Reinforcement Learning
Project 2

T-504-ITML, Machine Learning, 2022-1

Reykjavik University - School of Computer Science, Menntavegi 1, IS-101 Reykjavík, Iceland

Ágúst Þór Þrastarsson
agustt20@ru.is
Ragnar Smári Ómarsson
ragnaro20@ru.is

3. November 2022

1 The environment

Since all values for the attributes are integers in some interval, for example the velocity of the bird is an interval between -8 and 10 the environment itself is finite. In the environment we have a state, the state represents the state of the bird within the game each frame.
In each state we have total 4 values:

- player y: The current y position of the bird.
- next pipe top y: The top y position of the next gap
- next pipe dist to player: The horizontal distance between bird and next pipe
- player vel: The current vertical velocity

The min and max values for each attribute in a state:

- player y: min=0, max=387
- next pipe top y: min=25, max=192
- next pipe dist to player: min=0, max=283
- player vel: min=-8, max=10

The total number of states we can have from these values is:
388 * 168 * 284 * 19 = **351 732 864 number of states.**

This will heavily affect our training time so we can decrease the size by splitting each attribute (except the velocity) evenly into 15 intervals each. After these changes the total number of states we can have from these values is: 15 * 15 * 15 * 15 * 19 = **64 125 number of states.**

There was one problem with this approach, the initial value for next pipe dist to player was bigger than window width(283 px). So we had to create one if statement, that if the value of the next pipe dist to player was bigger than the max(283 px) it would be set to index 14.

1.1 Analyzing

Since we have talked about how the environment is structured and what values there are in each state but not about the environment itself. We can look at different things in the environment and tell if they are stochastic, deterministic, episodic etc. When looking at flappy bird, the only random variable in the game is the gap between the pipes, in which the bird has to go through, so the only stochastic variable in the environment is the gap since none of the other variables are stochastic or random, but the gap itself is always the same height just at different y position each time it spawns. All other variables or deterministic/constant, since they can be predicted.When the bird flaps, it always adds the same height to the y-position it is at that time, when the bird does not flap it always goes down for the same velocity and the speed of the bird is always the same.

The models were broken down into episodes, a single episode starts when the game start and ends when the bird dies, hits ground, top or pipes. So we can say that the environment is episodic but a single episode can be running for infinite amount of time.

1.2 Markov Decision

A Markov Decision process is defined by: State, Set of Actions(A), Probability, Reward, gamma. In flappy bird we have all of those where the set of actions are either to flap or to not flap (0, 1). We have the values for each state happening each frame so the process can be classified as a Markov Decision Process.

1.3 Which algorithms to use?

Monte Carlo, because all states are known, its not very sensitive to initial values, The distance between the bird and the first pipe is large and when the bird has gone through the first pipe the distance between the bird and the first pipe is a lot smaller than the initial distance between the bird and the first pipe. Itś also very simple.
The challenges are that Monte Carlo usually has high variance, which could be bad in this case. Monte Carlo must wait until end of episode before it starts to calculate the values. Could be slow.

