**Autumn Semester 2022/23**

**Autonomous Robotic Systems (COMP4082)**

**Coursework Report**

|  |  |
| --- | --- |
| Name | Tan Zhun Xian |
| Student ID | 20313854 |
| Lecturer | Tomas Maul |
| Date | 10/12/2022 |

Contents

[Introduction 3](#_Toc121601091)

[Background 3](#_Toc121601092)

[Methods 4](#_Toc121601093)

[Results 5](#_Toc121601094)

[Discussion 6](#_Toc121601095)

[References 7](#_Toc121601096)

[Appendices 8](#_Toc121601097)

[Cart Pole Innovation Results 8](#_Toc121601098)

[Cart Pole Symbols 8](#_Toc121601099)

[Cart Pole Single Q-Learning 8](#_Toc121601100)

[Cart Pole Double Q-Learning 9](#_Toc121601101)

[Cart Pole Triple Q-Learning 9](#_Toc121601102)

[Cart Pole Quadruple Q-Learning 10](#_Toc121601103)

[Cart Pole Quintuple Q-Learning 10](#_Toc121601104)

[Acrobot Innovation Results 11](#_Toc121601105)

[Acrobot Symbols 11](#_Toc121601106)

[Acrobot Single Q-Learning 11](#_Toc121601107)

[Acrobot Double Q-Learning 12](#_Toc121601108)

[Acrobot Triple Q-Learning 12](#_Toc121601109)

[Acrobot Quadruple Q-Learning 13](#_Toc121601110)

[Acrobot Quintuple Q-Learning 13](#_Toc121601111)

[Cart Pole Reward Functions Results 14](#_Toc121601112)

[Cart Pole Single Q-Learning 14](#_Toc121601113)

[Cart Pole Double Q-Learning 15](#_Toc121601114)

[Cart Pole Triple Q-Learning 16](#_Toc121601115)

[Cart Pole Quadruple Q-Learning 17](#_Toc121601116)

[Cart Pole Quintuple Q-Learning 18](#_Toc121601117)

[Acrobot Reward Functions Results 19](#_Toc121601118)

[Acrobot Single Q-Learning 19](#_Toc121601119)

[Acrobot Double Q-Learning 20](#_Toc121601120)

[Acrobot Triple Q-Learning 21](#_Toc121601121)

[Acrobot Quadruple Q-Learning 22](#_Toc121601122)

[Acrobot Quintuple Q-Learning 23](#_Toc121601123)

# Introduction

Reinforcement learning is a machine learning training method where desired behavior is rewarded (positive reinforcement) while undesirable behavior is penalized (negative reinforcement). A reinforcement learning agent can perceive and interpret its environment, act, and receive a reward or penalty based on the action. In this way, a reinforcement learning agent can learn through trial and error.

Open Ai Gym is an open-source Python library which facilitates the development and comparison of different reinforcement algorithms by providing a standard API to communicate between learning algorithms and environments.

In this paper, we will be focusing on the topic of reinforcement learning using Open Ai Gym. Two environments were chosen to be the focus of our research which are Cart Pole v0 and Acrobot v1 from the Classic Controls environment. Cart Pole was chosen to showcase positive reinforcement learning while Acrobot was chosen to showcase negative reinforcement learning.

The first task detailed in this paper is to innovate and improve existing reinforcement learning algorithms to improve the performance of the algorithm. The technique used is to combine different improvements of reinforcement learning algorithms and to compare the results. This allows us to ascertain the best combination which produces the best results.

The second task is to carry out simple research on the reinforcement learning agents. The research topic decided on was to study the effects of different reward functions on the performance of the reinforcement learning algorithms. The base reward function used for Cart Pole and Acrobot are uninformative. Several different reward functions have been proposed and will be tested to determine the best reward function for both problems.

The reinforcement algorithm used in this paper is Q-Learning. The base of the code used is adapted from medium.com[1](#_References) and was created by Matthew Chan. The code has been modified to run the chosen environment and collect results such as the mean number of episodes for completion and the standard deviation in completion episodes. The code also allows easier change in variables and reward functions to ease the process of result collection.

# Background

Before going further into the paper, research was done to discover any key papers that may be relevant to the topic. Research papers focusing on innovations to reinforcement learning algorithms and reward functions were prioritized.

For innovations in reinforcement learning algorithms, 6 research papers of interest were found. As shown by Jacopo Fior and Luca Cagliero, the granularity of the data influences the results of machine learning-based stock trading[2](#_References). H.R. Tizhoosh has also tried to introduce opposition-based learning as a new scheme for machine intelligence[3](#_References). Research into optimistic exploration value functions[4](#_References) has also been done by Michal Gregor and Juraj Spalek. Tom Zahavy, et al. have done research into action elimination with deep reinforcement learning[5](#_References). Research into discounting deep reinforcement learning[6](#_References) has also been done by Vincent François-Lavet, Raphael Fonteneau and Damien Ernst. Finally, Hado van Hasselt introduced Double Q-Learning algorithm which lacks the overestimation bias of Q-Learning algorithm[7](#_References).

For research into reinforcement learning reward functions, 3 research papers were found to be relevant. Stefan Šćepanović has tested and compared the training performance of a Double DQN with different reward functions[8](#_References). Reward functions based on the DDQN algorithm[9](#_References) have also been designed by Qianhao Xiao, Xin Zhang, et al. Finally, Jiexin Xie, Yue Li, et al. have optimized the reward functions of deep reinforcement learning for robotic trajectory planning[10](#_References).

# Methods

For both tasks, some parameters have been retained throughout the entire experiment. These parameters are:

1. Seed – 20313854
2. Max Discount Factor – 0.999
3. Min Learning Rate – 0.1
4. Min Explore Rate – 0.01

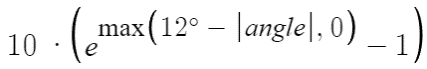
For the first task, 6 improvements were chosen based on the 6 research papers above on innovations in reinforcement learning algorithms. To facilitate easier testing and comparison between the results, improvements that span a range of values have been fixed such as granularity, discount factor and random initialization of initial Q-table values.

The 6 improvements to the reinforcement learning algorithm are:

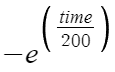
1. N-tuple Q-tables
   * Number of Q-tables range from 1 – 5
   * Action is selected based on the sum of all Q-tables values
   * First Q-table is randomly selected to be updated with reward
   * Second Q-table is randomly selected from remaining tables to get the best Q-value
2. Granularity of Q-tables
   * Granularity of Cart Pole is either (1, 1, 6, 3) or (1, 1, 6, 7)
   * Granularity of Acrobot is fixed at (1, 1, 1, 1, 10, 10)
3. Opposition Learning
   * Opposite action of a state is also acted upon, and the opposite reward added to the Q-table
4. Random Initialization of Q-tables
   * Q-tables can start with all 0s or with values drawn from a standard normal distribution
5. Discount Factor Discounting
   * Discount factor can be constant at 0.999 or start from 0.99 and increase by 0.001 each episode until 0.999
6. Action Limiting
   * Number of actions for Cart Pole is fixed at 2
   * Number of actions for Acrobot is 2 if do nothing action is removed or 3

For the second task of the paper, the best results from each category of the first task were chosen to have their reward functions changed so that new results can be generated and compared. Cart Pole received 5 new reward functions while 3 new reward functions were chosen for Acrobot.

