CS 440 MP 1 PART 2

MIKE WOJCIAK

1. In WorldWithoutThief

(a) In this case, the first episode returns a negative reward and then all other episodes return a reward of 0. The first episode is random since the robot has no prior knowledge. The other episodes then have no randomization so the robot will not walk over the 'slippery' tiles in order to reach the customers because they have a chance to return a negative reward if the robot drops a package.

(b) In this case, the episodes start out with a negative reward but they slowly grow and most of the episodes at the end are between 100 and 200. This is expected because the robot will slowly learn the best path to take as it takes a slightly different path each episode based on the ϵ value (0.1).

(c) With ϵ = 0.5 the episodes mostly have a negative reward mostly around -30 but ranging down to around -80. There does not seem to be much change in the reward from the early episodes and the later ones. The output is similar for higher ϵ values like 0.8. This makes sense because at this point the movement of the robot is mostly random and takes a random step more often than the suggested one.

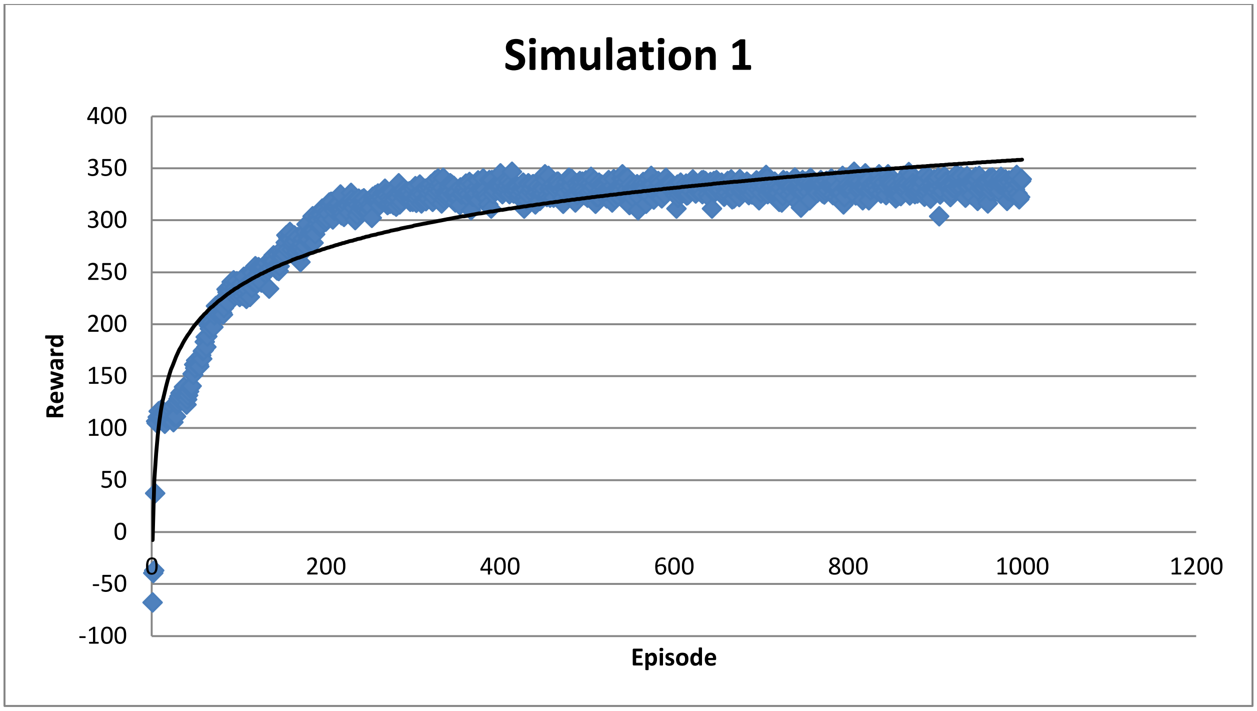
2. In WorldWithThief

(a) In this case know_thief is set as 'n' so the robot cannot avoid the thief. All the episodes have a reward of around -10. This makes sense because since the robot cannot avoid the thief, it will run into the thief more often and lose the packages therefore receiving a negative reward.

