Reinforcement Learning Project Document

Pourya Aliannejadi, Mona Saghafi

July 3, 2023

Introduction

In this document our goal is to analysis the result of application in different parameters.

Task 1

In this task we have to report the optimal policy and the reward of solving the problem. because the map is not slippery so we have only one optimal policy.

Optimal Policy:
0: LEFT, 1: RIGHT, 2: RIGHT, 3: RIGHT, 4: RIGHT, 5: RIGHT, 6: RIGHT, 7: LEFT.

Now we have to test the agent with different Gama(Discount Factor).

When Gama is equal to 1, the agent places equal importance on immediate and future rewards so because of the equality the agent can not decide which action to take so it will take infinite amount of time for the agent to make a choice so the problem will never get solved.

As the Gama goes towards the number 0 the agent starts to act more greedy and when Gama is 0, we have an greedy agent that only cares about immediate rewards.

So for the Gama of 1 the agent will never move but for the other Gamas that are less than 0 but not 0 the agent will solve the problem with 5 moves and the reward of 0.6.

Task 2

This task is like the last task but with a different map and different parameters.

The situation with Gama is just like before but the rewards and the moves change.
The agent solves the problem with 8 moves and the reward of 3.69 if Gama is $0 < Gama < 1$.

Task 3

In this task we have to report the optimal policy and the reward of solving the problem in both cases of map being slippery or not.

Not Slippery

In this case the agent solves the problem using the following policy:

Optimal Policy: 0: RIGHT, 1: DOWN, 2: DOWN, 3: DOWN, 4: DOWN, 5: LEFT, 6: DOWN, 7: DOWN, 8: DOWN, 9: DOWN, 10: LEFT, 11: DOWN, 12: DOWN, 13: DOWN, 14: DOWN, 15: LEFT, 16: DOWN, 17: DOWN, 18: DOWN, 19: DOWN, 20: LEFT, 21: LEFT, 22: LEFT, 23: LEFT, 24: LEFT.

in 6 moves with the reward of 2.5.

Slippery

In this case the problem is not deterministic anymore and there is chance that the agent might slip so the optimal policy changes and because of the non-determinism the moves required to solve the problem will be different each time with different rewards.

Optimal Policy: 0: UP, 1: RIGHT, 2: LEFT, 3: LEFT, 4: LEFT, 5: LEFT, 6: RIGHT, 7: DOWN, 8: DOWN, 9: DOWN, 10: LEFT, 11: RIGHT, 12: DOWN, 13: DOWN, 14: DOWN, 15: LEFT, 16: RIGHT, 17: DOWN, 18: DOWN, 19: DOWN, 20: LEFT, 21: DOWN, 22: LEFT, 23: LEFT, 24: DOWN.

Some examples of moves and rewards of the solved cases when slippery:
rewards: -1.0, moves: 13.
rewards: -20.0, moves: 51.
rewards: -12.0, moves: 35.
rewards: -4.5, moves: 20.

Task 4

Just like the last task but with a different map.

Not Slippery

In this case the agent solves the problem using the following policy:

Optimal Policy: 0: DOWN, 1: LEFT, 2: LEFT, 3: LEFT, 4: RIGHT, 5: DOWN, 6: DOWN, 7: DOWN, 8: LEFT, 9: LEFT, 10: LEFT, 11: LEFT, 12: DOWN, 13: DOWN, 14: RIGHT, 15: DOWN, 16: DOWN, 17: DOWN, 18: DOWN, 19: DOWN, 20: DOWN, 21: LEFT, 22: DOWN, 23: DOWN, 24: DOWN, 25: DOWN, 26: DOWN, 27: DOWN, 28: RIGHT, 29: LEFT, 30: LEFT, 31: LEFT, 32: LEFT, 33: LEFT, 34: LEFT.

in 8 moves with the reward of 1.5.

Slippery

In this case the problem is not deterministic anymore and there is chance that the agent might slip so the optimal policy changes and because of the non-determinism the moves required to solve the problem will be different each time with different rewards.