The 5 reward functions for Cart Pole are:

1. Termination Penalty
   * Actions that lead to termination are given a penalty of -1
2. Time-Based Reward
   * 
   * Reward increases exponentially with time
3. Uniform Angle Reward
   * 
   * Reward varies linearly with angle
4. Exponential Angle Reward
   * 
   * Reward increases more rapidly as angle decreases
5. Logarithmic Angle Reward
   * 
   * Reward decreases more rapidly as angle increases

The 3 reward functions for Acrobot are:

1. Time-Based Penalty
   * 
   * Reward decreases exponentially with time
2. Velocity-Based Penalty
   * 
   * Reward decreases when velocity decreases or when both velocities are in opposite directions
3. Height-Based Penalty
   * 
   * Reward varies linearly with height

A reward of 0 is still given if Acrobot reaches threshold and episode terminates.

# Results

For Cart Pole, a run is solved when the average time steps for the last 100 episodes is greater or equal to 195.0.

A run in Acrobot is solved when the average time steps for termination for the last 100 episodes is less than or equal to 195.0.

For the experiment, 50 runs were tested with each run lasting 500 episodes. Runs that fail to complete within the allocated 500 episodes were considered failures. The experiment was considered a success if 30 runs were successful.

Results for the 30 runs such as the mean number of episodes for completion, standard deviation in completion episodes, median, interquartile range, min, max and number of failed runs were collected. Experiments that fail the 30-run requirement have their results left empty.

The results were sorted in descending order based on the expected difference between the base sample and current sample average. The t-test at a one-tailed 99% confidence level was used to calculate the difference. A result of 0 indicates that there is no statistically significant difference between the two samples. If ties occur, the results were then sorted ascendingly by mean.

For the innovations of both [Cart Pole](#_Cart_Pole_Innovation) and [Acrobot](#_Acrobot_Innovation_Results), results generally improve when the number of Q-tables increases. The addition of random initialization of Q-table values and/or discount factor discounting produces unpredictable results. A combination of more than 3 improvements also produces worse results.

For the innovation of [Cart Pole](#_Cart_Pole_Innovation), choosing the correct granularity gives the largest improvement to results. At [low number of Q-tables,](#_Cart_Pole_Single) opposition learning causes the experiments with wrong granularity to fail. However, the absence of opposition learning at [high number of Q-tables](#_Cart_Pole_Quadruple) produces terrible results and failed experiments. The [best and most stable combination](#_Cart_Pole_Quintuple) seems to be 5 Q-tables with the correct granularity and opposition learning.

For the innovation of [Acrobot](#_Acrobot_Innovation_Results), limiting the action space gives the largest improvement to results. However, opposition learning in [Acrobot](#_Acrobot_Innovation_Results) causes failure in all cases. The [best and most stable combination](#_Acrobot_Quintuple_Q-Learning) seems to be 5 Q-tables with action limiting.

For the reward functions of [Cart Pole](#_Cart_Pole_Reward), time-based reward generally performs worse while all other reward functions perform better than the base reward function. For the 3 angle-based reward functions, a definitive best reward function cannot be determined. The results of the termination penalty reward function are also comparable to the 3 angle-based reward functions.

For the reward functions of [Acrobot](#_Acrobot_Reward_Functions), time-based penalty and height-based penalty reward functions generally perform better than the base reward function with time-based penalty being the clear winner. The results of velocity-based reward function were too volatile to make a clear conclusion.

# Discussion

Results generally improve when the number of Q-tables increases. In Double Q-learning, having two Q-tables reduces the overestimation bias in the best Q-value. By increasing the number of Q-tables, we may be able to reduce the overestimation bias even further and get better results.

Random initialization of Q-table values is used to encourage exploration in the early stages of the run yet produces unpredictable results. This is most likely due to the random nature of the improvement. If the Q-tables start with good values, then the run may reach convergence faster. If not, then the results of the run may be worse.

Discount factor discounting also produces unpredictable results. This may be because discount factor discounting was originally used in deep reinforcement learning over the course of thousands of time steps. The 200-time steps in Cart Pole and 500-time steps in Acrobot are too short of a time scale for the improvement to have any noticeable effect and may instead be detrimental.

Normally, improvements to the algorithm by themselves improve the results. Yet, when we combine more and more improvements, they may start to interfere with each other, and this causes worse results. So, including 3 improvements to the algorithm seems to be the limit.

Choosing the correct granularity is also important to getting the best results. If the granularity is too course, we may not be able to separate the state space effectively. If the granularity is too fine, we may get a lot of similar and redundant state spaces.

Choosing the correct actions is also important to getting good results. Actions like do nothing in Acrobot are useless and do not contribute anything meaningful to the results and should be removed.

Opposition learning only works if the reward function is informative or if the Q-tables are not saturated. The original reward functions for Cart Pole and Acrobot are uninformative and serve to only clutter up the Q-tables. For Cart Pole, saturation of the Q-table happens only on low numbers of Q-tables, so opposition learning works more quickly. For Acrobot, saturation of the Q-tables happens more easily. This is because a bad run in Cart Pole ends quickly while a bad run in Acrobot runs for 500-time steps.

For reward functions of [Cart Pole](#_Cart_Pole_Reward), time-based reward performs worse because time is not important as the run length is fixed at 200. Termination penalty is more informative as it differentiates between good and bad moves. The 3 angle-based rewards are also more informative yet the best among the 3 cannot be determined as the experiments are too simple.

For reward functions of [Acrobot](#_Acrobot_Reward_Functions), time-based penalty is the most informative as it penalizes a later move more heavily. Results of velocity-based penalty were volatile as the only information received is the velocity while termination happens randomly. Height-based penalty performs better than the base reward as it is more informative yet worse than time-based penalty as the system cannot directly influence the height.

In conclusion, different reinforcement learning problems have different conditions and parameters. So, care must be taken to choose the best combinations of improvements and the best reward function to obtain the best results. However, the most important factor is still the granularity of the state space and the size of the action space.

For future works, other improvements such as variable discount factor can be explored. Reward functions that change based on different conditions can also be researched.

# References

1. <https://medium.com/@tuzzer/cart-pole-balancing-with-q-learning-b54c6068d947>
2. 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.
3. Tizhoosh, Hamid. (2005). Opposition-Based Learning: A New Scheme for Machine Intelligence. IEEE. 1. 695 - 701. 10.1109/CIMCA.2005.1631345.
4. 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.
5. 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.
6. Francois, Vincent & Fonteneau, Raphael & Ernst, Damien. (2015). How to Discount Deep Reinforcement Learning: Towards New Dynamic Strategies.
7. Hasselt, H. (2010). Double Q-learning. Advances in neural information processing systems, 23.
8. 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.
9. Q. Xiao, X. Zhang, L. Jiang and M. Wang, "Design of Reward Functions Based on The DDQN Algorithm," 2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2022, pp. 600-604, doi: 10.1109/ICMTMA54903.2022.00125.
10. J. Xie, Z. Shao, Y. Li, Y. Guan and J. Tan, "Deep Reinforcement Learning With Optimized Reward Functions for Robotic Trajectory Planning," in IEEE Access, vol. 7, pp. 105669-105679, 2019, doi: 10.1109/ACCESS.2019.2932257.

