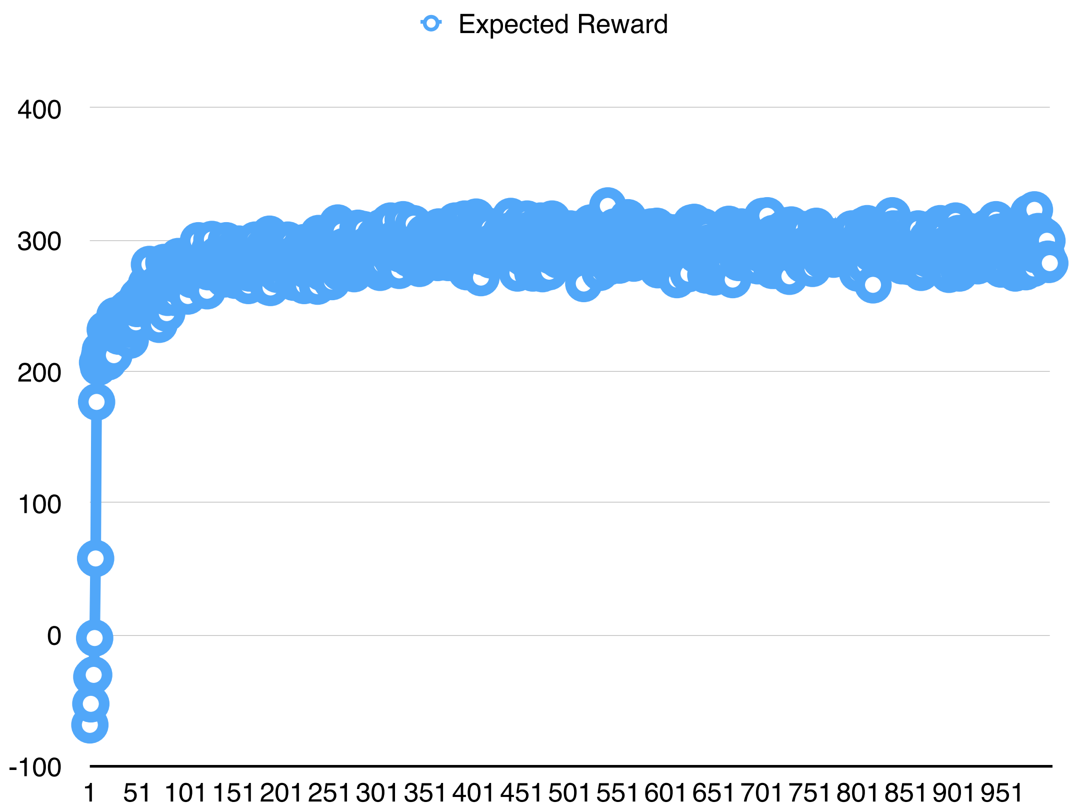
David Servose servose1 CS440 MP1 Part 2

1. In WorldWithoutThief
   A. The reward for each episode increases in just a few episodes from a negative value up to 0. However, without the random action, the agent is performing greedily and does not seek a higher reward.
   B. The reward for each episodes generally increases from a negative value up to about 200. However, compared to the last test, the reward increases much more quickly. There is much more variance in the rewards for each episode, some of the later episodes even have a negative reward. This is due to the randomness introduced and is expected.
   C. With a larger epsilon of 0.5, the reward for each episode is mostly negative, with an average reward of around -33, and does not generally increase. There is some variation due to the high randomness, but most episodes have a reward that is close to the value of the original unlearned agent.