Optimal Policy: 0: LEFT, 1: UP, 2: UP, 3: UP, 4: UP, 5: RIGHT, 6: DOWN, 7: LEFT, 8: LEFT, 9: LEFT, 10: LEFT, 11: LEFT, 12: RIGHT, 13: DOWN, 14: UP, 15: DOWN, 16: DOWN, 17: DOWN, 18: DOWN, 19: DOWN, 20: DOWN, 21: LEFT, 22: RIGHT, 23: DOWN, 24: LEFT, 25: LEFT, 26: LEFT, 27: DOWN, 28: DOWN, 29: LEFT, 30: LEFT, 31: LEFT, 32: LEFT, 33: LEFT, 34: DOWN.

Some examples of moves and rewards of the solved cases when slippery:
rewards: -33.5, moves: 78.
rewards: -41.0, moves: 93.
rewards: -8.5,moves: 28.
rewards: -25.0, moves: 61.

Task 5

In this task we have to examine the changes made when we change the reward of each move. Because the map is not slippery the only thing that changes is the reward and end of each episode.

Optimal Policy: 0: LEFT, 1: RIGHT, 2: RIGHT, 3: RIGHT, 4: RIGHT, 5: RIGHT, 6: RIGHT, 7: LEFT.

in 5 moves and with different rewards.
When move reward is -4 the reward will be -9.
When move reward is -2 the reward will be -1.
When move reward is 0 the reward will be 7.
When move reward is 2 the reward will be 15.

Task 6

Just like the last task but with a different map.

Optimal Policy: 0: LEFT, 1: RIGHT, 2: RIGHT, 3: RIGHT, 4: RIGHT, 5: RIGHT, 6: RIGHT, 7: LEFT, 8: LEFT, 9: RIGHT, 10: RIGHT, 11: RIGHT, 12: RIGHT, 13: RIGHT, 14: RIGHT, 15: UP, 16: LEFT, 17: RIGHT, 18:

RIGHT, 19: RIGHT, 20: RIGHT, 21: RIGHT, 22: RIGHT, 23: UP.

in 5 moves and with different rewards.
When move reward is -4 the reward will be -9.
When move reward is -2 the reward will be -1.
When move reward is 0 the reward will be 7.
When move reward is 2 the reward will be 15.

Task 7

In this task, after finding the right policy we have to use Monte Carlo to do some state value evaluation.

Optimal Policy: 0: DOWN, 1: RIGHT, 2: RIGHT, 3: LEFT, 4: UP, 5: UP, 6: UP, 7: RIGHT, 8: LEFT, 9: LEFT, 10: LEFT, 11: LEFT, 12: RIGHT, 13: LEFT, 14: DOWN, 15: LEFT, 16: DOWN, 17: RIGHT, 18: LEFT, 19: LEFT, 20: DOWN, 21: DOWN, 22: RIGHT, 23: DOWN, 24: LEFT.

Because the map is slippery it will take different moves to solve the problem, some examples: rewards: -19.0, moves: 70.
rewards: -1.0, moves: 52.
rewards: 32.0, moves: 19.
rewards: -16.0, moves: 67.

Monte Carlo

The reason why Monte Carlo produces different result even thought Gamma has not changed is because of the stochastic nature of the environment so each time it will be different.

Task 8

In this task we have to change the way we implemented our polices and use new policies and compare them to see which one is better.

And as it can be seen in the plots, the Down policy is better because the states near the goal have more values and it is more likely for the agent to find the answer faster.