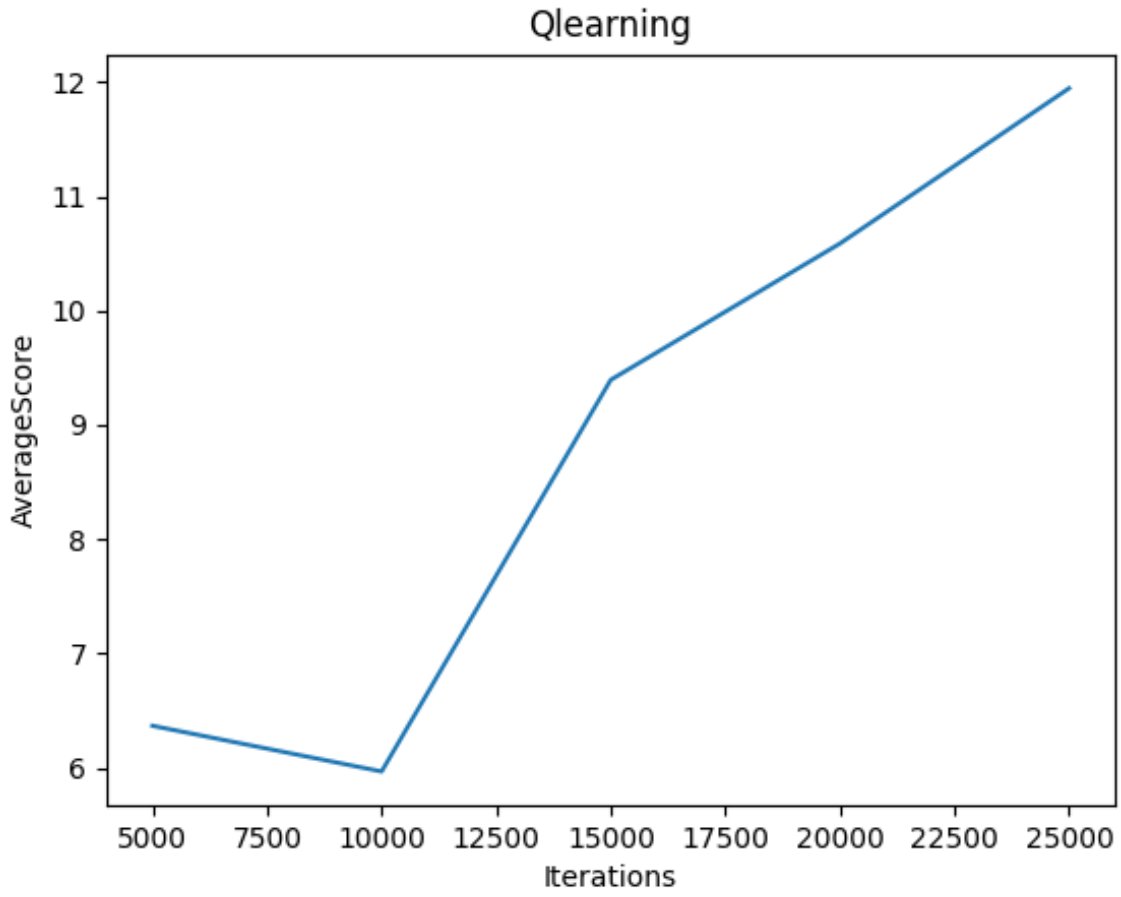
Sarsa and Q-learning can be used since it´s a Temporal difference algorithm like Q-learning and they are often quicker than for example Monte Carlo.
The challenges could be to tune and find the correct parameters so it will perform well.

2 Train and Test

2.1 Q Learning

Variables used:

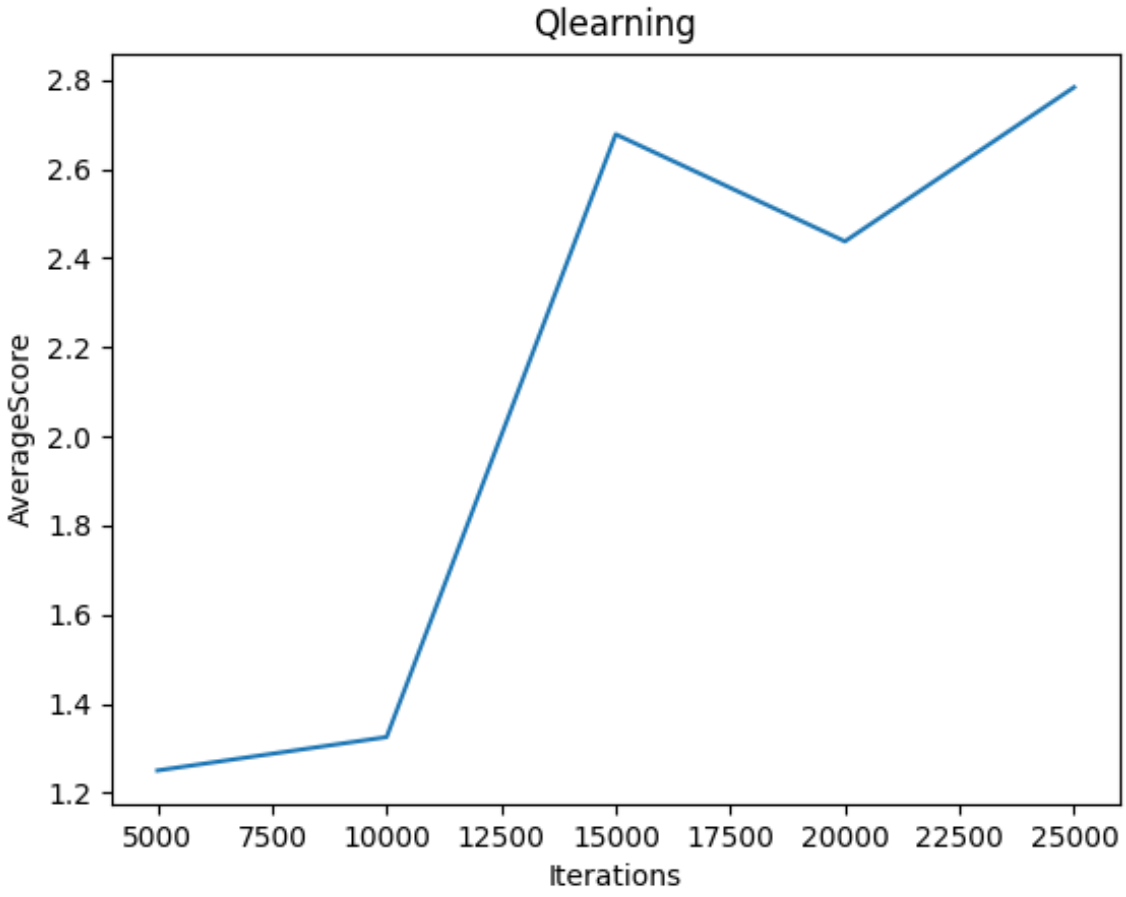
- Epsilon: 0.1
- Alpha: 0.1
- Learning Rate: 0.1
- Discounting = 1



2.2 Q Learning: Test 1

Added another state variable 'next_pipe_bottom_y' which represent the bottom pipe in the gap and its y position. So the number of states we can reach were * 15. Variables used:

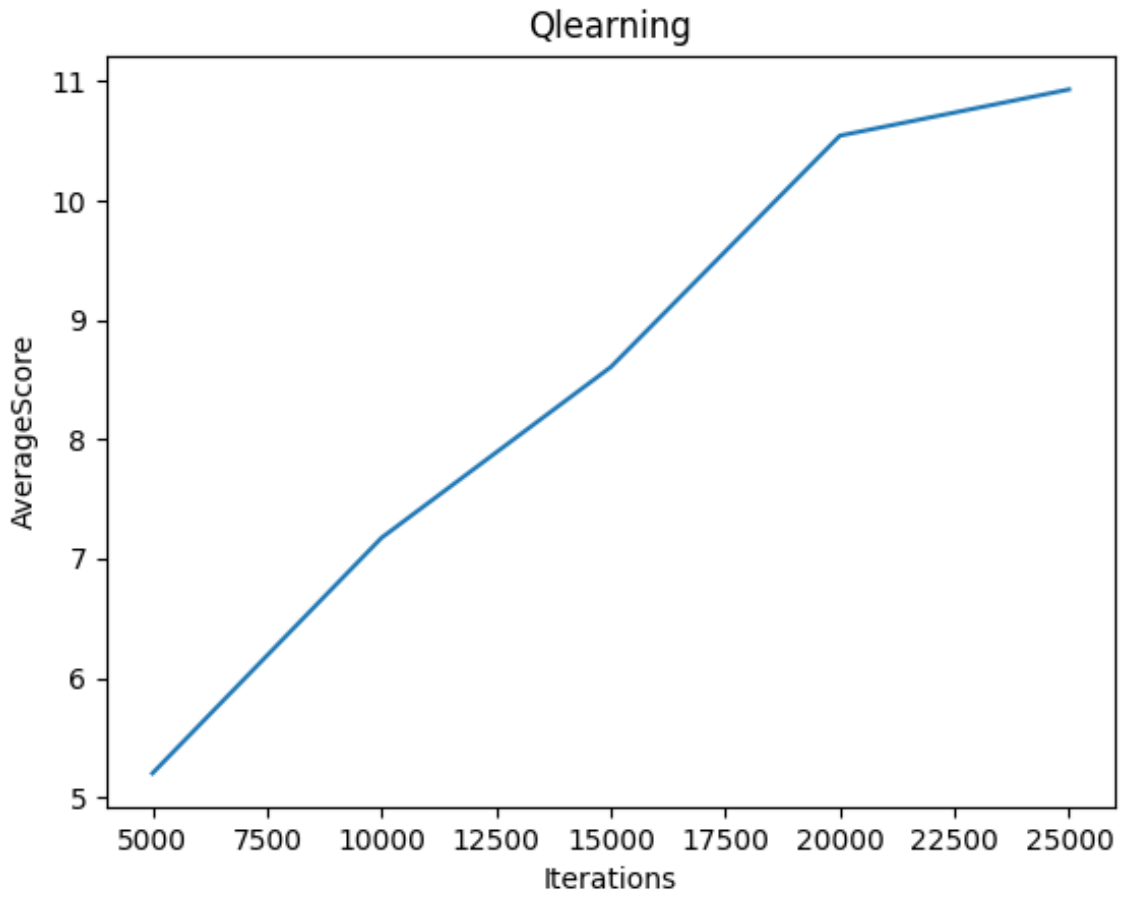
- Epsilon: 0.01
- Alpha: 0.01
- Learning Rate: 0.8
- Discounting = 1