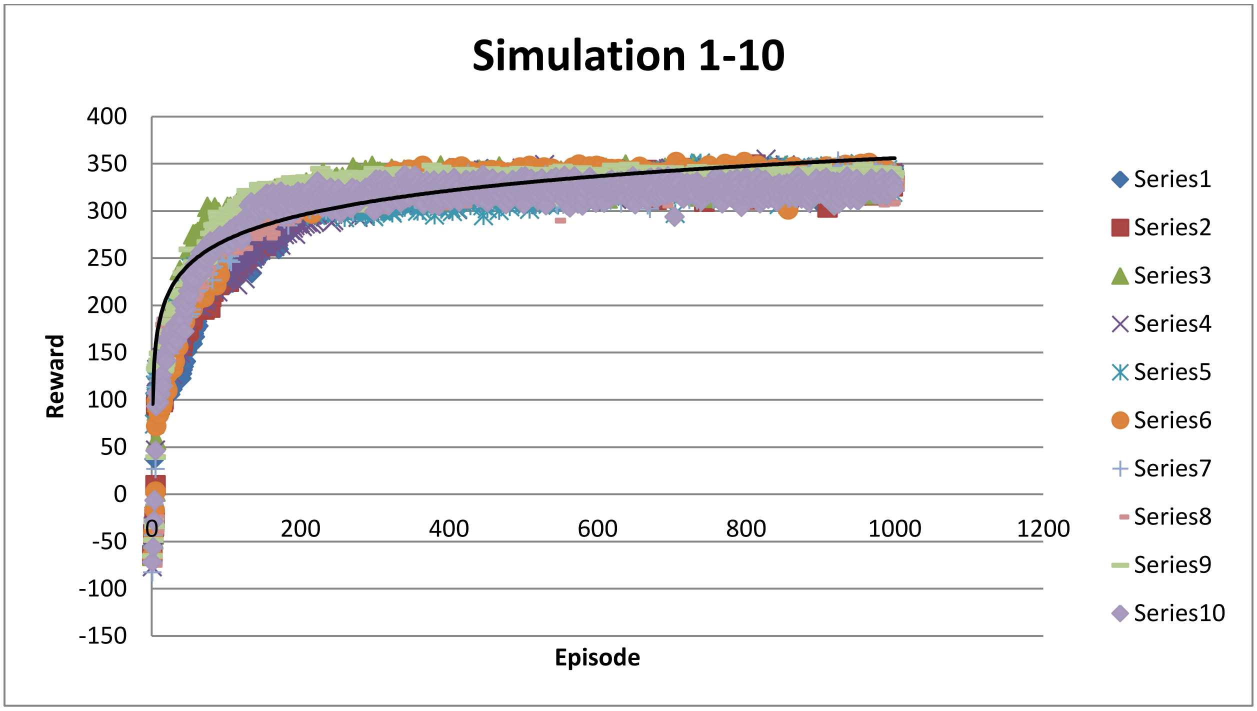
(b) If knows_thief is set as 'y' then the robot knows the thief and can avoid it. The episodes start by returning a negative reward but they quickly become positive and end up around 280. This makes sense because the early episodes will be more random in order to learn the best path and then it will improve on that path by making a slightly different path each time based on the ϵ value.

(c) The best learning rate I found was 0.1. To find this I tried various learning rates and graphed the outcome. 0.1 was the one that resulted in the highest rewards in the later episodes. The best ϵ value I found was 0.01. Higher ϵ's would plateau quicker but would result in lower rewards in later episodes when compared to 0.01. Therefore the best combination is a learning rate of 0.1 and an ϵ value of 0.01.

3. Graphs



In this graph the agent improves a bit slower (at around episode 300) but it reaches a plateau between 300 and 350.



This graph is very similar to the previous one, it simply has more data from each simulation. There is very little variation in each simulation which is due to the low ϵ value. Again the graph starts to plateau around episode 300 and the reward it plateaus around is between 300 and 350.