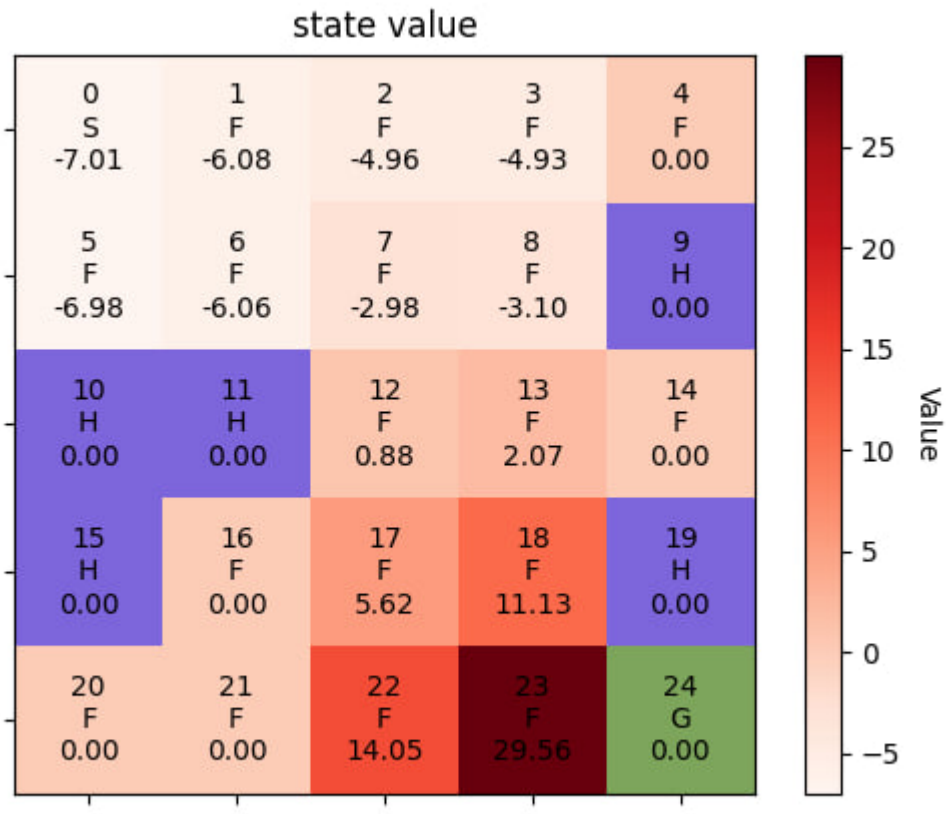

Figure 1: First Visit Monte Carlo with 500 episodes.

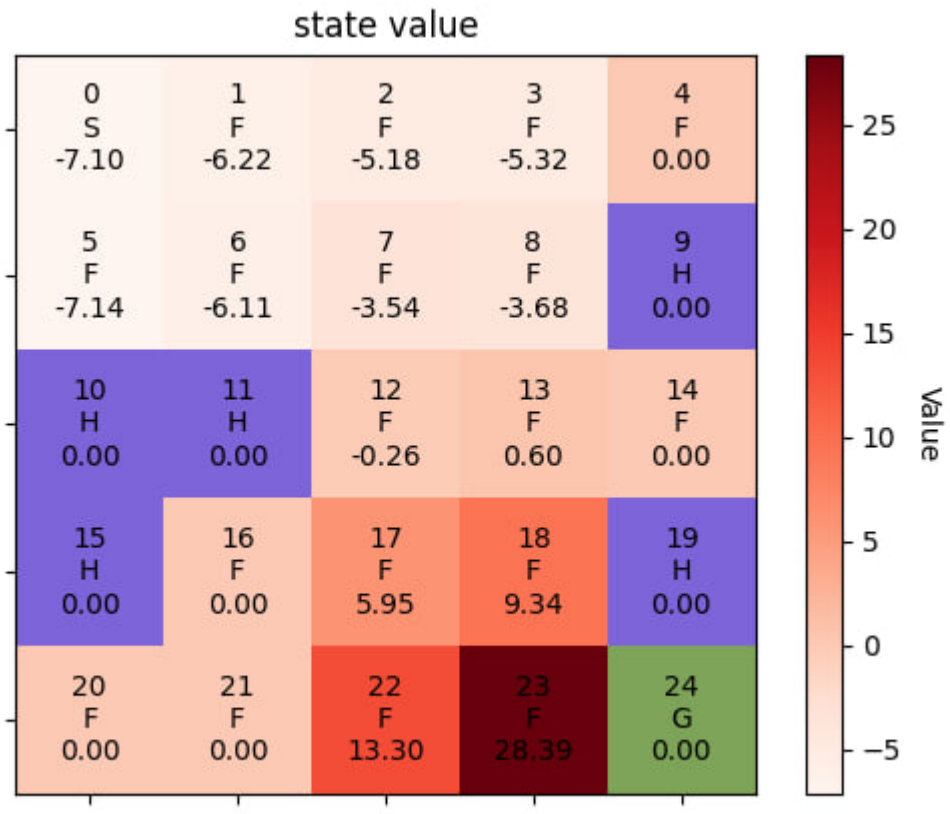

Figure 2: First Visit Monte Carlo with 5000 episodes.

