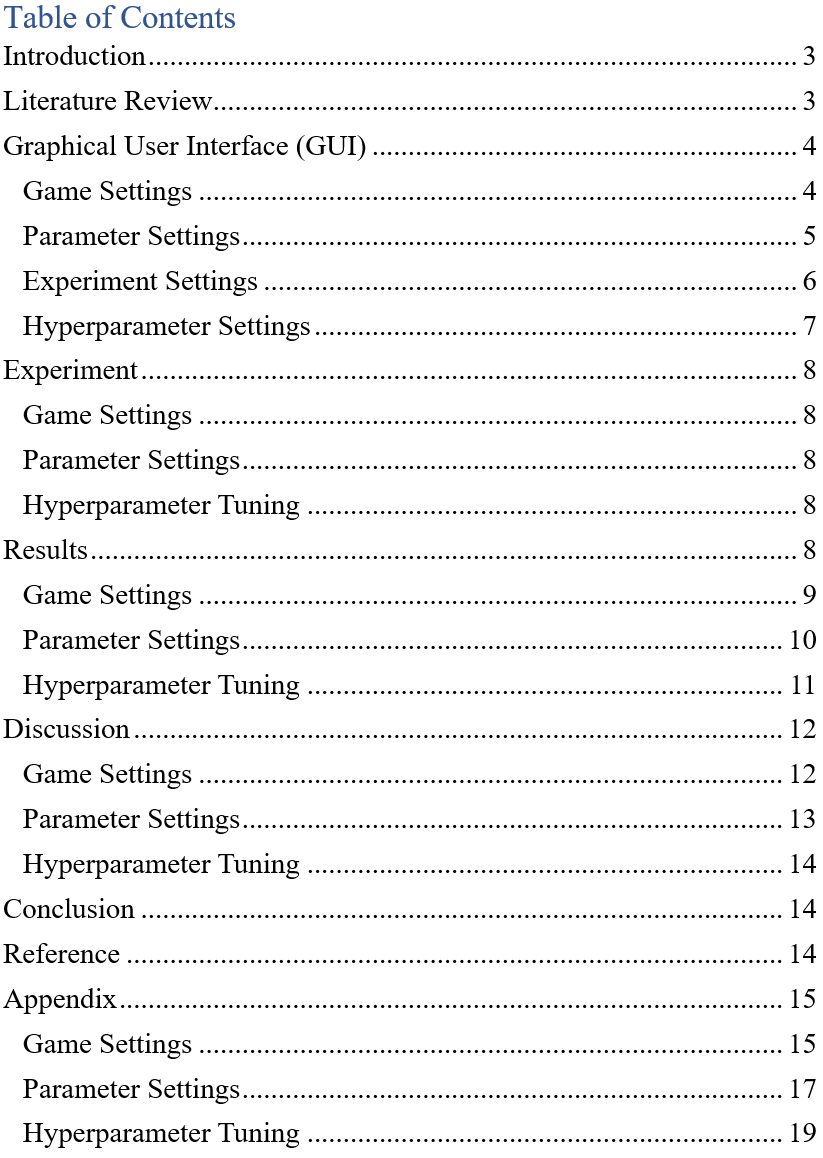


**Q-Learning:**

**Effects of Algorithmic Improvements and Different Settings on Brick Breaker Game Performance**





# Introduction

Reinforcement learning is a machine learning training method where desired behavior is rewarded while undesirable behavior is penalized. These are called positive and negative reinforcement respectively. A reinforcement learning agent is capable of observing and interpreting its environment, acting then receiving a reward or penalty based on its action. A reinforcement learning agent is said to learn through trial and error.

In this paper, we will be focusing on the topic of positive reinforcement by using the game Brick Breaker as the environment. In Brick Breaker, a player bounces a ball to destroy bricks by moving a paddle left and right. The player loses if the ball leaves the screen and wins if all bricks are destroyed.

Q-Learning is the reinforcement algorithm utilized in this paper. The foundation of the code was written by Matthew Chan and is available at medium.com[1](#_Reference). The code has been modified to work with the game.

This paper aims to study the effects of different improvements of Q-Learning on the agent performance. Combinations of these improvements will then be tested to determine the best combination and get the best possible results. The effects of varying the game settings on the agent performance will also be studied and documented.

