CSE 574 PROJECT 4(Bonus)

Vishnu Varshath Harishankar
Person Number 50291399
ubit vharisha

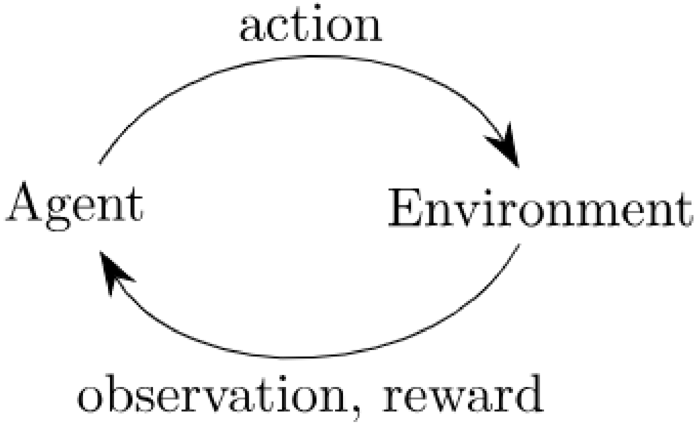
1 Gym Open AI

Gym Provides a toolkit for developing and comparing reinforcement learning algorithms. The most significant power of gym is to adapt to different kinds of agents (ie) it does not depend what type of agent is being used. The Gym library provides a set of environments which is a test problem upon which we can use Reinforcement Learning Algorithms such as DQN. The environments offered by gym expose themselves to the same interface making it easy to implement our algorithms.

2 Terminologies

2.1 Environment

The Environment is important because the agent interacts with the environment by first observing the state of the environment. Based on this observation the agent changes the environment by performing an action. Open Gym provides a lot of environments to work with



2.2 Agent

A reinforcement learning agent interacts with its environment in discrete time steps.At each time t, the agent receives an observation o_t, which typically includes the reward r_t. It then chooses an action a_t from the set of available actions, which is subsequently sent to the environment. The environment moves to a new state s_{t+1} and the reward r_{t+1} associated with the transition (s_t, a_t, s_{t+1}) is determined.

2.3 Observations

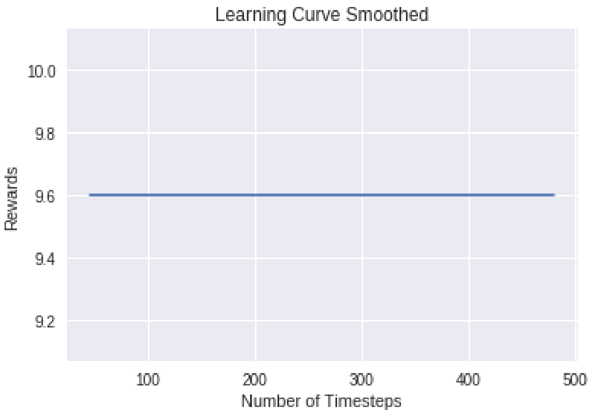
Observations are those provided by the environment to the agent. Based on the observations, the agent can decide to perform a certain action. Observation is something that the Environment provides to the agent as the result of some previous action taken by the agent.

3 Bonus Task-1

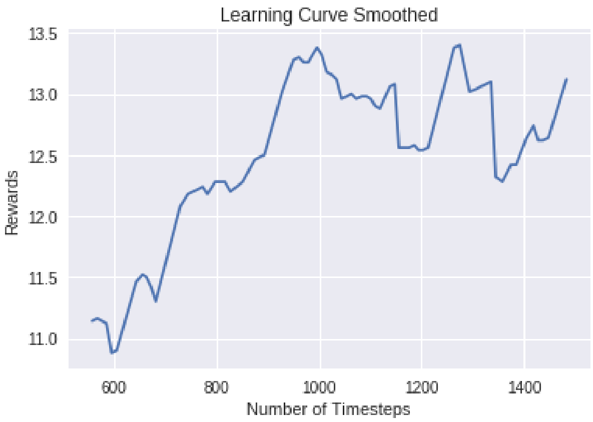
3.1 Implementation

The Environment chosen is CartPole-v0. The task is to balance the pole for as much time as possible by moving left,right or staying at the same place. By interacting with the environment it learns which direction to move when the pole starts to slant. We use the DQN algorithm which one of the many popular Algorithms to train the agent.