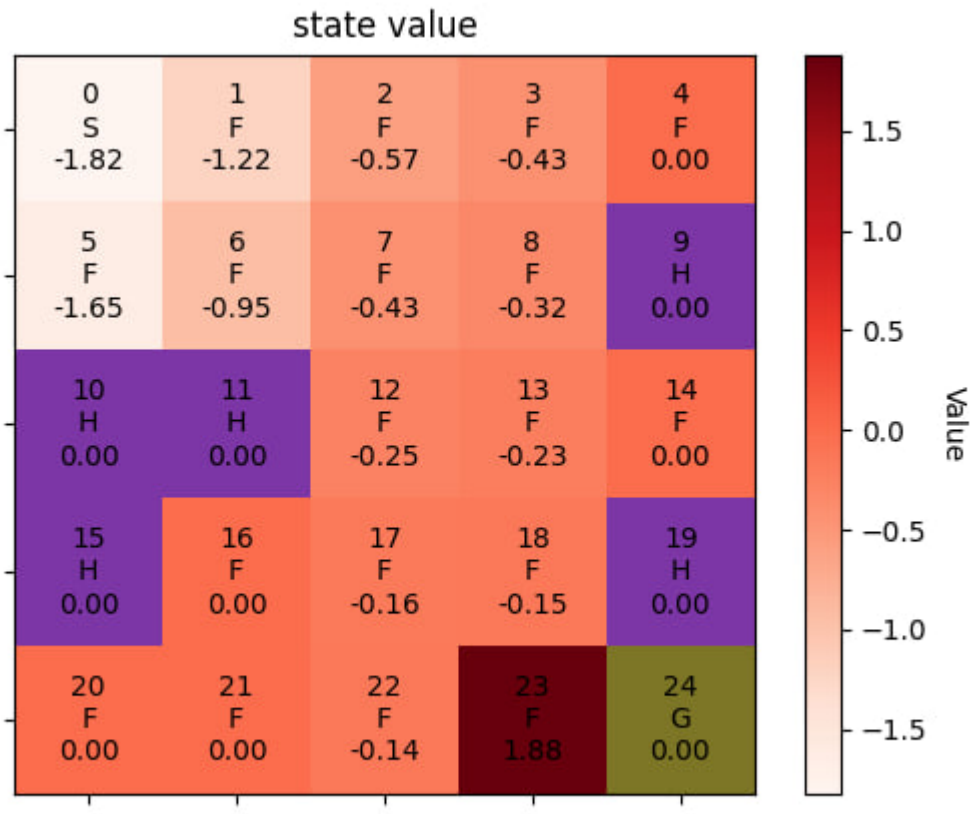

Figure 3: Every Visit Monte Carlo with 500 episodes.