Genetic algorithm will also be employed in this paper and is used to determine the hyperparameters of the Q-Learning agent. Different techniques such as single/double crossover and tournament/roulette selection are used to ensure more variety during the evolution process to produce better results.

A GUI was also created for users to more easily change the settings of the project and streamline the process of hyperparameter tuning and data collection.

# Literature Review

The main innovation in this paper is to combine existing improvements of the Q-Learning agent to produce a new, better solution. In the end, 7 research papers of interest were identified and listed below.

Firstly, **genetic algorithm** was used by Wicaksono, A. S., & Supianto, A. A. for hyperparameter optimization of the machine learning methods used in online news popularity prediction[2](#_Reference).

Secondly, H.R. Tizhoosh has introduced **opposition-based learning** as a new scheme for machine intelligence[3](#_Reference).

Thirdly, the **granularity of the state space** was shown to affect the results by Jacopo Fior and Luca Cagliero in their study of machine learning-based stock trading[4](#_Reference).

Furthermore, Michal Gregor and Juraj Spalek have also done research on optimistic exploration value functions[5](#_Reference). This has been adapted in this project as the **random initialization of the Q-tables**.

Moreover, research into action elimination with deep reinforcement learning[6](#_Reference) has also been done by Tom Zahavy, et al. This has been adapted to varying the **size of the action space** in the project.

Apart from that, research into **different reward functions** on the training performance of a Double DQN[7](#_Reference) has also been done by Stefan Šćepanović.

Finally, Double Q-Learning was introduced by Hado van Hasselt which lacks the overestimation bias of the Q-Learning algorithm[8](#_Reference). This has adapted to **N-tuple Q-Learning** which is another innovation in this paper where N Q-tables were created for each reinforcement learning agent.

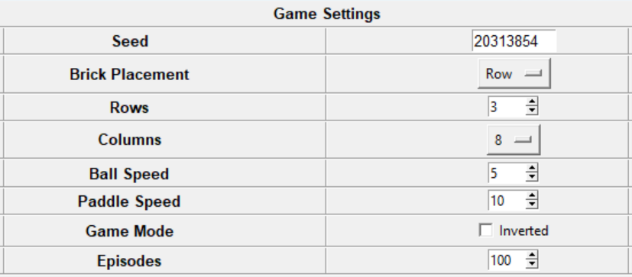
Thus, we can study the effects of combining multiple improvements and ascertain the best combination with the best results.

# Graphical User Interface (GUI)

As mentioned above, users can easily change the settings of the project using the GUI created. The different settings of the GUI will be explained below.

## Game Settings

The game settings allow users to customize the different aspects of the game. The different options available to the users are shown below in Figure 1.



***Figure 1: Game Settings***

* **Seed**

|  |  |
| --- | --- |
| Range | 0 – 4294967295 |
| Default | 20313854 |
| Description | Integer used to initialize the pseudorandom number generator to ensure that results are reproducible |

* **Brick Placement**

|  |  |
| --- | --- |
| Options | Row, Column, Random |
| Default | Row |
| Description | Controls how the 3 brick types are arranged in game.  **Row:** by row  **Column:** by column  **Random:** random |

* **Rows**

|  |  |
| --- | --- |
| Range | 1 – 10 |
| Default | 3 |
| Description | Controls the number of brick rows in the game. |

* **Columns**

|  |  |
| --- | --- |
| Options | 1,2,3,4,5,6,8,10,12,15 |
| Default | 8 |
| Description | Controls the number of bricks in each row in the game. Can only be a factor of 600 up to 15. |

* **Ball Speed**

|  |  |
| --- | --- |
| Range | 1 – 10 |
| Default | 5 |
| Description | Controls the speed of the ball in the game. |

* **Paddle Speed**

|  |  |
| --- | --- |
| Range | 5 – 15 |
| Default | 10 |
| Description | Controls the speed of the paddle in the game. |

* **Game Mode**

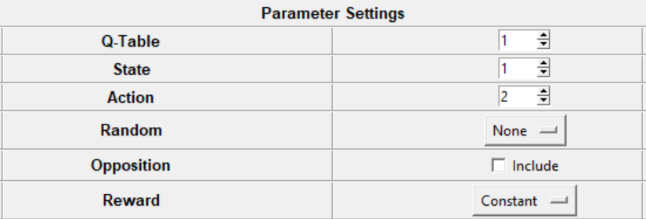
|  |  |
| --- | --- |
| Options | False, True |
| Default | False |
| Description | Allows inversion of the position of the bricks and paddle in the game. |

* **Episodes**

|  |  |
| --- | --- |
| Range | 1 – 200 |
| Default | 100 |
| Description | Controls the number of failed episodes before the game is reset and recorded as a failed run. |

## Parameter Settings

The parameter settings allow users to control the different settings of the Q-Learning agent. The different options available are shown below in Figure 2.