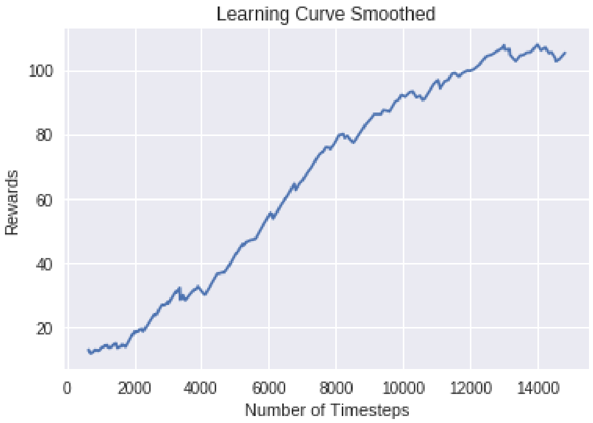
The number of time steps were varied from 500 to 1500000 and the results were plotted.



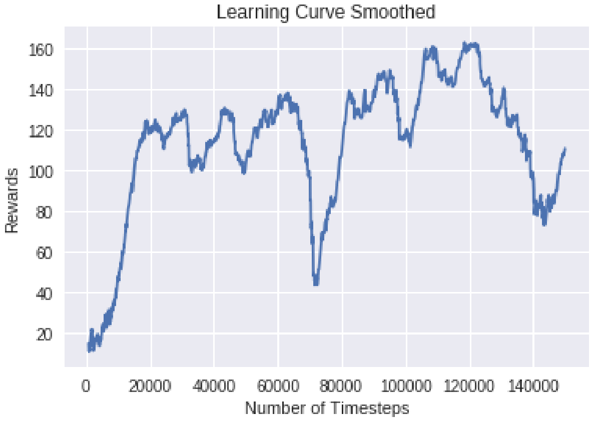
The above graph shows that the agent has not learned anything significant while interacting the environment within the number of time steps (ie) 500.



The above graph shows that the agent has started to learn something while interacting with the environment but can still learn more. There are a lot of fluctuations in the graph. The timesteps chosen here was 1500.



The above graph shows that the agent has started to consistently learn something from interacting with the environment and has the potential to learn more. The timesteps chosen for training is 15000.

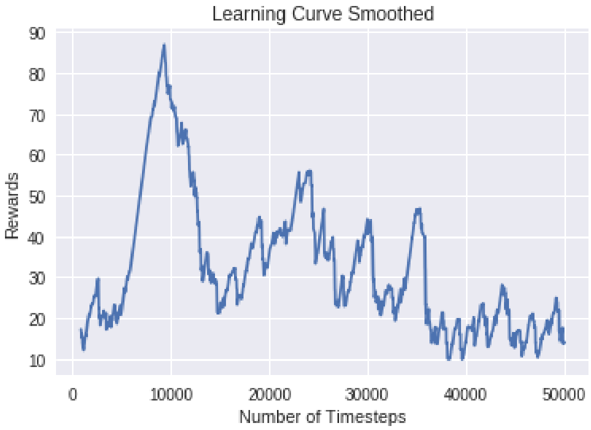


The above graph shows that around timesteps=120000 the model got the maximum rewards which is about 160 while interacting with the environment.

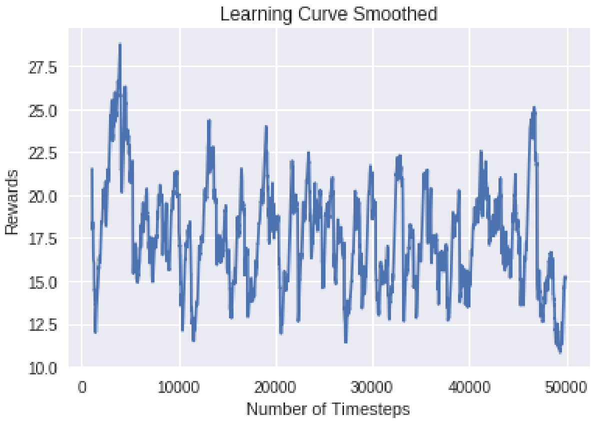
3.2 Tweaking Learning Rate

Learning rate tells the magnitude of step that is taken towards the solution.

