**COSC-4117EL: Assignment 2 Report**

**Group Number: 2**

**Group Member:**

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **STUDENT#** | **EMAIL** | **CONTRIBUTION** |
| **Haoliang Sheng** | 0441916 | hsheng@laurentian.ca | 60% |
| **Zihao Zhou** | 0429993 | zzhou3@laurentian.ca | 20% |
| **Jiazhou Ye** | 0426609 | Jye1@laurentian.ca | 20% |

1. **Abstract**

In this study, the objective was to compare and contrast the performances of two reinforcement learning methods: Markov Decision Processes (MDP) and Q-Learning, in navigating a robot within a GridWorld environment. We employed known parameters for MDP, such as the transition matrix T and reward function R, while Q-Learning operated without this explicit knowledge. Through extensive experimentation, we observed that MDP generally offers faster and more consistent convergence. However, Q-Learning, with its flexibility in parameters, still manages to produce policies of consistently good quality.

1. **Introduction**

The task of navigating a robot within an environment, while seemingly straightforward, poses intricate challenges when we aim for efficient and optimal paths. This assignment focuses on the GridWorld environment, a simple yet effective representation that challenges algorithms to find the best paths in a grid-like structure. The significance of this problem lies in its foundational role in robotics and AI navigation systems. By optimizing pathfinding in such controlled scenarios, we can build more complex and efficient real-world systems. Our group aimed to leverage two popular reinforcement learning methods, MDP and Q-Learning, to discern their efficacy in tackling this problem. Through systematic experimentation and analysis, we aimed to unearth insights into their convergence behaviors, policy quality, and overall performance.

1. **Methodology**

To ensure a comprehensive analysis, our experimentation followed a structured approach:

**Parameter Selection:**

* **Default Parameters:**
  + **Grid Constants:** The grid's size is defined by **`GRID\_SIZE`** (10x10 cells) with each cell being **`CELL\_SIZE`** (60 pixels) in width and height, resulting in a display size of **`SCREEN\_WIDTH x SCREEN\_HEIGHT`**.
  + **Reward/Penalty Constants:** In our GridWorld, collecting gold offers a `**GOLD\_REWARD`** of 10 points, while stepping into a trap incurs a `**TRAP\_PENALTY`** of -10 points. Reaching the goal grants a `**GOAL\_REWARD`** of 200 points, while every step taken costs a `**LIVING\_PENALTY`** of -1, incentivizing faster goal attainment.
  + **Colors:** Different entities in the GridWorld are represented with distinct colors. The robot is visualized with `**ROBOT\_COLOR`** (blue), the goal with GOAL\_COLOR (green), traps with `**TRAP\_COLOR`** (red), and so on.
  + **Actions:** The robot can perform actions defined in the `**ACTIONS`** list, allowing it to move up, down, left, or right.
  + **Algorithm Constants:** The discount factor for future rewards is set to `**GAMMA`** (0.9). The `**CONVERGE\_THRESHOLD`** (0.0001) determines the point of negligible change in the value function, indicating convergence.
* For both MDP and Q-Learning, multiple parameters were experimented with. This included varying the learning rate and exploration probability ϵ for Q-Learning.

**Randomized GridWorld Generation:** For each parameter combination, we generated 10 distinct GridWorld environments using different random seeds. This approach allowed us to evaluate the consistency and adaptability of each method across varied environments.

**Algorithm Execution:** Both MDP and Q-Learning algorithms were run on each of these GridWorlds, and the results were recorded.

**Results Analysis:** The outcomes from the algorithms were analyzed using the mean and standard deviation across the 10 different GridWorld scenarios for each parameter combination. This provided insights into the average performance and variability of the methods.

By following this methodology, combined with the default parameters, our experiments aimed to offer a detailed understanding of the performance nuances of MDP and Q-Learning in the GridWorld environment.

1. **Instructions for Using the Script**

**Script Execution:**

* Navigate to the directory containing the script.
* Execute the command: **`python COSC\_4117EL\_A2\_G2.py`**.

**Choosing the Method:**

Upon startup, choose between MDP and Q-Learning by entering 1 for MDP or 2 for Q-Learning.

**Inputting Parameters for Q-learning:**

If Q-learning is selected, you'll be prompted to:

* Input the learning rate , which defines how much new Q-value estimates overwrite previous ones.
* Specify the exploration probability ϵ, determining the chance of the agent selecting a random action over following the current policy.

