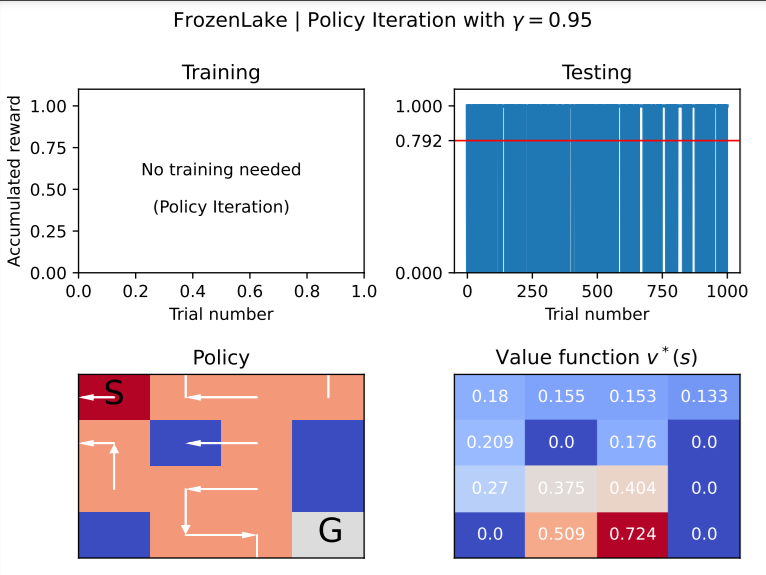
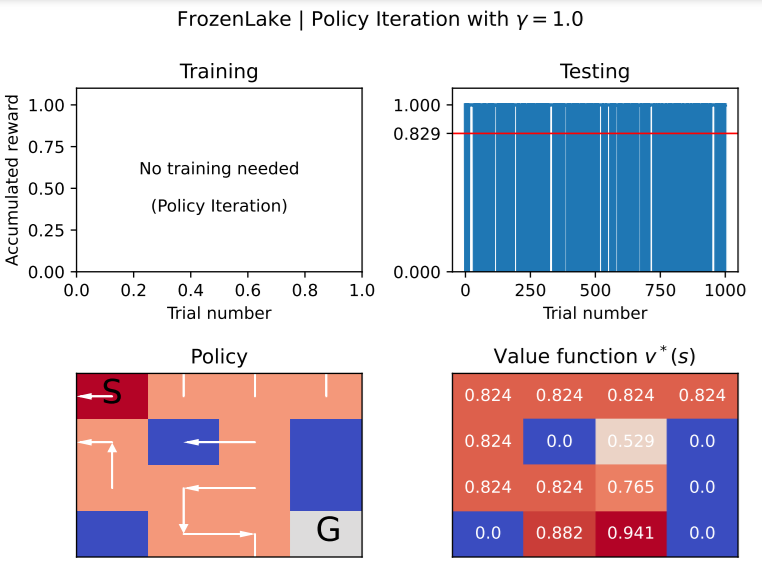
**Assignment 2**

Reinforcement Learning KU, WS 2024/25

|  |  |  |
| --- | --- | --- |
| **Team members** | | |
| Last name | First name | Matriculation Number |
| Srdjan | Stjepanovic | 12430077 |
|  |  |  |

2.) Planning in a Grid World





1. Policy
2. Gamma 0.95

The policy seems **more conservative**, especially in states where the risk of falling into a hole is higher.

1. Gamma 1.0

The policy is more consistent and aggressive, with higher action values even near the starting position. This reflects the policy's confidence in accounting for long-term rewards and achieving the goal.

1. Value function
2. Gamma 0.95

Values near the start state are small, which suggests the policy does not assign significant weight to **long-term rewards.** The value function increases more significantly near the goal state (0.509, 0.724), showing that the policy still recognizes the importance of reaching the goal but discounts future steps.

