Matt Hoffman (mjhoffm2) CS 440 MP1 Part 2

1. a) After incurring a -8.5 penalty on the first trial, the Agent consistently earned zero reward for the next 999 episodes. The Agent clearly learned that any action outside of a select few were likely to incur a negative penalty, so it simply looped forever outside of the slippage areas to avoid them.

b) Setting epsilon to 0.10, the Agent was able to consistently earn positive rewards after about 50-100 episodes, but the reward received varied greatly even at the end of the full 1000 episodes. In the last 50 trials, the reward for the Agent varied from 20 to over 200 at random. This indicates to me that an epsilon value of 10% is already too high.

c) Setting epsilon to 0.5 causes the Agent to only receive a positive reward on rare occasion, due to the fact that it does not follow its own policy. Anything higher than that is pointless. I had some success with an epsilon value of 0.2, which was just as consistent as an epsilon value of 0.1 but with a more restricted range of reward (60 - 144 as opposed to 0 - 220) and a lower average reward. Setting epsilon to 0.05 gave some interesting results. Occasionally the Agent would take a long time to converge, and had a rare chance to not converge at all possibly due to some bad luck in the first couple episodes, but on average this gave the best rewards. It even seems that there is a direct relationship between how long it took to converge and how good on average the rewards were by the end. An epsilon value below 0.05 tended to still have many negative episodes even near the end of 1000 trials.

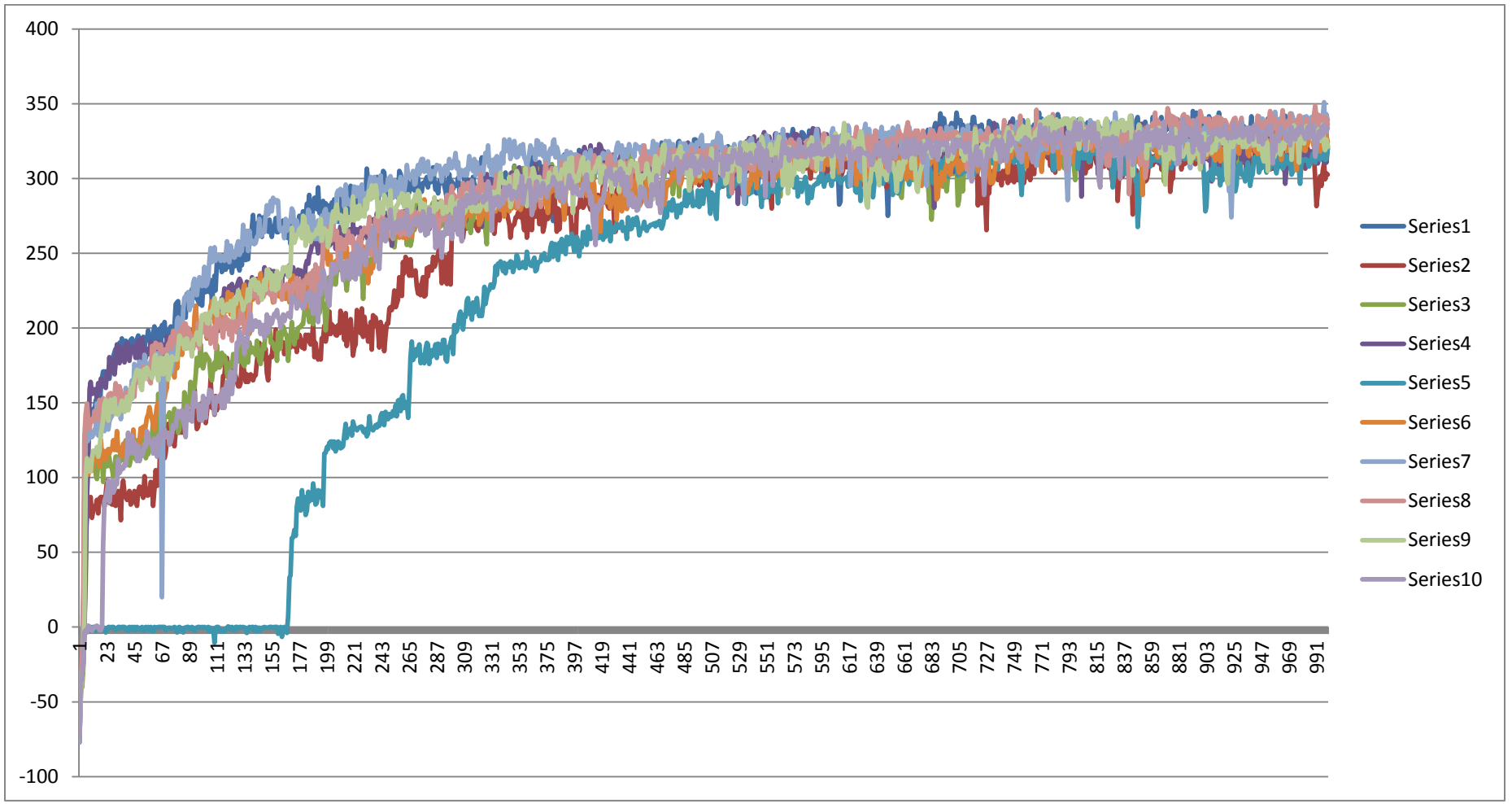
2. a) With an epsilon value of 0.05, the Agent was never able to incur a positive reward. This is likely due to the challenge being too difficult to ever find a solution randomly. By the end, the Agent had been able to reduce its negative reward from about -35 to a range of -7 to -16

b) With information about the thief, the Agent was able to consistently earn rewards between 260 and 300. On top of that, the Policy for the Agent started to converge after as few as 5 episodes.

