**Assignment 2**

1. (1)

The maximum sum of rewards a single episode can achieve is 4.1, with the trajectory: , , , , , , , , , , , , , , , .

2. (1)

The advantage of experience replay is to remove correlation problem in consecutive frames in deep Q learning; we can also keep the history of some rare experience in which the agent can learn more.

(2)

The benefit of using target network is to stabilization during training period, since if the parameter keeps updating for each iteration, the target is jiggling which leads to learning bad parameters.

(3)

To help vectorization while computation, so it could reduce complexity.

3. (1)

The equation 2 can be rewritten as:

where and is the corresponding state-action pair element of vector . Since is a one-hot vector, where only the corresponding state-action pair element is 1 (to be updated), and 0 otherwise, we only update parameter in the position of state-action pair. Hence, the equation 2 is exactly the same when is represented to .

(2)

(4)

图片包含 游戏机

描述已自动生成

4. (2)

图片包含 游戏机, 男人

描述已自动生成

The performance is the same as linear approximation, and it has a longer training time.

5. (1)

We want our agent to learn a good strategy in a continuous space. Otherwise, if only a single frame is input into deep Q network, the agent does not know which direction and velocity of the pong. Another benefit is we do not need to output an action for every frame in the game, so it reduces the frequency of decisions.

(2)

The number of parameters of deep Q network:

The number of parameters of linear approximation:

6. (1)

The term is not computable, because there is no actions available in approximated value function . In other words, we only specify features for all states rather than all state-action pairs, so we cannot compare which action gains higher value.

(2)

We can either sample an action based on state then use the approximated Q value to determine the value of , which is SARSA, or take an action which maximizes Q value: , i.e. Q learning method.

(3)

No, since the target network and evaluation network has the same parameter , each time the parameter is updated, the target may oscillate from time to time, which results in unstable training. One of solution is use another parameter to specify target network and update for a fixed period of time.