# Appendices

## Cart Pole Innovation Results

### Cart Pole Symbols

|  |  |  |  |
| --- | --- | --- | --- |
| Symbol | Name | Excluded | Included |
| G | Granularity | (1, 1, 6, 3) | (1, 1, 6, 7) |
| O | Opposition Learning | No | Yes |
| R | Random Initialization | Zeros | Normal |
| D | Discount Factor | 0.999 | (0.99, 100) |

### Cart Pole Single Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 257.27 | 14.94 | 256.0 | 25.0 | 288 | 235 | 0 | 0 |
| One | | | | | | | | |
| G | 237.03 | 12.74 | 236.0 | 17.0 | 269 | 217 | 0 | 11.66 |
| D | 257.97 | 15.28 | 259.0 | 21.0 | 294 | 230 | 0 | 0 |
| R | 266.47 | 17.92 | 263.5 | 25.0 | 305 | 235 | 0 | 0 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| G + R | 231.10 | 9.32 | 231.5 | 12.0 | 250 | 216 | 0 | 18.48 |
| G + D | 238.30 | 14.59 | 236.5 | 16.0 | 273 | 217 | 0 | 9.85 |
| G + O | 256.23 | 12.85 | 252.0 | 16.0 | 287 | 239 | 0 | 0 |
| D + R | 260.77 | 16.24 | 259.5 | 18.0 | 313 | 237 | 0 | 0 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| G + D + R | 234.57 | 13.34 | 232.0 | 17.0 | 268 | 216 | 0 | 13.95 |
| G + O + D | 250.03 | 8.18 | 249.0 | 10.0 | 275 | 236 | 2 | 0 |
| G + O + R | 252.80 | 10.53 | 254.0 | 14.0 | 276 | 229 | 0 | 0 |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| G + O + D + R | 247.80 | 7.06 | 247.0 | 12.0 | 261 | 237 | 0 | 2.25 |

### Cart Pole Double Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 254.90 | 19.99 | 252.5 | 29.0 | 303 | 223 | 1 | 0 |
| One | | | | | | | | |
| G | 230.03 | 16.98 | 224.5 | 23.0 | 275 | 206 | 1 | 17.36 |
| D | 249.83 | 17.32 | 252.0 | 22.0 | 296 | 215 | 4 | 0 |
| R | 252.77 | 19.85 | 251.0 | 25.0 | 328 | 221 | 0 | 0 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| G + D | 229.50 | 19.48 | 223.5 | 16.0 | 283 | 207 | 0 | 17.05 |
| G + O | 243.77 | 17.88 | 239.0 | 24.0 | 299 | 221 | 0 | 3.32 |
| G + R | 238.63 | 38.79 | 229.0 | 24.0 | 418 | 201 | 0 | 0.48 |
| D + R | 256.47 | 19.06 | 255.0 | 29.0 | 298 | 216 | 4 | 0 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| G + D + R | 232.57 | 13.84 | 230.5 | 19.0 | 265 | 205 | 0 | 15.80 |
| G + O + R | 246.97 | 18.29 | 243.5 | 23.0 | 296 | 217 | 0 | 0 |
| G + O + D | 249.60 | 15.99 | 248.0 | 22.0 | 290 | 218 | 1 | 0 |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| G + O + D + R | 245.97 | 13.43 | 244.5 | 18.0 | 285 | 228 | 0 | 2.53 |

### Cart Pole Triple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 242.50 | 18.16 | 237.5 | 15.0 | 304 | 220 | 3 | 4.50 |
| One | | | | | | | | |
| G | 217.13 | 10.23 | 215.5 | 10.0 | 248 | 205 | 0 | 32.23 |
| D | 242.37 | 13.40 | 242.5 | 16.0 | 274 | 218 | 5 | 6.13 |
| R | 242.03 | 19.36 | 236.5 | 21.0 | 296 | 212 | 9 | 4.56 |
| O | 291.23 | 31.16 | 283.5 | 27.0 | 409 | 256 | 0 | 0 |
| Two | | | | | | | | |
| G + R | 218.13 | 14.51 | 215.0 | 16.0 | 255 | 199 | 1 | 30.04 |
| G + D | 220.33 | 13.83 | 217.5 | 15.0 | 254 | 201 | 0 | 28.05 |
| G + O | 225.00 | 9.20 | 223.0 | 15.0 | 246 | 210 | 0 | 24.61 |
| D + R | 238.33 | 12.83 | 234.5 | 18.0 | 272 | 221 | 6 | 10.34 |
| O + R | 284.07 | 18.16 | 278.5 | 25.0 | 324 | 257 | 0 | 0 |
| O + D | 294.67 | 33.95 | 284.5 | 40.0 | 418 | 254 | 1 | 0 |
| Three | | | | | | | | |
| G + O + D | 221.77 | 5.95 | 221.5 | 7.0 | 236 | 209 | 0 | 28.48 |
| G + D + R | 222.03 | 12.89 | 220.5 | 17.0 | 251 | 204 | 1 | 26.62 |
| G + O + R | 224.77 | 11.15 | 223.0 | 9.0 | 262 | 206 | 0 | 24.36 |
| O + D + R | 283.00 | 11.83 | 282.5 | 20.0 | 305 | 261 | 0 | 0 |
| Four | | | | | | | | |
| G + O + D + R | 224.40 | 9.05 | 227.0 | 14.0 | 238 | 206 | 0 | 25.24 |

### Cart Pole Quadruple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 233.60 | 15.50 | 231.0 | 21.0 | 275 | 211 | 13 | 14.27 |
| One | | | | | | | | |
| G | 223.77 | 20.00 | 219.0 | 17.0 | 299 | 194 | 13 | 22.60 |
| D | 234.53 | 12.92 | 232.0 | 11.0 | 276 | 213 | 13 | 14.11 |
| R | 234.73 | 16.79 | 231.5 | 18.0 | 282 | 213 | 7 | 12.72 |
| O | 275.10 | 13.73 | 274.0 | 17.0 | 323 | 256 | 0 | 0 |
| Two | | | | | | | | |
| G + O | 217.23 | 7.31 | 216.0 | 12.0 | 234 | 207 | 0 | 32.78 |
| G + D | 218.53 | 13.39 | 215.5 | 15.0 | 262 | 202 | 9 | 29.98 |
| G + R | 220.30 | 21.51 | 212.0 | 15.0 | 295 | 202 | 12 | 25.53 |
| D + R | 242.60 | 24.30 | 235.5 | 39.0 | 297 | 217 | 18 | 2.21 |
| O + D | 274.17 | 15.20 | 271.0 | 13.0 | 329 | 254 | 0 | 0 |
| O + R | 279.23 | 29.03 | 274.0 | 18.0 | 410 | 244 | 0 | 0 |
| Three | | | | | | | | |
| G + O + D | 218.20 | 6.28 | 217.5 | 7.0 | 233 | 204 | 0 | 31.99 |
| G + O + R | 219.23 | 8.09 | 219.0 | 9.0 | 238 | 205 | 0 | 30.62 |
| G + D + R | 226.57 | 23.56 | 220.5 | 37.0 | 301 | 197 | 15 | 18.51 |
| O + D + R | 279.07 | 17.36 | 275.0 | 20.0 | 316 | 249 | 0 | 0 |
| Four | | | | | | | | |
| G + O + D + R | 219.03 | 10.13 | 218.0 | 12.0 | 249 | 203 | 0 | 30.35 |

