Daniel Huber 661113485

AI MP1 Part 2

1)a)
While there are some random non-zero (1 or -0.5) rewards at the beginning,
the learned path will result in a final reward of 0 after a little amount of episodes.
The Agent follows this path without looking left or right, because the Q values on this path are the highest.
Therfore, the Agent can not find a better way.

b)
There are more non-zero values than before. I noticed that non-zero final rewards are almost never alone.
Therefore, if a random action that leads to non-optimal behaviour occurs,
the algorithm needs some episodes to find a new "optimal" path that it stays on until the next random action.
The majority of episodes still ends with a reward of 0.
Because the discount value and the learning rate are too small,
the effect of positive rewards is not strong enough to really improve the Q-values.

c)
The higher the epsilon-value, the higher the chance of non-zero final rewards. I tried epsilon values from 0.3 to 0.7.
The final rewards may vary between -2 and 2.
There is too much randomness, so the Q-values are rarely relevant for choosing the next step.
While the agent may be better trained than before,
it does not follow the algorithm's suggestions often enough to profit from the gained knowledge.

2)a)
At the beginning, there is a longer learning phase than in earlier experiments, because epsilon is smaller.
After several hundred episodes, there is no real learning anymore, as there is no possible way that the agent could prevent a possible collision with the thief.
The risk of meeting the thief can only be minimized to one state on the path,
because the Agent doesn't know where the thief is. As soon as the learned path evades all slippery states and only visits one state on the path of the thief, it can't improve any more.

b)
When the Agent knows where the thief is, it can evade him. Therefore, the reward improves compared to the experiments. The reward is 1 most of the time.
The Agent can only adjust the Q values to what it is able to observe.The number of states in the main learning phase stays the same,
because epsilon stays the same. After this phase, improvements are less and less common, but do, of course, still happen,
as we are still training our agent.(Same for every experiment with epsilon>0)

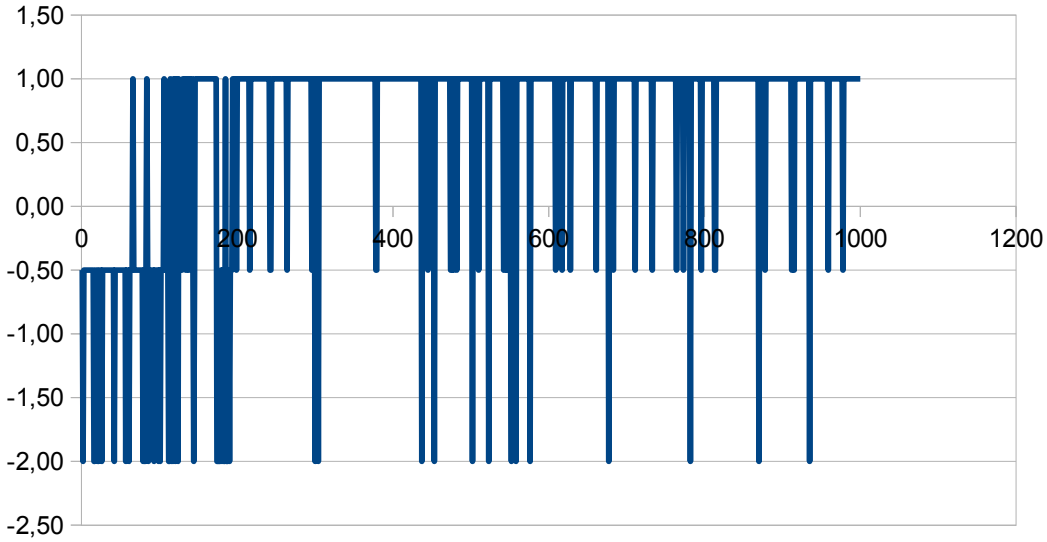
c)
epsilon= 0.01
rate=0.35
These values for the variables provide very good results for this specific task.

Daniel Huber 661113485

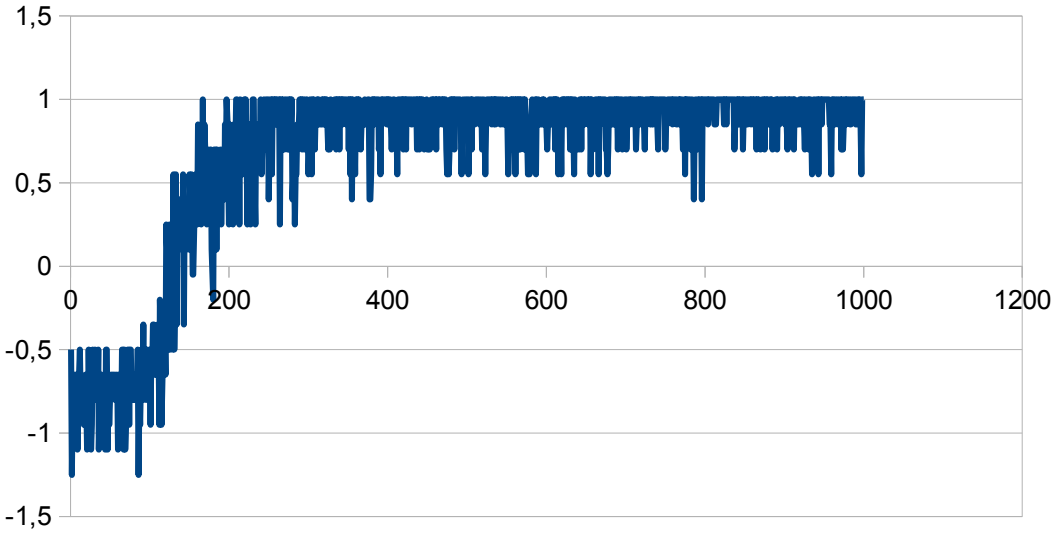
The low epsilon ensures that there are not too many random events at the later stages of the training, while there are still enough random actions to find a good path at the beginning.
0.35 as a learning rate supports a short learning phase at the beginning, while it does not react too strongly to
random negative actions in later stages.
This combination of epsilon and learning rate is very efficient, because the values for the variables provide positive effects,
while evening out the negative effects of the other variable (eg. relative strong reaction to random actions evens out the rare occurence of random actions)

3)

Plot of the discounted rewards of 1000 episodes:



Plot of the average discounted rewards of 10 games of 1000 episodes:



The trend of improvement is more obvious, when we take the average discounted rewards of 10 games. The average makes sure, that single episodes are not shown that explizit. Therefore the

general learning-trend is better shown. Still, a single episode can always be a bad choice, as it may be founded on random acts.