1) WorldWithoutThief

(a) The reward is consistently 0 for every episode. Without the chance of randomly choosing an action the robot has no chance to "experiment" and so it never performs any risky actions and is essentially stuck on the left side of the board because of the slippery tiles.

(b) The reward is consistenly high (in the 150 – 200 range). This is because the robot has enough randomness that it will take actions that don't seem to be the immediate best action, but will behave based on its immediate best interest most of the time. This leads to a robot that will discover the best actions and then usually take them.

(c) The robot consistently does very badly, getting negative rewards most of the time. This is because the actions are essentially too random. There is a 50% chance that the robot will take a random action which makes it behave erratically and essentially cancels any benefits the random "exploration" gives.

2) WorldWithThief

(a) The robot consistently gets negative rewards. This is because without knowing the location of the thief, the robot tries to find the best action as if it were in WorldWithoutThief which causes it to run into the thief a lot and lose a lot of points that way.

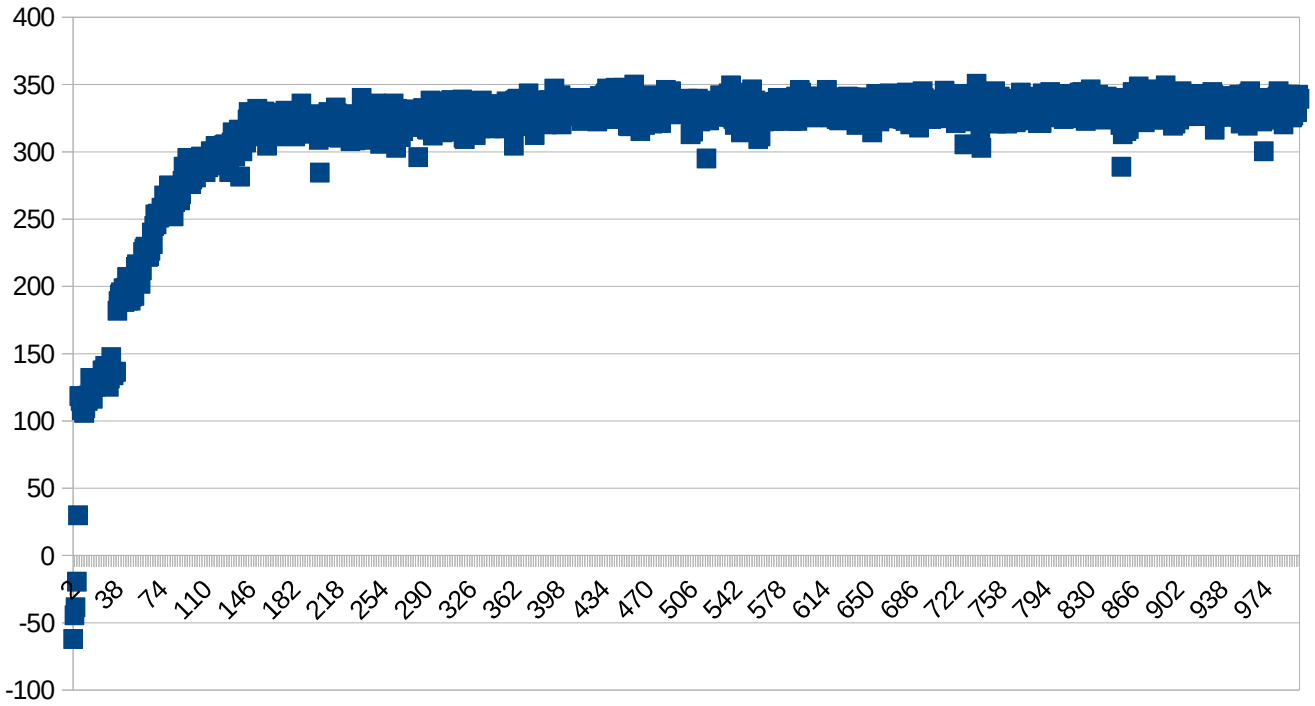
(b) The robot gets rewards in the high 200's range. This is because it knows where the thief is and can avoid it. Obviously being able to take this into account makes a big difference since without knowing where the thief is the robot would be getting punished for having the package stolen without knowing why which would lead to unpredictable behavior.