2.3 Q Learning: Test 2

Variables used:

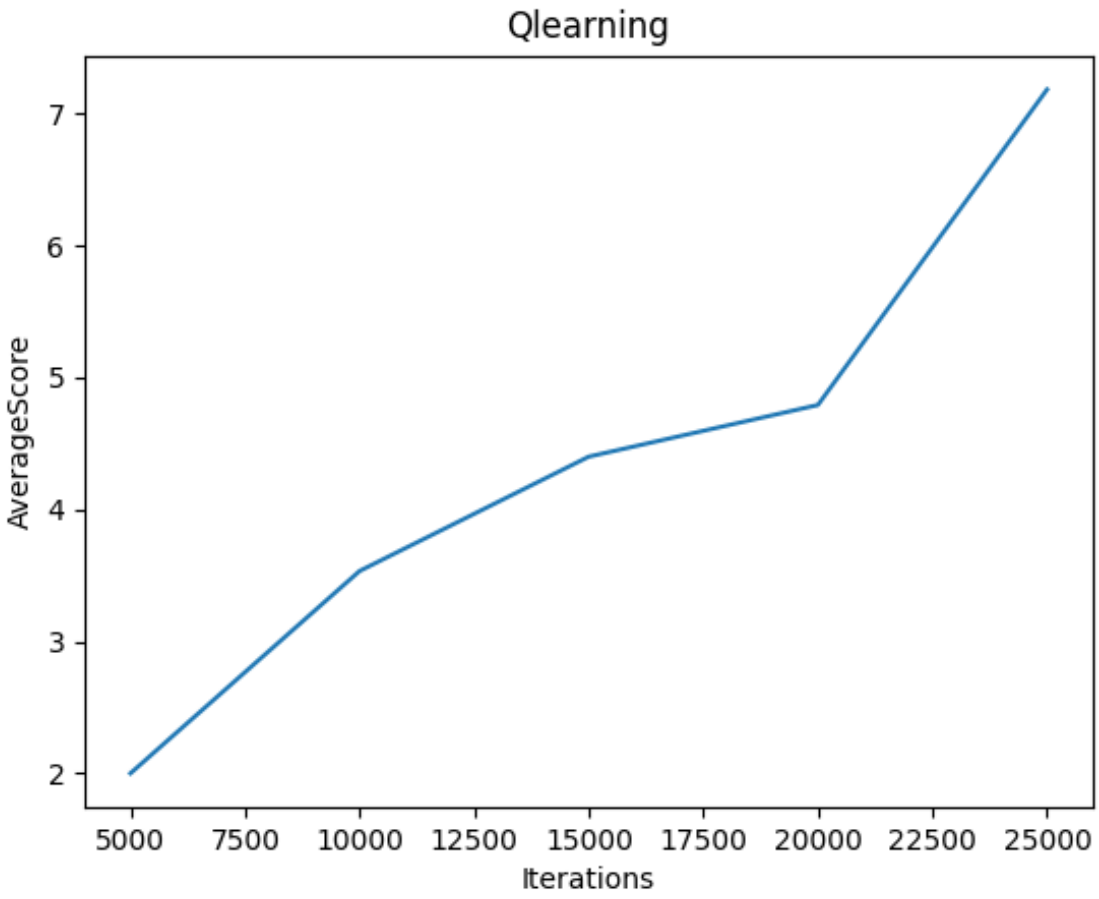
- Epsilon: 0.1
- Alpha: 0.1
- Learning Rate: 0.1
- Discounting = 0.8



