Figure 1 shows a policy where the agent interacts with the environment 500 times per episode and starts to find the goal for episodes above 100. However, reaching the optimal policy starts around 500 episodes, to prove that the policy starts to converge at that point we increased the episodes to 1000 and we observed that the policy got much better, which can be seen on Figure 2. The reason why the policy starts the converge toward the optimal policy such late is due to the amount of interaction steps of the agent with the environment per episode, so in order to find the optimal policy in fewer episode we increased the interaction step per episode to 1000. The improvement can be seen on Figure 3. With the configuration the agent starts to reach the goal between 0 and 100. Furthermore, the policy starts to converge toward the optimal policy for episodes higher than 100, which is due to the higher exploration possibilities of the agent. The agent spends more time to explore the environment, which leads to finding the optimal policy in much fewer episodes. In order to show that the optimal policy is found we increased the episodes to 1000 and we see that the policy doesn’t increase further, Figure 4. Another interesting fact is that the new policy finds the goal in fewer steps, where the best average reward is less than -0.2, in contrast to the policy in Figure 2, where the best average reward is around -0.3.

|  |  |
| --- | --- |
| Figure 1 Tile SARSA – Average rewards per episode with  500 steps for 500 episodes | Figure 2 Tile SARSA – Average rewards per episode  with 500 steps for 1000 episodes |

|  |  |
| --- | --- |
| Figure 3 Tile SARSA – Average rewards per episode  with 1000 steps for 500 episodes | Figure 4 Tile SARSA – Average rewards per episode  with 1000 steps for 1000 episodes |

In Figure 5, the agent interacts with the environment 500 times per episode. The agent starts to find the goal shortly after the start and after 100 episodes it finds the goal regularly, so the policy gets better. However, we can observe that the agent fluctuates to find the goal in a continuative sequence of episodes. We increase the environment interaction of the agent, which leads the agent to find the goal more often in the first 100 episodes, as can be seen in Figure 6. Again, we can observe that the agent fluctuates in finding the goal in a continuative sequence of episodes. An interesting fact is, that the policy in Figure 6 has a higher average reward (best avg reward around -0.2) than the policy in Figure 5 (best avg reward is around -0.3), which is probably due to the higher exploration possibility.

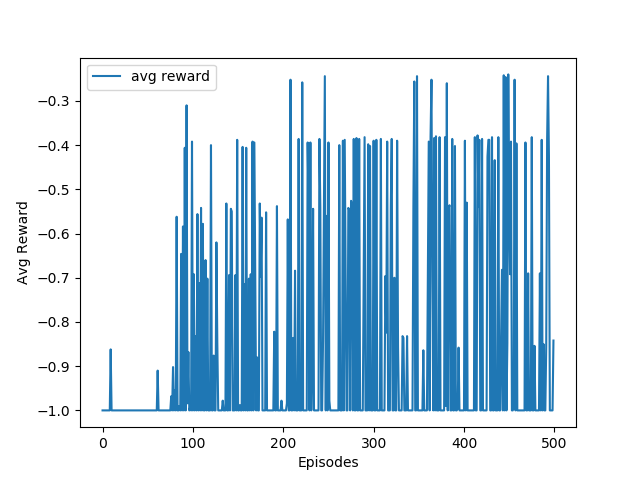


Figure 5 Neural Network SARSA – Average rewards

per episode with 500 steps for 500 episodes

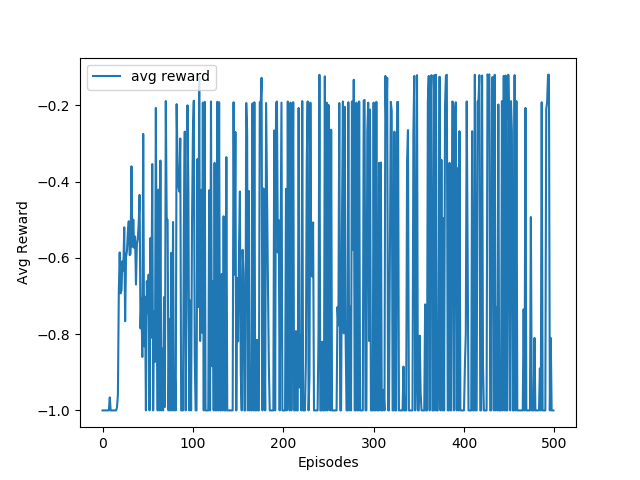


Figure 6 Neural Network SARSA – Average reward per episode

with 1000 steps for 1000 episodes