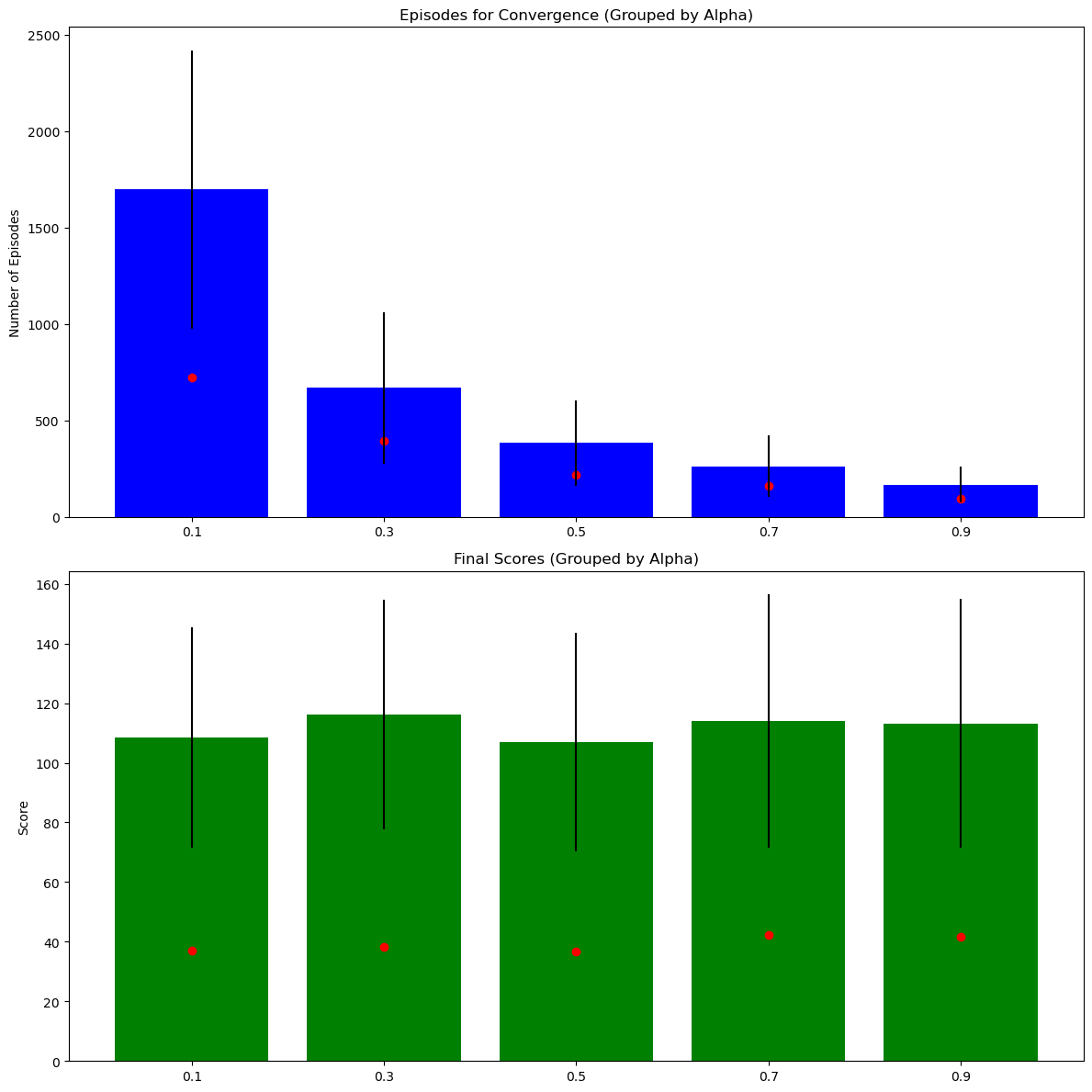
**Viewing the Results:**

The Pygame window will showcase the GridWorld environment, visualizing the robot's journey based on the acquired policy. Interactions and the concluding goal-reaching action will be displayed, culminating with the achieved score.

**Understanding Convergence with the `evaluate\_policy` Function:**

The script incorporates the **`evaluate\_policy`** function to assess the robot's efficacy. If the robot takes more steps than the grid cells total (greater than **`size \* size`** steps), it might indicate non-convergence. In such cases, a message stating "The method can't converge within time" will appear, implying that the robot may be stuck below the convergence threshold (set at 0.0001) or the chosen method/parameters might not be ideal for the GridWorld configuration.

1. **Experimental Results**



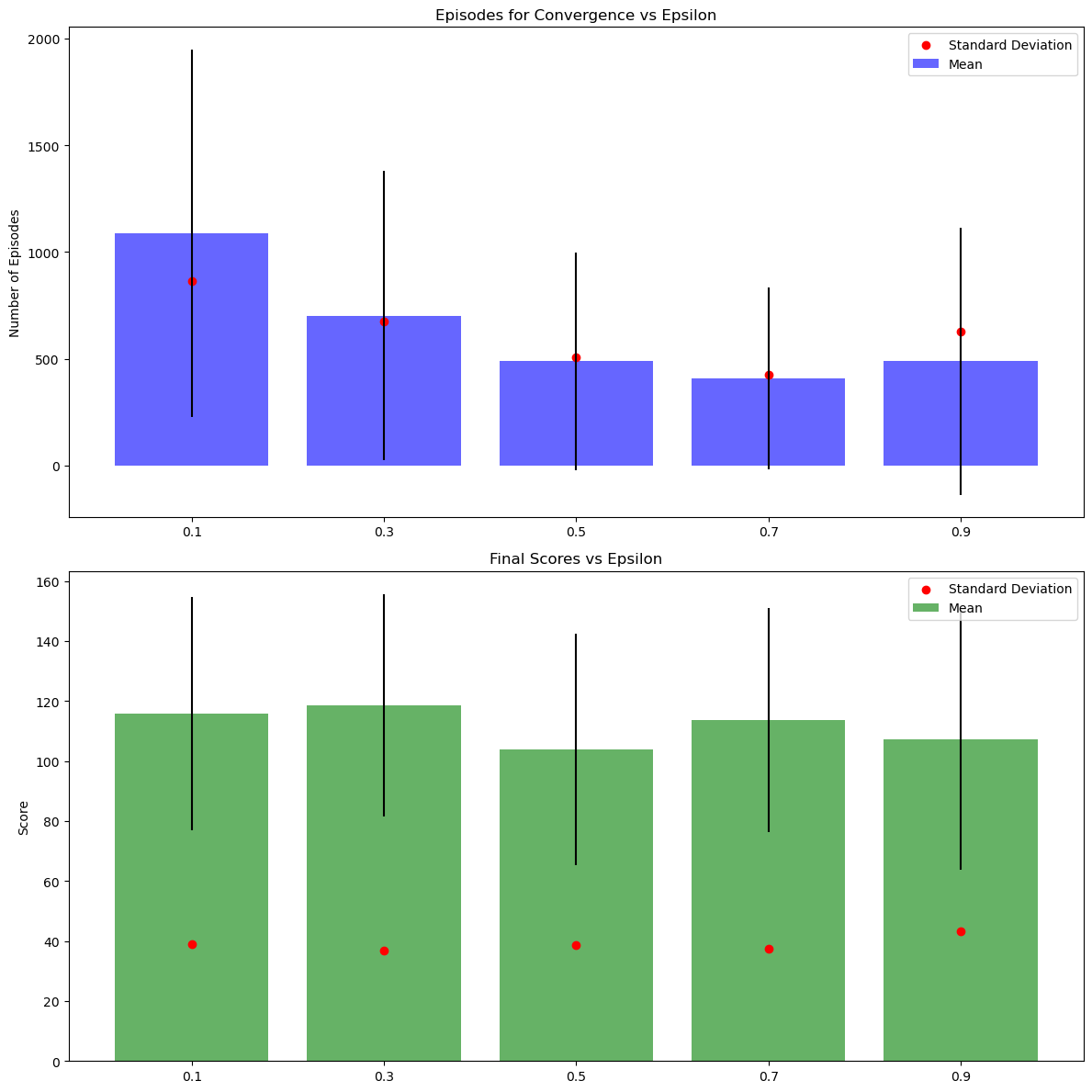
Here are the combined visualizations for both the mean and standard deviation values, as they relate to (learning rate):