***Figure 2: Parameter Settings***

* **Q-Table**

|  |  |
| --- | --- |
| Range | 1 – 5 |
| Default | 1 |
| Description | Controls the number of Q-tables generated for each learning agent. |

* **State**

|  |  |
| --- | --- |
| Range | 1 – 8 |
| Default | 1 |
| Description | Controls the granularity of the state space of the Q-tables. State is assigned based on distance between the ball and the paddle. |

* **Action**

|  |  |
| --- | --- |
| Options | 2, 3 |
| Default | 2 |
| Description | Controls the size of the action space. The 2 default actions are for the paddle to move left and right. The 3rd action is to do nothing. |

* **Random**

|  |  |
| --- | --- |
| Options | None, Normal, Uniform |
| Default | None |
| Description | Controls the initial values of the Q-tables.  **None:** 0s  **Normal:** -10 – 10  (Normal distribution)  **Uniform:** -10 – 10  (Uniform distribution) |

* **Opposition**

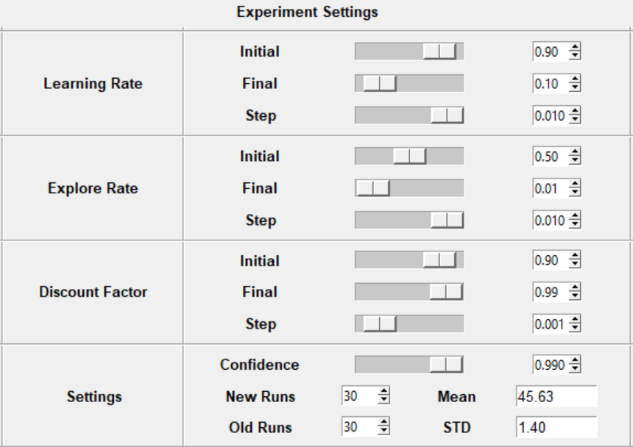
|  |  |
| --- | --- |
| Options | False, True |
| Default | False |
| Description | Allows the agent to also take the opposite action and update the Q-table. The do-nothing action has no opposite action and is only updated once. |

* **Reward**

|  |  |
| --- | --- |
| Options | Constant, Time-Based,  X-Distance,  X-Distance (Center),  XY-Distance |
| Default | Constant |
| Description | Controls the reward function used by the agent. **Constant:** 1 if run not ended else 0  **Time-Based:** turn count  **X-Distance:** 6 – distance (ball.mid.x, paddle.mid.x)  **X-Distance (Center):**  6 – minimum distance (ball.mid.x, paddle.x)  **XY-Distance:** 7.5 – dist (ball.mid, paddle.mid) |

## Experiment Settings

The experiment settings allow users to control the process of comparing different parameter settings. It is accessed by pressing the experiment button. The options available are shown below in Figure 3.



***Figure 3: Experiment Settings***

* **Learning Rate**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 / 0.000 – 0.010 |
| Default | 0.90 / 0.10 / 0.010 |
| Description | Controls how much the new learned value is used. |

* **Explore Rate**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 / 0.000 – 0.010 |
| Default | 0.50 / 0.01 / 0.010 |
| Description | Controls the chance that a random action is taken instead of the best action. |

* **Discount Factor**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 / 0.000 – 0.010 |
| Default | 0.90 / 0.99 / 0.001 |
| Description | Determines the importance of future rewards. |

* **Confidence**

|  |  |
| --- | --- |
| Range | 0.500 – 0.999 |
| Default | 0.990 |
| Description | Controls the confidence level of the t-test used to evaluate the results of the different experiments. |

* **New Runs**

|  |  |
| --- | --- |
| Range | 30 – 100 |
| Default | 30 |
| Description | Controls the number of successful runs needed for results collection. |