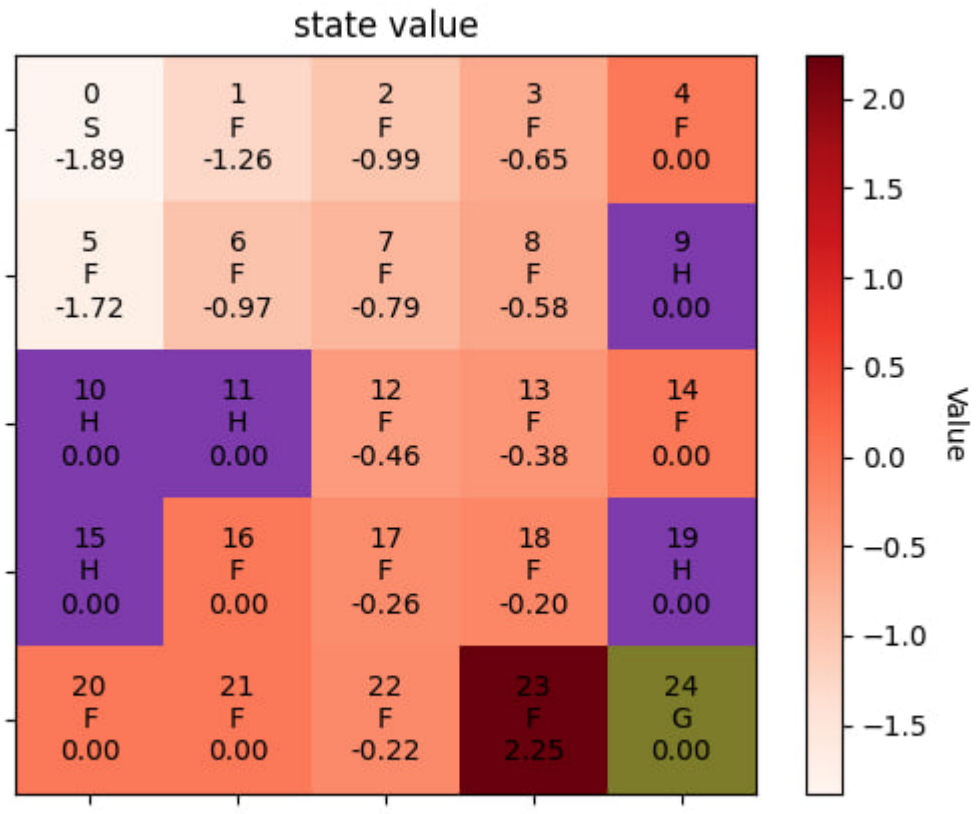

Figure 4: Every Visit Monte Carlo with 5000 episodes.

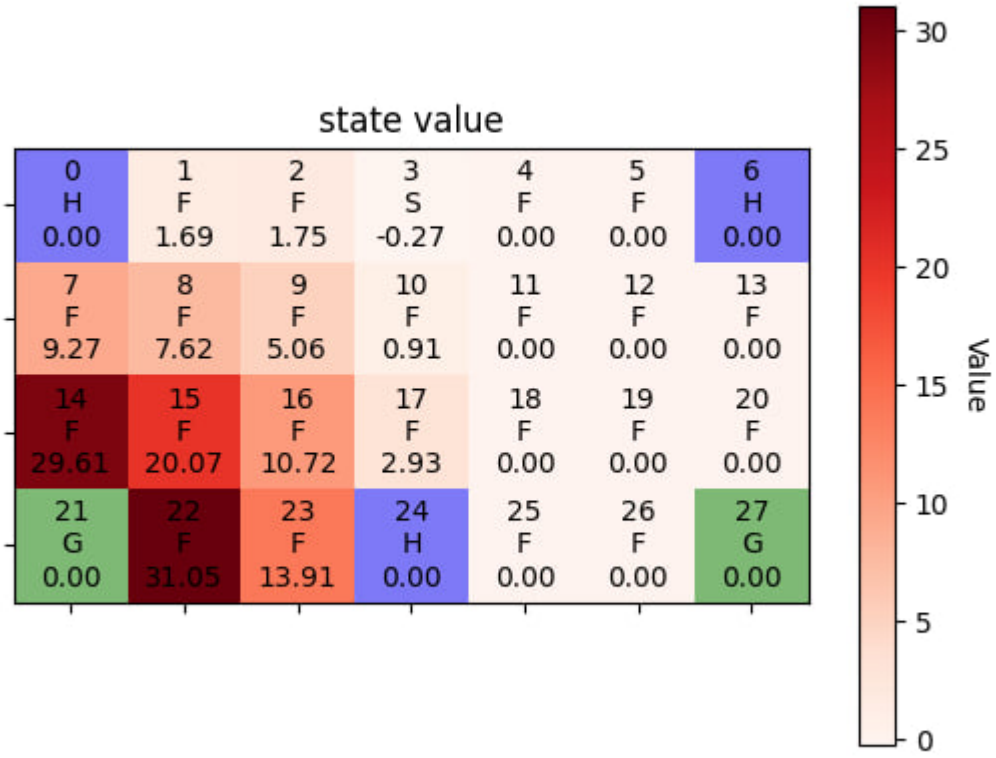

Figure 5: First Visit Monte Carlo with Left policy.