**Episodes for Convergence vs. :**

* The top graph showcases the mean number of episodes required for convergence for each value, with error bars indicating one standard deviation.
* Additionally, the red scatter points represent the standard deviation values for each .
* We observe a decreasing trend in both the mean number of episodes and their variability (standard deviation) as increases. This suggests that with a higher learning rate, Q-learning tends to converge faster and more consistently.

**Final Scores vs. :**

* The bottom graph displays the mean final scores achieved by the robot for each value, with error bars indicating one standard deviation.
* The red scatter points represent the standard deviation values for each The mean final scores appear to be stable across different values, indicating consistent policy quality. However, the variability in scores (standard deviation) remains relatively stable across different values.



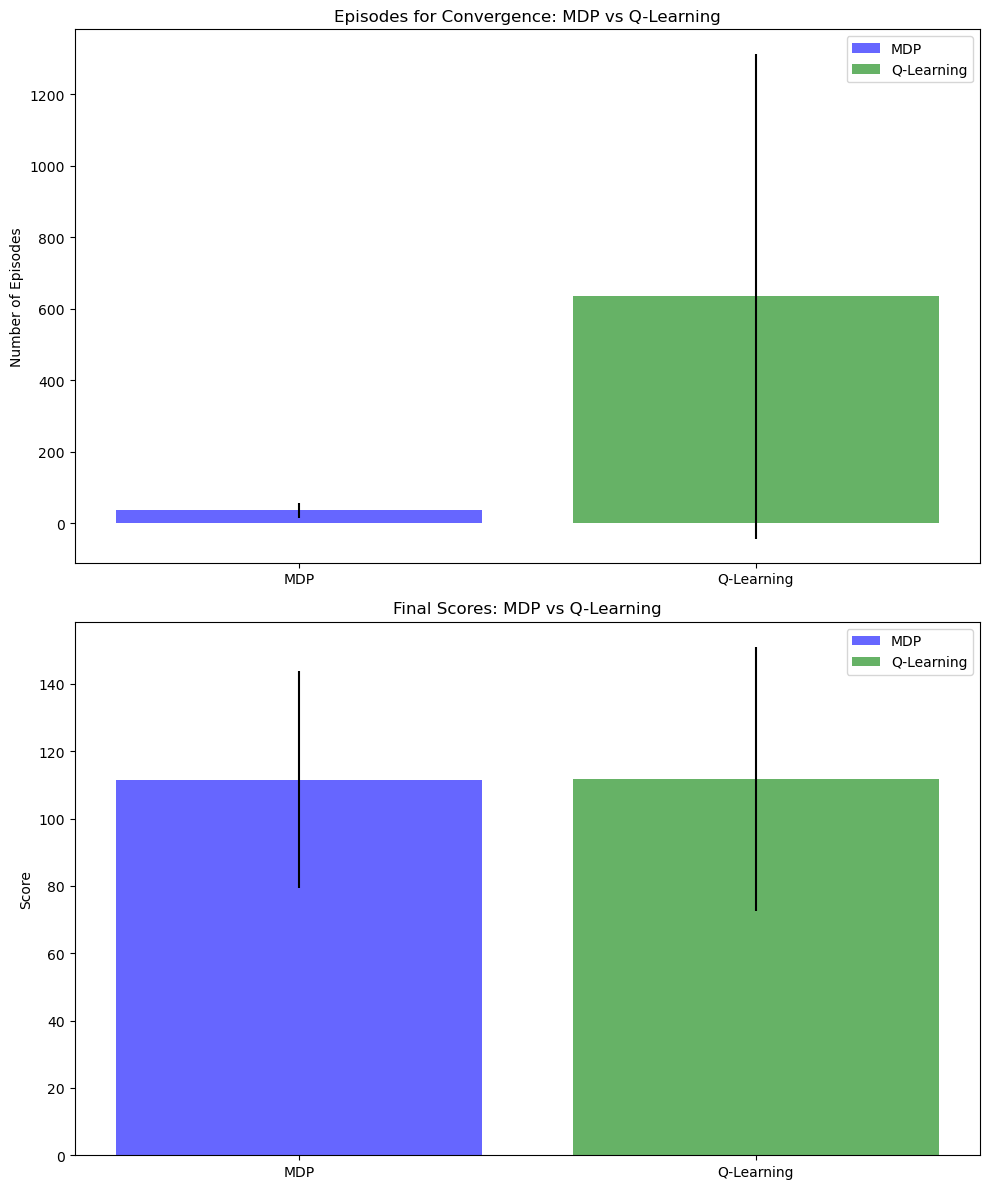
The combined visualizations for both the mean and standard deviation values, as they relate to ϵ (exploration probability), are as follows:

