DM887 Reinforcement Learning Jiawei Zhao

Assignment 1: Q-Learning for GridWorld

1. I designed two grids.
   The grid of game A uses default immediate costs (i.e. negative rewards) for each type of state and the default conditions of termination (an agent). The agent is supposed to learn the diagonal path from the top-left corner to the goal at the bottom of the grid.
   The grid of game B is not configured with the default immediate costs and conditions of termination. That is, the cost (negative immediate reward) of falling into a hole was set at 5 instead of 100, to lure the agent to enter a hole. In addition, a hole is set as a non-terminal state as opposed to the default configuration of GridWorld. Such novel holes were distributed along the shortest diagonal paths and formed soft mazes. Therefore, the agent has to learn to circumvent the holes and optimally reach the goal along the less immediately rewardable versatile paths consisting of normal states.

2. Training was performed using a learning rate *α*=0.2, *γ*=0.9, and a maximum episode number of 500. The action at each time step per episode was selected by an ε-greedy algorithm given the policy table of the Q-learner class within the first 1,000 time steps of an episode. If the time step at an episode exceeds 1,000, the agent would be considered as being trapped into a circular path consisting of non-terminal normal states that would never terminate the episode. Therefore, the action at each time step above 1000 was chosen with a completely equal probabibility of each possible action given a state to maximize stochasticity to inspire the agent to leave a circular path.

3. As per the plot "*learning-curve-game_a.png*", we may observe that, with regards to game A, a higher *ε* worsens the reduction of mean cumulative cost across episodes, while a larger *N* improves the reduction of mean cumulative cost. Per my personal assumption, an ε higher than 0.1 adds excessive stochasticity in decision of the action per time step, thus increasing the probability of termination of an episode at a hole by an action deviated from the Q-table-based policy table. Additionally, a higher *N* enables the agent to update the Q-table based on the immediate costs and minimum cumulative costs at a larger sub-grid, resulting in more sophisticated exploitation.
   As per the plot "*learning-curve-game_b.png*", we may observe that, with regards to game B, a higher *N* improves the reduction of mean cumulative cost slightly due to the aforementioned cause. However, a higher *ε* improves the reduction of mean cumulative cost across episodes at game B, since only one terminal state exists at the grid of game B. Therefore, higher stochasticity by a higher ε enables more complete exploration by the agent, thus improving the update of Q-tables.