### Cart Pole Quintuple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | - | - | - | - | - | - | - | - |
| One | | | | | | | | |
| O | 263.10 | 14.37 | 260.5 | 12.0 | 298 | 239 | 0 | 0 |
| G | - | - | - | - | - | - | - | - |
| D | - | - | - | - | - | - | - | - |
| R | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| G + O | 215.53 | 7.88 | 214.0 | 11.0 | 231 | 199 | 0 | 34.36 |
| O + D | 267.60 | 21.01 | 263.5 | 25.0 | 330 | 237 | 0 | 0 |
| O + R | 268.70 | 16.30 | 267.0 | 19.0 | 305 | 248 | 0 | 0 |
| G + D | - | - | - | - | - | - | - | - |
| G + R | - | - | - | - | - | - | - | - |
| D + R | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| G + O + R | 214.07 | 9.53 | 214.5 | 12.0 | 231 | 196 | 0 | 35.46 |
| G + O + D | 219.63 | 7.65 | 218.5 | 10.0 | 239 | 205 | 1 | 30.31 |
| O + D + R | 273.13 | 24.76 | 267.5 | 25.0 | 379 | 245 | 0 | 0 |
| G + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| G + O + D + R | 217.07 | 6.33 | 216.5 | 9.0 | 231 | 208 | 0 | 33.12 |

## Acrobot Innovation Results

### Acrobot Symbols

|  |  |  |  |
| --- | --- | --- | --- |
| Symbol | Name | Excluded | Included |
| A | Action Limiting | 3 | 2 |
| O | Opposition Learning | No | Yes |
| R | Random Initialization | Zeros | Normal |
| D | Discount Factor | 0.999 | (0.99, 100) |

### Acrobot Single Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 293.87 | 36.64 | 286.5 | 43.0 | 414 | 252 | 0 | 0 |
| One | | | | | | | | |
| A | 268.47 | 29.96 | 260.0 | 23.0 | 359 | 237 | 0 | 4.73 |
| R | 286.70 | 30.39 | 278.0 | 41.0 | 360 | 244 | 0 | 0 |
| D | 297.63 | 46.78 | 274.5 | 72.0 | 443 | 242 | 0 | 0 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| A + D | 266.87 | 29.70 | 259.5 | 20.0 | 384 | 235 | 0 | 6.40 |
| A + R | 270.67 | 32.62 | 259.0 | 32.0 | 369 | 236 | 0 | 1.77 |
| D + R | 303.80 | 36.48 | 300.0 | 39.0 | 398 | 255 | 0 | 0 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| A + O | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| A + D + R | 274.57 | 41.31 | 263.5 | 29.0 | 443 | 243 | 0 | 0 |
| A + O + D | - | - | - | - | - | - | - | - |
| A + O + R | - | - | - | - | - | - | - | - |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| A + O + D + R | - | - | - | - | - | - | - | - |

### Acrobot Double Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 258.03 | 16.76 | 259.0 | 21.0 | 302 | 233 | 2 | 18.24 |
| One | | | | | | | | |
| A | 243.83 | 17.35 | 245.0 | 26.0 | 293 | 214 | 1 | 32.33 |
| D | 274.63 | 54.31 | 257.0 | 40.0 | 478 | 219 | 0 | 0 |
| R | 283.13 | 56.42 | 268.0 | 43.0 | 481 | 225 | 2 | 0 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| A + D | 250.43 | 22.90 | 245.5 | 20.0 | 322 | 219 | 0 | 24.57 |
| A + R | 263.13 | 49.91 | 244.5 | 32.0 | 430 | 224 | 1 | 3.70 |
| D + R | 277.57 | 41.34 | 263.5 | 38.0 | 395 | 224 | 2 | 0 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| A + O | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| A + D + R | 270.80 | 67.56 | 247.0 | 42.0 | 500 | 225 | 0 | 0 |
| A + O + D | - | - | - | - | - | - | - | - |
| A + O + R | - | - | - | - | - | - | - | - |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| A + O + D + R | - | - | - | - | - | - | - | - |

### Acrobot Triple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 286.97 | 67.94 | 262.0 | 68.0 | 496 | 226 | 3 | 0 |
| One | | | | | | | | |
| R | 263.93 | 37.28 | 251.5 | 37.0 | 414 | 227 | 2 | 7.11 |
| A | 257.60 | 66.80 | 238.5 | 26.0 | 466 | 210 | 2 | 2.99 |
| D | 278.47 | 48.87 | 263.0 | 45.0 | 428 | 227 | 2 | 0 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| A + D | 245.97 | 39.84 | 234.0 | 22.0 | 417 | 213 | 1 | 24.26 |
| A + R | 244.37 | 50.33 | 232.0 | 19.0 | 500 | 217 | 0 | 22.31 |
| D + R | 270.83 | 49.80 | 258.0 | 36.0 | 479 | 229 | 1 | 0 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| A + O | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| A + D + R | 236.90 | 17.40 | 234.5 | 17.0 | 291 | 216 | 1 | 39.25 |
| A + O + D | - | - | - | - | - | - | - | - |
| A + O + R | - | - | - | - | - | - | - | - |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| A + O + D + R | - | - | - | - | - | - | - | - |

### Acrobot Quadruple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 263.23 | 51.20 | 250.0 | 53.0 | 491 | 216 | 4 | 3.14 |
| One | | | | | | | | |
| A | 233.00 | 38.63 | 226.0 | 18.0 | 428 | 207 | 0 | 37.61 |
| R | 247.20 | 25.84 | 240.5 | 32.0 | 342 | 212 | 2 | 27.09 |
| D | 266.73 | 34.05 | 264.0 | 62.0 | 353 | 222 | 1 | 5.29 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| A + D | 232.97 | 23.27 | 223.5 | 23.0 | 308 | 209 | 1 | 41.94 |
| A + R | 235.00 | 32.52 | 225.5 | 34.0 | 358 | 200 | 0 | 37.47 |
| D + R | 260.53 | 28.06 | 258.5 | 38.0 | 318 | 221 | 2 | 13.18 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| A + O | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| A + D + R | 241.37 | 41.27 | 232.0 | 23.0 | 423 | 210 | 0 | 28.39 |
| A + O + R | - | - | - | - | - | - | - | - |
| A + O + D | - | - | - | - | - | - | - | - |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| A + O + D + R | - | - | - | - | - | - | - | - |

### Acrobot Quintuple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| None | | | | | | | | |
| - | 251.47 | 34.26 | 243.5 | 40.0 | 343 | 213 | 4 | 20.49 |
| One | | | | | | | | |
| A | 222.37 | 10.50 | 221.5 | 12.0 | 261 | 208 | 0 | 54.85 |
| R | 250.70 | 26.15 | 242.5 | 41.0 | 318 | 221 | 2 | 23.51 |
| D | 258.30 | 43.75 | 255.0 | 35.0 | 450 | 212 | 0 | 10.64 |
| O | - | - | - | - | - | - | - | - |
| Two | | | | | | | | |
| A + D | 228.77 | 19.21 | 225.5 | 18.0 | 285 | 202 | 1 | 47.03 |
| A + R | 233.30 | 52.16 | 225.5 | 21.0 | 495 | 202 | 1 | 32.73 |
| D + R | 249.80 | 25.25 | 249.5 | 35.0 | 299 | 211 | 1 | 24.63 |
| O + D | - | - | - | - | - | - | - | - |
| O + R | - | - | - | - | - | - | - | - |
| A + O | - | - | - | - | - | - | - | - |
| Three | | | | | | | | |
| A + D + R | 222.67 | 10.28 | 223.0 | 14.0 | 250 | 206 | 0 | 54.58 |
| A + O + R | - | - | - | - | - | - | - | - |
| A + O + D | - | - | - | - | - | - | - | - |
| O + D + R | - | - | - | - | - | - | - | - |
| Four | | | | | | | | |
| A + O + D + R | - | - | - | - | - | - | - | - |