**Episodes for Convergence vs. ϵ:**

* + - The top graph showcases the mean number of episodes required for convergence for each ϵ value, with error bars indicating one standard deviation.
    - Additionally, the red scatter points represent the standard deviation values for each ϵ.
    - We observe that as ϵ increases, the mean number of episodes for convergence tends to decrease, suggesting faster learning with more exploration. Additionally, the variability in the number of episodes (standard deviation) reduces as ϵ increases, leading to more consistent convergence rates.

**Final Scores vs. ϵ:**

* + - The bottom graph displays the mean final scores achieved by the robot for each ϵ value, with error bars indicating one standard deviation.
    - The red scatter points represent the standard deviation values for each ϵ.
    - The mean final scores remain relatively consistent across different ϵ values. However, there are slight variations in scores across different ϵ values, suggesting that certain exploration-exploitation balances might be marginally more effective than others. The variability in scores (standard deviation) remains relatively stable across different ϵ values.



Here's the visual comparison between MDP (Value Iteration) and Q-Learning:

**Episodes for Convergence: MDP vs. Q-Learning:**

* The top graph compares the mean number of episodes required for convergence for both MDP and Q-Learning, with error bars indicating one standard deviation.
* We observe that MDP generally requires fewer episodes to converge compared to Q-Learning. Additionally, the variability (standard deviation) in the number of episodes is lower for MDP compared to Q-Learning.

**Final Scores: MDP vs. Q-Learning:**

* The bottom graph compares the mean final scores achieved by the robot using both MDP and Q-Learning, with error bars indicating one standard deviation.
* The mean final scores are comparable between MDP and Q-Learning, suggesting that both methods result in similarly effective policies. However, Q-Learning has a slightly wider spread (higher standard deviation), implying that the scores can vary more across different runs or parameter combinations.

**4. Discussion**

The visual and statistical analyses provide a comprehensive view of the performance of MDP and Q-learning in the GridWorld environment:

**Inherent Knowledge in MDP:** MDP has a distinct advantage in that its parameters, like the transition matrix T and the reward function R, are pre-defined and known. In contrast, Q-Learning operates without explicit knowledge of the transition matrix and reward function. This inherent knowledge in MDP typically allows it to converge faster and with greater stability.

**Convergence Speed:** Given the known parameters in MDP, it tends to converge faster and more consistently than Q-Learning. The exploration probability (ϵ) in Q-Learning influences its convergence speed, with higher ϵ values leading to quicker learning. Similarly, an increased learning rate () in Q-Learning accelerates its convergence.

**Policy Quality:** Both MDP and Q-Learning yield policies of comparable quality. However, despite the uncertainties in Q-Learning due to the unknown transition matrix and reward function, it still manages to produce policies of consistently good quality. Q-Learning displays more variability in its outcomes, potentially due to the randomness introduced by exploration and different parameter combinations.

**Exploration vs. Exploitation in Q-Learning:** The trade-off between exploration and exploitation in Q-Learning is evident. More exploration (higher ϵ) leads to faster learning but might introduce variability in the outcomes. On the other hand, a higher learning rate () facilitates faster and consistent learning without affecting the quality of the policy significantly.

In conclusion, while MDP offers rapid and consistent learning in the GridWorld environment due to its known parameters, Q-Learning provides flexibility with its parameters, allowing for potential optimizations based on the specific needs and characteristics of the environment, even without explicit knowledge of the system dynamics.

**Appendix**

**Mdp stats**

|  |  |  |
| --- | --- | --- |
|  | **Episodes for Convergence** | **Final Scores** |
| **mean** | 36.1 | 111.5098 |
| **std** | 20.73885 | 32.20729 |
| **min** | 21 | 51.77186 |
| **max** | 89 | 169 |