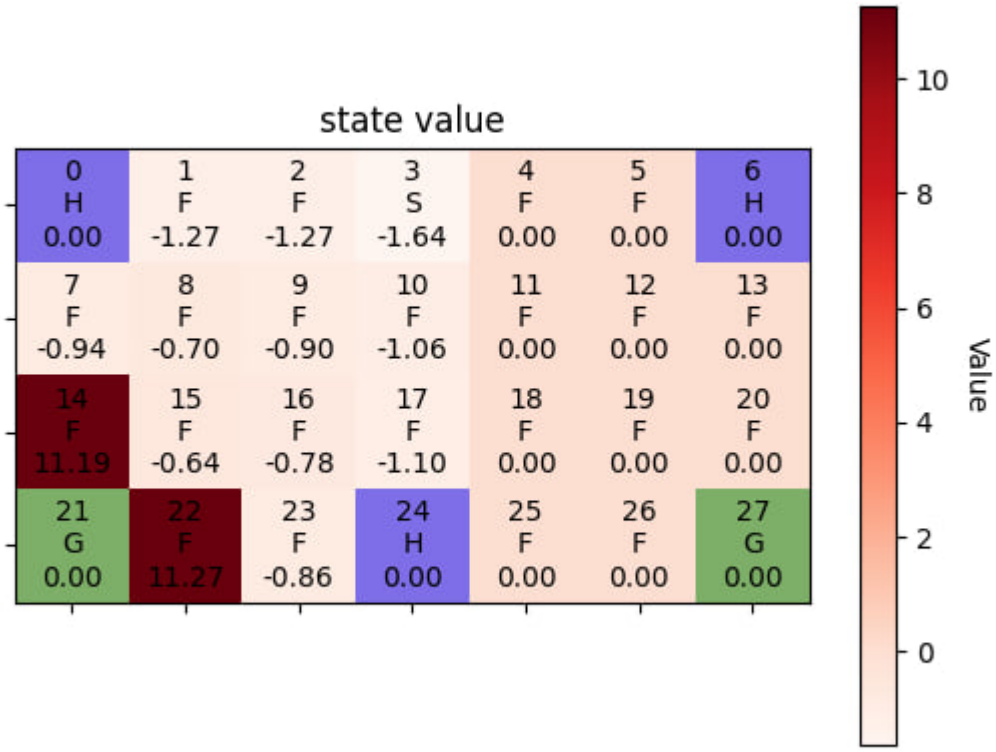

Figure 6: Every Visit Monte Carlo with Left policy.

