Project 4

Qiyu Chen

April 15, 2023

Abstract

This project explores regularity discovery using PRML techniques. Specifically, I choose the Reinforcement Learning-Based Regularity Detection option, which explores Q-learning and Deep Q-learning in maze solving.

Contents

1 Investigation of the Baseline Method

In this section, q-learning and deep q-learning that are used in maze-solving problems are discussed. Their performance on sample maze configurations are also evaluated.

1.1 Q-learning

To apply Q-learning in reinforcement learning, several components are needed: Agent, Environment, Actions, and Awards. In our maze-solving setting, the environment is a simple 2D maze. The agent is a blue dot which needs to find its way from the top left corner (blue square) to the bottom right corner (red square). The actions space is {up, down, right, left}, and it the way is blocked, the agent will remain at the same location. A reward of 1 is given when the agent reaches the goal. For every step in the maze, the agent receives a reward of -0.1/(number of cells). Once the agent reaches the goal, the maze will be reset. There is a Q-table at each state where each cell corresponds to a state-action pair value, which is the overall expected reward achieved by the agent if adopts this action. Initially, the Q-table is useless. But after many explorations from the agent, the Q-table will be updated. Eventually, the agent can utilize Q-table for decision making. Q-learning is such a model-free reinforcement learning algorithm that helps agent learn the Q-table at each possible state.

1.2 Deep Q-learning

Deep Q-learning is a deep neural network approximation of traditional Q-learning approach. In deep q neural network, the state of the agent is the input, and outputs of DQN are Q-values corresponding to different actions taken at this state. The environment and actions in Deep Q-learning experiment is same to the setting in Q-learning, except that we simply places a wall on grid position instead of removing the wall in certain direction (e.g., north, east, south, west). In addition, the agent received a reward of -1 on each non-goal grid, and a reward of 0 on the goal grid. In exploration mode, the agent takes random action at that state. In exploit mode, the agent adopts the action with the largest q-value output by DQN.

1.3 Performance evaluation

Q-learning
One 3x3, 5x5, and 10x10 sample maze configurations are experimented in this section (figure 1-3). Evaluation are based on their rewards obtained per episode. The policy is considered "learnt" if the agent can achieve 10 streaks.

From figures 4-6, we can see that all of them can achieve relatively great performance in terms of reward received after learning the policy. However, we can see that the episodes required to learn the policy increases non-linearly when the size of maze get increased, meaning the difficulty of solving maze increases non-linearly as its size increases. This is validated by the fact that 5x5 maze only uses few more episodes to learn the policy compared to the 3x3 setting, but 10x10 maze needs 4x more episodes than 5x5 maze did.