It should not be too big a number as it may continuously oscillate around the minima and it should not be too small of a number else it will take a lot of time and iterations to reach the minima.



The above graph depicts a learning rate of 0.005. Notice the inconsistencies of the agent trying learn from observations. Since the learning rate is high it was not able to find the minima



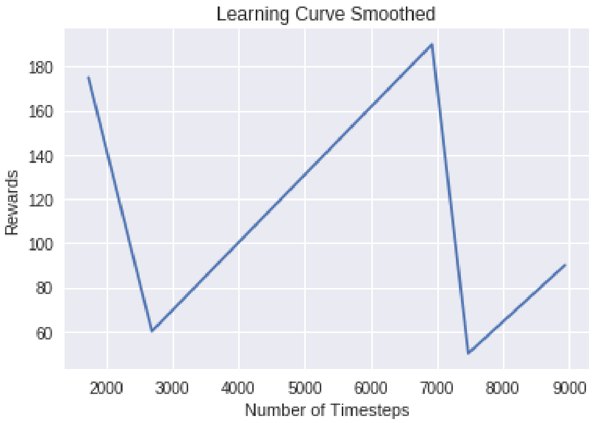
The above graph depicts a learning rate of 0.05. There are very great oscillations and the agent is not able to get maximum rewards(in comparison to the default learning rate and learning rate=0.005)

The agent was able to learn to balance the pole and the results are downloaded in the form of a video from the code.

4 Bonus Task-2

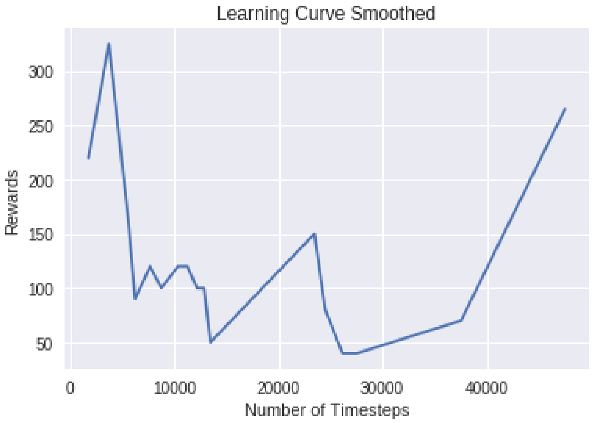
The Atari game that was chosen Demon Attack. The goal of the game is to maximize the points by killing all the demons while surviving the attack from demons.

The model was run with time steps of 9000.



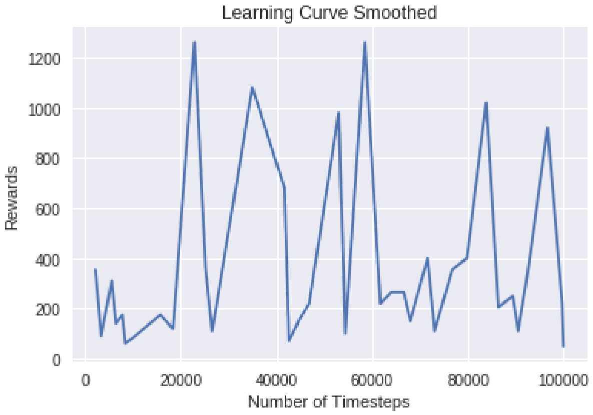
The above graph shows the sharp increase followed by decrease of the rewards obtained by the agent. It was not able to learn in the given timesteps.

The model was run with time steps of 40000.



The above graph shows variations of the rewards obtained by the agent on interacting with the environment. The maximum reward obtained was near 300. Notice in the end that the rewards started to increase. Therefore it requires more time steps to learn properly.

The model was run with time steps of 100000.



The above graph shows variations of the rewards obtained by the agent on interacting with the environment. The maximum reward obtained was near 1200.