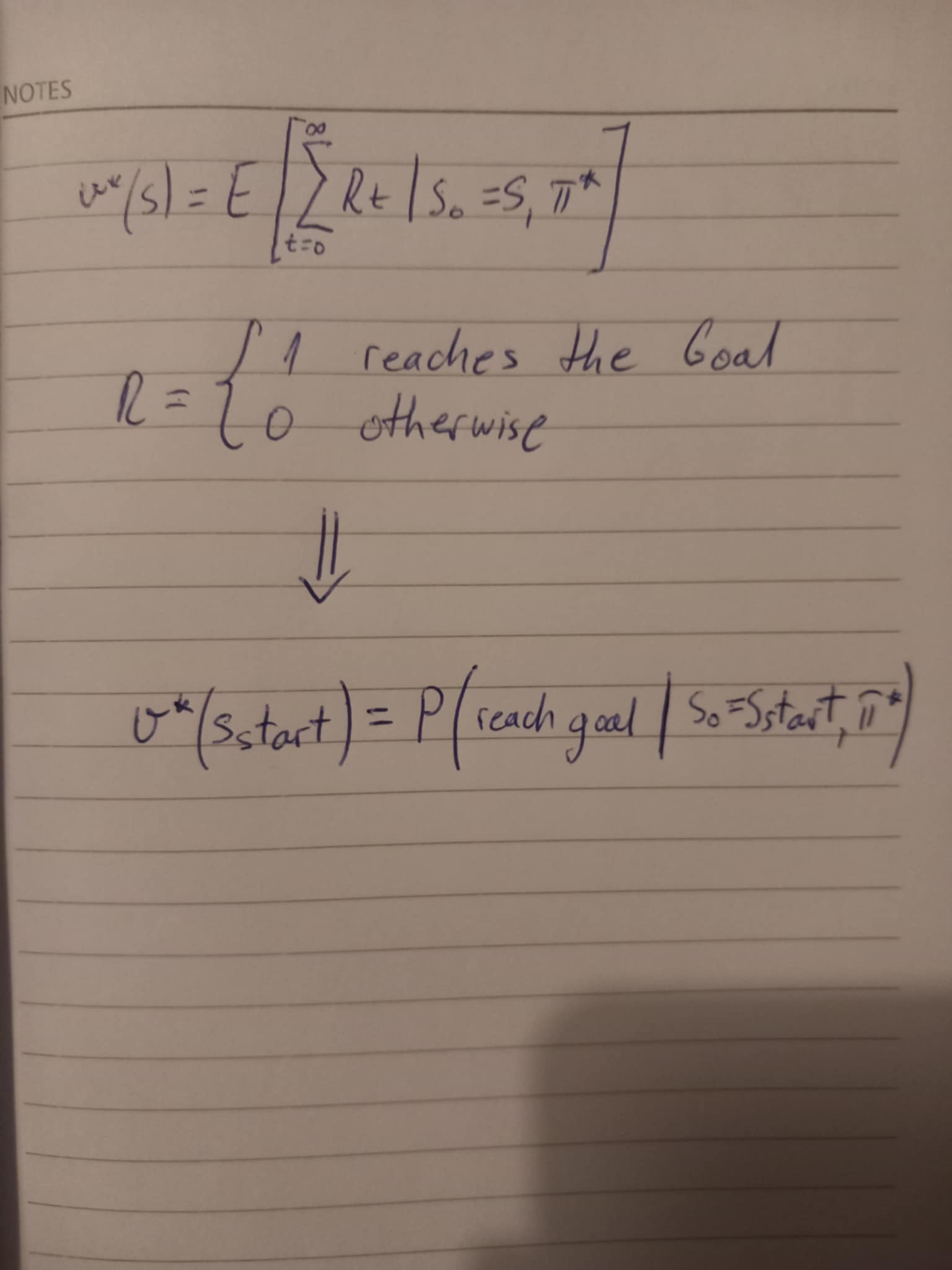
1. Gamma 1.0

State values are significantly higher across the, which reflects the undiscounted future rewards. Near the goal state, values like 0.882 and 0.941 indicate that the policy assigns high importance to reaching the goal.

1. Accumulated reward

Gamma = **1.0** results in **better performance** with a higher accumulated reward during both training and testing. This is because the algorithm accounts for the **full future reward**, which is critical in long-term decision-making environments.

b)



* For small N=10:

The mean reward is 0.9, which is slightly higher due to randomness in such a small number of episodes. With fewer episodes, fluctuations are common because there isn’t enough data to smooth out the randomness.

* For N=100:

The mean reward drops to 0.79, as the randomness starts to average out, but it still hasn’t fully stabilized.

* For N=1000 and N=10000:

The mean reward stabilizes around 0.829 and 0.8272, respectively. These values are close to the true v, which represents the probability of reaching the goal under the optimal policy.