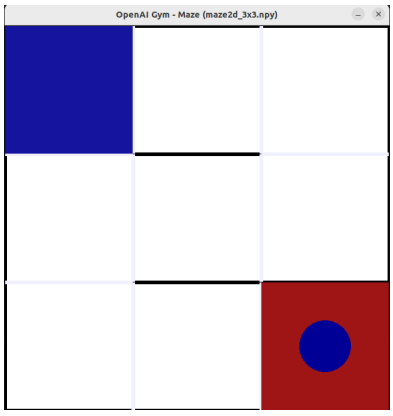

Figure 1: Sample 3x3 maze configuration

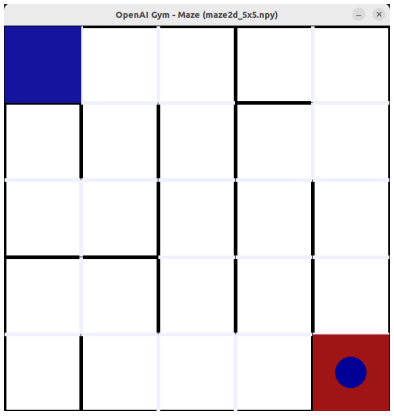

Figure 2: Sample 5x5 maze configuration

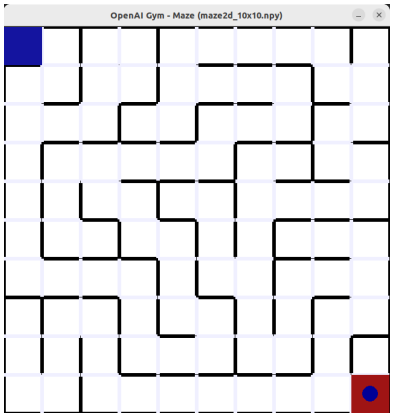

Figure 3: Sample 10x10 maze configuration

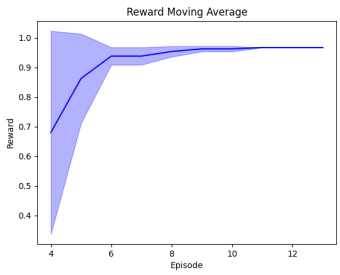

Figure 4: Reward obtained per episode on the sample 3x3 maze

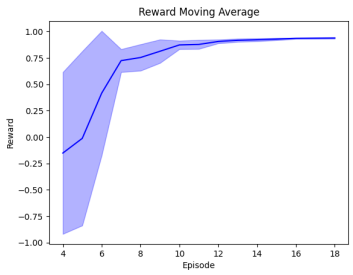

Figure 5: Reward obtained per episode on the sample 5x5 maze

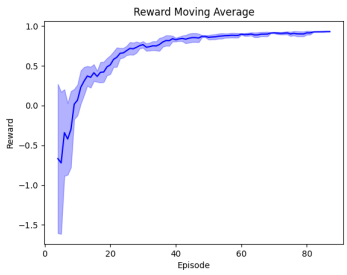

Figure 6: Reward obtained per episode on the sample 10x10 maze

Deep Q-learning

Similarly, one 3x3, 5x5, and 10x10 sample maze configurations are experimented in this section (figure 7-9). Evaluation are based on their rewards obtained per episode. But the policy is considered "learnt" if the agent can achieve 5 streaks.

Figures 10-12 are rewards obtained per episode in Deep Q-learning. We can see that the Deep Q-learning still learns the policy in both 3x3 and 5x5 setting using similar episodes. However, it fails on learning the 10x10 maze configuration, where the learning stops simply because it reaches maximum of 5000 episodes. Also, the fluctuation of rewards on this "hard" maze is apparently larger in mazes which are learnable. One possible explanation for this is that in the sample 10x10 maze configuration, there are much more loops which misdirect the agent from reaching the goal, making learning procedure harder.