2. In WorldWithThief
   A. The performance of this experiment is the worst yet, with an average reward of around -70. Lower rewards are expected due to the addition of the thief. Due to the high level of randomness, the scores do not noticeably improve with an increase in the number of episodes.
   B. The performance where the learning agent knows the position of the thief is even worse than not knowing where the thief is. The rewards start off similar to part A, around -70. The rewards then quickly drop to around -200. This is due to the learning agent where the thief is, but the thief moves so the agent makes misinformed decisions which leads to worse scores.
   C. The best learning rate is 0.04 and the best epsilon is also 0.04. The rewards decreased quickly as randomness increased. This is expected as randomness is used to escape from greedy situations, and a high randomness leads to poor moves. Rewards decrease as the learning rate increases as well. A higher learning rate means more reliance on past moves for expected reward, and since the thief moves, relying so heavily on past moves does not guarantee a good score.

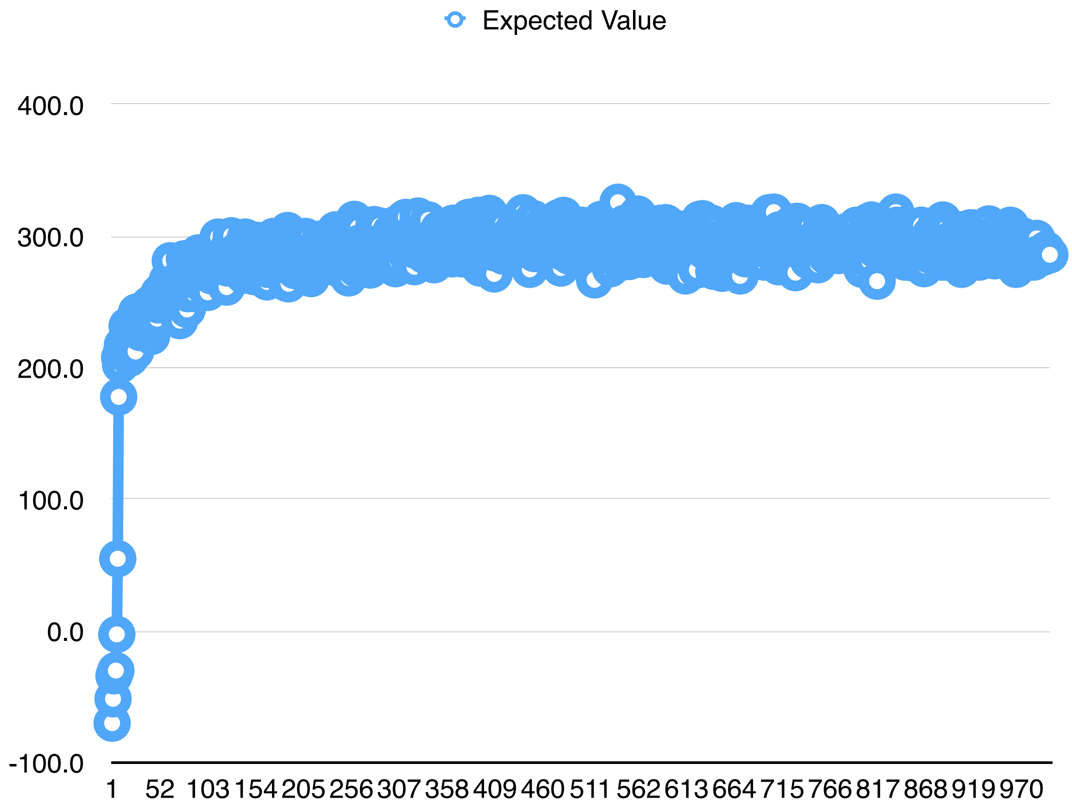
David Servose servose1 CS440 MP1 Part 2

3.



From the plot above we can see that the expected reward increases exponentially until it tapers off to the final expected value of 286. This is expected as the initial episodes use an untrained agent, which then learns and increases the expected reward.

David Servose servose1 CS440 MP1 Part 2



The plot above shows the expected discounted reward averaged over 10 simulations. The plot looks similar to the previous plot of only one simulation of the expected discounted reward. This is to be expected. The difference is a smaller variance.