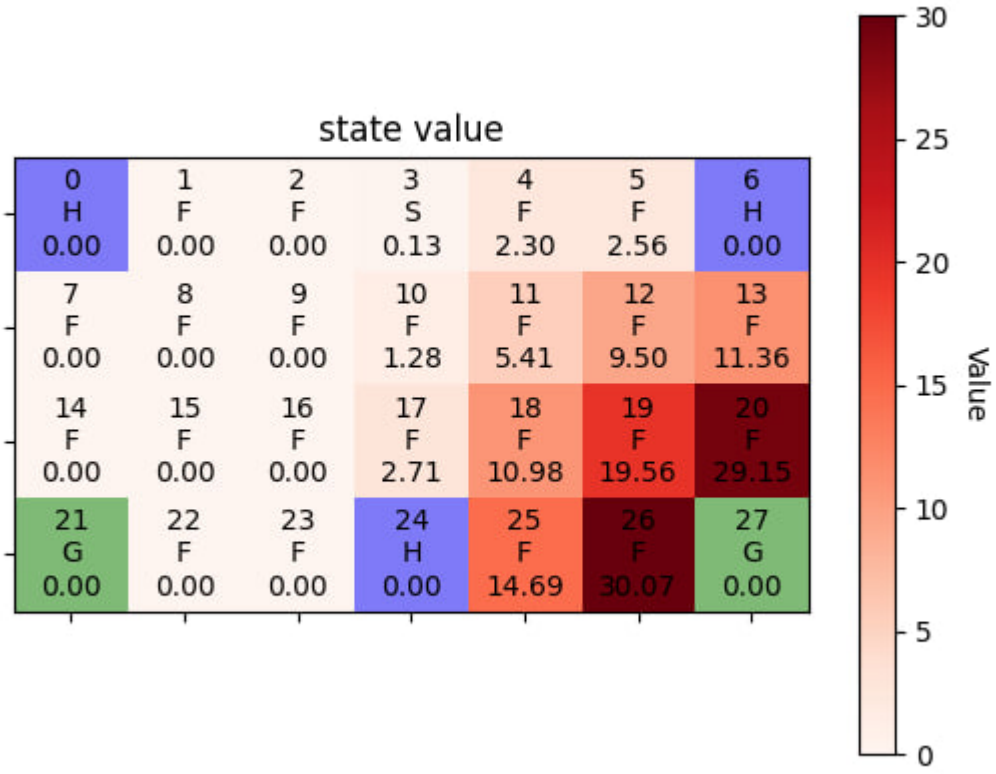

Figure 7: First Visit Monte Carlo with Right policy.

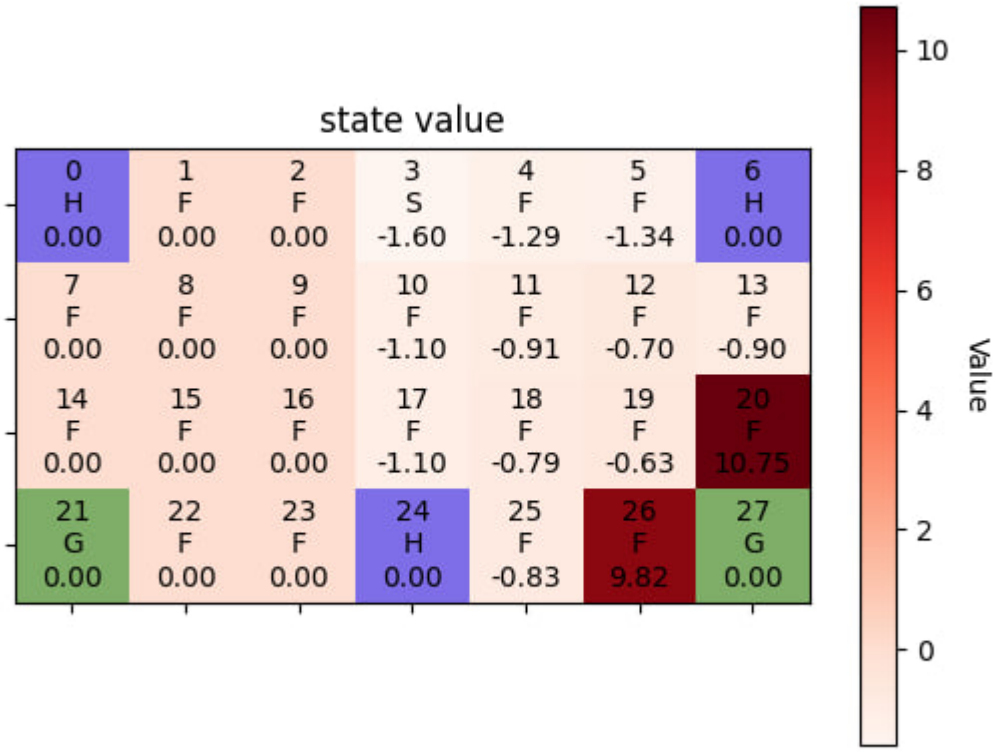

Figure 8: Every Visit Monte Carlo with Right policy.