* **Old Runs**

|  |  |
| --- | --- |
| Range | 30 – 100 |
| Default | 30 |
| Description | Controls the size of the old sample to be compared. |

* **Mean**

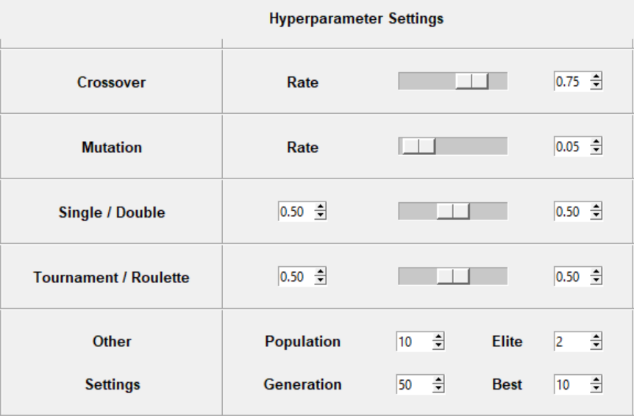
|  |  |
| --- | --- |
| Range | 1.00 – 200.00 |
| Default | 45.63 |
| Description | Controls the mean of the old sample to be compared. |

* **STD**

|  |  |
| --- | --- |
| Range | 0.00 – 140.71 |
| Default | 1.40 |
| Description | Controls the standard deviation of the old sample to be compared. |

## Hyperparameter Settings

The hyperparameter settings allow users to control the process of hyperparameter tuning of the agent. It is accessed by pressing the tuning button. The options available are shown below in Figure 4.



***Figure 4: Hyperparameter Settings***

* **Crossover Rate**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 |
| Default | 0.75 |
| Description | Controls the percentage of chromosomes that undergo the process of crossover. |

* **Mutation Rate**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 |
| Default | 0.05 |
| Description | Controls the percentage of genes that undergo the process of mutation. |

* **Single / Double**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 |
| Default | 0.50 / 0.50 |
| Description | Controls the percentage of single and double crossover when successful crossover occurs. |

* **Tournament / Roulette**

|  |  |
| --- | --- |
| Range | 0.00 – 1.00 |
| Default | 0.50 / 0.50 |
| Description | Controls the percentage of tournament and roulette selection for the selection of the next generation. |

* **Population**

|  |  |
| --- | --- |
| Range | 10 – 100 (±2) |
| Default | 10 |
| Description | Controls the population size of each generation. |

* **Elite**

|  |  |
| --- | --- |
| Range | 0 – 10 (±2) |
| Default | 2 |
| Description | Controls the number of fittest chromosomes that are guaranteed to survive. |

* **Generation**

|  |  |
| --- | --- |
| Range | 1 – 100 |
| Default | 50 |
| Description | Controls the number of iterations of genetic algorithm to be run. |

* **Best**

|  |  |
| --- | --- |
| Range | 1 – 10 |
| Default | 10 |
| Description | Controls the number of fittest chromosomes saved and displayed at the end. |

# Experiment

The experiments to be run using the GUI are listed below. The results of the default settings are used as the baseline for comparison. The experiments are run for at least 30 times for the Central Limit Theorem to hold and to ensure statistical significance of the results.

## Game Settings

Firstly, each **game** setting is varied individually and the results recorded to study their effects and determine the significance of each game setting to the results of the reinforcement learning agent.

## Parameter Settings

Secondly, each **parameter** setting is varied individually and the results recorded to study their effects and determine the significance of each parameter setting to the results of the reinforcement learning agent.

Different parameter settings are then combined to determine which combination produces the best results before moving onto hyperparameter tuning.

## Hyperparameter Tuning

The genetic algorithm is run using the best parameter combination found above. It is run for 50 generations with a population size of 10 with default settings. The fittest 10 chromosomes are recorded and shown.

The fitness value of each chromosome is determined by running the experiment only once for the sake of time. To reduce the overestimation bias, if the same chromosome is evaluated more than once, the worst result is taken as the fitness value.

Finally, the fittest chromosomes are then combined with the best parameter combination. The results are recorded to determine the best hyperparameter settings which produces the best results.

