Report

I used a shared Q table instead of individual Q tables per sweeper to speed up convergence. This was because they all share the same environment and therefore the same Q values for getting rewards. The Q learning parameter values I chose were purely to promote exploration in the beginning of the simulation and then exploitation towards the end of the 50 iterations. The discount factor is kept at 0.9 to prioritize rewards in the distant future; this was because I decided to take into account the deterministic behavior that arises due to exploitation becoming the agent's priority near the end of the simulation. The learning rate is initialized at 0.9 as the agent is going to be exploring in the beginning of the simulation, and therefore more value will be placed on new experiences. The learning rate decays over the span of the iteration at a constant 0.02, as the agent is going to prioritize exploitation, and less value will be placed on new experiences.

I used both the decaying learning rate and E-greedy strategies to create a balance in the exploration exploitation tradeoff. The epsilon value was set at 1 so that the agent has a greater chance of exploring his environment in the beginning; then as the epsilon value decays the agent has a greater chance of exploiting his environment.

Can check the output.txt file for the results of a test.

<u>For Test 1 the results were:</u>

Most mines gathered: 23

Average mines gathered: 14.74

Average mines deaths: 0

The mines never die since there are no supermines.

<u>For Test 2 the results were:</u>

Most mines gathered: 20

Average mines gathered: 10.18

Average mines deaths: 0.3

The mines stop dying at around 20 – 25 iterations.

For Test 3 the results were:

Most mines gathered: 2

Average mines gathered: 0.32

Average mines deaths: 1