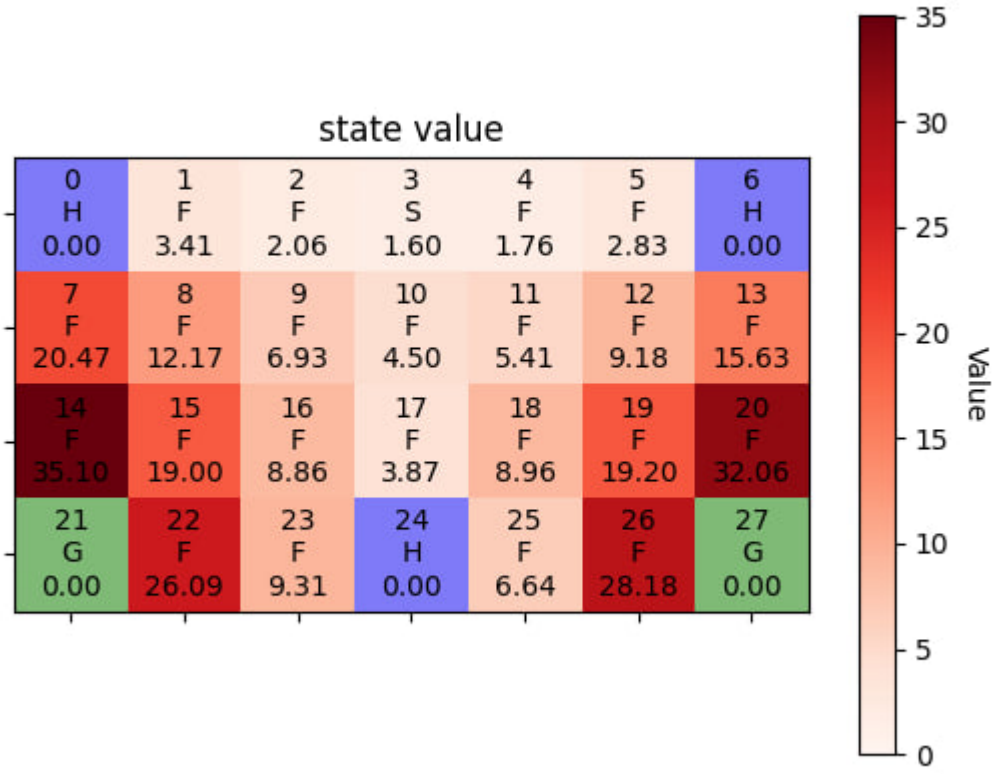

Figure 9: First Visit Monte Carlo with Down policy.

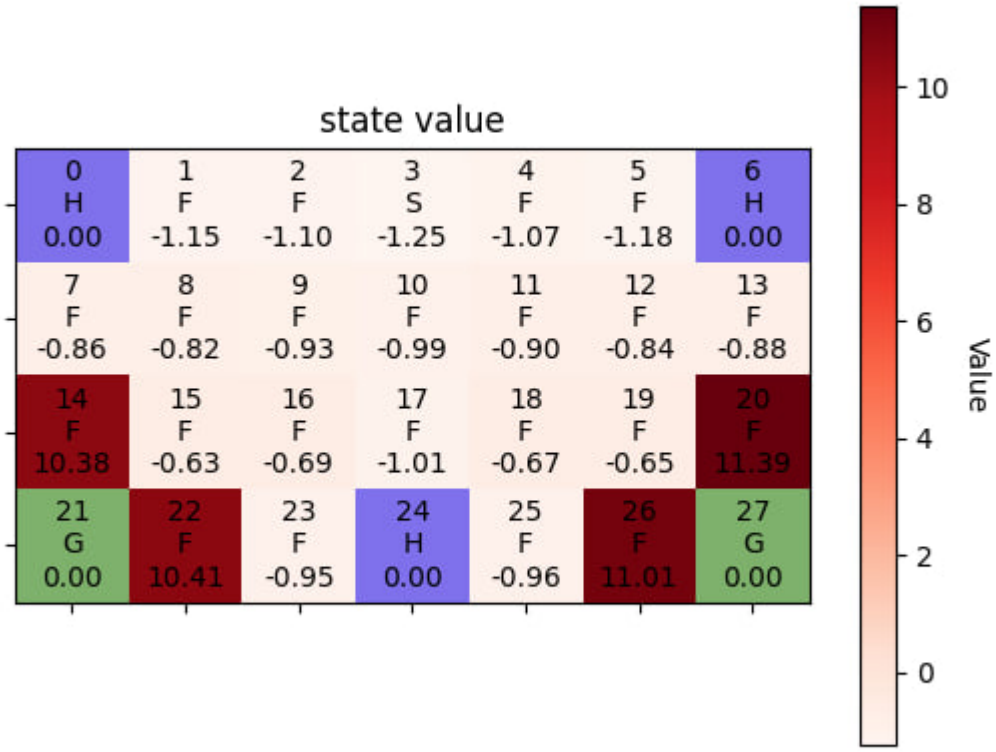

Figure 10: Every Visit Monte Carlo with Down policy.