# Results

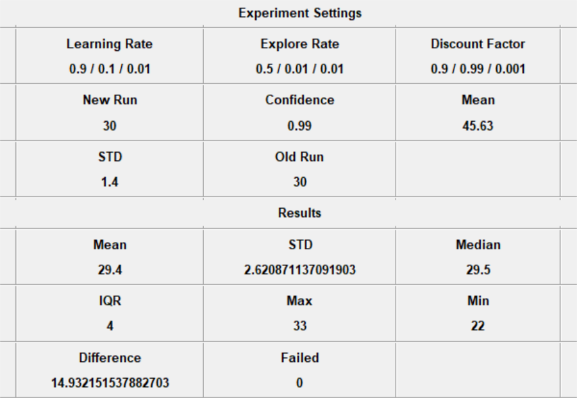
Metrics such as the mean, median, STD, IQR, max, min, difference and failed are recorded. Only relevant metrics are shown but the full list can be seen in the appendix.

The most important metric is the difference. It records the expected difference in mean between the 2 samples. It is calculated using the mean and STD with a t-test with 99% confidence level. A higher difference means the new sample is objectively better.

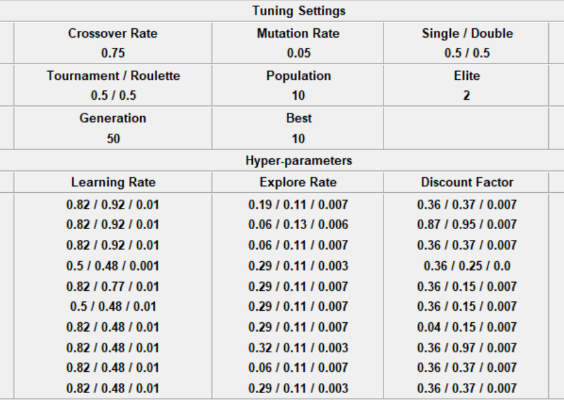
Examples of the result screens are given below in Figure 5, 6 and 7.



***Figure 5: Initial Settings***



***Figure 6: Experiment Results***



***Figure 7: Hyperparameter Results***

## Game Settings

The results of varying the game settings are shown below. We only study the mean as the difference is not relevant here.

* [**Brick Placement**](#_Game_Settings)

There are only minor differences in the results of different brick placements.

* [**Rows**](#_Game_Settings)

As the number of row increases, the mean of the results increases.

* [**Columns**](#_Game_Settings)

As the number of columns increases, the mean of the results also increases.

* [**Ball Speed**](#_Game_Settings)

As the ball speed increases, the mean of the results decreases before increasing again.

* [**Paddle Speed**](#_Game_Settings)

After an initial high, the mean of the results remains relatively stable.

* [**Game Mode**](#_Game_Settings)

Both game modes have identical results. Inverting the bricks and paddle position seems to have no effect on the results.

## Parameter Settings

The results of varying the parameter settings are shown below. We study the difference to discover which parameters has the greatest improvements on the results.

* [**Q-Table**](#_Parameter_Settings)

As the number of Q-table increases, the difference also increases.

* [**State**](#_Parameter_Settings)

As the number of state space increases, the difference also increases.

* [**Action**](#_Parameter_Settings)

When the action space increases, there is minimal difference between the results.

* [**Random**](#_Parameter_Settings)

There is no difference between initializing the Q-tables with 0s or random values.

* [**Opposition**](#_Parameter_Settings)

Opposition learning fails to work with the **Constant** reward function.

* [**Reward**](#_Parameter_Settings)

Only one of the reward function, **X-Distance (Center)** shows a significant difference while the rest have none.

|  |  |
| --- | --- |
| Symbol | Function |
| C | Constant |
| XC | X-Distance (Center) |
| T | Time-Based |
| X | X-Distance |
| XY | XY-Distance |

* [**Combination**](#_Parameter_Settings)

**Random** was excluded from combination as there was no difference in the results. All other parameters were included.

|  |  |
| --- | --- |
| Symbol | Parameter Settings |
| R | Reward - XC |
| S | State - 8 |
| O | Opposition - True |
| A | Action - 3 |
| Q | Q-Table - 5 |

We choose **R** as the first parameter. By combining **R** with the other 4, we find that **RS, RO** > **R** > **RA,** **RQ** Thus, we then combine **RS** and **RO** into **RSO**. However, we find that **RS** > **RSO** > **RO**. So, we conclude that **RS** is the best combination.

