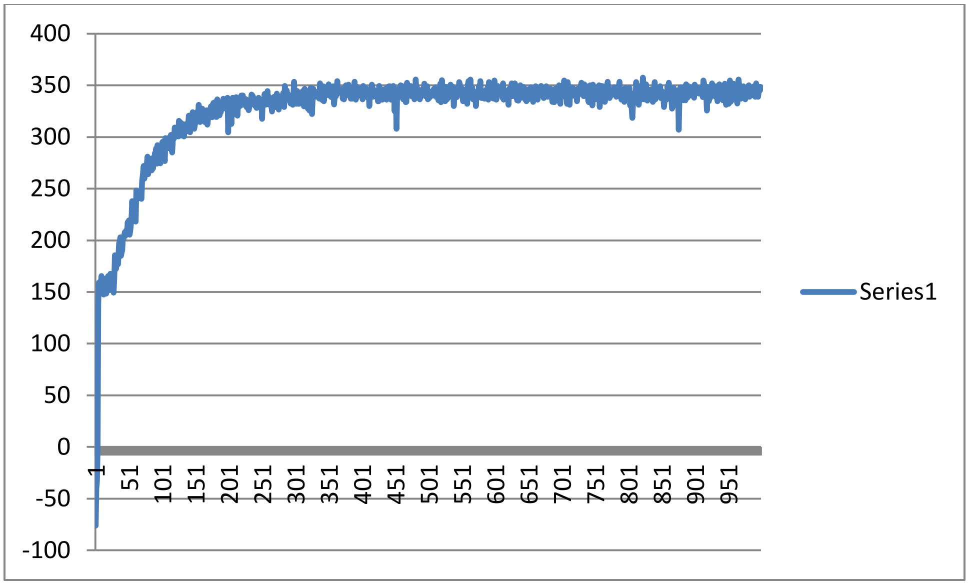
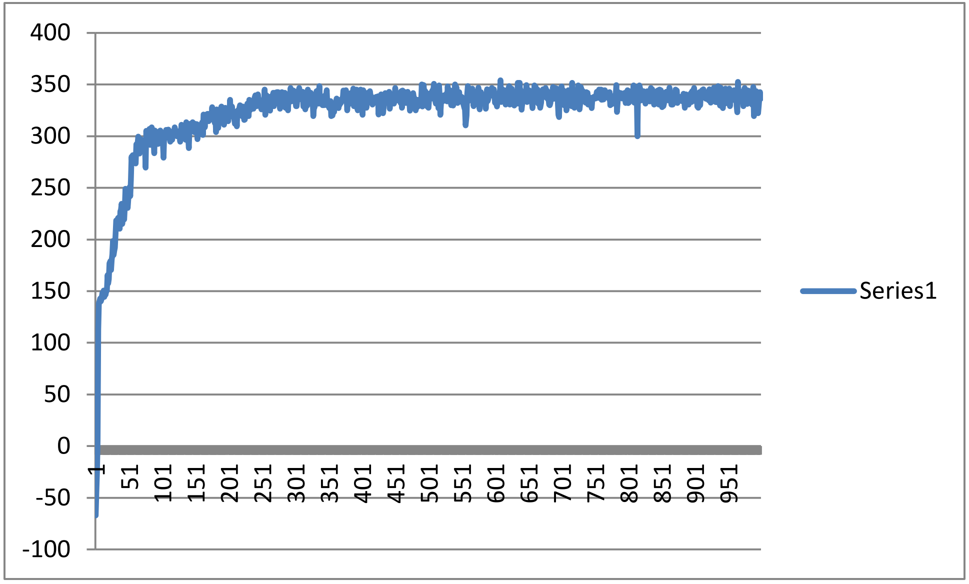
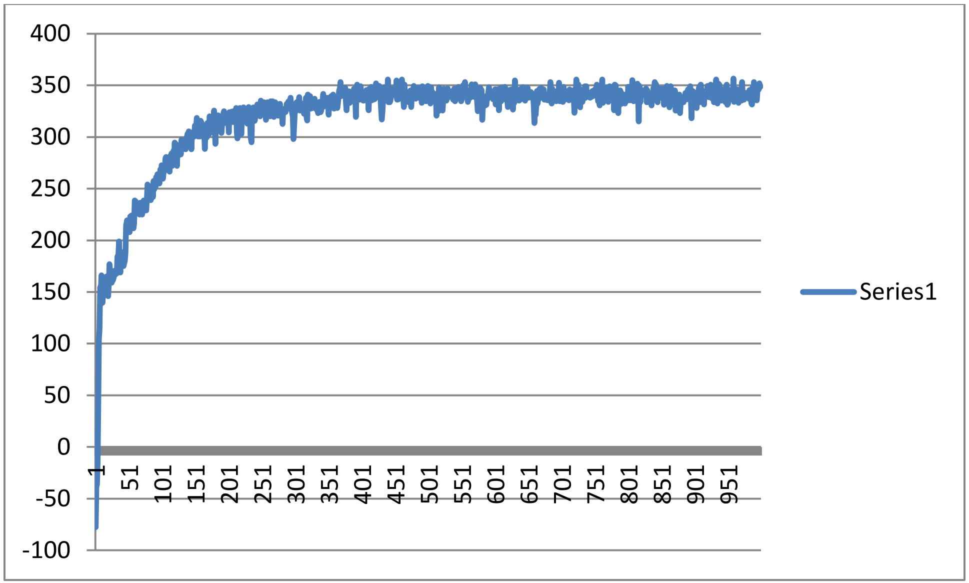
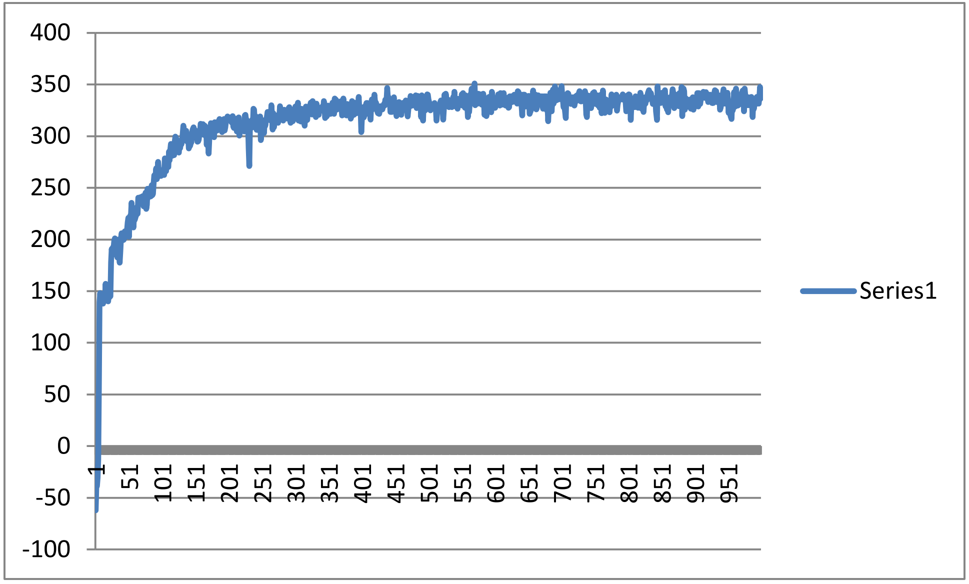
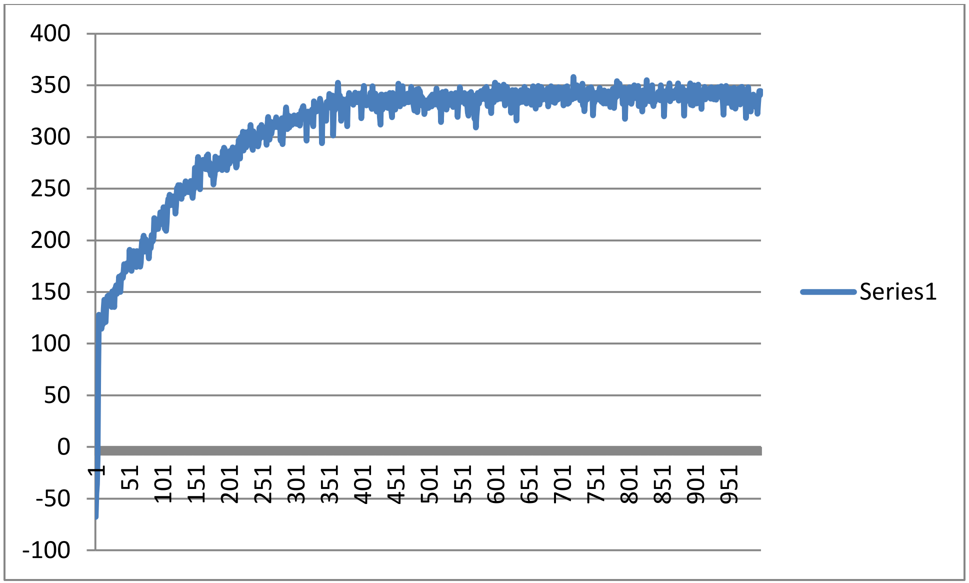
1.
    a. The reward on average was always 0. Because he doesn't explore, he is missing out on possible opportunities to increase his reward.
    b. The reward on average was a little over 150. This little chance of exploration allows the agent to find better rewards, but isn't so high that his actions are completely random.
    c. The reward on average was about -30. This is the other extreme. If the agent performs too many random actions, this is most likely going to make him lose. An analogous example is playing chess. Even if you're a complete novice, you'd be able to beat a player who is just making random moves.
2.
    a. The reward on average was about -15. Since the agent doesn't know about the thief, he never tries to avoid the thief and keeps getting his packages stolen.
    b. The reward skyrockets to over 250. Now that the agent can tell where the thief is, over time he will learn to avoid the thief, thus minimizing losses.
    c. It seems like the best learning rate is about .15 and the best epsilon is about .005. The learning rate is relatively low because you want to rely mostly on your past experience. The epsilon is extremely low because again you want to rely mostly on your past experience. It isn't 0 though because like we established in question 1, the agent needs the possibility of finding better rewards than the ones he already knows about.
3. Here are the plots of the end rewards after each episode for each simulation. X axis is trial number and Y axis is reward.

