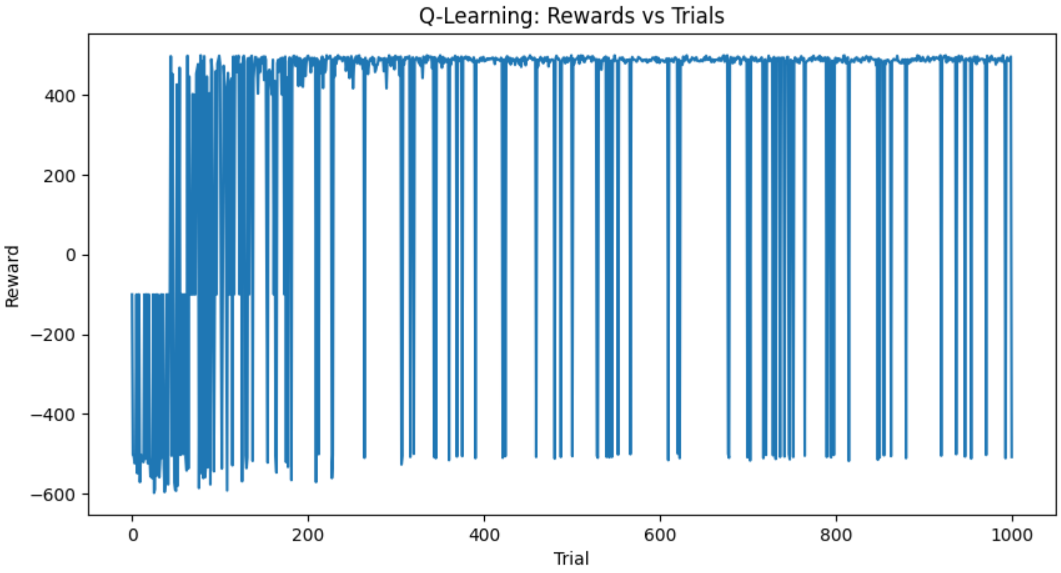
16-662: Robot Autonomy
Homework 4: Q-Learning

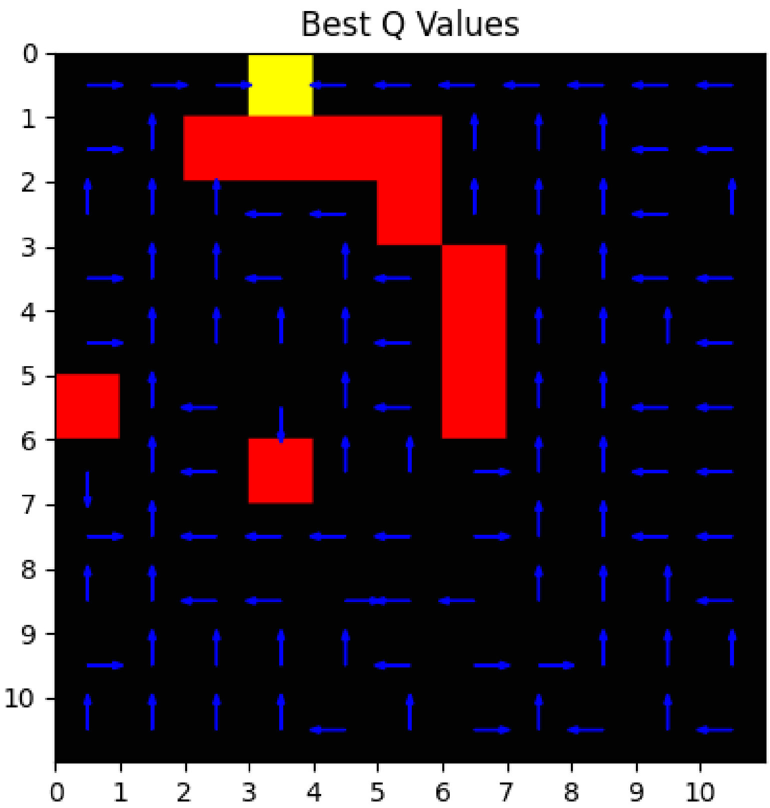
Abhinandan Vellanki – abhinanv

1. Q/A
    a. The random dips in reward while training is most likely caused due to the stochasticity in the environment which maintains a chance of encountering low reward terminal states.
    b. The Q-values make sense, they represent the highest future reward yielding action at each state.
    c. The Q-agent follows the optimal path in the videos, but it might not always do so due to the stochasticity in the environment.

