Argha Boksi & Jashaswimalya Acharjee
CS21D407 & CS22E005
February 24, 2023

CS 6700: Reinforcement Learning
Programming Assignment: 1
Course Instructor : Prof. Balaraman Ravindran

1 Tasks

- Implement **SARSA** and **Q-Learning**.
- For each algorithm, run experiments with `wind=False` and `wind=True`; two different `start states:` $(0, 4)$, $(3, 6)$; two values of p $(1.0, 0.7)$; and two types of exploration strategies (`$\varepsilon$-greedy and softmax`), making it **16** different configurations in total.
- For each of the 16 configurations, determine the best set of hyperparameters (ε in ε-greedy exploration, temperature β or τ in softmax exploration, learning rate α, and discount factor γ) and plot the following:
  - Reward curves and the number of steps to reach the goal in each episode (during the training phase with the best hyperparameters).
  - Heatmap of the grid with state visit counts, i.e., the number of times each state was visited throughout the training phase.
  - Heatmap of the grid with Q values after training is complete, and optimal actions for the best policy.
- For each of the algorithm, provide a written description of the policy learnt, explaining the behavior of the agent, and your choice of hyperparameters. This description should also provide information about how the behavior changes with and without the wind, for different levels of stochasticity and for different start states.

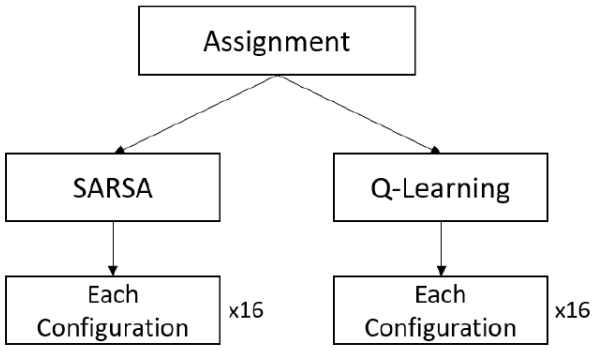

Figure 1: Organization of the Experiments

2 Solution

For each of the 32 configurations we found the best set of hyperparameters. These are the values we tried for finding the best set of hyperparameters - $\epsilon = \{0.001, 0.01, 0.1\}$, $\beta = \{0.01, 0.1, 1.0, 2.0\}$, $\gamma = \{0.7, 0.8, 0.9, 1.0\}$, $\alpha = \{0.001, 0.01, 0.1, 1.0\}$.

2.1 Algorithm: Q-Learning

2.1.1 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heat-map it can be seen that the agent visits the upper left corner very frequently. Also from the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high.

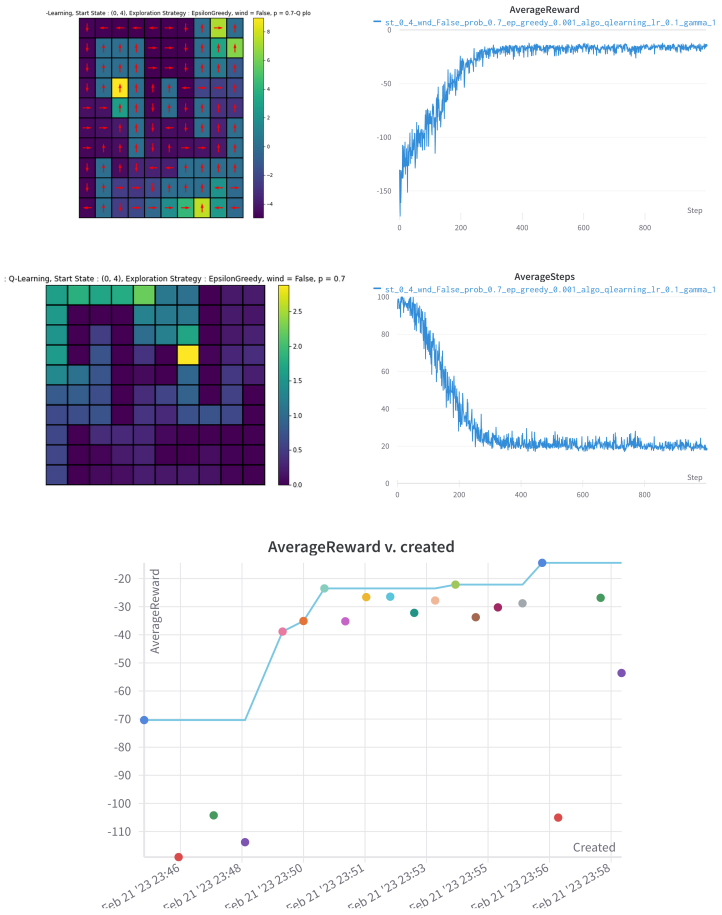

Figure 2: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` False, `P:` 0.7, `Strategy:` ϵ-`greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 1

2.1.2 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = False, p = 1

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** From the state-visit count heat-map it can be seen that the agent visits the upper left corner very frequently. It has learnt to reach the upper left corner goal state.Also from the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high.

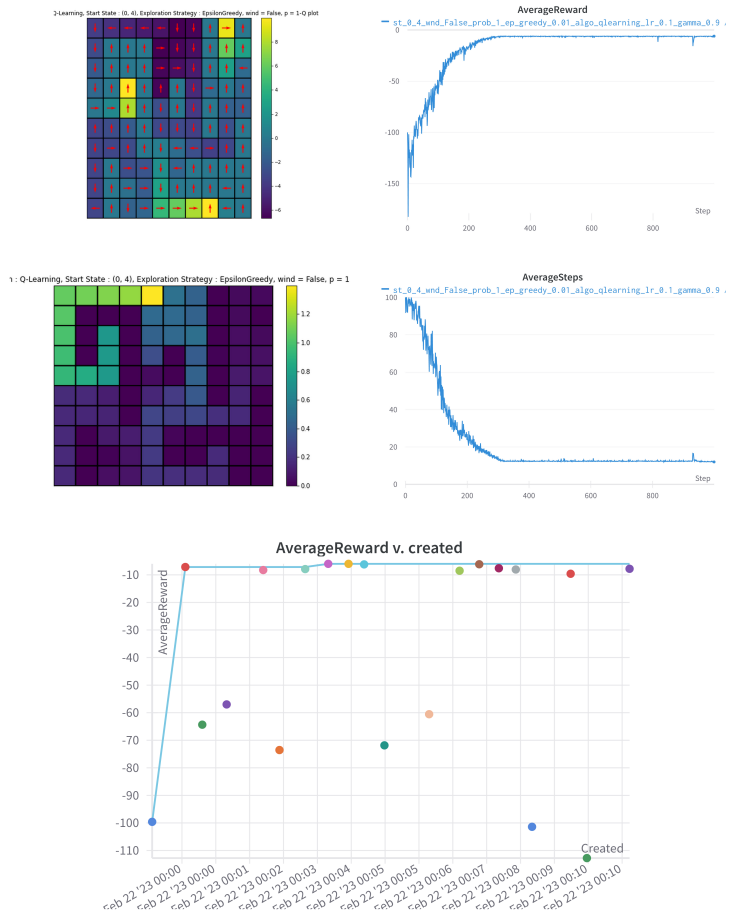

Figure 3: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` False, `P:` 1, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 0.9

2.1.3 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high. From the state-visit count heat-map it can be seen that the agent has learnt to reach bottom right goal state. It does not visit the other two goal states that often.

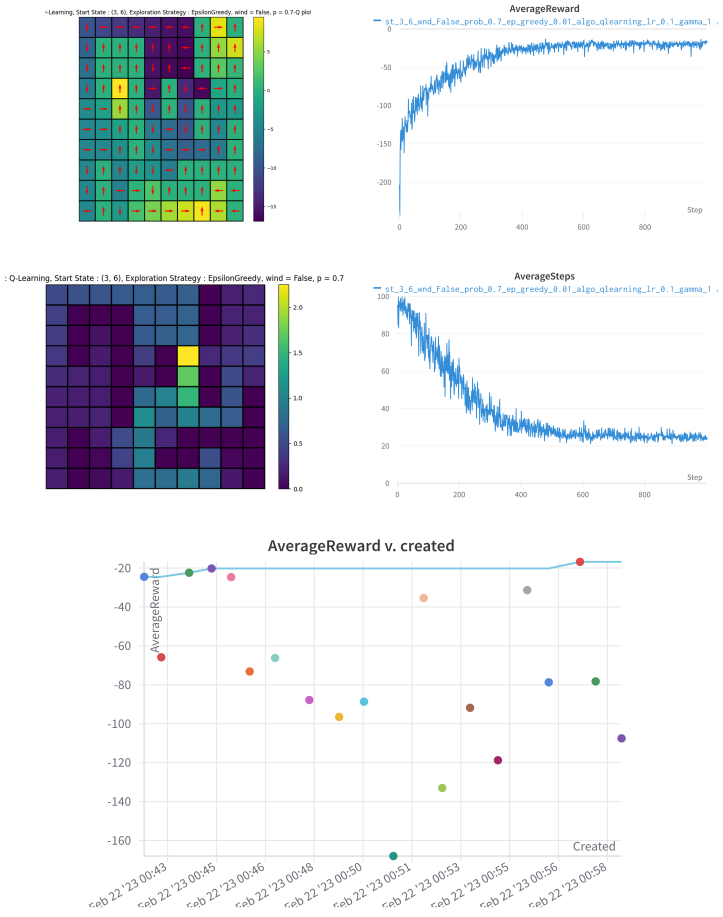

Figure 4: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` False, `P:` 0.7, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 1

2.1.4 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = False, p = 1

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 1, γ: 1
- **Policy Learnt:** From the state-visit count plot it is clear that the agent has learnt to reach the upper right corner goal state. Also from the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high.