**Ql\_stats**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Episodes for Convergence** | | | | **Final Scores** | | | |
|  |  | **mean** | **std** | **min** | **max** | **mean** | **std** | **min** | **max** |
| **Alpha** | **Epsilon** |  |  |  |  |  |  |  |  |
| **0.1** | **0.1** | 2560.2 | 608.7103 | 1398 | 3393 | 100.5932 | 34.30032 | 55.08058 | 141.8 |
| **0.3** | 1962.2 | 295.6743 | 1789 | 2782 | 114.6383 | 37.17608 | 60.8595 | 168 |
| **0.5** | 1346.6 | 495.6229 | 1147 | 2753 | 118.7865 | 42.07496 | 62.51866 | 199 |
| **0.7** | 1131.1 | 360.1489 | 949 | 2148 | 114.9278 | 33.26522 | 77.09344 | 199 |
| **0.9** | 1488.5 | 741.1431 | 999 | 2962 | 93.78336 | 38.20082 | 53.53369 | 178 |
| **0.3** | **0.1** | 1255.5 | 301.8588 | 743 | 1642 | 119.5798 | 42.03906 | 70.36938 | 178 |
| **0.3** | 693.5 | 124.982 | 616 | 1041 | 130.0333 | 27.3679 | 94.269 | 178 |
| **0.5** | 490.8 | 199.3979 | 394 | 1054 | 95.62424 | 32.36526 | 47.9047 | 150.8 |
| **0.7** | 430.5 | 241.8784 | 337 | 1118 | 117.5964 | 44.14026 | 64.43102 | 199 |
| **0.9** | 469.9 | 329.0209 | 328 | 1403 | 117.8354 | 42.5566 | 56.87707 | 178 |
| **0.5** | **0.1** | 764.6 | 153.9417 | 474 | 1028 | 103.464 | 39.79913 | 45.26394 | 178 |
| **0.3** | 416.2 | 81.39315 | 357 | 640 | 113.3562 | 44.66993 | 54.88032 | 199 |
| **0.5** | 266.7 | 13.22498 | 249 | 293 | 102.1582 | 27.11553 | 67.4841 | 136.22 |
| **0.7** | 235.7 | 63.81057 | 193 | 414 | 96.65255 | 26.98266 | 69.55938 | 150.8 |
| **0.9** | 228.2 | 18.65952 | 205 | 268 | 118.8121 | 44.13239 | 51.75566 | 199 |
| **0.7** | **0.1** | 534.6 | 56.0004 | 428 | 615 | 135.3211 | 37.30526 | 77.09344 | 178 |
| **0.3** | 264.7 | 14.51474 | 244 | 297 | 109.6846 | 33.39012 | 48.26888 | 159 |
| **0.5** | 220.7 | 144.182 | 151 | 625 | 102.8343 | 49.34829 | 0 | 178 |
| **0.7** | 139.3 | 10.16585 | 126 | 152 | 114.2631 | 37.51044 | 55.08058 | 169 |
| **0.9** | 150.4 | 11.6352 | 129 | 166 | 108.1024 | 52.2707 | 0 | 178 |
| **0.9** | **0.1** | 322.1 | 85.25315 | 223 | 494 | 119.6237 | 37.6906 | 61.93842 | 188 |
| **0.3** | 169.1 | 18.1503 | 150 | 213 | 125.0848 | 43.19273 | 70.73753 | 199 |
| **0.5** | 118.1 | 11.00959 | 105 | 138 | 99.36188 | 42.21158 | 50.3812 | 151.8 |
| **0.7** | 114.1 | 58.12907 | 88 | 278 | 124.7671 | 43.84187 | 63.16679 | 188 |
| **0.9** | 111.7 | 36.42969 | 86 | 208 | 97.10165 | 40.65917 | 40.07119 | 159.9 |

