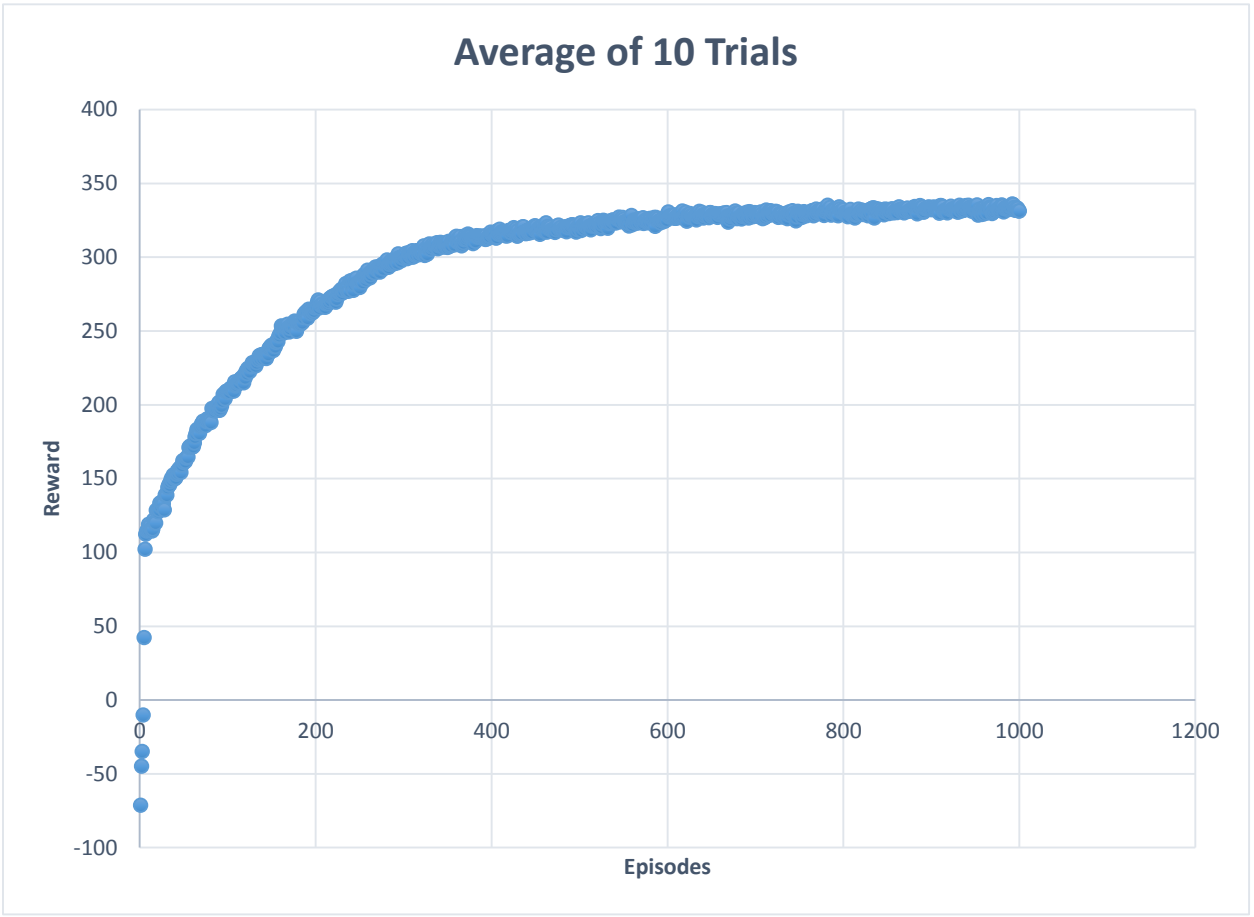
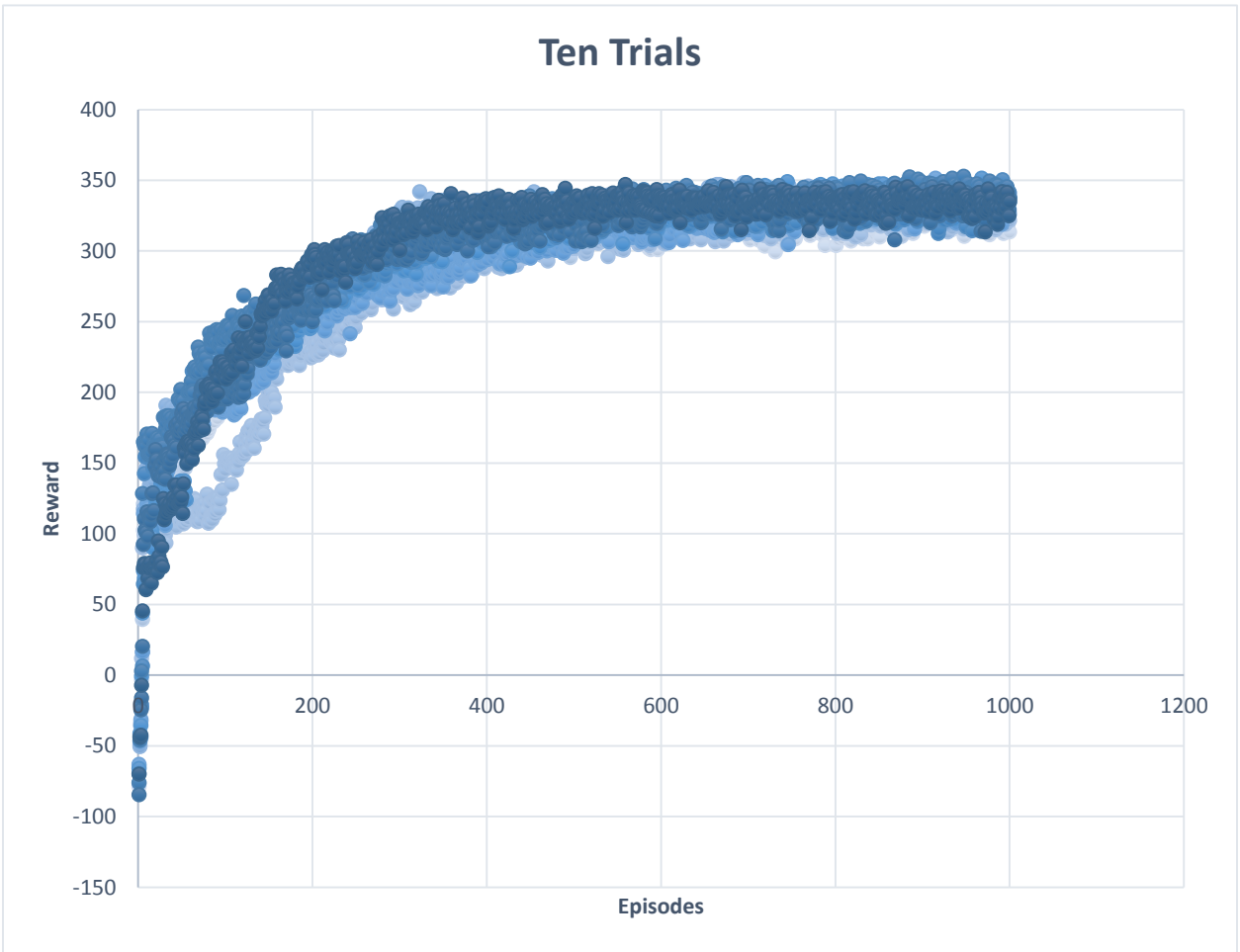
Soham Waychal
waychal2
CS440 - Fall 2014