## Cart Pole Reward Functions Results

### Cart Pole Single Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Termination | 256.90 | 17.24 | 252.5 | 17.0 | 309 | 233 | 1 | 0 |
| Original | 257.27 | 14.94 | 256.0 | 25.0 | 288 | 235 | 0 | 0 |
| Logarithmic | 257.70 | 13.68 | 254.0 | 22.0 | 288 | 228 | 2 | 0 |
| Uniform | 258.93 | 14.02 | 262.5 | 20.0 | 287 | 231 | 0 | 0 |
| Time | 263.07 | 14.03 | 259.5 | 16.0 | 293 | 235 | 0 | 0 |
| Exponential | 265.43 | 18.49 | 265.5 | 28.0 | 310 | 230 | 1 | 0 |
| Granularity | | | | | | | | |
| Termination | 231.27 | 14.10 | 229.0 | 12.0 | 268 | 213 | 0 | 17.03 |
| Time | 233.80 | 13.44 | 233.0 | 18.0 | 278 | 214 | 0 | 14.69 |
| Exponential | 234.50 | 12.40 | 235.0 | 15.0 | 263 | 212 | 0 | 14.29 |
| Logarithmic | 233.23 | 17.75 | 230.5 | 33.0 | 267 | 210 | 0 | 13.90 |
| Original | 237.03 | 12.74 | 236.0 | 17.0 | 269 | 217 | 0 | 11.66 |
| Uniform | 237.47 | 17.54 | 237.0 | 17.0 | 312 | 215 | 0 | 9.74 |
| Granularity + Random | | | | | | | | |
| Exponential | 229.00 | 12.64 | 227.0 | 17.0 | 269 | 212 | 0 | 19.72 |
| Original | 231.10 | 9.32 | 231.5 | 12.0 | 250 | 216 | 0 | 18.48 |
| Termination | 230.70 | 11.56 | 227.0 | 17.0 | 257 | 212 | 0 | 18.32 |
| Time | 232.97 | 11.81 | 235.0 | 19.0 | 259 | 215 | 0 | 15.98 |
| Uniform | 233.43 | 14.10 | 232.0 | 15.0 | 269 | 209 | 0 | 14.86 |
| Logarithmic | 236.27 | 16.21 | 232.0 | 18.0 | 282 | 213 | 0 | 11.38 |
| Granularity + Discount + Random | | | | | | | | |
| Uniform | 231.83 | 13.58 | 230.5 | 17.0 | 262 | 203 | 0 | 16.62 |
| Termination | 232.20 | 12.31 | 231.0 | 23.0 | 253 | 211 | 0 | 16.61 |
| Original | 234.57 | 13.34 | 232.0 | 17.0 | 268 | 216 | 0 | 13.95 |
| Exponential | 235.13 | 12.39 | 232.0 | 20.0 | 258 | 214 | 0 | 13.66 |
| Logarithmic | 236.13 | 16.68 | 232.5 | 22.0 | 279 | 211 | 0 | 11.36 |
| Time | 238.70 | 13.00 | 234.5 | 18.0 | 274 | 225 | 0 | 9.92 |
| Granularity + Opposition + Discount + Random | | | | | | | | |
| Exponential | 237.30 | 7.58 | 238.5 | 11.0 | 253 | 224 | 0 | 12.65 |
| Logarithmic | 238.00 | 7.71 | 237.5 | 12.0 | 253 | 220 | 0 | 11.93 |
| Uniform | 239.53 | 9.50 | 237.5 | 10.0 | 265 | 221 | 0 | 10.00 |
| Termination | 246.00 | 9.43 | 245.0 | 15.0 | 262 | 231 | 0 | 3.55 |
| Original | 247.80 | 7.06 | 247.0 | 12.0 | 261 | 237 | 0 | 2.25 |
| Time | 256.67 | 10.50 | 254.5 | 15.0 | 278 | 241 | 3 | 0 |

### Cart Pole Double Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Exponential | 249.73 | 14.71 | 247.5 | 20.0 | 285 | 223 | 0 | 0 |
| Logarithmic | 251.20 | 19.17 | 247.5 | 22.0 | 300 | 222 | 0 | 0 |
| Termination | 252.10 | 20.25 | 247.0 | 29.0 | 319 | 222 | 3 | 0 |
| Time | 252.17 | 15.92 | 248.0 | 16.0 | 292 | 225 | 5 | 0 |
| Original | 254.90 | 19.99 | 252.5 | 29.0 | 303 | 223 | 1 | 0 |
| Uniform | 260.43 | 26.31 | 254.0 | 34.0 | 325 | 225 | 0 | 0 |
| Granularity | | | | | | | | |
| Uniform | 223.83 | 13.51 | 224.0 | 17.0 | 255 | 203 | 0 | 24.64 |
| Exponential | 224.60 | 14.63 | 222.5 | 18.0 | 258 | 208 | 0 | 23.54 |
| Termination | 230.60 | 13.49 | 228.5 | 16.0 | 260 | 203 | 2 | 17.88 |
| Original | 230.03 | 16.98 | 224.5 | 23.0 | 275 | 206 | 1 | 17.36 |
| Time | 231.33 | 13.95 | 229.0 | 20.0 | 254 | 202 | 0 | 17.01 |
| Logarithmic | 233.23 | 26.62 | 225.0 | 24.0 | 308 | 208 | 0 | 10.70 |
| Granularity + Discount | | | | | | | | |
| Exponential | 225.40 | 19.71 | 219.5 | 20.0 | 276 | 201 | 0 | 21.07 |
| Uniform | 226.73 | 18.56 | 220.0 | 24.0 | 277 | 202 | 0 | 20.13 |
| Logarithmic | 227.87 | 17.58 | 223.5 | 19.0 | 265 | 204 | 0 | 19.33 |
| Time | 227.97 | 19.94 | 221.5 | 29.0 | 297 | 203 | 0 | 18.42 |
| Original | 229.50 | 19.48 | 223.5 | 16.0 | 283 | 207 | 0 | 17.05 |
| Termination | 231.87 | 24.47 | 223.0 | 21.0 | 322 | 208 | 0 | 12.88 |
| Granularity + Discount + Random | | | | | | | | |
| Uniform | 226.50 | 14.31 | 225.0 | 19.0 | 276 | 208 | 0 | 21.73 |
| Time | 231.27 | 14.83 | 227.0 | 22.0 | 265 | 206 | 1 | 16.81 |
| Logarithmic | 230.50 | 18.93 | 231.0 | 27.0 | 292 | 202 | 0 | 16.24 |
| Original | 232.57 | 13.84 | 230.5 | 19.0 | 265 | 205 | 0 | 15.80 |
| Exponential | 232.43 | 28.16 | 225.0 | 19.0 | 335 | 203 | 0 | 10.91 |
| Termination | 235.07 | 28.15 | 225.0 | 27.0 | 351 | 206 | 1 | 8.28 |
| Granularity + Opposition + Discount + Random | | | | | | | | |
| Termination | 236.93 | 9.95 | 237.0 | 15.0 | 257 | 220 | 8 | 12.50 |
| Uniform | 238.53 | 14.58 | 235.0 | 21.0 | 274 | 215 | 0 | 9.62 |
| Exponential | 239.40 | 11.92 | 238.0 | 22.0 | 262 | 223 | 0 | 9.52 |
| Logarithmic | 238.67 | 17.55 | 235.5 | 20.0 | 298 | 211 | 0 | 8.54 |
| Original | 245.97 | 13.43 | 244.5 | 18.0 | 285 | 228 | 0 | 2.53 |
| Time | 253.47 | 17.51 | 249.5 | 22.0 | 293 | 224 | 0 | 0 |