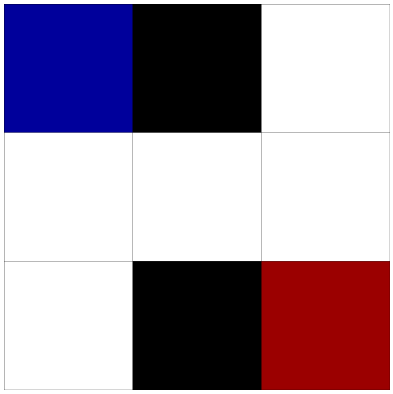

Figure 7: Sample 3x3 maze configuration

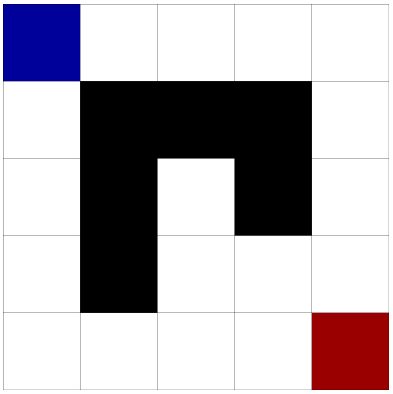

Figure 8: Sample 5x5 maze configuration

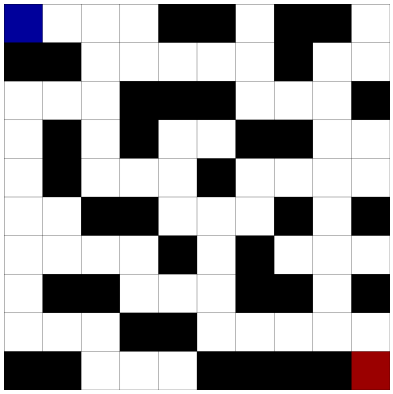

Figure 9: Sample 10x10 maze configuration

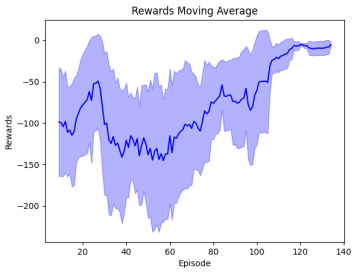

Figure 10: Reward obtained per episode on the sample 3x3 maze

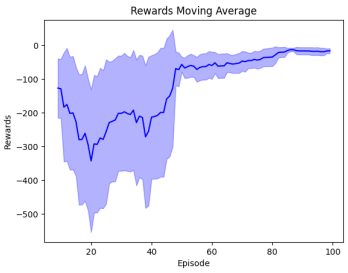

Figure 11: Reward obtained per episode on the sample 5x5 maze

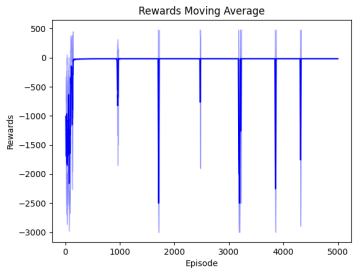

Figure 12: Reward obtained per episode on the sample 10x10 maze

2 Exploring Regularity in RL-based Learning

2.1 New method in maze solving

Except Q-learning and Deep Q-learning, I also investigate another method that is applicable in maze solving: Double Deep Q-Network (DDQN)[1].

In high-level idea, Double Deep Q-Network is a variant of Deep Q-learning. In standard DQN, the Q-values are estimated using single network. One drawback of this approach is that it can sometimes result in overestimation of the Q-values. To address this issue, DDQN adopts two separate networks: one network is used to select the action to take, while the other is used to evaluate the Q-value of that action. That is, the overestimation is reduced by decoupling the action selection and the Q-value evaluation.

2.2 Performance comparison on different maze configurations

Q-learning
One new 3x3, 5x5, and 10x10 mazes are created for the q-learning experiment in comparison with samples ones.

Figure 13 is the new 3x3 maze created for comparison. As we can see, the maze is more straightforward compared with the sample 3x3 maze configuration. As expected, figure 14 shows that the agent uses less episodes to learn the policy.

New 5x5 maze configuration in figure 15 adds one slightly longer loop in first two columns, which might be the fact that it takes few more episodes to learn the policy compared to the sample one, as observed in figure 16.

One interesting observation is noticed in the new 10x10 maze configuration described in figure 17. As we can see, though the new maze does not add more loops into the maze, but it constructs long "dead ends" which drags the reward of the agent obtained in early learning episodes. Comparing figure 6 and 18, we see that the agent uses much less episodes to learn the policy, but the reward obtained in early episodes fluctuate greatly, and this might be explained by the introduction of long "dead end".

Deep Q-learning
For the Deep Q-learning, only one new 3x3 and 5x5 mazes are created for comparison, as it failed on learning the sample 10x10 mazes.

The new 3x3 maze configuration in figure 19 only leaves one possible path from start position to the goal position. Comparing figure 10 and 20, though both of them use similar number of episodes to learn the policy, but the agent has greater average rewards obtained in the new maze configuration.