## Hyperparameter Tuning

The **RSO** parameter combination is used in genetic algorithm. The fittest chromosomes are listed below arranged in the order Initial / Final / Step and are given in percentages.

* **Final Chromosomes**

These are the final 10 fittest chromosomes.



* [**Chromosome Results**](#_Hyperparameter_Tuning)

These are the results of combining the chromosome hyperparameters with **RSO**.

* [**Chromosome Results Sorted**](#_Hyperparameter_Tuning)

These are the chromosome results sorted by the difference in descending order. Most of the results have a difference greater than 40 with the **best** result having a **difference** of 43.12 and a **mean** of only 1.83 episodes.

* **Final Chromosomes Sorted**

These are the final chromosomes sorted based on the sorted chromosome results.



# Discussion

The results of the experiments above will now be discussed so that we may gain better understanding and insight into the different settings of the experiments.

## Game Settings

First, we discuss the effects of varying the game settings on the results.

* **Brick Placement**

As seen above, there is minimal difference in the results of different brick placements. This is because the overall objective of the game is still the same which is to break 3 rows of bricks without losing the ball. The difference in results comes from the positioning of the 3 different types of bricks.

* **Rows**

As the number of row increases, the mean of the results also increases. When the number of row increases, the number of bricks also increases. This increases the length of the game and gives the agent more chances to commit mistakes and lose the game. So, the number of episodes increases.

However, the rate of increase of the episodes slows down as the number of rows increases. As the length of the game increases, the agent becomes a better player and makes less mistakes thus fewer extra episodes are needed to beat the game.

* **Columns**

As the number of columns increases, the mean of the results also increases. The reasoning for this is similar to the one above where increasing the number of bricks increases the length of the game and gives the agent more chances to lose. The rate of increases also slows down as the agent slowly becomes a better player at the game.

* **Ball Speed**

As the ball speed increases, the mean of the results decreases before increasing again along with the number of failed runs.

A reinforcement learning agent needs to win and lose to learn the correct moves. When the ball is slow, it takes more turns to move and lose. The agent takes longer to learn the correct moves and the number of episodes needed is higher. So, increasing the ball speed decreases the episode mean.

After a certain point, the ball moves and loses so quickly that the agent has no time to learn. This causes the failure rate to increase as the ball speed increases.

* **Paddle Speed**

After an initial high along with a high number of failed runs, the mean of the results remains relatively stable.

For the paddle, its movement speed is not very important. It just has to be fast enough to catch the ball. So, increasing the paddle speed does not affect the mean very much.

However, issues arise when the paddle is too slow. The paddle may not be able to traverse the screen to catch the ball and this causes failure rates to increase. Thus, any paddle speed above 5 is usable.

* **Game Mode**

Both game modes have identical results. Inverting the bricks and paddle position seems to have no effect on the results. This is due to the fact that the agent is training against the same 3 rows of bricks and thus the only difference is to invert the controls.

## Parameter Settings

Next, we discuss the effects of varying the parameter settings on the results.

* **Q-Table**

As the Q-table increases, the difference also increases along with the failed runs.

As mentioned above, Double Q-Learning was able to reduce the overestimation bias in the best Q-value. By increasing the number of Q-tables, the overestimation bias is reduced further and leads to better results.

However, the increase in Q-tables has also increased the number of state spaces that the agent has to fill. If the state spaces are correctly filled, we can expect better results. If not, the agent will fail to choose the correct move. As the number of state spaces increase, the chances that they are not filled correctly and thus the failure rate increases.

* **State**

As the number of state space increases, the difference also increases. The granularity of the state space is also important in getting the best results. If the state space is too coarse, it may not be effectively separated. If it is too fine, we end up with redundant state spaces. For the agent, it is clear that the state space is too coarse and thus we get better results as we increase the state space.

* **Action**

When the action space increases, there is minimal difference between the results. Choosing the correct action space is also important in getting good results. In this case, the 3rd action of do nothing is meaningless and does not contribute to the results. However, a more meaningful action may actually lead to better results.

* **Random**

There is no difference between initializing the Q-tables with 0s or random values. The simplicity and short length of the game may be to blame. Exploration is promoted in the early stages with random values. A longer and complex game may see better results.

* **Opposition**

Opposition learning fails to work with the **Constant** reward function because it works by also taking the opposite action so there is twice the information available. The issue is the **Constant** reward function gives equal results for both actions. Thus, a different reward function may see better results.