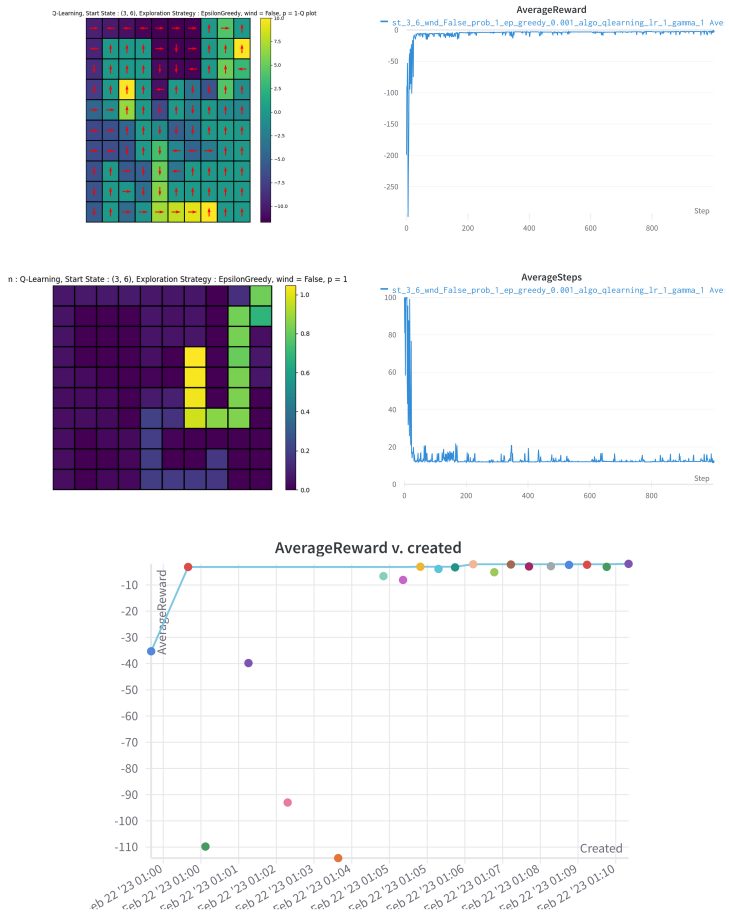

Figure 5: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` False, `P:` 1, `Strategy:` ϵ`-greedy`, ϵ`:` 0.001, `Learning Rate:` 1, γ`:` 1

2.1.5 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count plot it is clear that the agent visits the top left corner very often. But it does not visit the top left goal state that often.

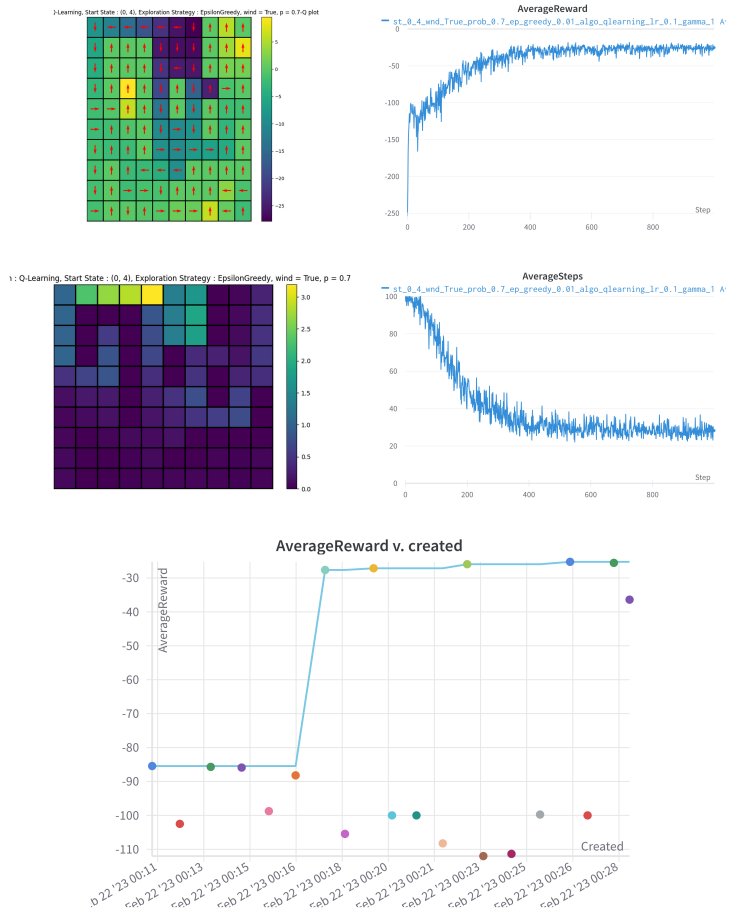

Figure 6: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` True, `P:` 0.7, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 1

2.1.6 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = True, p = 1

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count plot it is clear that the agent visits the upper left corner frequently. Also from the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high.

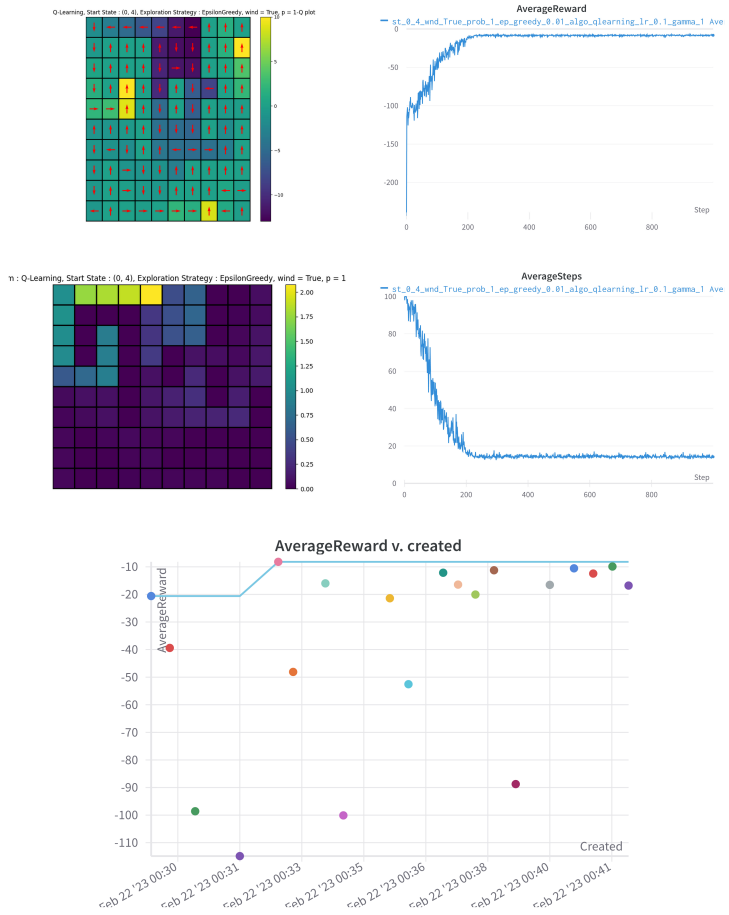

Figure 7: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` True, `P:` 1, `Strategy:` ϵ-`greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 1

2.1.7 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.1$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** From the state-visit count plot it can be seen that the agent visits both the top left cells and the middle region of the grid very often. However it does not visit any of the goal states frequently.

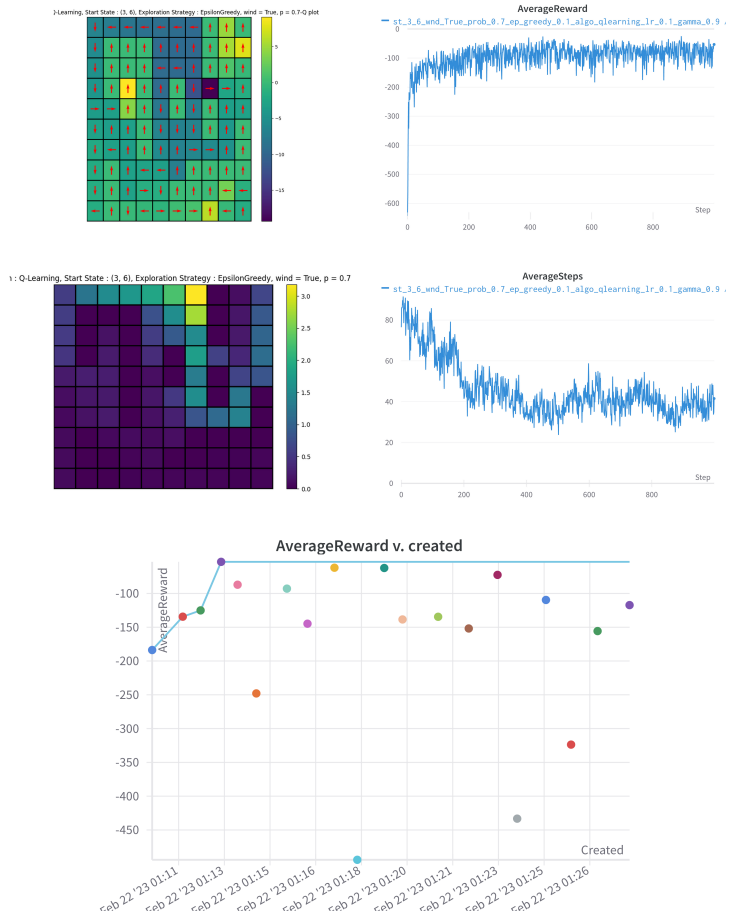

Figure 8: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` True, `P:` 0.7, `Strategy:` ϵ-`greedy`, ϵ`:` 0.1, `Learning Rate:` 0.1, γ`:` 0.9

2.1.8 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = True, p = 1

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** From the state-visit count plot it can be seen that the agent visits the top right corner frequently and it has learnt to reach the top right corner goal state.