In contrast, the new 5x5 maze configuration in figure 21 splits the maze into two parts,

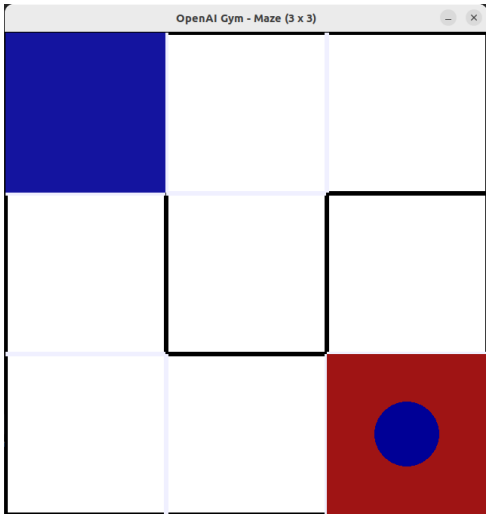

Figure 13: New 3x3 maze configuration

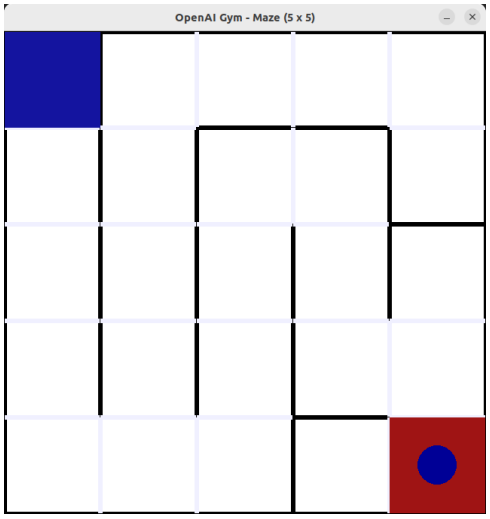

Figure 15: New 5x5 maze configuration

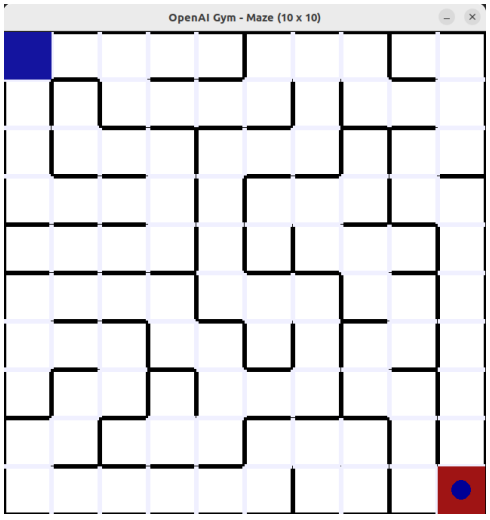

Figure 17: New 10x10 maze configuration

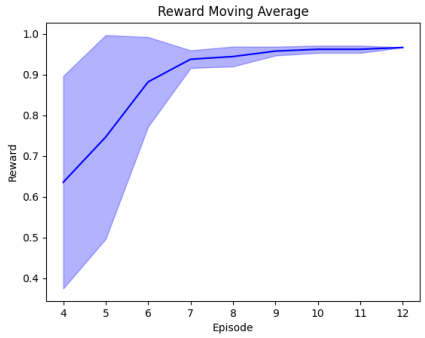

Figure 14: Reward obtained per episode on new 3x3 maze configuration

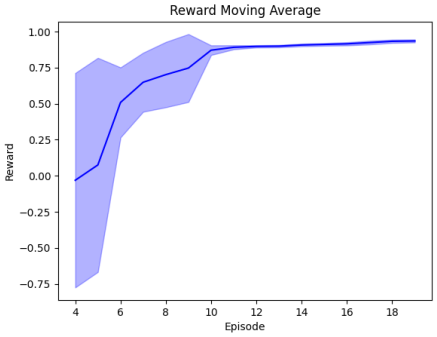

Figure 16: Reward obtained per episode on new 5x5 maze configuration

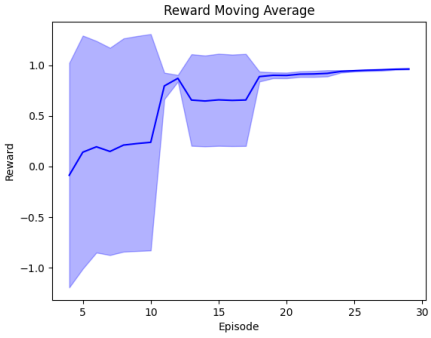

Figure 18: Reward obtained per episode on new 10x10 maze configuration

