provide reasonable prediction of future actions. Also, each policy or action has a probability to be picked, and the learning algorithm specify how the agent's policy varies base on its previous experience. We use $q\pi$ to represent action-value function for policy π, use $v\pi(s)$ to represent values of policy π under state 's'.

The way to choose the optimal policy is basically finding the best policy that can bring the best amount of reward over a long run. It is similar to the concept mentioned last chapter, optimal action-value function, denoted q*, the only difference is that MDPs here has environment states and reward to be counted rather than just reward. However, the limitation on the optimality of the MDPs is obvious, such as memory available may not enough to build up the approximation on value functions, policies and models. And the problem can be approached in very different ways depend on the assumptions and knowledges available to the agents. Therefore, in reinforcement learning, the optimal solution may not be found but it can be approximated in some ways.