**Raw\_results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Alpha** | **Epsilon** | **Episodes for Convergence** | **Final Scores** |
| MDP | - | - | 21 | 118.659 |
| MDP | - | - | 51 | 126.22 |
| MDP | - | - | 27 | 51.77186 |
| MDP | - | - | 41 | 141.8 |
| MDP | - | - | 89 | 94.94938 |
| MDP | - | - | 27 | 99.2882 |
| MDP | - | - | 24 | 169 |
| MDP | - | - | 24 | 122.098 |
| MDP | - | - | 25 | 106.5782 |
| MDP | - | - | 32 | 84.73297 |
| Q-Learning | 0.1 | 0.1 | 2841 | 97.06885 |
| Q-Learning | 0.1 | 0.1 | 2433 | 141.8 |
| Q-Learning | 0.1 | 0.1 | 2570 | 65.74947 |
| Q-Learning | 0.1 | 0.1 | 1398 | 141.8 |
| Q-Learning | 0.1 | 0.1 | 2748 | 108.2882 |
| Q-Learning | 0.1 | 0.1 | 3361 | 141.8 |
| Q-Learning | 0.1 | 0.1 | 3393 | 113.9492 |
| Q-Learning | 0.1 | 0.1 | 2709 | 77.65938 |
| Q-Learning | 0.1 | 0.1 | 2186 | 62.73632 |
| Q-Learning | 0.1 | 0.1 | 1963 | 55.08058 |
| Q-Learning | 0.1 | 0.3 | 1897 | 168 |
| Q-Learning | 0.1 | 0.3 | 1789 | 107.22 |
| Q-Learning | 0.1 | 0.3 | 1825 | 118.12 |
| Q-Learning | 0.1 | 0.3 | 1838 | 117.5592 |
| Q-Learning | 0.1 | 0.3 | 1983 | 67.4841 |
| Q-Learning | 0.1 | 0.3 | 1825 | 141.8 |
| Q-Learning | 0.1 | 0.3 | 1826 | 60.8595 |
| Q-Learning | 0.1 | 0.3 | 1873 | 141.8 |
| Q-Learning | 0.1 | 0.3 | 1984 | 149.9 |
| Q-Learning | 0.1 | 0.3 | 2782 | 73.64059 |
| Q-Learning | 0.1 | 0.5 | 1260 | 99.2882 |
| Q-Learning | 0.1 | 0.5 | 1147 | 168 |
| Q-Learning | 0.1 | 0.5 | 1153 | 120.198 |
| Q-Learning | 0.1 | 0.5 | 2753 | 101.5104 |
| Q-Learning | 0.1 | 0.5 | 1207 | 135.22 |
| Q-Learning | 0.1 | 0.5 | 1155 | 62.51866 |
| Q-Learning | 0.1 | 0.5 | 1219 | 132.8 |
| Q-Learning | 0.1 | 0.5 | 1172 | 99.41047 |
| Q-Learning | 0.1 | 0.5 | 1232 | 199 |
| Q-Learning | 0.1 | 0.5 | 1168 | 69.91964 |
| Q-Learning | 0.1 | 0.7 | 1016 | 99.53479 |
| Q-Learning | 0.1 | 0.7 | 1055 | 120.198 |
| Q-Learning | 0.1 | 0.7 | 991 | 199 |
| Q-Learning | 0.1 | 0.7 | 992 | 77.09344 |
| Q-Learning | 0.1 | 0.7 | 949 | 111.0738 |
| Q-Learning | 0.1 | 0.7 | 1038 | 122.1392 |
| Q-Learning | 0.1 | 0.7 | 1091 | 95.75938 |
| Q-Learning | 0.1 | 0.7 | 965 | 123.51 |
| Q-Learning | 0.1 | 0.7 | 1066 | 112.098 |
| Q-Learning | 0.1 | 0.7 | 2148 | 88.87147 |
| Q-Learning | 0.1 | 0.9 | 999 | 105.1931 |
| Q-Learning | 0.1 | 0.9 | 1079 | 100.0983 |
| Q-Learning | 0.1 | 0.9 | 1219 | 79.95 |
| Q-Learning | 0.1 | 0.9 | 1016 | 128.659 |
| Q-Learning | 0.1 | 0.9 | 1220 | 101.9738 |
| Q-Learning | 0.1 | 0.9 | 1185 | 62.54985 |
| Q-Learning | 0.1 | 0.9 | 1252 | 67.05212 |
| Q-Learning | 0.1 | 0.9 | 1148 | 53.53369 |
| Q-Learning | 0.1 | 0.9 | 2962 | 178 |
| Q-Learning | 0.1 | 0.9 | 2805 | 60.82369 |
| Q-Learning | 0.3 | 0.1 | 1642 | 93.38344 |
| Q-Learning | 0.3 | 0.1 | 743 | 159 |
| Q-Learning | 0.3 | 0.1 | 934 | 149.9 |
| Q-Learning | 0.3 | 0.1 | 1479 | 70.36938 |
| Q-Learning | 0.3 | 0.1 | 1604 | 178 |
| Q-Learning | 0.3 | 0.1 | 1402 | 78.53329 |
| Q-Learning | 0.3 | 0.1 | 1418 | 178 |
| Q-Learning | 0.3 | 0.1 | 988 | 106.5782 |
| Q-Learning | 0.3 | 0.1 | 1167 | 88.46938 |
| Q-Learning | 0.3 | 0.1 | 1178 | 93.56428 |
| Q-Learning | 0.3 | 0.3 | 655 | 94.269 |
| Q-Learning | 0.3 | 0.3 | 624 | 150.8 |
| Q-Learning | 0.3 | 0.3 | 689 | 96.65938 |
| Q-Learning | 0.3 | 0.3 | 616 | 141.8 |
| Q-Learning | 0.3 | 0.3 | 682 | 159 |
| Q-Learning | 0.3 | 0.3 | 670 | 120.198 |
| Q-Learning | 0.3 | 0.3 | 654 | 107.4374 |
| Q-Learning | 0.3 | 0.3 | 622 | 125.949 |
| Q-Learning | 0.3 | 0.3 | 1041 | 178 |
| Q-Learning | 0.3 | 0.3 | 682 | 126.22 |
| Q-Learning | 0.3 | 0.5 | 420 | 96.8492 |
| Q-Learning | 0.3 | 0.5 | 438 | 47.9047 |
| Q-Learning | 0.3 | 0.5 | 431 | 112.9492 |
| Q-Learning | 0.3 | 0.5 | 483 | 76.96239 |
| Q-Learning | 0.3 | 0.5 | 1054 | 150.8 |