MP1 Part 2 – Experimentation

1. **A)** The reward for each episodes converges to zero quickly because the robot learns quickly that the second column is slippery and that it has a high chance of dropping the package. This leads it to change the qValue those tiles to a negative value. Since it will always choose the greatest qValue and there are no random movements, the robot will never cross over to deliver the package.
   B) If epsilon is set to 0.1, the robot can randomly move over to the other side of the slippery column even if all the tiles in the slippery column have a negative qValue. Once on the other side, it can deliver the package and add a positive reward to those states. This would lead to the robot picking that path to deliver the package since it resulted in a net gain.
   C) If epsilon is set to 0.5, the robot does a lot of random movements (half of its actions become random) and does not follow the path with the highest utility. This leads to a lower reward because the robot must diverges from path of highest utility fairly often.
2. **A)** If the robot does not know there is a thief, it treats the middle block in the third column (the thief column) as a 100% failure square. It never crosses the slippery column to get to the other side because as far as it is concerned all that lies that way is negative rewards. This leads it to stay on the left side because the qValue of all the other tiles is negative since it has not experienced any reward yet.
   B) Since the robot knows that there is a thief, it does not treat the middle block in the third column as impenetrable and tries to explore the world. It learns where the rewards are and tries to avoid the thief. This results in very high positive rewards, and the robot delivering the packages properly while avoiding the slippery tiles.
   C) A range of values work optimally for learning rate and epsilon. For learning rate, the best most effective values lie in the range from 0.1 to 0.25. The best epsilon values lie in the range from 0.004 to 0.005. To figure out these values, we tried a wide range of values for both the parameters and noted down the ones that gave the highest rewards. We then used the same method on a smaller range and fine-tuned our results.
3. For this question, we used values of 0.1 for learning rate and 0.0045 for epsilon. The rewards per episode start negative since the policy hasn't been set properly. After the first 10 or so episodes, the policy knows where the rewards are but does not know what the optimal path to get there is and keeps updating itself until the best path (or almost the best path) is found, and it converges. The most optimal path in the world with thief gives a reward of around 335-340 per episode. There is a lot of static in the ten trials because the policy finds several sub-optimal but very close to optimal paths that are a result of the policy noting that changing the path to an even more optimal solution would require decreasing the total reward. Therefore, it converges to a very close to optimal solution (sometimes it might even converge to the best one) but produces a lot of static because there are a lot of close to optimal solutions.


Average of 10 Trials
400
350
300
250
200
150
100
50
0
-50
-100
Reward
0
200
400
600
800
1000
1200
Episodes


Ten Trials
400
350
300
250
200
150
100
50
0
-50
-100
-150
Reward
0
200
400
600
800
1000
1200
Episodes