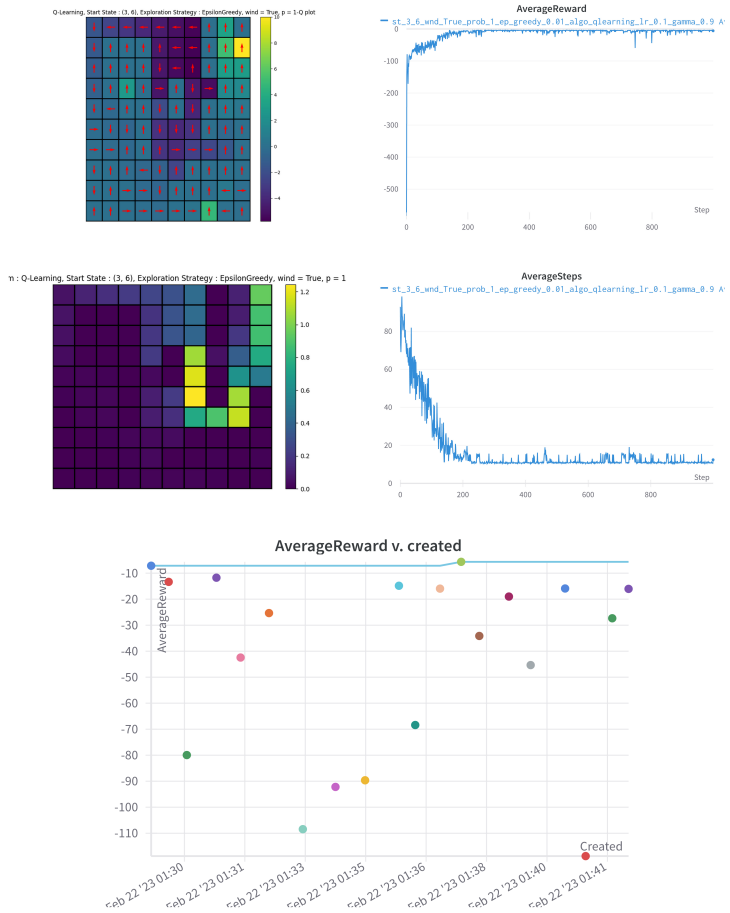

Figure 9: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` True, `P:` 1, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 0.9

2.1.9 Strategy = softmax, start_state = $(0, 4)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\beta = 0.01$, Learning Rate: 0.1, γ: 0.8
- **Policy Learnt:** From the state-visit count plot it can be seen that the agent visits the top left corner frquently. So it has learnt to reach the upper left corner goal state. Also from the Q-value heat-map it is clear that the Q-values near the goal states(for the optimal actions) are high.

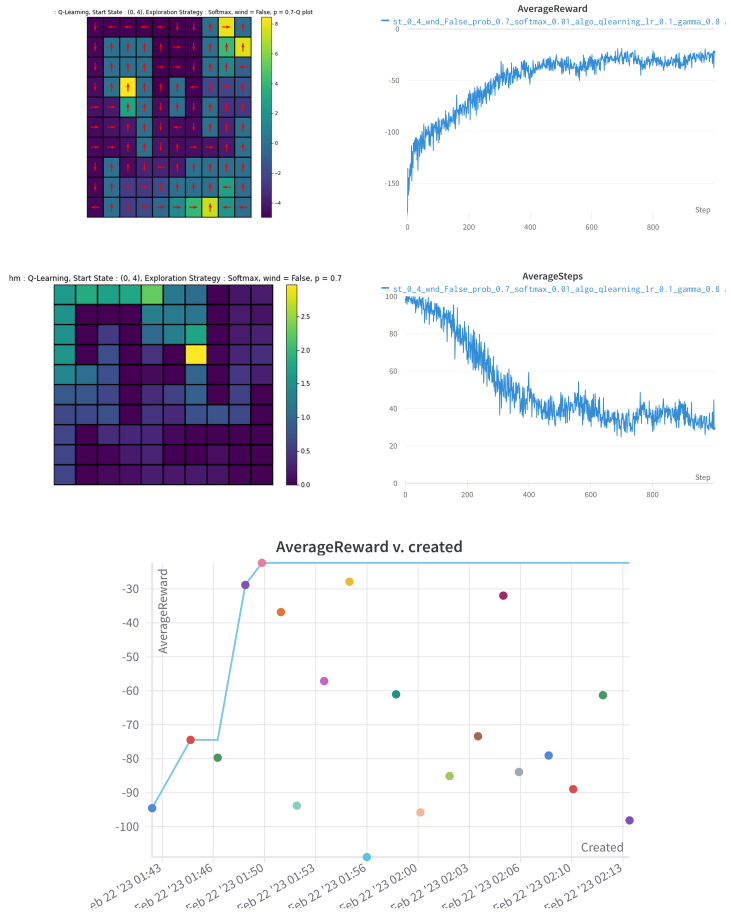

Figure 10: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` False, `P:` 0.7, `Strategy:` Softmax, τ: 0.01, `Learning Rate:` 0.1, γ: 0.8

2.1.10 Strategy = Softmax, start_state = $(0, 4)$, Wind = False, p = 1

- **Best Hyperparameters:** $\beta = 1$, Learning Rate: 1, γ: 1
- **Policy Learnt:** The agent visits the upper left corner very often. The agent has learnt to reach the upper left corner goal state. It also visits the bottom right goal state quite often.

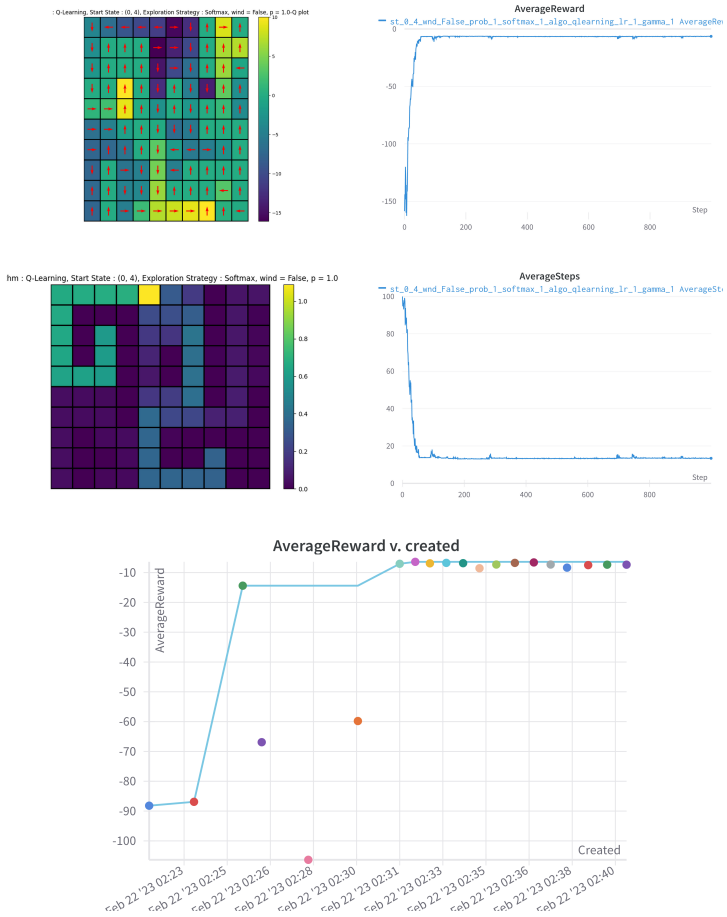

Figure 11: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` False, `P:` 1, `Strategy:` `Softmax`, τ`:` 1, `Learning Rate:` 1, γ`:` 1

2.1.11 Strategy = Softmax, start_state = $(3, 6)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\beta = 0.1$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count plot it is clear that the agent visits the bottom right cells of the grid very often. It has learnt to reach the bottom right corner goal state.

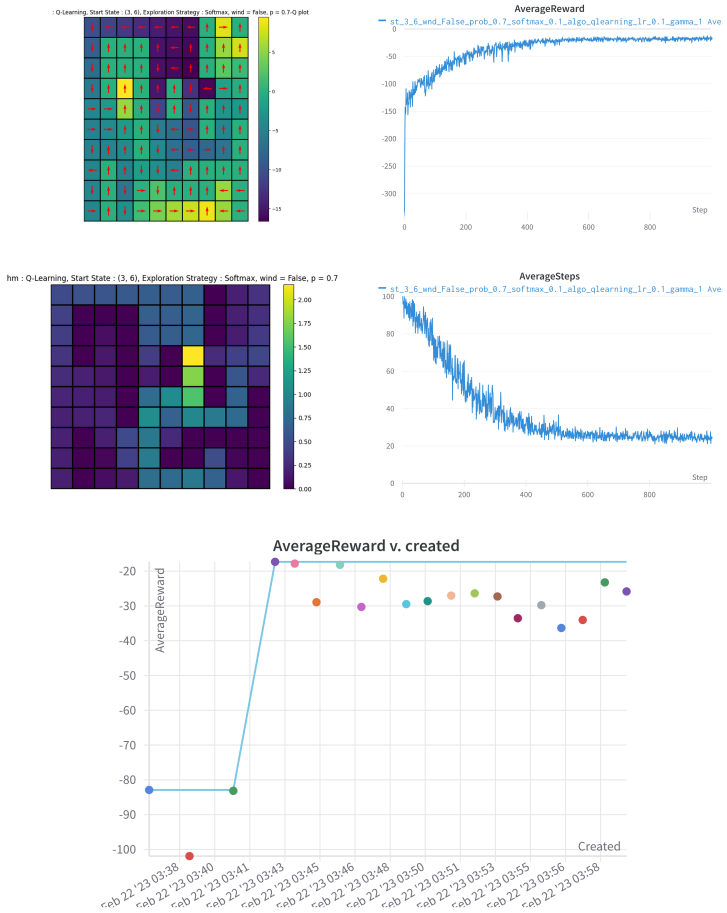

Figure 12: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` False, `P:` 0.7, `Strategy:` Softmax, τ`:` 0.1, `Learning Rate:` 0.1, γ`:` 1

2.1.12 Strategy = Softmax, start_state = $(3, 6)$, Wind = False, p = 1

- **Best Hyperparameters:** $\beta = 1$, Learning Rate: 1, γ: 1
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent visits bottom right goal state very frequently. It has also learnt to reach top right corner goal state.