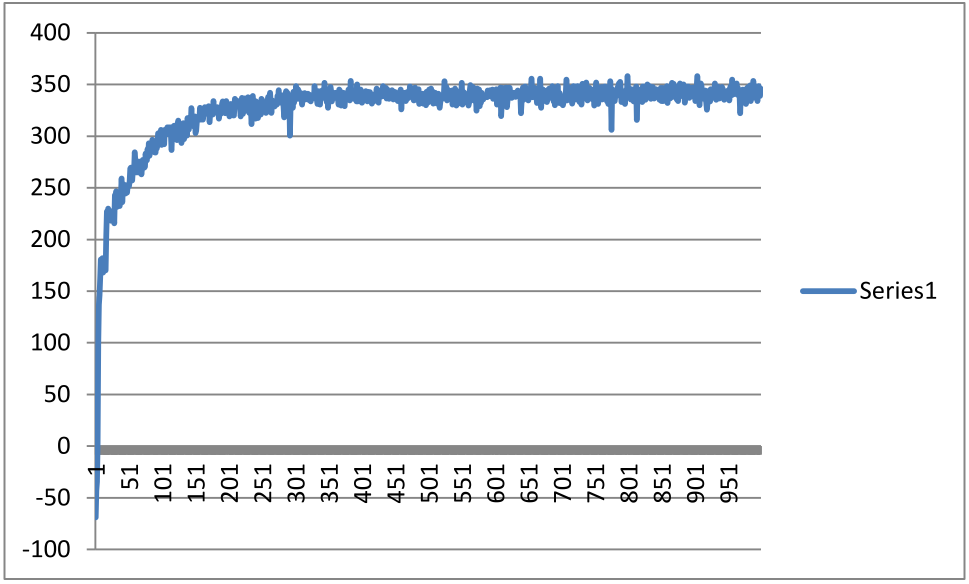

400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1

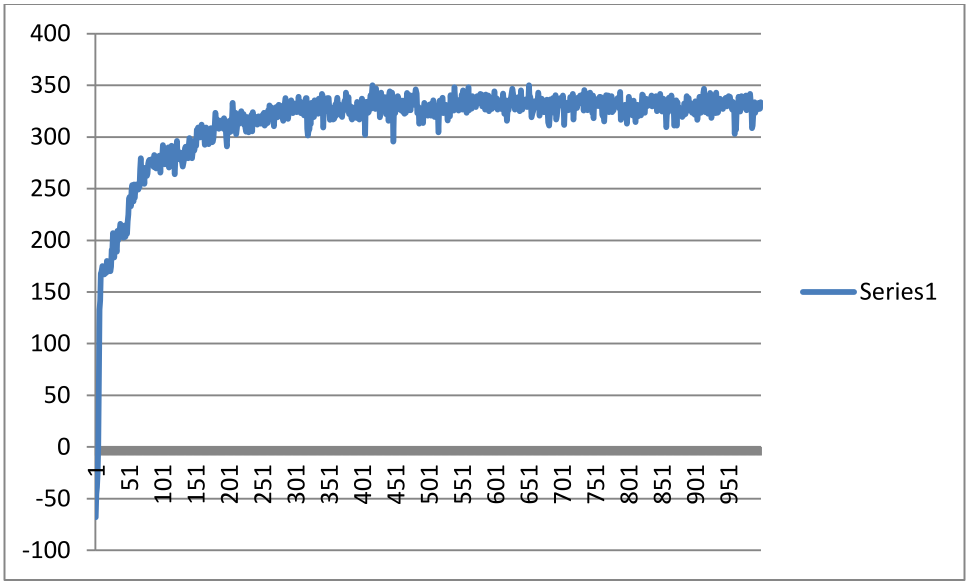

400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1

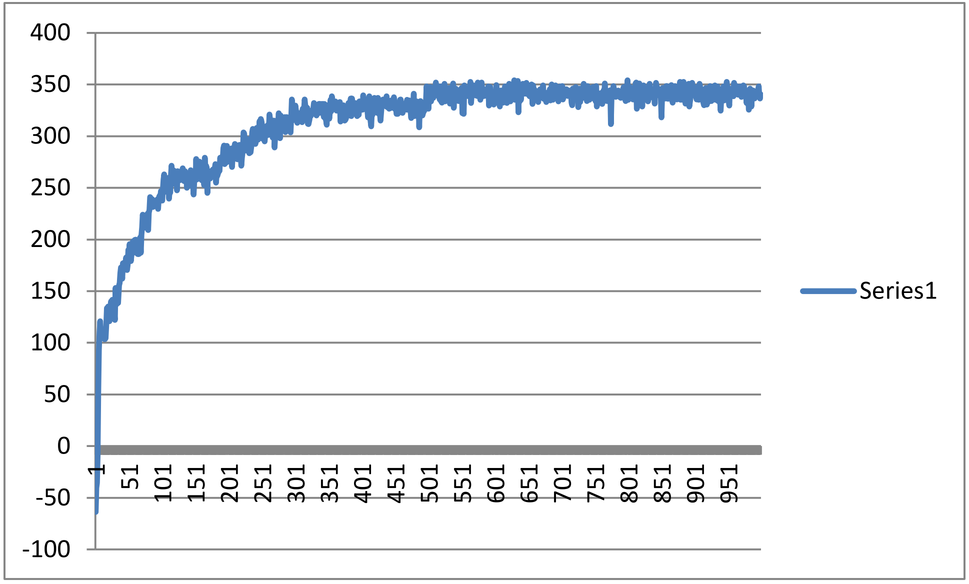

400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1

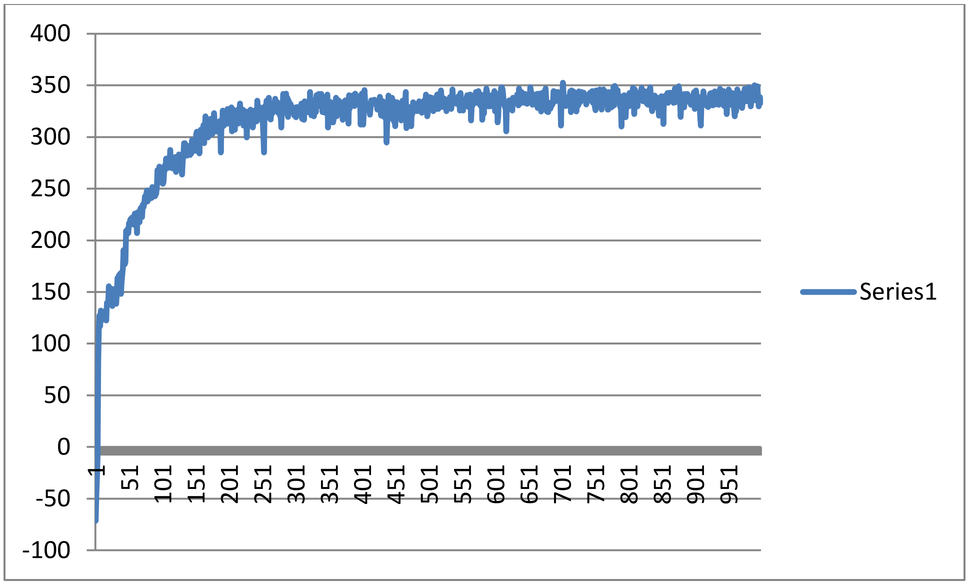

400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1

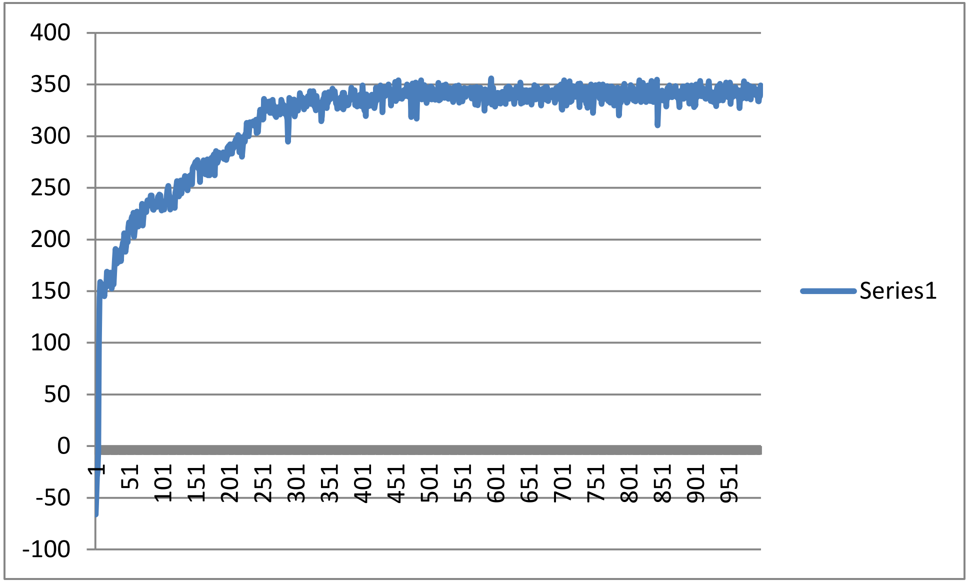

400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1


400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1


400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1


400
350
300
250
200
150
100
50
0
-50
-100
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Series1





In the first few trials of every simulation, the agent performs very poorly, mainly because of lack of experience. However, he learns very quickly, reaching the maximum reward capacity at about 300 trials. But because there is a limit on the reward even with an ideal policy, the agents best is capped at around 350 reward points. The graphs look somewhat like a shifted -1/x graph.