| Q-Learning | 0.3 | 0.5 | 394 | 121.827 |
| Q-Learning | 0.3 | 0.5 | 429 | 122.098 |
| Q-Learning | 0.3 | 0.5 | 411 | 99.2882 |
| Q-Learning | 0.3 | 0.5 | 405 | 67.4841 |
| Q-Learning | 0.3 | 0.5 | 443 | 60.07966 |
| Q-Learning | 0.3 | 0.7 | 367 | 67.03543 |
| Q-Learning | 0.3 | 0.7 | 358 | 98.2882 |
| Q-Learning | 0.3 | 0.7 | 337 | 112.098 |
| Q-Learning | 0.3 | 0.7 | 340 | 159 |
| Q-Learning | 0.3 | 0.7 | 1118 | 64.43102 |
| Q-Learning | 0.3 | 0.7 | 351 | 134.41 |
| Q-Learning | 0.3 | 0.7 | 338 | 95.46089 |
| Q-Learning | 0.3 | 0.7 | 364 | 87.24 |
| Q-Learning | 0.3 | 0.7 | 361 | 199 |
| Q-Learning | 0.3 | 0.7 | 371 | 159 |
| Q-Learning | 0.3 | 0.9 | 328 | 136.22 |
| Q-Learning | 0.3 | 0.9 | 375 | 56.87707 |
| Q-Learning | 0.3 | 0.9 | 338 | 178 |
| Q-Learning | 0.3 | 0.9 | 341 | 99.50785 |
| Q-Learning | 0.3 | 0.9 | 393 | 120.198 |
| Q-Learning | 0.3 | 0.9 | 1403 | 178 |
| Q-Learning | 0.3 | 0.9 | 374 | 151.8 |
| Q-Learning | 0.3 | 0.9 | 374 | 83.65444 |
| Q-Learning | 0.3 | 0.9 | 420 | 72.09838 |
| Q-Learning | 0.3 | 0.9 | 353 | 101.9982 |
| Q-Learning | 0.5 | 0.1 | 655 | 149.9 |
| Q-Learning | 0.5 | 0.1 | 750 | 109.2882 |
| Q-Learning | 0.5 | 0.1 | 790 | 64.08623 |
| Q-Learning | 0.5 | 0.1 | 908 | 178 |
| Q-Learning | 0.5 | 0.1 | 744 | 100.1253 |
| Q-Learning | 0.5 | 0.1 | 474 | 76.71356 |
| Q-Learning | 0.5 | 0.1 | 752 | 97.31275 |
| Q-Learning | 0.5 | 0.1 | 660 | 45.26394 |
| Q-Learning | 0.5 | 0.1 | 1028 | 87.19082 |
| Q-Learning | 0.5 | 0.1 | 885 | 126.759 |
| Q-Learning | 0.5 | 0.3 | 360 | 85.70243 |
| Q-Learning | 0.5 | 0.3 | 428 | 199 |
| Q-Learning | 0.5 | 0.3 | 403 | 99.2882 |
| Q-Learning | 0.5 | 0.3 | 399 | 63.74059 |
| Q-Learning | 0.5 | 0.3 | 381 | 54.88032 |
| Q-Learning | 0.5 | 0.3 | 398 | 102.5643 |
| Q-Learning | 0.5 | 0.3 | 357 | 149.9 |
| Q-Learning | 0.5 | 0.3 | 402 | 120.198 |
| Q-Learning | 0.5 | 0.3 | 394 | 159 |
| Q-Learning | 0.5 | 0.3 | 640 | 99.2882 |
| Q-Learning | 0.5 | 0.5 | 283 | 119.388 |
| Q-Learning | 0.5 | 0.5 | 255 | 118.659 |
| Q-Learning | 0.5 | 0.5 | 260 | 124.8492 |
| Q-Learning | 0.5 | 0.5 | 266 | 82.31047 |
| Q-Learning | 0.5 | 0.5 | 261 | 70.9231 |
| Q-Learning | 0.5 | 0.5 | 274 | 67.4841 |
| Q-Learning | 0.5 | 0.5 | 249 | 90.2882 |
| Q-Learning | 0.5 | 0.5 | 261 | 77.1402 |
| Q-Learning | 0.5 | 0.5 | 265 | 134.32 |
| Q-Learning | 0.5 | 0.5 | 293 | 136.22 |
| Q-Learning | 0.5 | 0.7 | 222 | 91.48344 |
| Q-Learning | 0.5 | 0.7 | 232 | 69.55938 |
| Q-Learning | 0.5 | 0.7 | 215 | 150.8 |
| Q-Learning | 0.5 | 0.7 | 224 | 91.9982 |
| Q-Learning | 0.5 | 0.7 | 228 | 72.99834 |
| Q-Learning | 0.5 | 0.7 | 193 | 72.26707 |
| Q-Learning | 0.5 | 0.7 | 414 | 119.8541 |
| Q-Learning | 0.5 | 0.7 | 220 | 113.2931 |
| Q-Learning | 0.5 | 0.7 | 202 | 112.4831 |
| Q-Learning | 0.5 | 0.7 | 207 | 71.78877 |
| Q-Learning | 0.5 | 0.9 | 225 | 99.2882 |
| Q-Learning | 0.5 | 0.9 | 232 | 99.63251 |
| Q-Learning | 0.5 | 0.9 | 205 | 178 |
| Q-Learning | 0.5 | 0.9 | 237 | 130.198 |
| Q-Learning | 0.5 | 0.9 | 268 | 77.69367 |
| Q-Learning | 0.5 | 0.9 | 213 | 136.22 |
| Q-Learning | 0.5 | 0.9 | 240 | 108.0443 |
| Q-Learning | 0.5 | 0.9 | 205 | 199 |
| Q-Learning | 0.5 | 0.9 | 229 | 108.2882 |
| Q-Learning | 0.5 | 0.9 | 228 | 51.75566 |
| Q-Learning | 0.7 | 0.1 | 615 | 151.61 |
| Q-Learning | 0.7 | 0.1 | 614 | 159 |
| Q-Learning | 0.7 | 0.1 | 544 | 178 |
| Q-Learning | 0.7 | 0.1 | 544 | 133.51 |
| Q-Learning | 0.7 | 0.1 | 505 | 77.09344 |
| Q-Learning | 0.7 | 0.1 | 562 | 96.0031 |
| Q-Learning | 0.7 | 0.1 | 508 | 136.22 |
| Q-Learning | 0.7 | 0.1 | 428 | 84.7741 |
| Q-Learning | 0.7 | 0.1 | 531 | 178 |
| Q-Learning | 0.7 | 0.1 | 495 | 159 |
| Q-Learning | 0.7 | 0.3 | 297 | 75.15212 |
| Q-Learning | 0.7 | 0.3 | 263 | 126.22 |