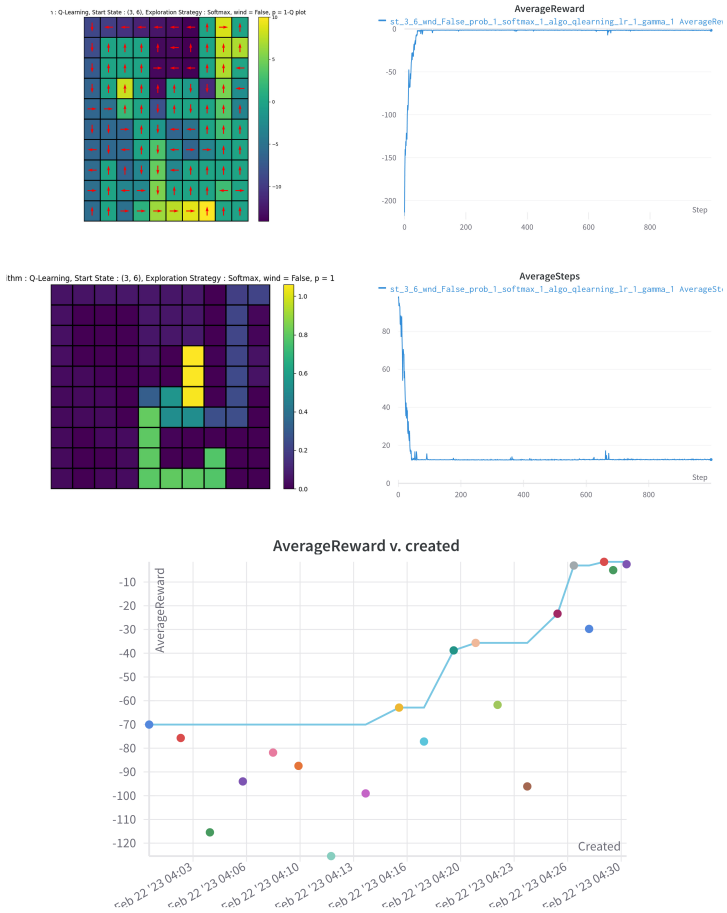

Figure 13: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` False, `P:` 1, `Strategy:` Softmax, τ: 1, `Learning Rate:` 1, γ: 1

2.1.13 Strategy = Softmax, start_state = $(0, 4)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\beta = 2$, Learning Rate: 1, γ: 1
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent does not visit any of the goal states very frequently. Although it seems it visits the top right corner goal relatively more often than the other two goal states.

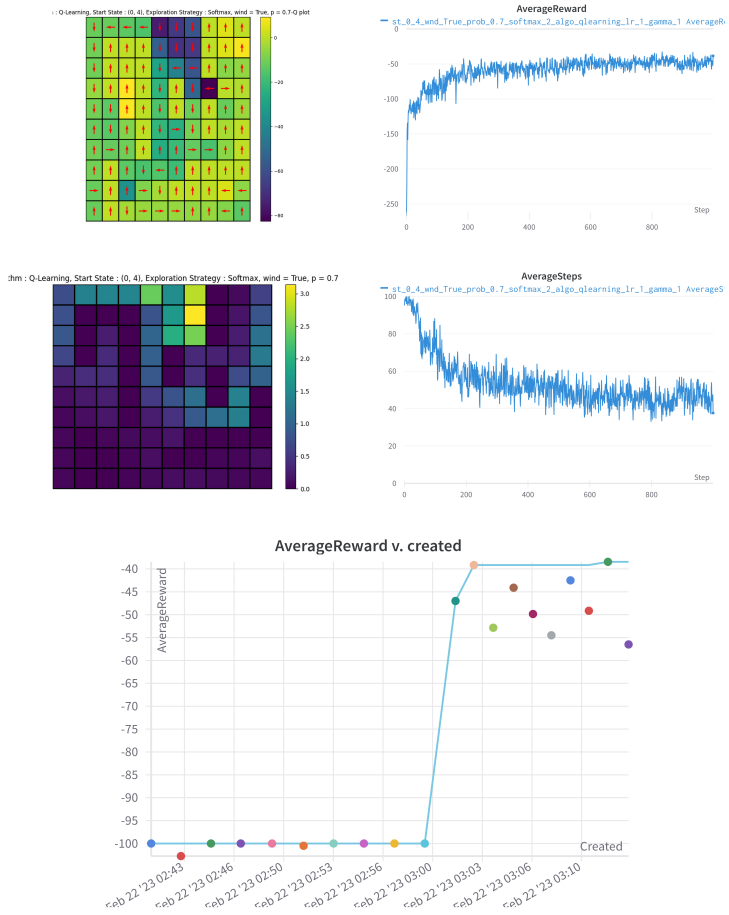

Figure 14: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` True, `P:` 0.7, `Strategy:` Softmax, τ`:` 2, `Learning Rate:` 1, γ`:` 1

2.1.14 Strategy = Softmax, start_state = $(0, 4)$, Wind = True, p = 1

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state- visit count heat-map it is clear that the agent visits the top left corner very frequently. It has learnt to reach the upper left corner goal state.

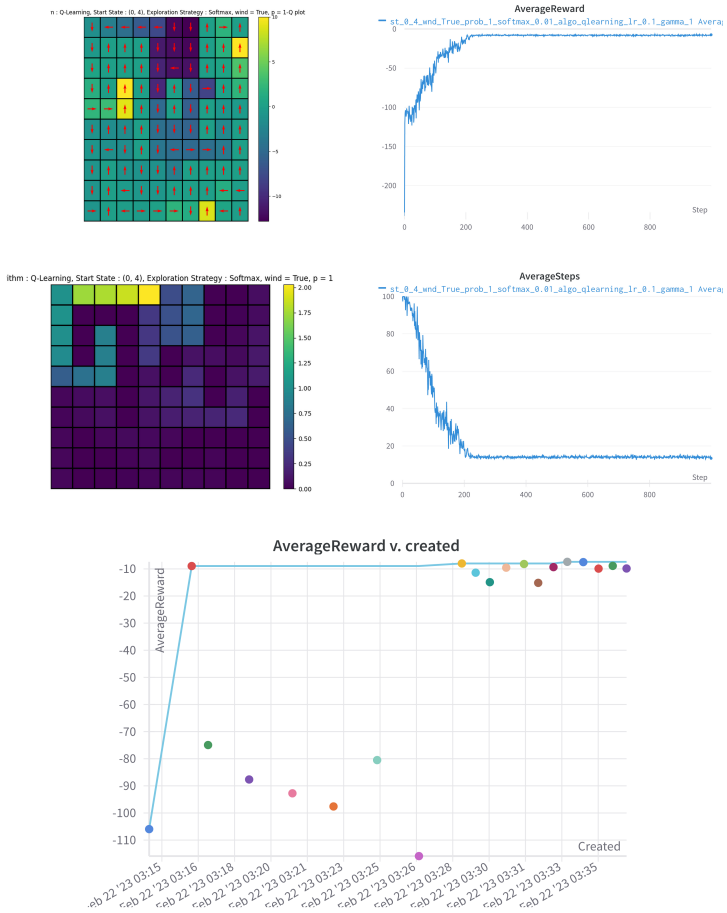

Figure 15: `Algorithm:` Q-Learning, `State:` $(0, 4)$, `Wind:` True, `P:` 1, `Strategy:` Softmax, τ: 0.01, `Learning Rate:` 0.1, γ: 1

2.1.15 Strategy = Softmax, start_state = $(3, 6)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\beta = 1$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count plot it is clear that the agent has learnt to visit the top right corner goal state. It hardly visits the other two goal states.

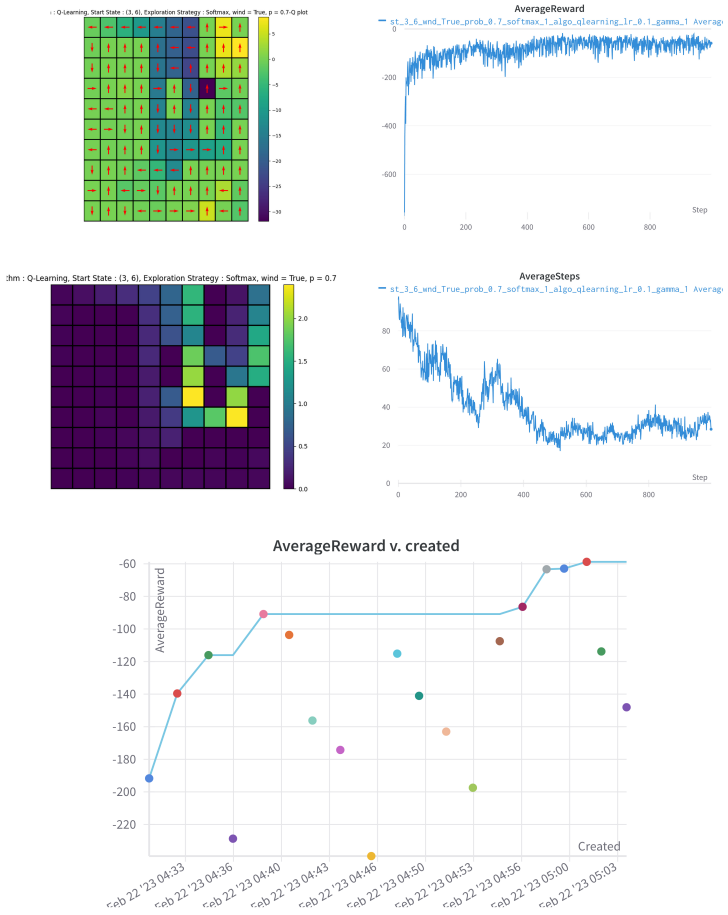

Figure 16: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` True, `P:` 0.7, `Strategy:` Softmax, τ: 1, `Learning Rate:` 0.1, γ: 1

2.1.16 Strategy = Softmax, start_state = $(3, 6)$, Wind = True, p = 1

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** The agent has learnt to visit the top right corner goal state. It rarely visits the other two goal states.

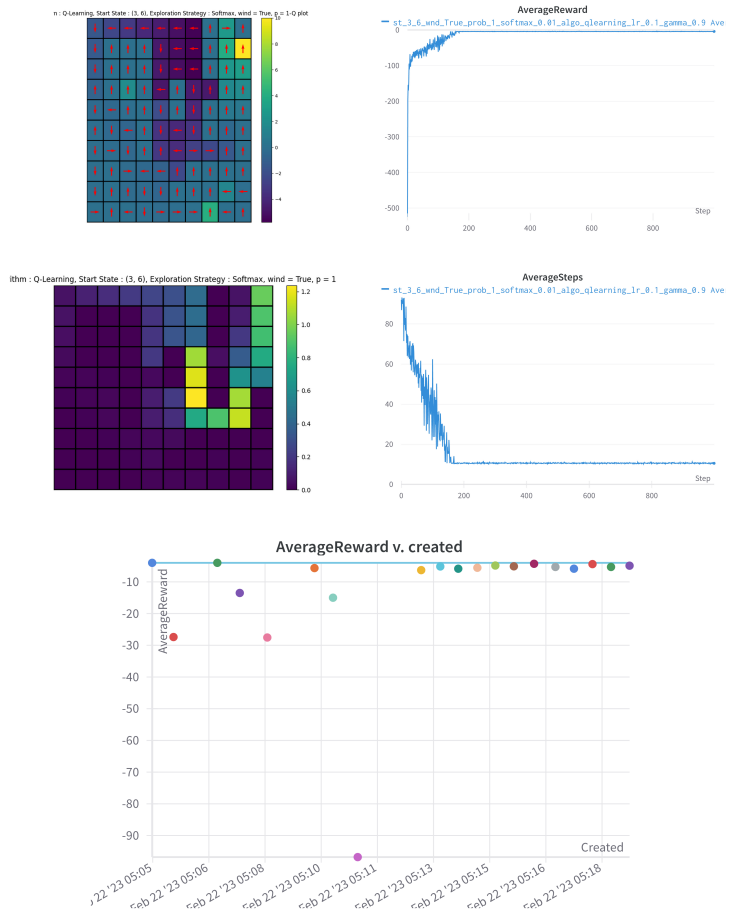

Figure 17: `Algorithm:` Q-Learning, `State:` $(3, 6)$, `Wind:` True, `P:` 1, `Strategy:` Softmax, τ: 0.01, `Learning Rate:` 0.1, γ: 0.9

2.2 Algorithm: SARSA

2.2.1 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 0.8
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent visits the upper left corner very often. It has learnt to reach the top left corner goal state. It rarely visits the other two goal states.

