CS6700 Programming Assignment-1

K Kamalesh Kumar(CE20B054), Kamalesh M(CE20B055)

24th February 2023

In this report, we compare and analyze the performance of two Temporal Difference learning methods, SARSA and Q learning, in a grid world environment considering multiple variations with respect to the start states, environment stochasticities and variations of the dynamics function p.

In general, we observed Q-learning to achieve better final performance and sharper convergence than SARSA, and SARSA learned a more conservative policy and was less likely to take a lower return path when compared to Q-learning.

On a finer granularity, the presence of wind has an effect on the frequency of visits to different states the agents makes. This is clear from analyzing the heatmap of the state visit counts in both the cases. When starting from the state $(3, 6)$, the agent is more likely to visit the goal state $(0, 9)$. Whereas, when the starting state is $(0, 4)$, the agent, in most configurations, learns a policy that drives it to the goal state $(2, 2)$. The dynamics function p also affects the extent of exploration that occurs during training. The setting $p = 1$,limits the state visit counts of the majority of the states, and thus requires more exploration through increasing β for softmax in some cases, but no such changes were required when $\epsilon - greedy$ was used. Since the rewards are mostly negative in the majority of the states, setting γ closer to 1 aided the agent in learning a policy that preferred a shorter path. With respect to the learning rate α, lower values ensured stable learning of the value functions and thus better convergence.

1 SARSA

In this section, we present the results of the SARSA algorithm in the grid world under various environment conditions.

1.1 With Wind

1.1.1 Strategy - Epsilon Greedy

When the exploration policy is $\epsilon - greedy$,the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$

The above-mentioned hyperparameters were also consistent with a grid-search (Fig 1) that was performed with different configurations.

1.1.2 Strategy - SoftMax

When the exploration policy is Softmax, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$

1.2 Without Wind

1.2.1 Strategy - Epsilon Greedy

For the without wind case, when the exploration policy is $\epsilon - greedy$, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \epsilon = 0.01$

1.2.2 Strategy - SoftMax

For the without wind case, when the exploration policy is softmax, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 0.1$

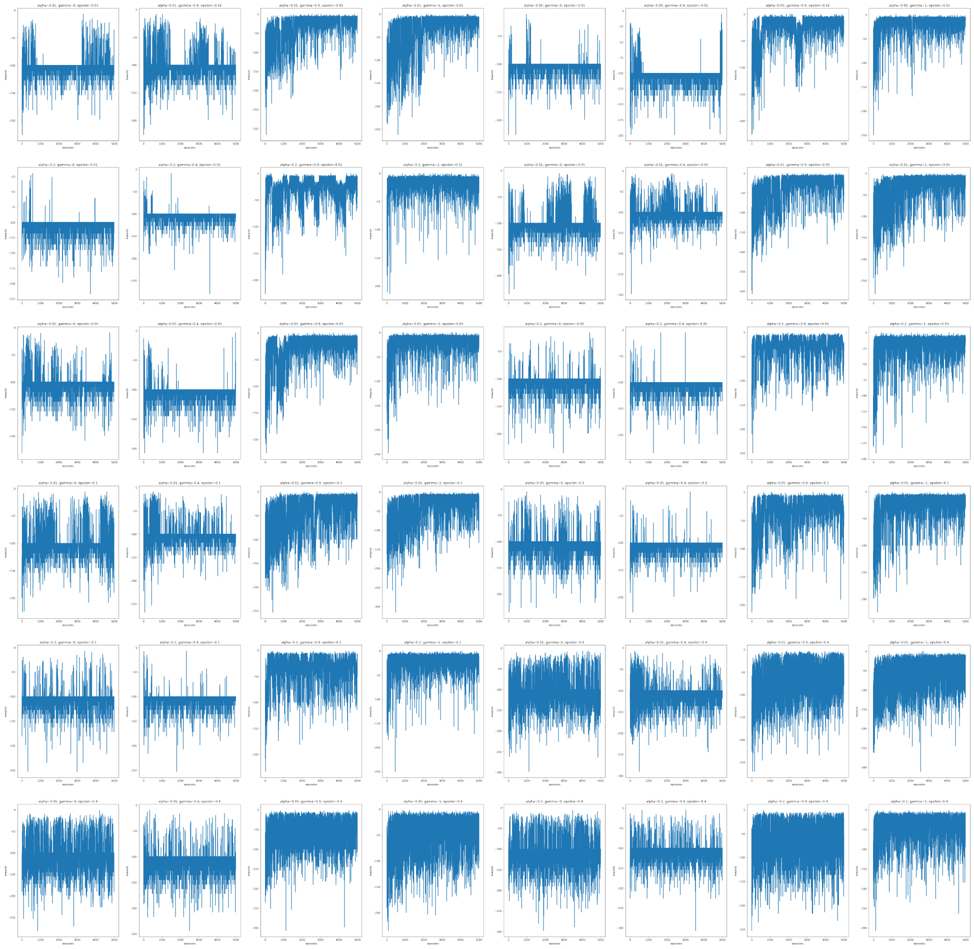

Figure 1: Example for one of the SARSA Grid search results conducted for a configuration with wind, $p = 0.7$ and start state as $(3, 6)$.