### Cart Pole Triple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Logarithmic | 236.20 | 14.04 | 237.0 | 24.0 | 260 | 211 | 0 | 12.12 |
| Termination | 236.93 | 16.85 | 235.0 | 19.0 | 304 | 214 | 0 | 10.50 |
| Exponential | 237.63 | 15.66 | 233.5 | 22.0 | 269 | 215 | 0 | 10.18 |
| Original | 242.50 | 18.16 | 237.5 | 15.0 | 304 | 220 | 3 | 4.50 |
| Time | 244.13 | 16.85 | 243.0 | 29.0 | 279 | 215 | 0 | 3.30 |
| Uniform | 248.03 | 21.34 | 244.5 | 27.0 | 301 | 215 | 0 | 0 |
| Granularity | | | | | | | | |
| Exponential | 213.47 | 7.91 | 213.0 | 10.0 | 228 | 196 | 0 | 36.42 |
| Logarithmic | 214.73 | 11.86 | 213.0 | 16.0 | 241 | 199 | 0 | 34.21 |
| Uniform | 215.50 | 10.78 | 214.5 | 11.0 | 243 | 199 | 0 | 33.72 |
| Original | 217.13 | 10.23 | 215.5 | 10.0 | 248 | 205 | 0 | 32.23 |
| Termination | 220.77 | 12.16 | 220.5 | 17.0 | 239 | 196 | 0 | 28.09 |
| Time | 221.17 | 13.28 | 221.0 | 17.0 | 263 | 203 | 3 | 27.37 |
| Granularity + Random | | | | | | | | |
| Exponential | 217.27 | 11.30 | 215.5 | 15.0 | 242 | 200 | 0 | 31.82 |
| Logarithmic | 217.77 | 11.05 | 215.0 | 15.0 | 239 | 198 | 0 | 31.39 |
| Original | 218.13 | 14.51 | 215.0 | 16.0 | 255 | 199 | 1 | 30.04 |
| Termination | 219.53 | 10.96 | 218.0 | 13.0 | 245 | 189 | 0 | 29.65 |
| Uniform | 219.87 | 10.34 | 220.0 | 15.0 | 238 | 202 | 0 | 29.47 |
| Time | 223.07 | 11.47 | 219.5 | 13.0 | 248 | 204 | 1 | 25.97 |
| Granularity + Opposition + Discount | | | | | | | | |
| Logarithmic | 215.00 | 12.00 | 214.5 | 12.0 | 257 | 198 | 0 | 33.90 |
| Exponential | 216.93 | 8.73 | 216.5 | 13.0 | 237 | 204 | 0 | 32.78 |
| Uniform | 216.90 | 10.52 | 217.0 | 13.0 | 242 | 200 | 0 | 32.39 |
| Termination | 219.20 | 8.04 | 219.0 | 11.0 | 237 | 207 | 0 | 30.66 |
| Original | 221.77 | 5.95 | 221.5 | 7.0 | 236 | 209 | 0 | 28.48 |
| Time | 225.87 | 10.36 | 227.0 | 17.0 | 250 | 205 | 0 | 23.46 |
| Granularity + Opposition + Discount + Random | | | | | | | | |
| Exponential | 214.93 | 10.32 | 212.0 | 15.0 | 234 | 198 | 0 | 34.40 |
| Logarithmic | 216.37 | 8.74 | 215.0 | 13.0 | 235 | 201 | 0 | 33.34 |
| Termination | 217.23 | 6.75 | 217.0 | 11.0 | 234 | 204 | 0 | 32.88 |
| Uniform | 217.50 | 9.61 | 217.0 | 13.0 | 237 | 195 | 0 | 32.01 |
| Original | 224.40 | 9.05 | 227.0 | 14.0 | 238 | 206 | 0 | 25.24 |
| Time | 227.70 | 12.24 | 226.5 | 12.0 | 264 | 210 | 0 | 21.13 |

### Cart Pole Quadruple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Termination | 226.73 | 11.93 | 224.5 | 20.0 | 253 | 208 | 0 | 22.19 |
| Logarithmic | 227.60 | 11.25 | 227.5 | 15.0 | 264 | 211 | 0 | 21.50 |
| Uniform | 230.63 | 12.45 | 228.5 | 16.0 | 268 | 214 | 0 | 18.14 |
| Exponential | 233.43 | 14.66 | 231.5 | 21.0 | 268 | 203 | 0 | 14.69 |
| Original | 233.60 | 15.50 | 231.0 | 21.0 | 275 | 211 | 13 | 14.27 |
| Time | 236.97 | 16.11 | 233.5 | 18.0 | 295 | 217 | 16 | 10.70 |
| Granularity | | | | | | | | |
| Uniform | 210.80 | 10.12 | 207.0 | 11.0 | 237 | 197 | 0 | 38.59 |
| Exponential | 212.53 | 8.31 | 212.5 | 11.0 | 235 | 198 | 0 | 37.27 |
| Logarithmic | 212.13 | 10.26 | 212.0 | 12.0 | 231 | 195 | 0 | 37.22 |
| Termination | 212.67 | 9.85 | 211.5 | 15.0 | 234 | 199 | 1 | 36.78 |
| Original | 223.77 | 20.00 | 219.0 | 17.0 | 299 | 194 | 13 | 22.60 |
| Time | - | - | - | - | - | - | - | - |
| Granularity + Opposition | | | | | | | | |
| Uniform | 209.73 | 6.67 | 209.0 | 8.0 | 228 | 196 | 0 | 40.39 |
| Exponential | 210.73 | 6.95 | 212.0 | 10.0 | 223 | 196 | 0 | 39.34 |
| Logarithmic | 211.50 | 9.35 | 208.0 | 13.0 | 238 | 200 | 0 | 38.07 |
| Termination | 215.23 | 8.72 | 213.0 | 8.0 | 241 | 201 | 0 | 34.48 |
| Original | 217.23 | 7.31 | 216.0 | 12.0 | 234 | 207 | 0 | 32.78 |
| Time | 220.83 | 7.13 | 221.5 | 8.0 | 237 | 205 | 0 | 29.21 |
| Granularity + Opposition + Discount | | | | | | | | |
| Uniform | 208.97 | 5.84 | 209.0 | 9.0 | 220 | 199 | 0 | 41.30 |
| Logarithmic | 210.20 | 9.50 | 208.5 | 12.0 | 236 | 197 | 0 | 39.34 |
| Exponential | 211.47 | 7.07 | 210.0 | 9.0 | 233 | 199 | 0 | 38.58 |
| Termination | 214.03 | 7.04 | 213.5 | 8.0 | 231 | 200 | 0 | 36.03 |
| Original | 218.20 | 6.28 | 217.5 | 7.0 | 233 | 204 | 0 | 31.99 |
| Time | 222.67 | 11.39 | 221.0 | 13.0 | 258 | 204 | 0 | 26.40 |
| Granularity + Opposition + Discount + Random | | | | | | | | |
| Logarithmic | 209.70 | 8.44 | 207.5 | 9.0 | 232 | 195 | 0 | 40.08 |
| Uniform | 210.23 | 6.93 | 210.0 | 10.0 | 225 | 201 | 0 | 39.84 |
| Exponential | 212.37 | 10.20 | 212.5 | 7.0 | 239 | 190 | 0 | 37.00 |
| Termination | 216.37 | 9.82 | 216.0 | 13.0 | 247 | 200 | 0 | 33.09 |
| Original | 219.03 | 10.13 | 218.0 | 12.0 | 249 | 203 | 0 | 30.36 |
| Time | 224.77 | 13.01 | 222.5 | 21.0 | 266 | 207 | 0 | 23.85 |