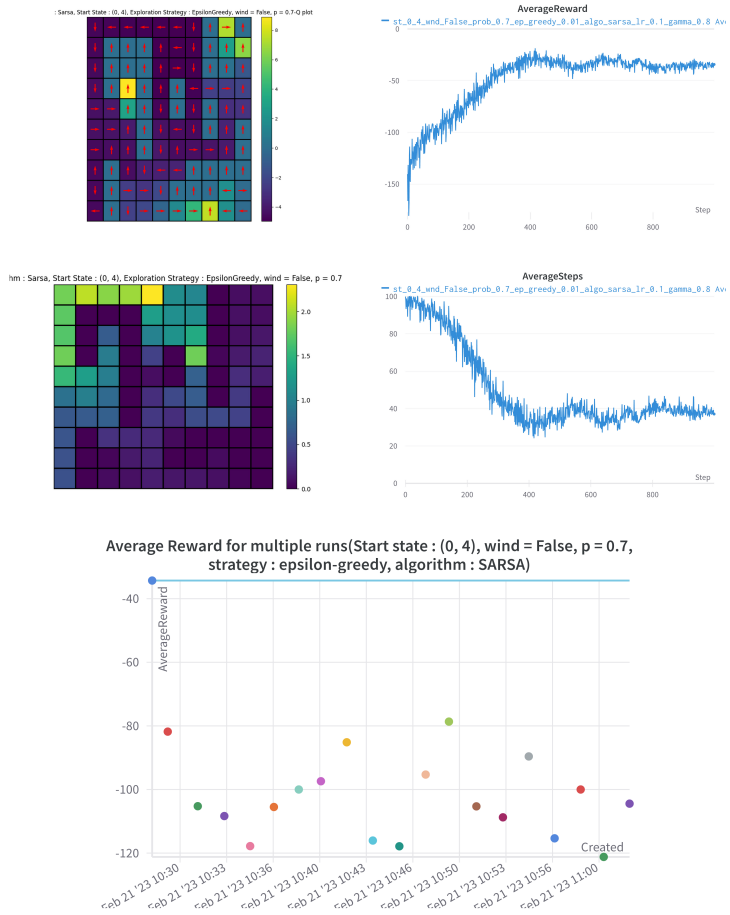

Figure 18: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` False, `P:` 0.7, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 0.8

2.2.2 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = False, p = 1

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent visits the upper left corner very often. It has learnt to reach the top left corner goal state. It rarely visits the other two goal states.

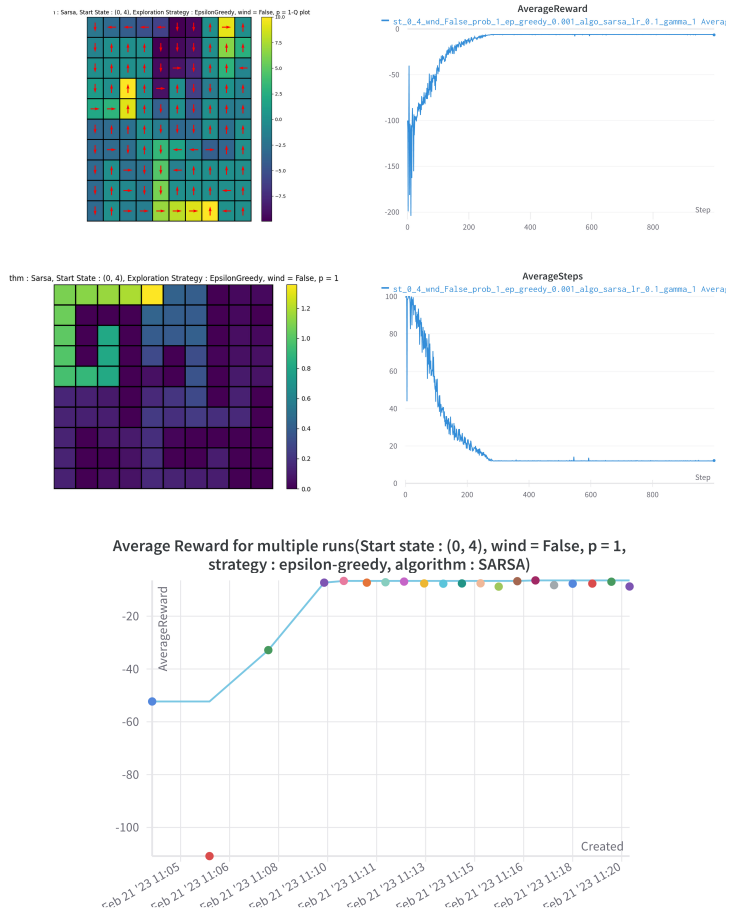

Figure 19: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` False, `P:` 1, `Strategy:` ϵ-`greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 1

2.2.3 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent does not visit any of the three goal states very frequently. However it does visit the bottom right goal state more times than the other two goal states.

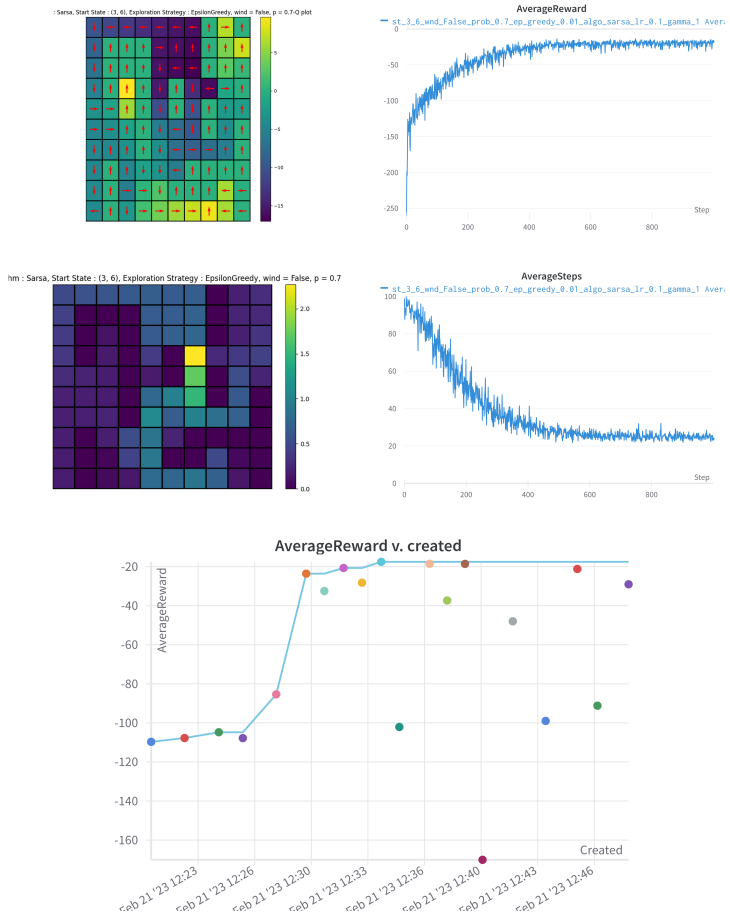

Figure 20: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` False, `P:` 0.7, `Strategy:` ϵ-`greedy`, ϵ: 0.01, `Learning Rate:` 0.1, γ: 1

2.2.4 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = False, p = 1

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** From the state-visit count heat-map it is clear that the agent visits the bottom right corner goal state much more frequently than the other two goal states. So it has learnt to reach the bottom right goal state.

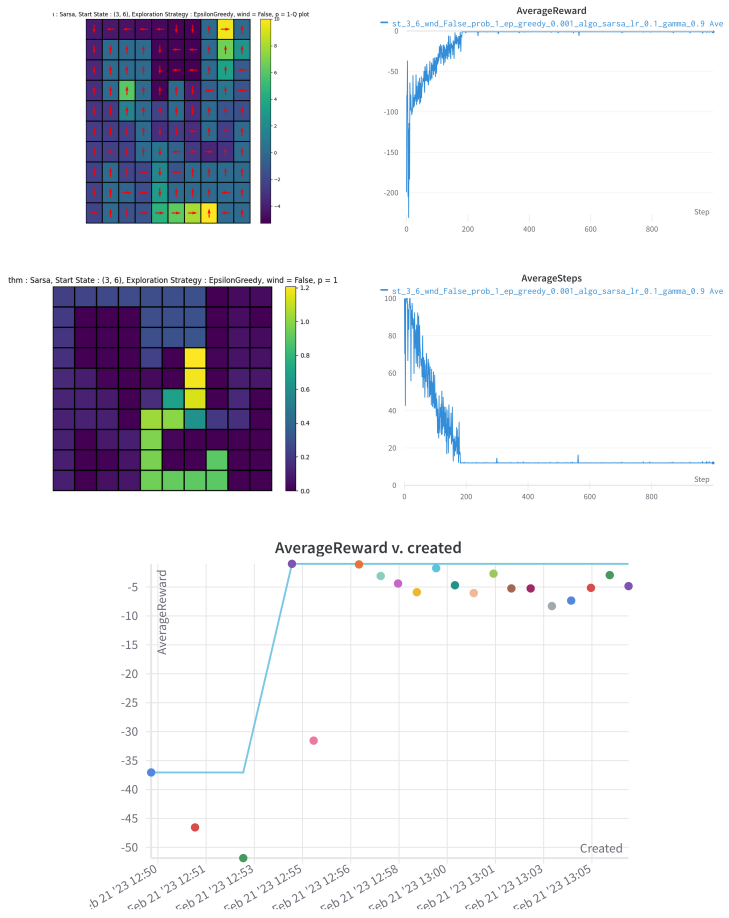

Figure 21: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` False, `P:` 1, `Strategy:` ϵ`-greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 0.9

2.2.5 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits upper left corner very often. However it does not visit any of the three goal states frequently.

