Haoming Liang
1430396

Reinforcement Learning methods can be divided into two categories: model-based (example: Dynamic Programming, Heuristic Search) and model-free (example: Monte Carlo, Temporal-Difference). Model-based methods are mainly about planning, and model-free methods are mainly about learning. Now newly introduce Tabular Methods to perform planning and learning, which is developed by combining previously introduced algorithms, and study the similarities between them.

Model is defined by how the agent predict the environment in result of their actions, including the next state and the possible rewards. Distribution model refers to models with information of all possible results, and their probability of each result. Sample model refers to models of one single result sampled from all possibilities. Generally, distribution models are better than sample models since distribution models provide more information, however sample models are easier to obtain.

Planning, is the process of computation with model as the input. The usage of model here is to simulate an episode and produce experiences. From the simulated experiences, we can update the value functions and the policy. Learning, is to take real experiences from the environment as the input only. But many of the times, simulated experiences are nearly consistent to real experiences, and thus simulated experiences could also be applied to learning.

During the process of planning, the interacting experiences could change the model and thus interfere with planning. To deal with such issues, a new algorithm Dyna-Q is used. A general structure of Dyna-Q is as follow: interactions between agent and environment through real experiences, direct reinforcement learning from real experiences to update policy/value functions, model learning from real experiences, simulate experiences from a model and do planning update on policy/value functions.

The algorithm of Dyna-Q is evolved from Q-Learning, with addition of two new steps after the update of action value. Firstly, add the current reward and the next state into the model of a state-action pair. Secondly, repeat the following process for number of planning steps: randomly select a previous state and a random action taken at that state for a state-action pair, observe the reward and the next state from that state-action pair, then do a Q update with these values.

As observed in the graph, the learning curve of 0 planning step (direct reinforcement learning) is steep, and the learning process is slower than using planning step. With 5 planning steps, such issues are mostly solved and the learning is significantly improved. With 50 planning steps, the agent already gain knowledge of the optimal policy in about 3 episodes. From the above observations, we can already see the significance of using models to do planning updates.