### Cart Pole Quintuple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Termination | 219.53 | 10.37 | 217.0 | 14.0 | 242 | 205 | 0 | 29.80 |
| Uniform | 223.60 | 9.16 | 223.5 | 9.0 | 243 | 198 | 0 | 26.01 |
| Exponential | 228.37 | 13.07 | 227.0 | 19.0 | 252 | 202 | 0 | 20.23 |
| Logarithmic | 230.67 | 12.64 | 229.5 | 19.0 | 252 | 206 | 0 | 18.06 |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Opposition | | | | | | | | |
| Logarithmic | 241.17 | 10.73 | 238.5 | 16.0 | 263 | 224 | 0 | 8.07 |
| Exponential | 242.00 | 13.16 | 242.0 | 14.0 | 279 | 222 | 0 | 6.57 |
| Uniform | 242.83 | 11.81 | 242.0 | 14.0 | 271 | 225 | 0 | 6.12 |
| Termination | 252.13 | 10.63 | 255.0 | 14.0 | 271 | 228 | 0 | 0 |
| Original | 263.10 | 14.37 | 260.5 | 12.0 | 298 | 239 | 0 | 0 |
| Time | 272.50 | 18.85 | 269.0 | 23.0 | 319 | 243 | 0 | 0 |
| Granularity + Opposition | | | | | | | | |
| Exponential | 209.17 | 8.57 | 206.0 | 12.0 | 233 | 196 | 0 | 40.58 |
| Termination | 209.57 | 6.85 | 209.0 | 9.0 | 225 | 196 | 0 | 40.52 |
| Logarithmic | 209.50 | 7.69 | 208.5 | 8.0 | 230 | 197 | 0 | 40.43 |
| Uniform | 214.27 | 9.74 | 213.0 | 7.0 | 248 | 195 | 0 | 35.21 |
| Original | 215.53 | 7.88 | 214.0 | 11.0 | 231 | 199 | 0 | 34.36 |
| Time | 218.30 | 7.76 | 216.0 | 10.0 | 235 | 206 | 0 | 31.62 |
| Granularity + Opposition + Random | | | | | | | | |
| Exponential | 209.57 | 6.85 | 210.5 | 9.0 | 223 | 194 | 0 | 40.52 |
| Logarithmic | 210.57 | 8.13 | 210.5 | 11.0 | 227 | 199 | 0 | 39.27 |
| Termination | 211.50 | 6.19 | 211.5 | 8.0 | 226 | 202 | 0 | 38.71 |
| Uniform | 211.27 | 9.10 | 212.0 | 13.0 | 228 | 190 | 0 | 38.36 |
| Original | 214.07 | 9.53 | 214.5 | 12.0 | 231 | 196 | 0 | 35.46 |
| Time | 216.27 | 7.94 | 214.5 | 8.0 | 236 | 201 | 3 | 33.61 |
| Granularity + Opposition + Discount + Random | | | | | | | | |
| Uniform | 209.20 | 8.24 | 209.5 | 13.0 | 232 | 198 | 0 | 40.62 |
| Exponential | 210.53 | 8.02 | 208.0 | 8.0 | 234 | 200 | 0 | 39.33 |
| Termination | 212.23 | 8.39 | 213.5 | 17.0 | 224 | 199 | 0 | 37.56 |
| Logarithmic | 213.73 | 8.42 | 214.0 | 8.0 | 240 | 199 | 0 | 36.05 |
| Original | 217.07 | 6.33 | 216.5 | 9.0 | 231 | 208 | 0 | 33.11 |
| Time | 217.13 | 12.37 | 215.0 | 10.0 | 270 | 201 | 2 | 31.67 |

## Acrobot Reward Functions Results

### Acrobot Single Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Time | 255.87 | 26.25 | 247.0 | 20.0 | 325 | 228 | 0 | 18.32 |
| Velocity | 288.23 | 36.60 | 275.5 | 39.0 | 389 | 247 | 1 | 0 |
| Original | 293.87 | 36.64 | 286.5 | 43.0 | 414 | 252 | 0 | 0 |
| Height | 296.70 | 34.59 | 292.0 | 43.0 | 398 | 245 | 1 | 0 |
| Action | | | | | | | | |
| Time | 243.80 | 15.12 | 242.5 | 15.0 | 287 | 224 | 0 | 32.76 |
| Height | 259.43 | 20.08 | 254.0 | 22.0 | 314 | 233 | 0 | 16.19 |
| Velocity | 267.37 | 31.17 | 257.5 | 34.0 | 394 | 235 | 0 | 5.49 |
| Original | 268.47 | 29.96 | 260.0 | 23.0 | 359 | 237 | 0 | 4.73 |
| Action + Discount | | | | | | | | |
| Time | 238.93 | 14.74 | 236.0 | 10.0 | 284 | 220 | 0 | 37.69 |
| Height | 259.03 | 11.90 | 257.0 | 14.0 | 289 | 236 | 0 | 18.01 |
| Velocity | 265.43 | 24.13 | 260.0 | 14.0 | 376 | 246 | 0 | 9.27 |
| Original | 266.87 | 29.70 | 259.5 | 20.0 | 384 | 235 | 0 | 6.40 |
| Action + Discount + Random | | | | | | | | |
| Time | 239.60 | 15.89 | 236.5 | 15.0 | 308 | 222 | 0 | 36.83 |
| Height | 259.43 | 18.22 | 253.0 | 21.0 | 326 | 240 | 0 | 16.56 |
| Velocity | 261.23 | 21.41 | 255.0 | 12.0 | 328 | 239 | 0 | 14.10 |
| Original | 274.57 | 41.31 | 263.5 | 29.0 | 443 | 243 | 0 | 0 |
| Action + Opposition + Discount + Random | | | | | | | | |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Velocity | - | - | - | - | - | - | - | - |
| Height | - | - | - | - | - | - | - | - |

### Acrobot Double Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Original | 258.03 | 16.76 | 259.0 | 21.0 | 302 | 233 | 2 | 18.24 |
| Time | 262.73 | 43.64 | 250.0 | 29.0 | 438 | 233 | 3 | 6.25 |
| Height | 279.60 | 43.51 | 271.0 | 55.0 | 416 | 229 | 5 | 0 |
| Velocity | 289.33 | 50.32 | 278.0 | 39.0 | 477 | 223 | 1 | 0 |
| Action | | | | | | | | |
| Original | 243.83 | 17.35 | 245.0 | 26.0 | 293 | 214 | 1 | 32.33 |
| Time | 248.10 | 29.86 | 240.5 | 22.0 | 332 | 212 | 0 | 25.12 |
| Velocity | 256.23 | 41.28 | 244.5 | 16.0 | 436 | 222 | 1 | 13.53 |
| Height | 262.07 | 58.30 | 243.5 | 20.0 | 470 | 215 | 2 | 1.73 |
| Action + Discount | | | | | | | | |
| Time | 239.40 | 37.74 | 231.5 | 15.0 | 425 | 211 | 1 | 31.49 |
| Original | 250.43 | 22.90 | 245.5 | 20.0 | 322 | 219 | 0 | 24.56 |
| Height | 262.93 | 37.88 | 251.5 | 33.0 | 370 | 229 | 1 | 7.92 |
| Velocity | 261.37 | 45.48 | 248.0 | 25.0 | 460 | 230 | 2 | 6.99 |
| Action + Discount + Random | | | | | | | | |
| Time | 235.53 | 16.78 | 233.0 | 22.0 | 284 | 210 | 1 | 40.73 |
| Velocity | 245.43 | 21.21 | 240.5 | 32.0 | 305 | 214 | 1 | 29.95 |
| Height | 256.90 | 41.07 | 244.0 | 23.0 | 438 | 224 | 1 | 12.93 |
| Original | 270.80 | 67.56 | 247.0 | 42.0 | 500 | 225 | 0 | 0 |
| Action + Opposition + Discount + Random | | | | | | | | |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Velocity | - | - | - | - | - | - | - | - |
| Height | - | - | - | - | - | - | - | - |