2.4 Q Learning: Test 3

Variables used:

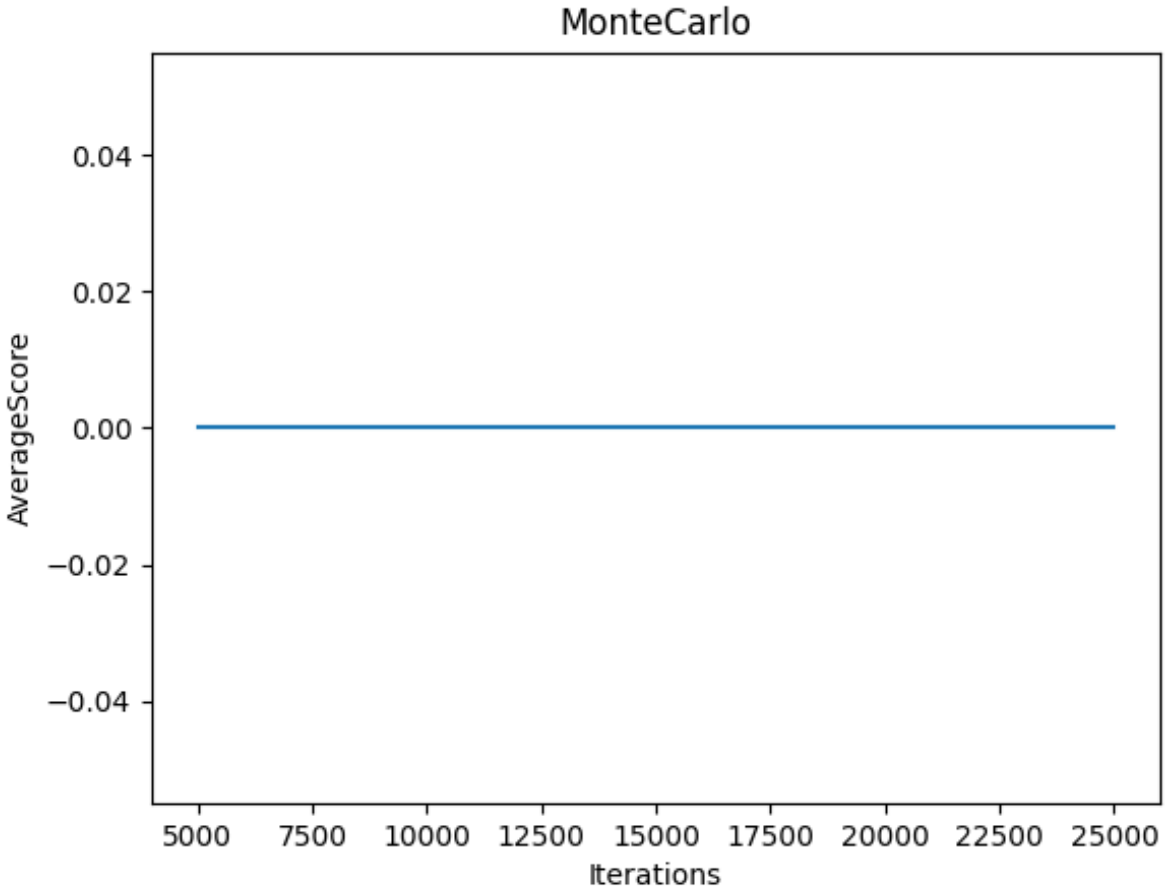
- Epsilon: 0.01
- Alpha: 0.1
- Learning Rate: 0.1
- Discounting = 0.8