in which bottom left part leads to large loop area and the top right part leads to goal. Comparing to the sample 5x5 maze, this new maze is much complex, leading to more episodes for learning the policy, as observed in figure 22.

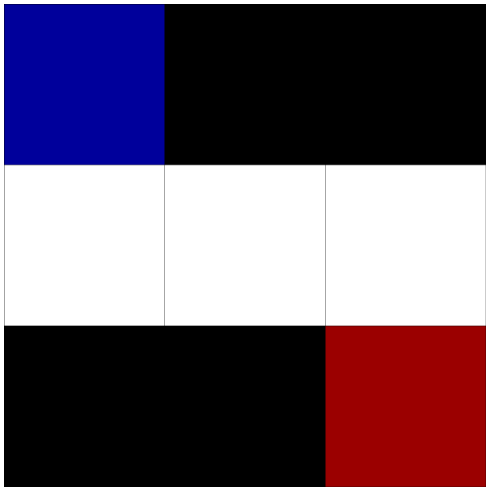

Figure 19: New 3x3 maze configuration (DQL)

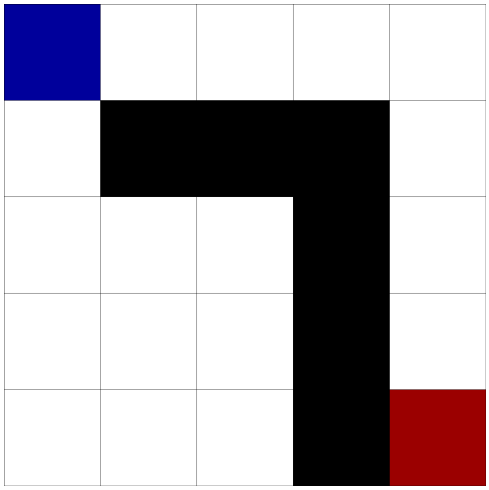

Figure 21: New 5x5 maze configuration (DQL)

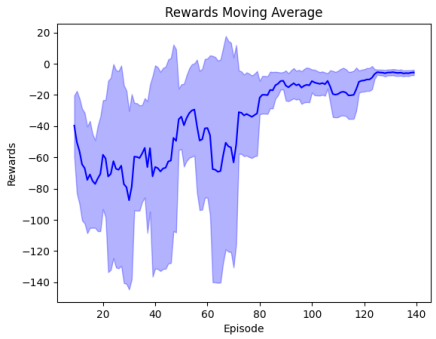

Figure 20: Reward obtained per episode on new 3x3 maze configuration (DQL)

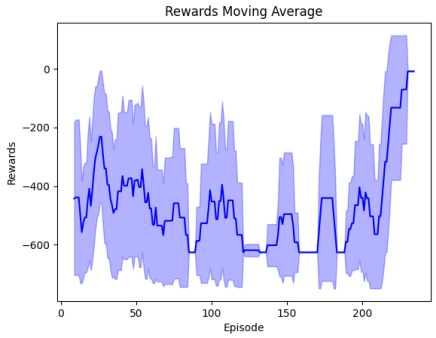

Figure 22: Reward obtained per episode on new 5x5 maze configuration (DQL)

2.3 Performance on mazes with regularity

Q-learning

For Q-learning, I construct two new 5x5 mazes with regularity for experiment, as noted in figure 23 and 25.