* **Reward**

Only one of the reward function, **X-Distance (Center)** shows a significant difference while the rest have none.

**Constant** and **Time-Based** both learn that a longer run is better but **Time-Based** is dominated by its later moves. **X-Distance** performs worse than **X-Distance (Center)** because the distance from the whole paddle is more important than just the midpoint. **XY-Distance** performs the worst as the vertical distance is irrelevant to the game.

* **Combination**

**State** and **Opposition** works well with the new **Reward** function while **Action** and **Q-Table** do not. However, the combination of **State** and **Opposition** produces worse results. This may be caused by interference between the different parameters. There is also a limit to how much the results can be improved which can only be breached by changing the hyperparameters.

## Hyperparameter Tuning

Finally, we discuss the results of hyperparameter tuning. The top 5 hyperparameters are given below.



Firstly, a high learning rate is preferred so the new learned value is fairly important.

Secondly, a moderately low explore rate is best. The explore rate needs to be low enough so the agent can still take the best action most of the time while high enough to encourage sufficient exploration.

Finally, the discount factor needs to be moderate. The agent should place more emphasis on immediate rewards but also put a little focus into future rewards.

# Conclusion

In conclusion, Q-Learning has been successfully used to teach an agent to play the game Brick Breaker. The experiments were successfully run using the GUI given.

In this paper, the effects of varying the game and parameter settings were studied. The best combination of parameters was also discovered which is to use the correct **Reward** function and a high **State** space.

**Genetic algorithm** was then used to tune the hyperparameters. It was discovered that the best hyperparameters were a high learning rate, a moderately low explore rate and a moderate discount factor.

The best results obtained by the agent is an episode **mean** of only 1.68 and a **difference** of 43.12 with a 99% confidence level.

A good extension to the project would be to adapt Q-Learning to other games.

# Reference

The list of references used are listed below.

1. <https://medium.com/@tuzzer/cart-pole-balancing-with-q-learning-b54c6068d947>
2. Wicaksono, A. S., & Supianto, A. A. (2018). Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction. *International Journal of Advanced Computer Science and Applications*, *9*(12).
3. Tizhoosh, Hamid. (2005). Opposition-Based Learning: A New Scheme for Machine Intelligence. IEEE. 1. 695 - 701. 10.1109/CIMCA.2005.1631345.
4. J. Fior and L. Cagliero, "Exploring the Use of Data at Multiple Granularity Levels in Machine Learning-Based Stock Trading," 2020 International Conference on Data Mining Workshops (ICDMW), 2020, pp. 333-340, doi: 10.1109/ICDMW51313.2020.00053.
5. M. Gregor and J. Spalek, "The optimistic exploration value function," 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES), 2015, pp. 119-123, doi: 10.1109/INES.2015.7329650.
6. Zahavy, T., Haroush, M., Merlis, N., Mankowitz, D. J., & Mannor, S. (2018). Learn what not to learn: Action elimination with deep reinforcement learning. Advances in neural information processing systems, 31.
7. S. Šćepanović, "Testing reward function choice influence on training performance of Double DQN," 2022 11th Mediterranean Conference on Embedded Computing (MECO), 2022, pp. 1-4, doi: 10.1109/MECO55406.2022.9797177.
8. Hasselt, H. (2010). Double Q-learning. Advances in neural information processing systems, 23.

# Appendix

## Game Settings

* [**Brick Placement**](#_Game_Settings_1)



* [**Rows**](#_Game_Settings_1)



* [**Columns**](#_Game_Settings_1)



* [**Ball Speed**](#_Game_Settings_1)



* [**Paddle Speed**](#_Game_Settings_1)



* [**Game Mode**](#_Game_Settings_1)



## Parameter Settings

* [**Q-Table**](#_Parameter_Settings_1)



* [**State**](#_Parameter_Settings_1)



* [**Action**](#_Parameter_Settings_1)



* [**Random**](#_Parameter_Settings_1)



* [**Opposition**](#_Parameter_Settings_1)



* [**Reward**](#_Parameter_Settings_1)



* [**Combination**](#_Parameter_Settings_1)



## Hyperparameter Tuning

* [**Chromosome Results**](#_Hyperparameter_Tuning_1)



* [**Chromosome Results Sorted**](#_Hyperparameter_Tuning_1)