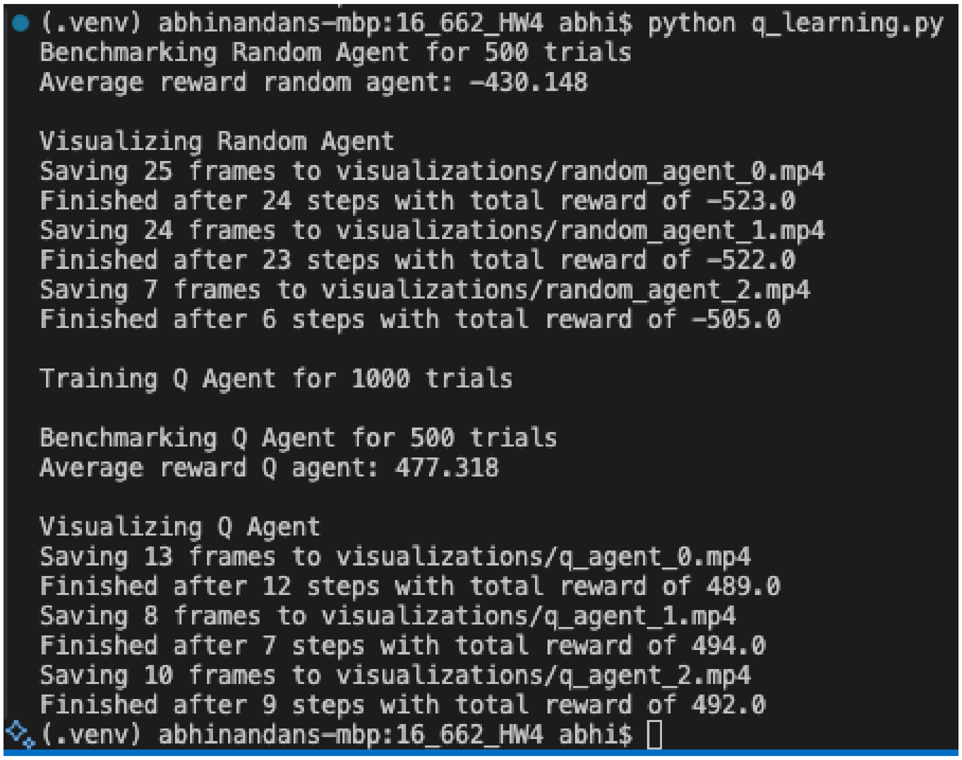
2.



3. With given hyperparameters:



4.



```
(.venv) abhinandans-mbp:16_662_HW4 abhi$ python q_learning.py
Benchmarking Random Agent for 500 trials
Average reward random agent: -430.148

Visualizing Random Agent
Saving 25 frames to visualizations/random_agent_0.mp4
Finished after 24 steps with total reward of -523.0
Saving 24 frames to visualizations/random_agent_1.mp4
Finished after 23 steps with total reward of -522.0
Saving 7 frames to visualizations/random_agent_2.mp4
Finished after 6 steps with total reward of -505.0

Training Q Agent for 1000 trials

Benchmarking Q Agent for 500 trials
Average reward Q agent: 477.318

Visualizing Q Agent
Saving 13 frames to visualizations/q_agent_0.mp4
Finished after 12 steps with total reward of 489.0
Saving 8 frames to visualizations/q_agent_1.mp4
Finished after 7 steps with total reward of 494.0
Saving 10 frames to visualizations/q_agent_2.mp4
Finished after 9 steps with total reward of 492.0
(.venv) abhinandans-mbp:16_662_HW4 abhi$
```

5. Role of hyperparameters:
    a. Q_ALPHA: Learning rate affects how much the q values are updated each time. Increasing this parameter increases the likelihood of convergence but also of overshoot. Decreasing the parameter smoothens the learning curve, increasing convergence time, but reducing oscillations.
    b. Q_GAMMA: Discount factor affects the importance given to future rewards. If reduced too much, the agent could begin to focus on immediate rewards and get stuck in local maxima, but this also decreases convergence time. If increased, the agent would be more incentivized to take short term sub-optimal actions en-route to an eventual reward, but this increases convergence time.
    c. Q_EPSILON: Exploration rate affects the likelihood of the agent to take random actions. Increasing it would increase convergence time but also nudge the agent out of local maxima. Decreasing it would decrease convergence time but increase the likelihood of sub-optimal policies.
    d. Q_RHO: Environment entropy affects the robustness of the agent. Increasing it would make things more difficult while training but increase robustness to noise, while decreasing it would stabilize the training process but risks overfitting the agent to this specific scenario.