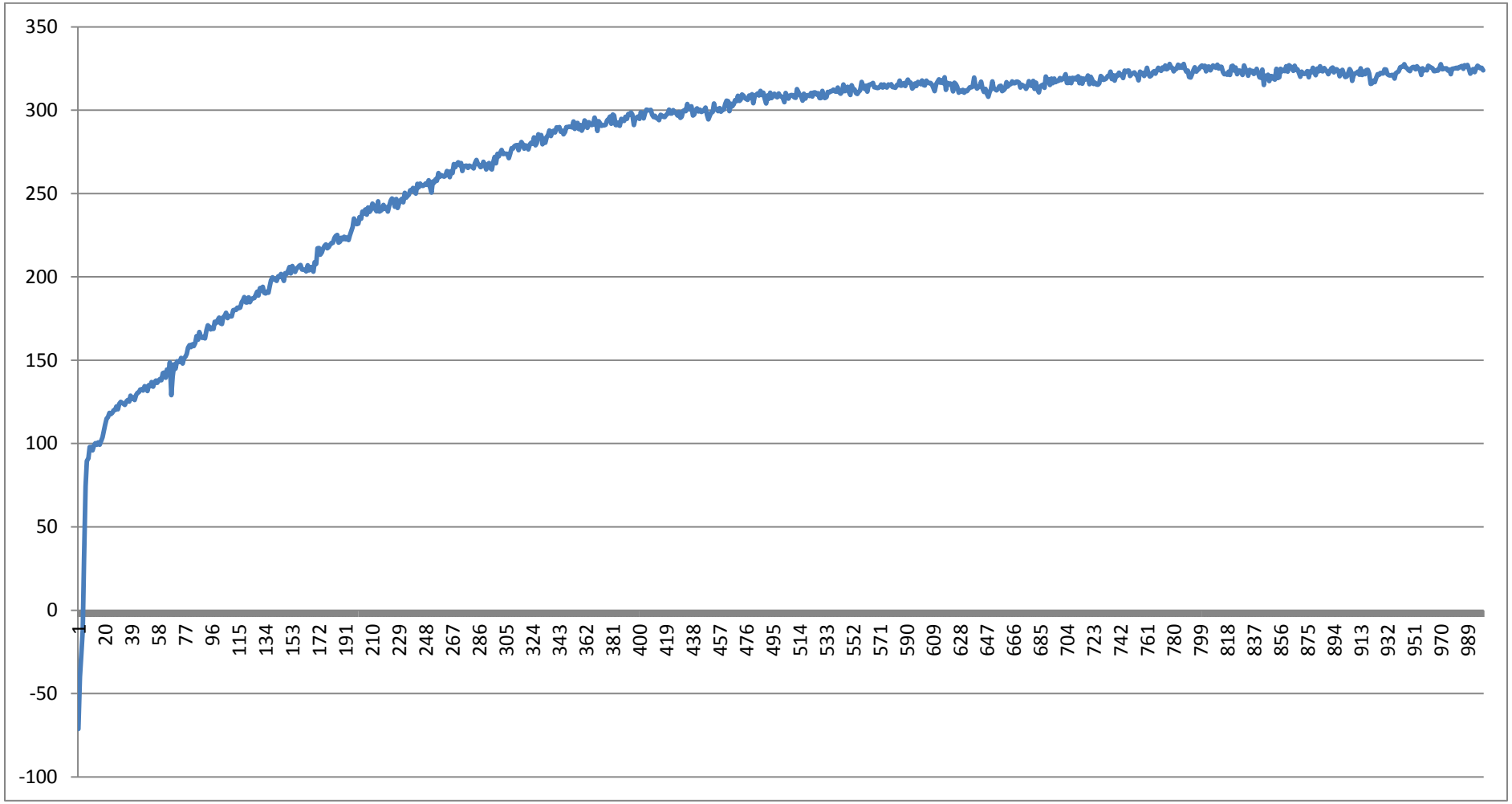
c) Keeping epsilon constant and adjusting rate, it seems that any value of rate below 0.5 has no significant impact at all as long as it is not zero. From 0.5 and above, the final reward diminishes. I found that the best value of rate for epsilon = 0.05 was as close to zero as possible (0.000001). However, higher values of rate turned out to be very effective if I made epsilon very small. I found epsilon = 0.001 and rate = 0.25 to work quite well, giving final rewards consistently between 320 and 345. It seems that the smaller the value of rate, the better the final

rewards, but the longer it takes to converge. I suspect that values of rate less than 0.25 would yield even better final results in more than 1000 episodes.

3. Plot of Episode vs Reward for each of the ten runs



Average of ten runs



Analysis:

In each of the ten runs, the Agent was able to achieve 300-350 reward consistently near the end of 1000 episodes. In most of the runs, the Agent was able to begin converging within the first twenty episodes. However, it is noticeable that in run number 5 the agent was not able to converge until over 150 episodes were completed. This is likely due to the Agent getting unlucky in the first several episodes and ending up with a very poor policy for quite some time. It would be ideal to begin epsilon at a higher value and decrease it over time to preserve accuracy but reduce the chance of this happening. It also looks like in most of the individual runs, there are a series of "jumps" that occur where an Agent is able to find a significant improvement to its policy during one of the episodes that results in a sharper increase in reward in comparison to the surrounding slope. These jumps or milestones are fairly pronounced in the individual plots but are not well reflected in the average of all ten.