### Acrobot Triple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Time | 251.30 | 49.85 | 236.5 | 32.0 | 466 | 211 | 4 | 15.55 |
| Height | 263.90 | 40.72 | 253.5 | 33.0 | 419 | 212 | 5 | 6.04 |
| Velocity | 267.03 | 55.20 | 254.0 | 28.0 | 468 | 222 | 3 | 0 |
| Original | 286.97 | 67.94 | 262.0 | 68.0 | 496 | 226 | 3 | 0 |
| Random | | | | | | | | |
| Time | 244.47 | 26.09 | 240.5 | 41.0 | 300 | 207 | 6 | 29.76 |
| Height | 262.27 | 31.48 | 250.0 | 57.0 | 344 | 225 | 1 | 10.50 |
| Original | 263.93 | 37.28 | 251.5 | 37.0 | 414 | 227 | 2 | 7.11 |
| Velocity | 268.97 | 38.63 | 264.0 | 33.0 | 413 | 222 | 3 | 1.65 |
| Action + Discount | | | | | | | | |
| Time | 227.67 | 18.21 | 224.0 | 17.0 | 278 | 199 | 0 | 48.33 |
| Height | 229.13 | 14.57 | 230.0 | 18.0 | 281 | 209 | 0 | 47.51 |
| Original | 245.97 | 39.84 | 234.0 | 22.0 | 417 | 213 | 1 | 24.26 |
| Velocity | 246.47 | 42.24 | 234.5 | 21.0 | 444 | 217 | 0 | 22.98 |
| Action + Discount + Random | | | | | | | | |
| Time | 224.60 | 19.44 | 223.5 | 21.0 | 289 | 202 | 2 | 51.15 |
| Original | 236.90 | 17.40 | 234.5 | 17.0 | 291 | 216 | 1 | 39.25 |
| Velocity | 237.43 | 16.23 | 236.0 | 19.0 | 271 | 211 | 0 | 38.93 |
| Height | 247.03 | 42.49 | 233.0 | 24.0 | 418 | 210 | 2 | 22.33 |
| Action + Opposition + Discount + Random | | | | | | | | |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Velocity | - | - | - | - | - | - | - | - |
| Height | - | - | - | - | - | - | - | - |

### Acrobot Quadruple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Time | 228.57 | 15.68 | 224.0 | 26.0 | 261 | 201 | 1 | 47.90 |
| Height | 251.10 | 30.52 | 241.5 | 35.0 | 328 | 218 | 0 | 21.94 |
| Velocity | 255.90 | 55.98 | 238.0 | 37.0 | 481 | 215 | 1 | 8.75 |
| Original | 263.23 | 51.20 | 250.0 | 53.0 | 491 | 216 | 4 | 3.14 |
| Action | | | | | | | | |
| Velocity | 227.07 | 18.42 | 222.5 | 23.0 | 274 | 204 | 1 | 48.89 |
| Original | 233.00 | 38.63 | 226.0 | 18.0 | 428 | 207 | 0 | 37.61 |
| Height | 235.13 | 40.59 | 229.0 | 27.0 | 426 | 200 | 1 | 34.85 |
| Time | 230.27 | 65.31 | 210.0 | 29.0 | 490 | 191 | 2 | 30.89 |
| Action + Discount | | | | | | | | |
| Time | 221.03 | 40.47 | 211.5 | 16.0 | 415 | 194 | 1 | 48.99 |
| Velocity | 233.57 | 16.98 | 230.5 | 26.0 | 267 | 209 | 0 | 42.66 |
| Height | 228.13 | 39.81 | 220.5 | 12.0 | 433 | 209 | 0 | 42.11 |
| Original | 232.97 | 23.27 | 223.5 | 23.0 | 308 | 209 | 1 | 41.94 |
| Action + Discount + Random | | | | | | | | |
| Time | 212.97 | 18.21 | 209.5 | 16.0 | 287 | 193 | 0 | 63.03 |
| Velocity | 232.57 | 23.78 | 226.5 | 22.0 | 321 | 206 | 0 | 42.22 |
| Height | 233.13 | 22.45 | 225.5 | 19.0 | 308 | 207 | 0 | 41.97 |
| Original | 241.37 | 41.27 | 232.0 | 23.0 | 423 | 210 | 0 | 28.39 |
| Action + Opposition + Discount + Random | | | | | | | | |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Velocity | - | - | - | - | - | - | - | - |
| Height | - | - | - | - | - | - | - | - |

### Acrobot Quintuple Q-Learning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | Mean | S.T.D | Median | I.Q.R | Max | Min | Failed | Difference |
| - | | | | | | | | |
| Time | 222.40 | 17.71 | 220.5 | 22.0 | 262 | 192 | 4 | 53.69 |
| Height | 244.67 | 25.30 | 238.5 | 27.0 | 346 | 212 | 0 | 29.76 |
| Original | 251.47 | 34.26 | 243.5 | 40.0 | 343 | 213 | 4 | 20.49 |
| Velocity | 256.50 | 39.75 | 247.0 | 46.0 | 374 | 210 | 1 | 13.76 |
| Action | | | | | | | | |
| Time | 208.70 | 16.84 | 204.0 | 10.0 | 274 | 190 | 1 | 67.56 |
| Height | 218.70 | 10.81 | 218.0 | 16.0 | 244 | 199 | 0 | 58.48 |
| Original | 222.37 | 10.50 | 221.5 | 12.0 | 261 | 208 | 0 | 54.85 |
| Velocity | 226.97 | 24.78 | 220.0 | 12.0 | 322 | 206 | 0 | 47.58 |
| Action + Discount | | | | | | | | |
| Time | 210.87 | 19.35 | 207.0 | 15.0 | 286 | 190 | 1 | 64.90 |
| Height | 217.73 | 16.88 | 212.5 | 15.0 | 277 | 198 | 0 | 58.52 |
| Original | 228.77 | 19.21 | 225.5 | 18.0 | 285 | 202 | 1 | 47.03 |
| Velocity | 231.10 | 38.19 | 221.0 | 17.0 | 418 | 205 | 0 | 39.65 |
| Action + Discount + Random | | | | | | | | |
| Time | 206.47 | 14.04 | 202.0 | 21.0 | 237 | 190 | 3 | 70.26 |
| Height | 219.67 | 13.76 | 218.5 | 21.0 | 246 | 193 | 0 | 57.11 |
| Original | 222.67 | 10.28 | 223.0 | 14.0 | 250 | 206 | 0 | 54.58 |
| Velocity | 229.77 | 16.91 | 228.5 | 25.0 | 262 | 203 | 0 | 46.48 |
| Action + Opposition + Discount + Random | | | | | | | | |
| Original | - | - | - | - | - | - | - | - |
| Time | - | - | - | - | - | - | - | - |
| Velocity | - | - | - | - | - | - | - | - |
| Height | - | - | - | - | - | - | - | - |