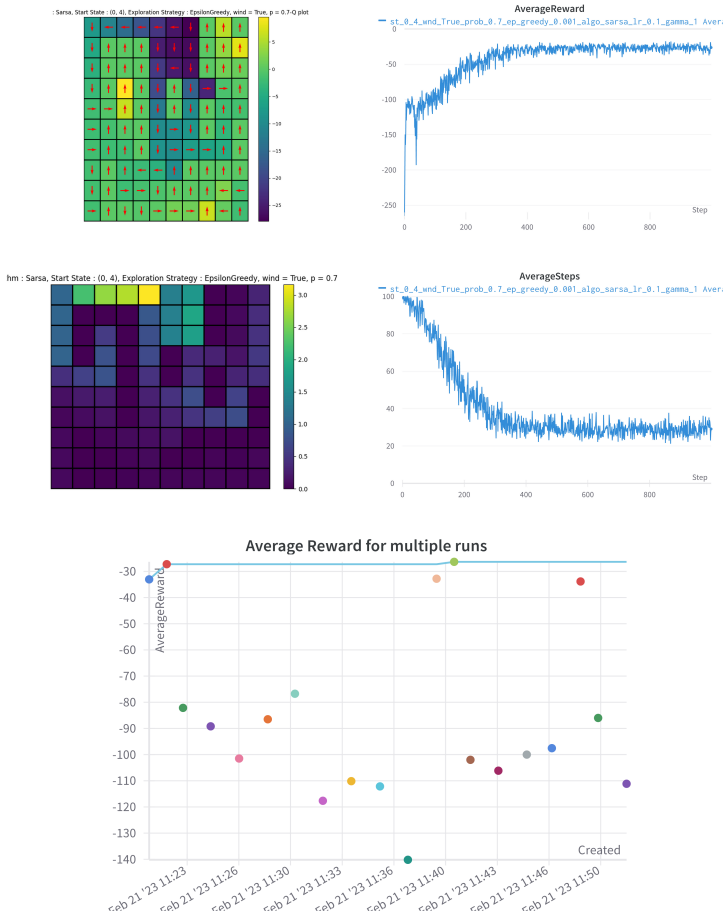

Figure 22: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` True, `P:` 0.7, `Strategy:` ϵ-`greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 1

2.2.6 Strategy = ϵ-greedy, start_state = $(0, 4)$, Wind = True, p = 1

- **Best Hyperparameters:** $\epsilon = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits the top left corner very frequently. It has learnt to reach the upper left corner goal state.

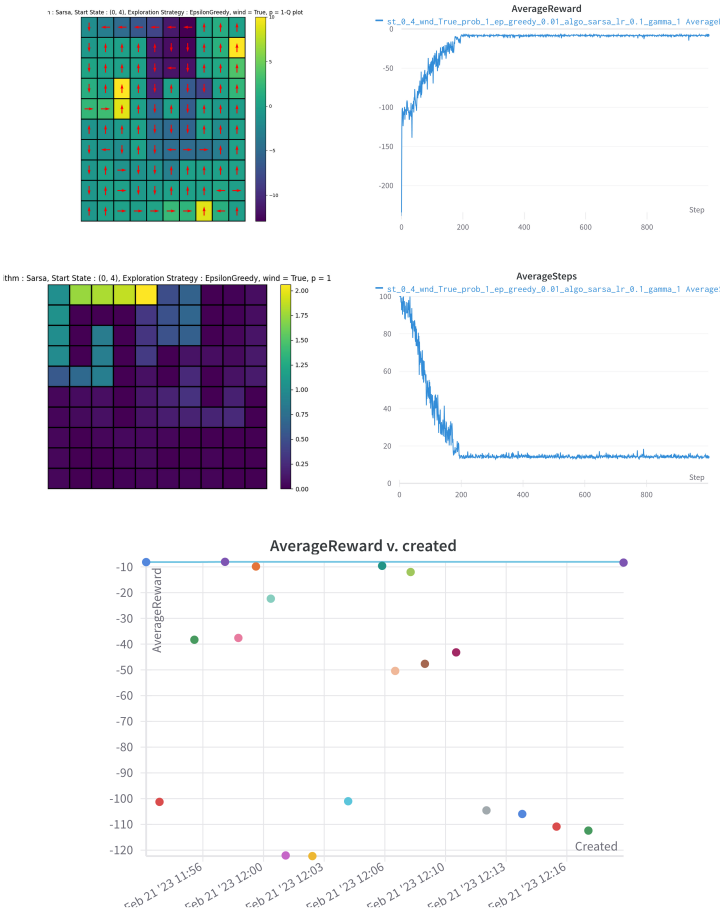

Figure 23: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` True, `P:` 1, `Strategy:` ϵ`-greedy`, ϵ`:` 0.01, `Learning Rate:` 0.1, γ`:` 1

2.2.7 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits the top right goal state more times than the other two goal states. However it does not visit any of the goal states as frequently as one would expect.

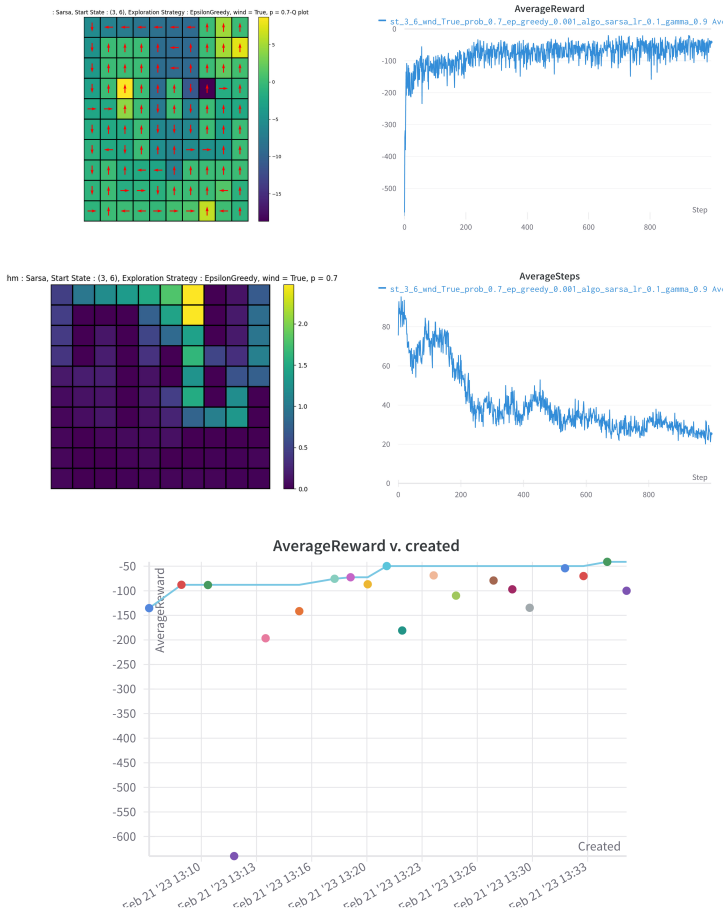

Figure 24: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` True, `P:` 0.7, `Strategy:` ϵ-`greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 0.9

2.2.8 Strategy = ϵ-greedy, start_state = $(3, 6)$, Wind = True, p = 1

- **Best Hyperparameters:** $\epsilon = 0.001$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits top right goal state much more frequently than the other two goal states. It rarely visits the other two goal states. So the agent has learnt to reach the upper right corner goal state.

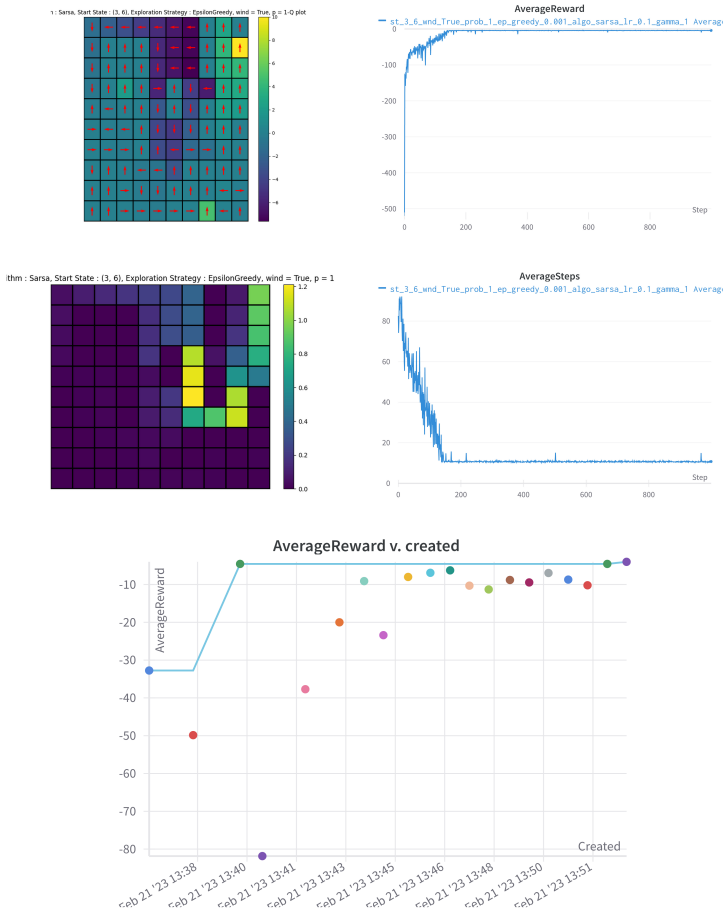

Figure 25: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` True, `P:` 1, `Strategy:` ϵ-`greedy`, ϵ`:` 0.001, `Learning Rate:` 0.1, γ`:` 1

2.2.9 Strategy = Softmax, start_state = $(0, 4)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits the upper left corner very frequently. It has learnt to reach top left goal state.