Before experiment, I expect that as both regular mazes consist of more loops than the sample configuration, they will behave poorly in comparison. However, evaluation on figure 24 and 26 both indicate that this kind of construction leads to better performance in terms of number of episodes required to learn the policy. I guess that this is caused by the many existed "dead ends" in sample configuration, which wastes agent lots of time on reaching the goal.

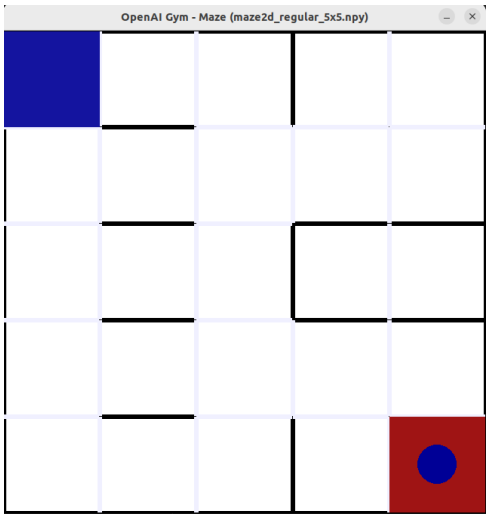

Figure 23: Regular 3x3 maze configuration

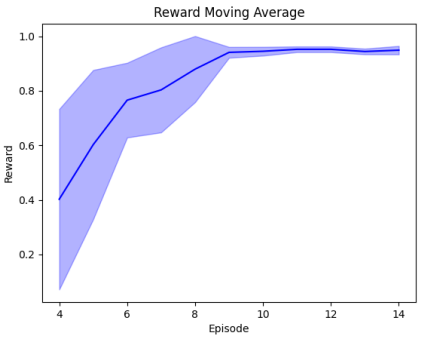

Figure 24: Reward obtained per episode on regular 3x3 maze configuration

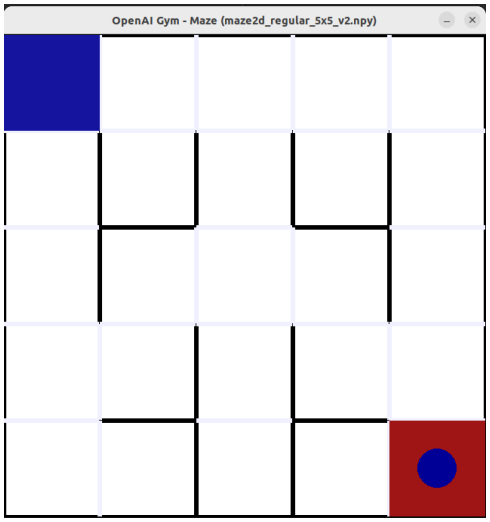

Figure 25: Regular 5x5 maze configuration

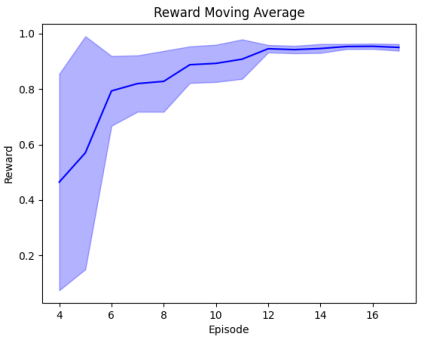

Figure 26: Reward obtained per episode on regular 5x5 maze configuration

Deep Q-learning

Similarly, two new 5x5 mazes with regularity are created for Deep Q-learning, as indicated in figure 27 and 29.

Comparing figure 11, 28, and 29, we notice that first regular maze is much "difficult" than the sample setting while the second one is comparable with it, in terms of the number of episodes used to learn the policy. Due to the existence of long "dead ends" in first regular maze, I think this further suggests prior assumption that the existence of long "dead ends" prevents agent from quickly learning the policy.