| Q-Learning | 0.7 | 0.3 | 272 | 121.098 |
| Q-Learning | 0.7 | 0.3 | 259 | 159 |
| Q-Learning | 0.7 | 0.3 | 254 | 108.2882 |
| Q-Learning | 0.7 | 0.3 | 258 | 79.28854 |
| Q-Learning | 0.7 | 0.3 | 263 | 126.22 |
| Q-Learning | 0.7 | 0.3 | 260 | 141.8 |
| Q-Learning | 0.7 | 0.3 | 244 | 111.5104 |
| Q-Learning | 0.7 | 0.3 | 277 | 48.26888 |
| Q-Learning | 0.7 | 0.5 | 237 | 90.94444 |
| Q-Learning | 0.7 | 0.5 | 151 | 82.1451 |
| Q-Learning | 0.7 | 0.5 | 156 | 0 |
| Q-Learning | 0.7 | 0.5 | 183 | 72.79851 |
| Q-Learning | 0.7 | 0.5 | 178 | 141.8 |
| Q-Learning | 0.7 | 0.5 | 159 | 112.098 |
| Q-Learning | 0.7 | 0.5 | 188 | 178 |
| Q-Learning | 0.7 | 0.5 | 166 | 112.098 |
| Q-Learning | 0.7 | 0.5 | 625 | 150.8 |
| Q-Learning | 0.7 | 0.5 | 164 | 87.65938 |
| Q-Learning | 0.7 | 0.7 | 152 | 91.99834 |
| Q-Learning | 0.7 | 0.7 | 129 | 169 |
| Q-Learning | 0.7 | 0.7 | 126 | 78.70341 |
| Q-Learning | 0.7 | 0.7 | 151 | 105.537 |
| Q-Learning | 0.7 | 0.7 | 134 | 169 |
| Q-Learning | 0.7 | 0.7 | 143 | 125.949 |
| Q-Learning | 0.7 | 0.7 | 127 | 112.098 |
| Q-Learning | 0.7 | 0.7 | 137 | 93.46444 |
| Q-Learning | 0.7 | 0.7 | 151 | 141.8 |
| Q-Learning | 0.7 | 0.7 | 143 | 55.08058 |
| Q-Learning | 0.7 | 0.9 | 129 | 100.1253 |
| Q-Learning | 0.7 | 0.9 | 154 | 159 |
| Q-Learning | 0.7 | 0.9 | 141 | 116.22 |
| Q-Learning | 0.7 | 0.9 | 158 | 0 |
| Q-Learning | 0.7 | 0.9 | 158 | 97.65938 |
| Q-Learning | 0.7 | 0.9 | 160 | 96.53741 |
| Q-Learning | 0.7 | 0.9 | 166 | 169 |
| Q-Learning | 0.7 | 0.9 | 136 | 84.38344 |
| Q-Learning | 0.7 | 0.9 | 151 | 80.09838 |
| Q-Learning | 0.7 | 0.9 | 151 | 178 |
| Q-Learning | 0.9 | 0.1 | 494 | 141.8 |
| Q-Learning | 0.9 | 0.1 | 291 | 188 |
| Q-Learning | 0.9 | 0.1 | 223 | 108.2882 |
| Q-Learning | 0.9 | 0.1 | 257 | 112.098 |
| Q-Learning | 0.9 | 0.1 | 314 | 108.2882 |
| Q-Learning | 0.9 | 0.1 | 429 | 61.93842 |
| Q-Learning | 0.9 | 0.1 | 256 | 66.83569 |
| Q-Learning | 0.9 | 0.1 | 358 | 150.8 |
| Q-Learning | 0.9 | 0.1 | 263 | 133.51 |
| Q-Learning | 0.9 | 0.1 | 336 | 124.6782 |
| Q-Learning | 0.9 | 0.3 | 163 | 106.5782 |
| Q-Learning | 0.9 | 0.3 | 167 | 159 |
| Q-Learning | 0.9 | 0.3 | 173 | 87.65938 |
| Q-Learning | 0.9 | 0.3 | 150 | 96.65938 |
| Q-Learning | 0.9 | 0.3 | 170 | 199 |
| Q-Learning | 0.9 | 0.3 | 213 | 126.22 |
| Q-Learning | 0.9 | 0.3 | 150 | 141.8 |
| Q-Learning | 0.9 | 0.3 | 160 | 70.73753 |
| Q-Learning | 0.9 | 0.3 | 181 | 178 |
| Q-Learning | 0.9 | 0.3 | 164 | 85.19344 |
| Q-Learning | 0.9 | 0.5 | 138 | 80.2882 |
| Q-Learning | 0.9 | 0.5 | 114 | 50.3812 |
| Q-Learning | 0.9 | 0.5 | 115 | 52.87759 |
| Q-Learning | 0.9 | 0.5 | 105 | 87.65938 |
| Q-Learning | 0.9 | 0.5 | 131 | 81.87641 |
| Q-Learning | 0.9 | 0.5 | 113 | 56.66702 |
| Q-Learning | 0.9 | 0.5 | 116 | 143.51 |
| Q-Learning | 0.9 | 0.5 | 105 | 151.8 |
| Q-Learning | 0.9 | 0.5 | 129 | 151.8 |
| Q-Learning | 0.9 | 0.5 | 115 | 136.759 |
| Q-Learning | 0.9 | 0.7 | 104 | 77.35477 |
| Q-Learning | 0.9 | 0.7 | 91 | 72.8337 |
| Q-Learning | 0.9 | 0.7 | 278 | 178 |
| Q-Learning | 0.9 | 0.7 | 113 | 112.098 |
| Q-Learning | 0.9 | 0.7 | 100 | 134.32 |
| Q-Learning | 0.9 | 0.7 | 91 | 141.8 |
| Q-Learning | 0.9 | 0.7 | 88 | 188 |
| Q-Learning | 0.9 | 0.7 | 88 | 121.098 |
| Q-Learning | 0.9 | 0.7 | 92 | 159 |
| Q-Learning | 0.9 | 0.7 | 96 | 63.16679 |
| Q-Learning | 0.9 | 0.9 | 133 | 50.71252 |
| Q-Learning | 0.9 | 0.9 | 98 | 159 |
| Q-Learning | 0.9 | 0.9 | 86 | 159.9 |
| Q-Learning | 0.9 | 0.9 | 95 | 85.2143 |
| Q-Learning | 0.9 | 0.9 | 86 | 80.89851 |
| Q-Learning | 0.9 | 0.9 | 108 | 80.61866 |
| Q-Learning | 0.9 | 0.9 | 107 | 40.07119 |
| Q-Learning | 0.9 | 0.9 | 95 | 85.2833 |
| Q-Learning | 0.9 | 0.9 | 101 | 126.22 |
| Q-Learning | 0.9 | 0.9 | 208 | 103.098 |