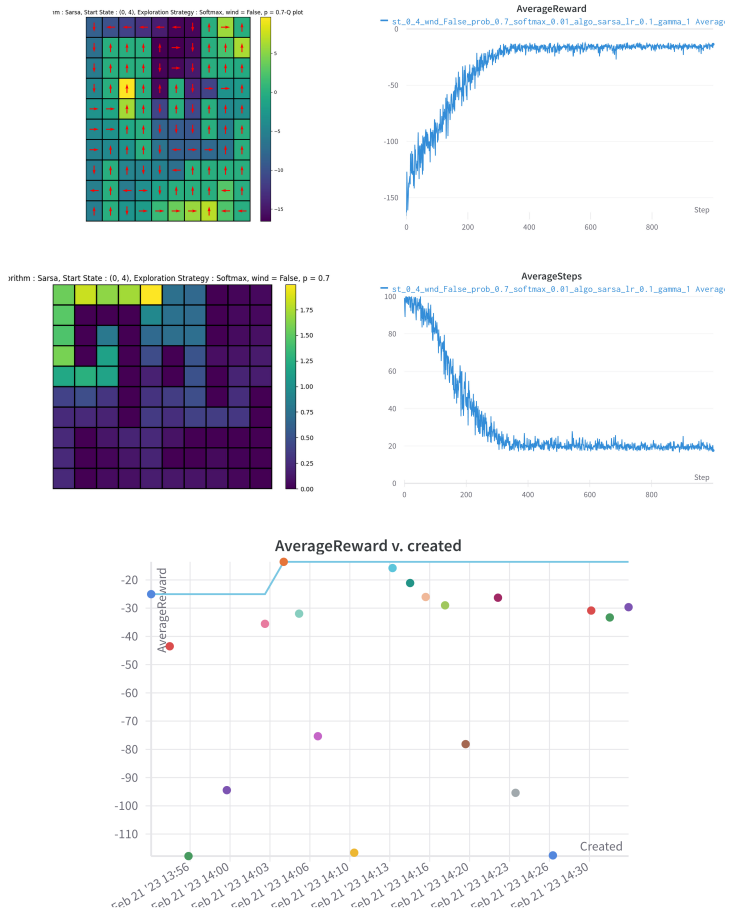

Figure 26: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` False, `P:` 0.7, `Strategy:` `Softmax`, τ: 0.01, `Learning Rate:` 0.1, γ: 1

2.2.10 Strategy = Softmax, start_state = $(0, 4)$, Wind = False, p = 1

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 0.7
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits the bottom right goal state very often. It has learnt to reach the bottom right goal state. It also visits the upper left corner of the grid.

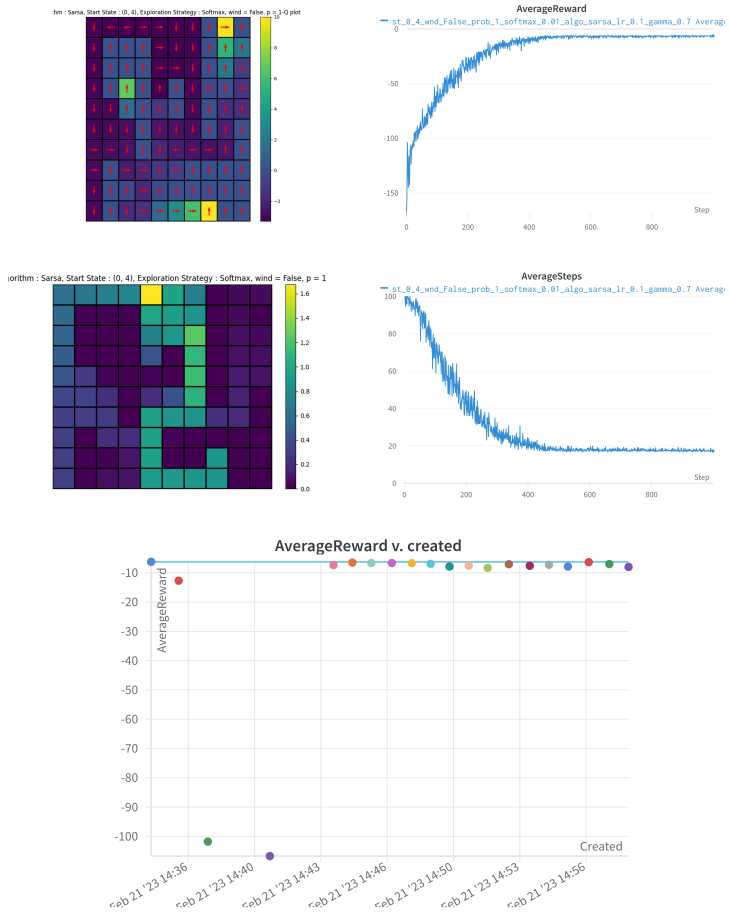

Figure 27: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` False, `P:` 1, `Strategy:` `Softmax`, τ`:` 0.01, `Learning Rate:` 0.1, γ`:` 0.7

2.2.11 Strategy = Softmax, start_state = $(3, 6)$, Wind = False, p = 0.7

- **Best Hyperparameters:** $\tau = 0.1$, Learning Rate: 1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent does not visit any of the goal states as frequently as one would expect. However it visits the top left and top right goal states more times than the bottom right goal state.

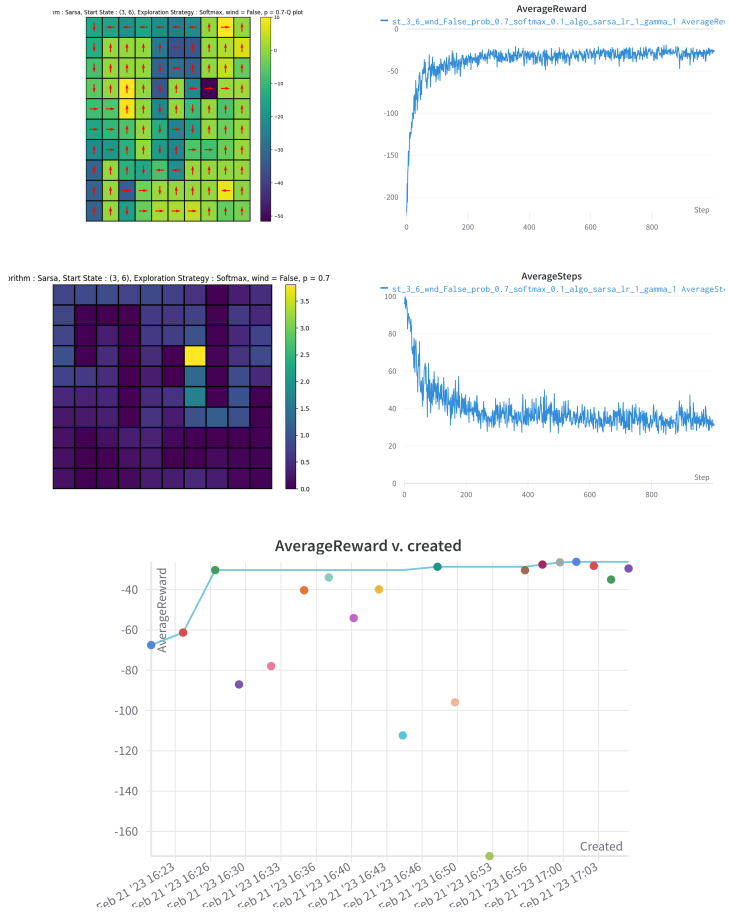

Figure 28: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` False, `P:` 0.7, `Strategy:` `Softmax`, τ: 0.1, `Learning Rate:` 1, γ: 1

2.2.12 Strategy = Softmax, start_state = $(3, 6)$, Wind = False, p = 1

- **Best Hyperparameters:** $\tau = 0.1$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent has learnt to reach the bottom right goal state.

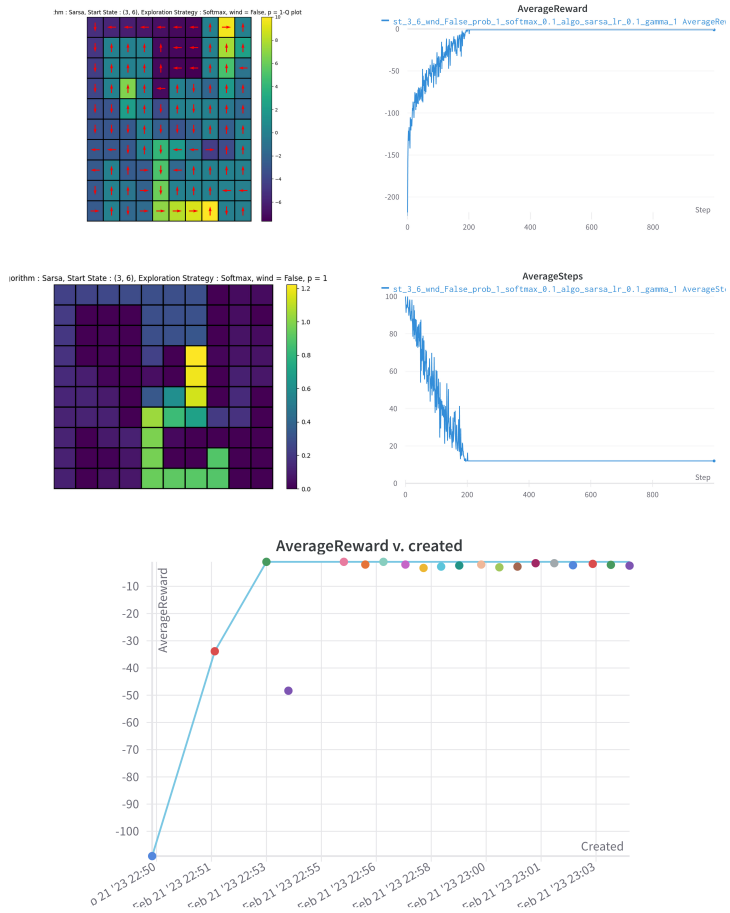

Figure 29: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` False, `P:` 1, `Strategy:` `Softmax`, τ`:` 0.1, `Learning Rate:` 0.1, γ`:` 1

2.2.13 Strategy = Softmax, start_state = $(0, 4)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** From the state-visit count heatmap it is clear that the agent visits the top left corner very frequently. However it does not visit the top left goal state very often.