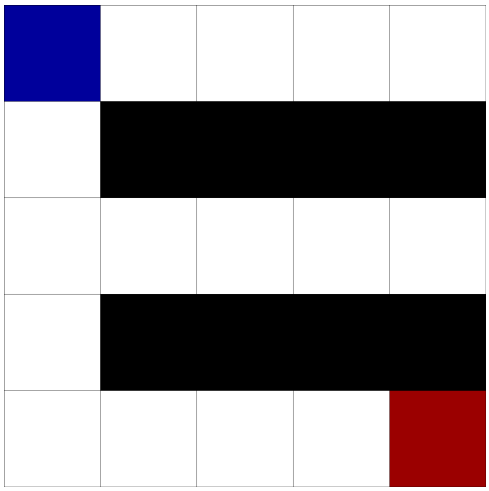

Figure 27: Regular 3x3 maze configuration

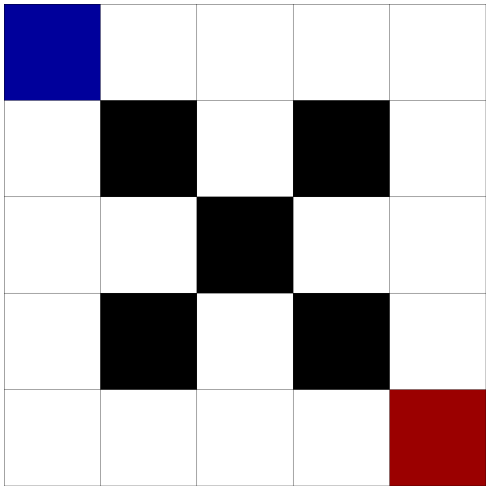

Figure 29: Regular 5x5 maze configuration

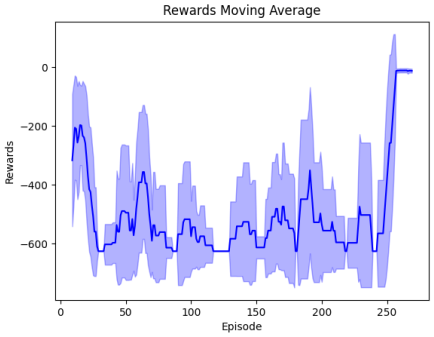

Figure 28: Reward obtained per episode on regular 3x3 maze configuration

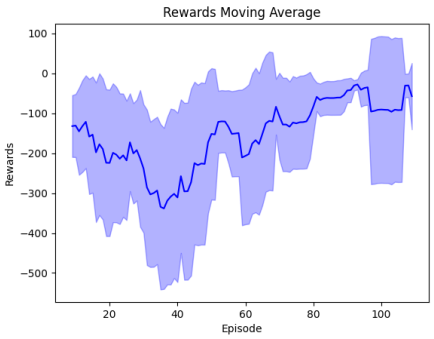

Figure 30: Reward obtained per episode on regular 5x5 maze configuration

2.4 Comparison between Q-learning and Deep Q-learning

Inspecting above evaluation results, I see that Q-learning allows agent to learn the policy quicker than Deep Q-learning does, in terms of number of episodes required to achieve enough strikes.

What's more, it is difficult for Deep Q-learning to learn the policy in certain maze configuration, which is supported by the fact that Deep Q-learning fails on handling the 10x10 sample maze configuration. But this conclusion is also dependent on the architecture of the neural network used. Changing the underlying architecture might resolve this problem.

Lastly, the fluctuation of reward obtained per episode is really high in Deep Q-learning, in contrast to the Q-learning approach. In most cases, rewards obtained per episode in Q-learning increases in gradual fashion, while the fluctuation of reward in Deep approach is extremely significant.

References

[1] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," *CoRR*, vol. abs/1509.06461, 2015. [Online]. Available: http://arxiv.org/abs/1509.06461 5