2 Q-Learning

In this section, we present the results of the Q-Learning algorithm in the grid world under various environment conditions.

2.1 With Wind

2.1.1 Strategy - Epsilon Greedy

With ϵ greedy, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$

2.1.2 Strategy - SoftMax

When the exploration policy is Softmax, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \beta = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \beta = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \beta = 1$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \beta = 1$

2.2 Without Wind

2.2.1 Strategy - Epsilon Greedy

With ϵ -greedy, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations, for the without wind case.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.2, \gamma = 1, \epsilon = 0.01$

2.2.2 Strategy - SoftMax

When the exploration policy is Softmax, for the case without wind, the following hyperparameters have been found experimentally best suited for the below-mentioned configurations.

- For $p = 0.7$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 0.01$
- For $p = 0.7$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 0.01$
- For $p = 1$ and start state $(3, 6)$: $\alpha = 0.05, \gamma = 1, \beta = 1$
- For $p = 1$ and start state $(0, 4)$: $\alpha = 0.05, \gamma = 1, \beta = 1$

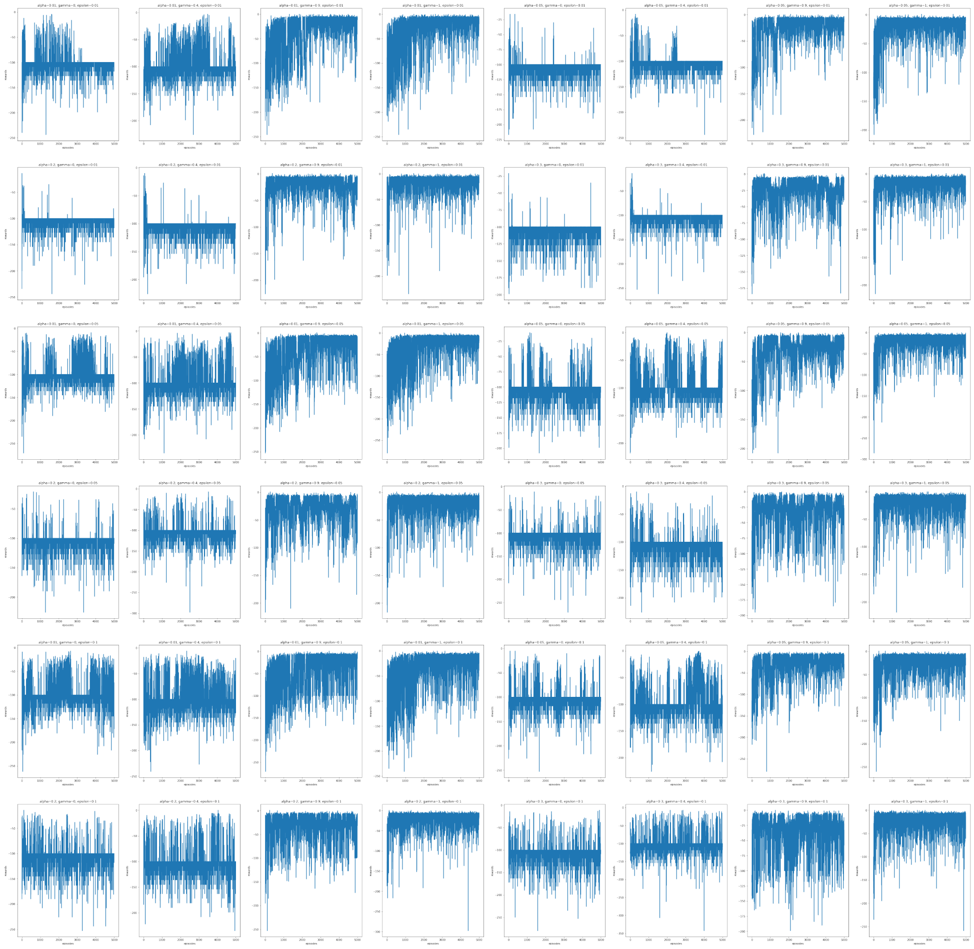

Figure 2: Example for one of the Q-learning Grid search results conducted for a configuration with wind, $p = 0.7$ and start state as $(3, 6)$.