2.5 Monte Carlo

Variables used:

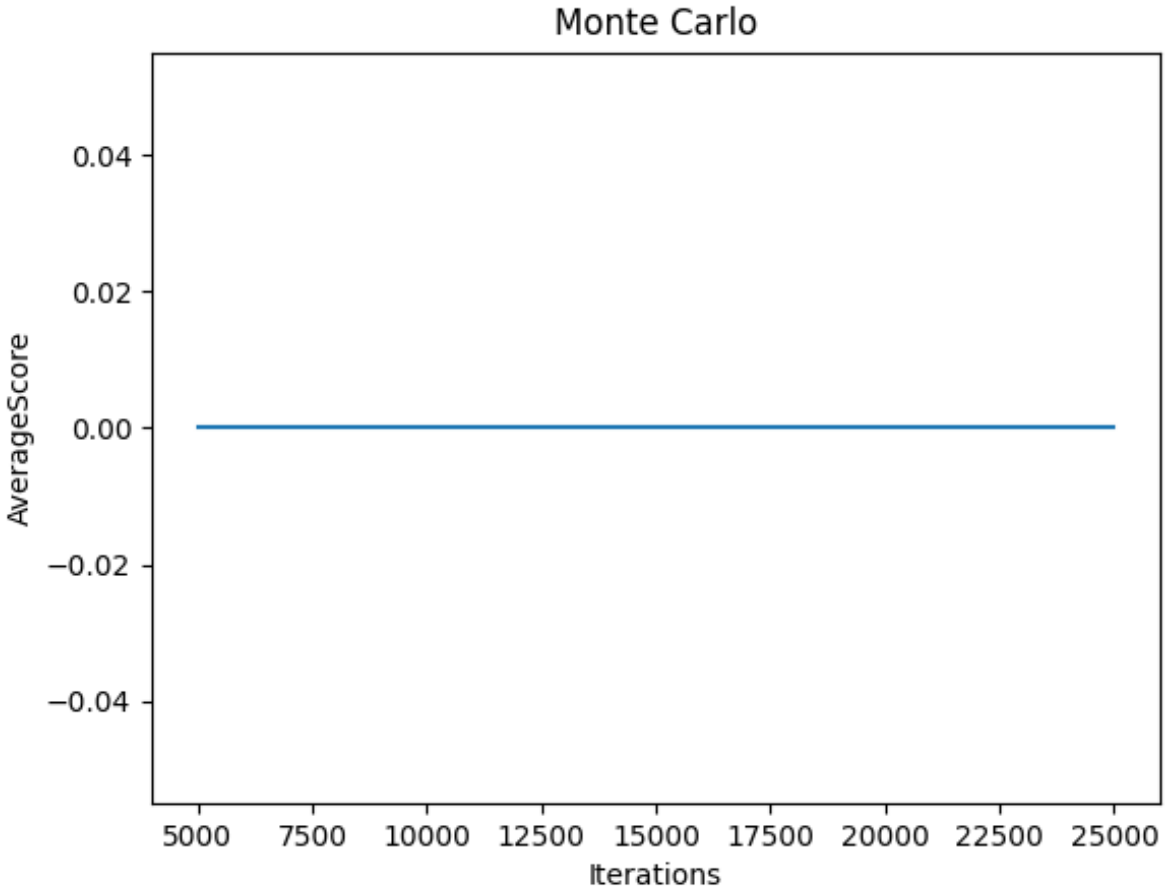
- Epsilon: 0.1
- Learning Rate: 0.1
- Discounting = 1



2.6 Monte Carlo: Test 1

Variables used:

- Epsilon: 0.01
- Learning Rate: 0.1
- Discounting = 1



2.7 Q-learning Neural network

We couldn't implement it

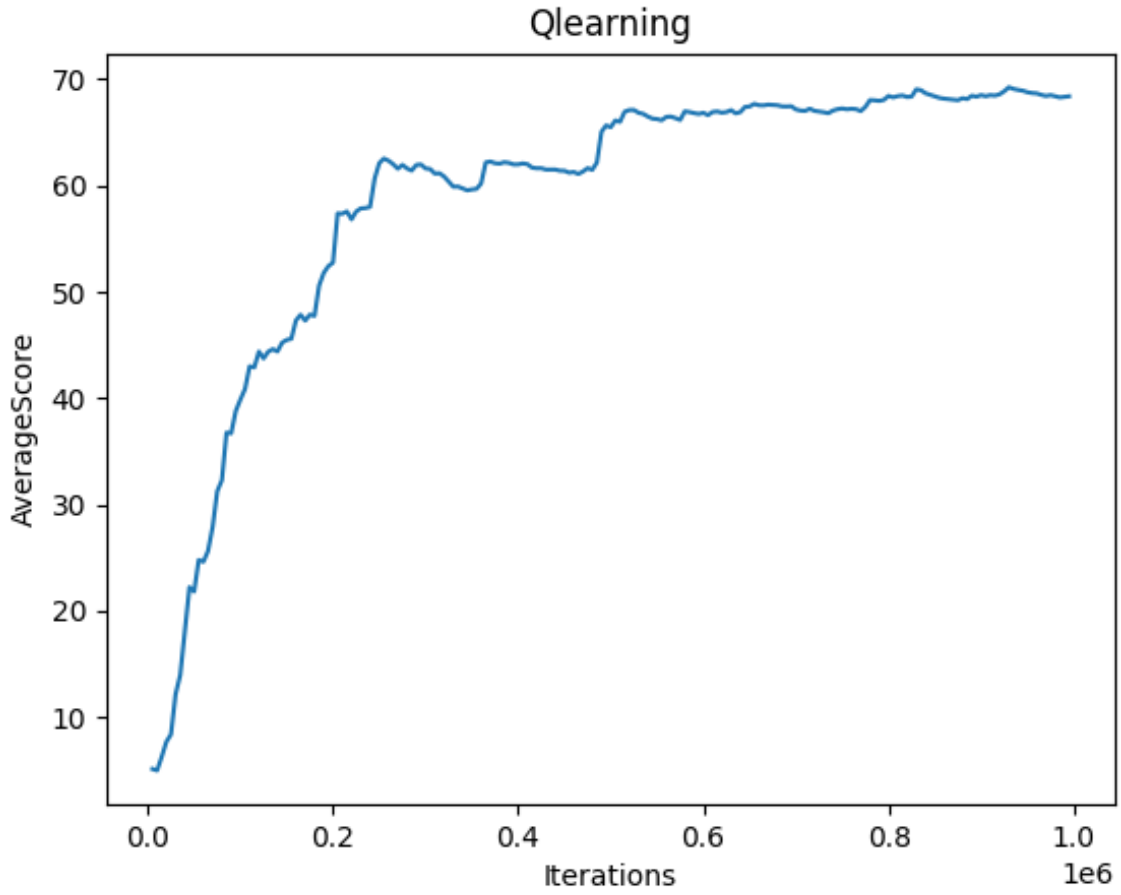
3 Results

Overall the project could be done better by finding better parameters for the algorithms and tune the Q-value factor for the Q-learning agent. As for Monte Carlo agent, it was very disappointing and could have been done better by adding the velocity of the bird and distance from the bird to the next top pipe y position to the Q-value factor so the agent will learn more after each episode. For the Neural-Network agent, we could not implement it but it could have been a fine agent because it learns in each episode on some batch of random 100 states and outputs Q-value if the bird is suppose to flap or not to flap.

4 Bonus

Here we talk about the results and which one was the best agent
Variables used:

- Epsilon: 0.1
- Alpha: 0.1
- Learning Rate: 0.1
- Discounting = 0.8



The agent we chose is the Q-learning agent where the discount factor is 0.8 as we can see from the plots above that it is the agent that is rising the most in average score for each iteration performed. In this plot below we can see that when we train around 500'000 iterations the agent doesn't improve after that and the average score would be around 70. This however can be improved by adding to the Q-value factor some calculation about the velocity of the bird and the distance to the next top pipe y position, the next top pipe y position and the next bottom pipe y position and get the middle point for that.