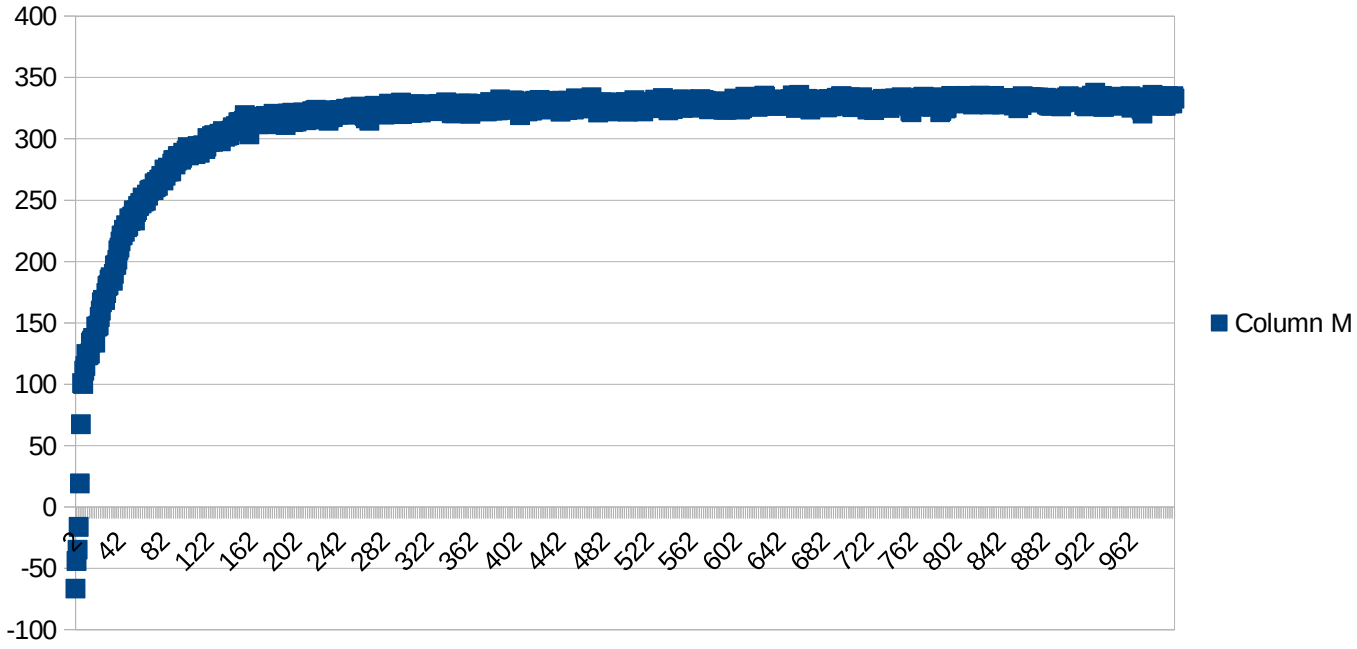
(c) The average reward seems to decrease if epsilon goes higher than 0.1 and smaller than 0.001 The sweet spot for epsilon seems to be between 0.005 and 0.02 and I saw the best performance at 0.0095 Learning rate gave good results with values between 0.05 and 0.3 and I seemed to have my best results with a value of 0.15

From this I would conclude that the robot benefits most from a small amount of randomness which lets it "explore" different actions but generally pick the best immediate one. Learning rate seems to benefit from being higher than epsilon but still fairly low. This is because there has to be a balance between the robot not learning at all and learning immediately. If the learning rate is too low than it will take too long for it to learn good actions, but if it is too high then it will weight the reward for whatever the new action was too highly and prefer it too much the next time.

3) For just one run of the Simulator it seems to quickly increase to a reward of around 350 within 150 episodes and then stay there for the rest of the episodes.

For the average value the graph looks very similar. It starts low and then reaches a max around the 150^{th} episode. The main difference is that the second graph lacks outliers like the first had.


400
350
300
250
200
150
100
50
0
-50
-100
38
74
110
146
182
218
254
290
326
362
398
434
470
506
542
578
614
650
686
722
758
794
830
866
902
938
974


400
350
300
250
200
150
100
50
0
-50
-100
42
82
122
162
202
242
282
322
362
402
442
482
522
562
602
642
682
722
762
802
842
882
922
962
Column M