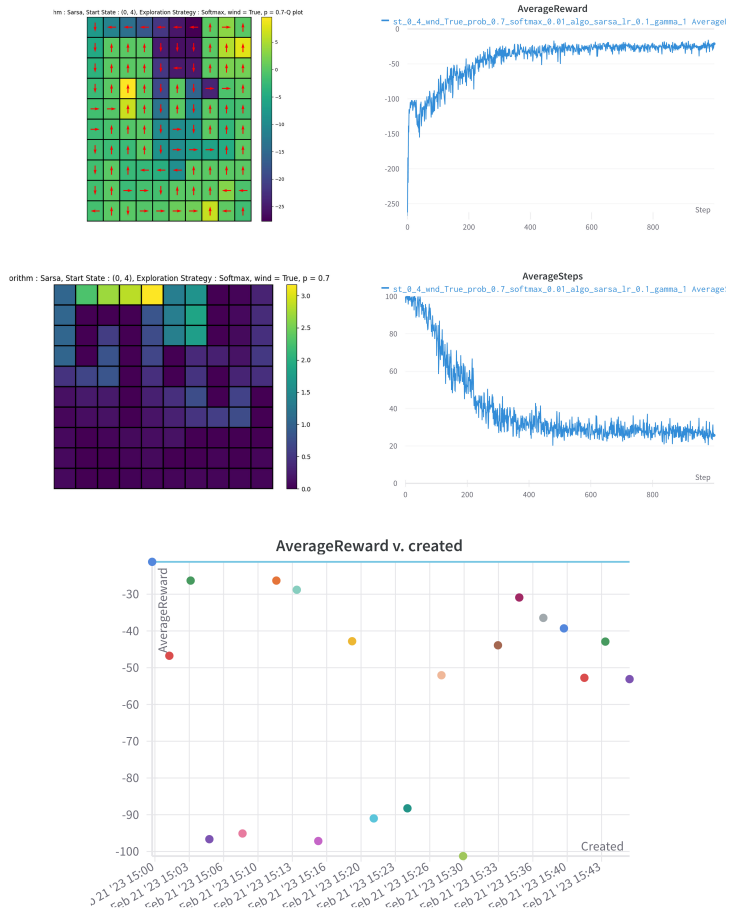

Figure 30: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` True, `P:` 0.7, `Strategy:` `Softmax`, τ`:` 0.01, `Learning Rate:` 0.1, γ`:` 1

2.2.14 Strategy = Softmax, start_state = $(0, 4)$, Wind = True, p = 1

- **Best Hyperparameters:** $\tau = 0.01$, Learning Rate: 0.1, γ: 0.9
- **Policy Learnt:** The agent visits the top left corner very frequently. It has learnt to reach the top left goal state.

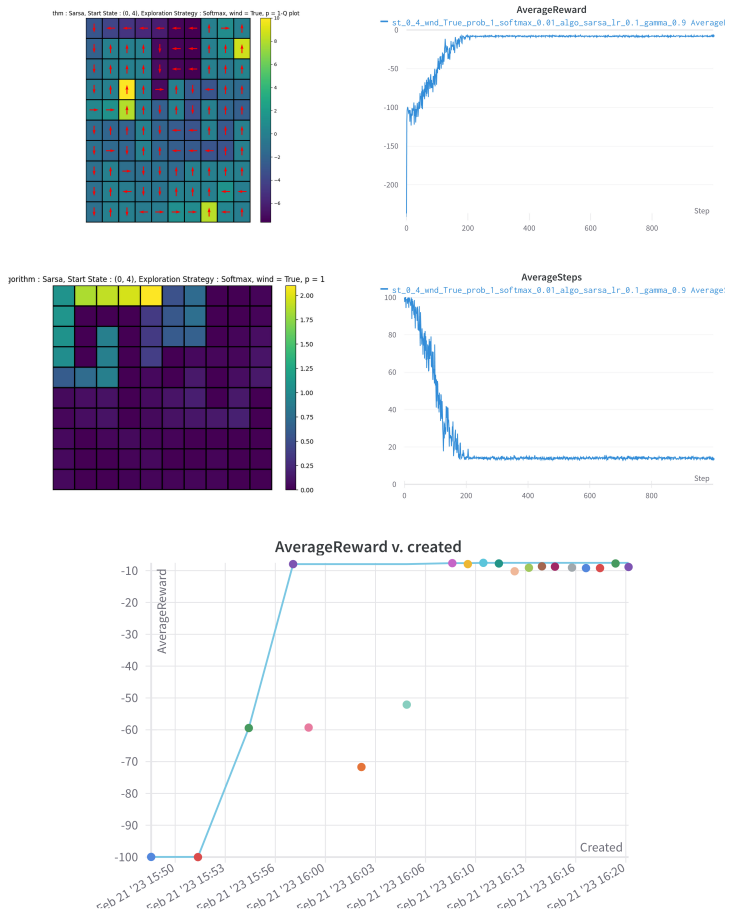

Figure 31: `Algorithm:` SARSA, `State:` $(0, 4)$, `Wind:` True, `P:` 1, `Strategy:` `Softmax`, τ: 0.01, `Learning Rate:` 0.1, γ: 0.9

2.2.15 Strategy = Softmax, start_state = $(3, 6)$, Wind = True, p = 0.7

- **Best Hyperparameters:** $\tau = 0.1$, Learning Rate: 0.1, γ: 1
- **Policy Learnt:** The agent has learnt to reach the top right corner goal state. It visits the top right goal state much more frequently than the other two goal states.

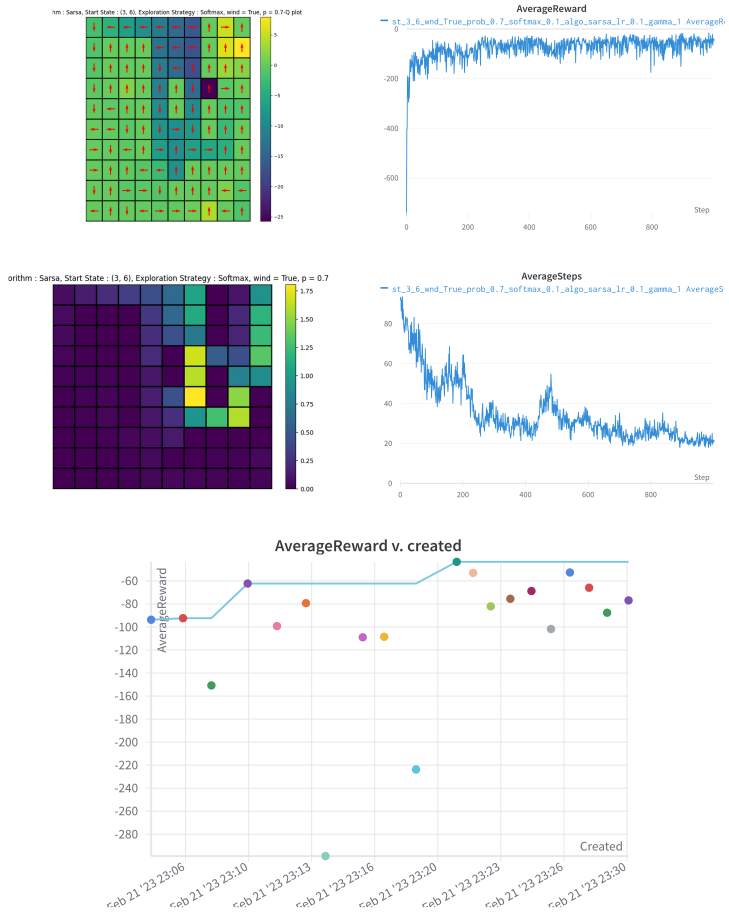

Figure 32: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` True, `P:` 0.7, `Strategy:` `Softmax`, τ`:` 0.1, `Learning Rate:` 0.1, γ`:` 1

2.2.16 Strategy = Softmax, start_state = $(3, 6)$, Wind = True, p = 1

- **Best Hyperparameters:** $\tau = 1$, Learning Rate: 1, γ: 1
- **Policy Learnt:** The agent has learnt to reach the top right goal state. Almost all the time it ends up visiting that state.

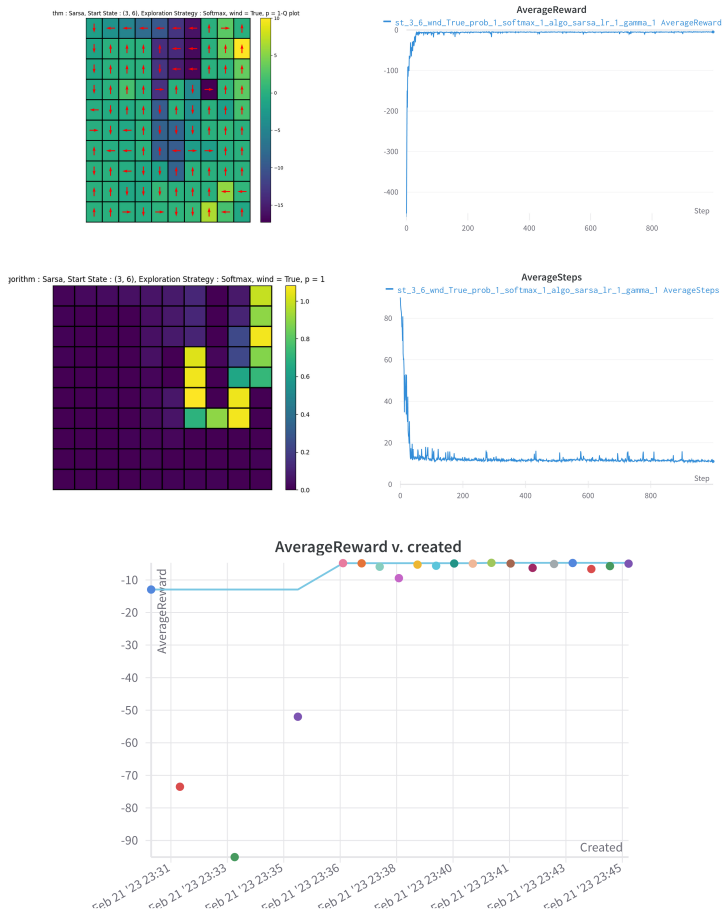

Figure 33: `Algorithm:` SARSA, `State:` $(3, 6)$, `Wind:` True, `P:` 1, `Strategy:` `Softmax`, τ: 1, `Learning Rate:` 1, γ: 1