1 a) The robot will go all the way up until it hits the boundary. Then it will stop there. A possible explanation might be that the rewards for each possible action all equal to zero (‘episodes.txt’), and the direction taken is all going north(‘policy.txt’). The agent is following purely greedy algorithm as epsilon is set to zero so there is no chances for exploration. So it will never cross to the right over the slippery areas. It is confusing though as for why the agent is stuck at the top rather than moving down or in another random direction.

1 b) The robot is able to deliver packages to both customers and return to pick up more. The episodes.txt file also shows an increase in rewards so the robot is trained well. Compared to 1 a) the epsilon is raised to 0.1 which allows a certain chance of exploration so the robot might cross the slipperiness.

1 c) The robot crosses more slipperiness on its way to deliver packages, making it more likely to drop them. This is because a larger epsilon is used here so the robot utilized less of the learned policy but explores more, so more chance to pick the slippery routes.

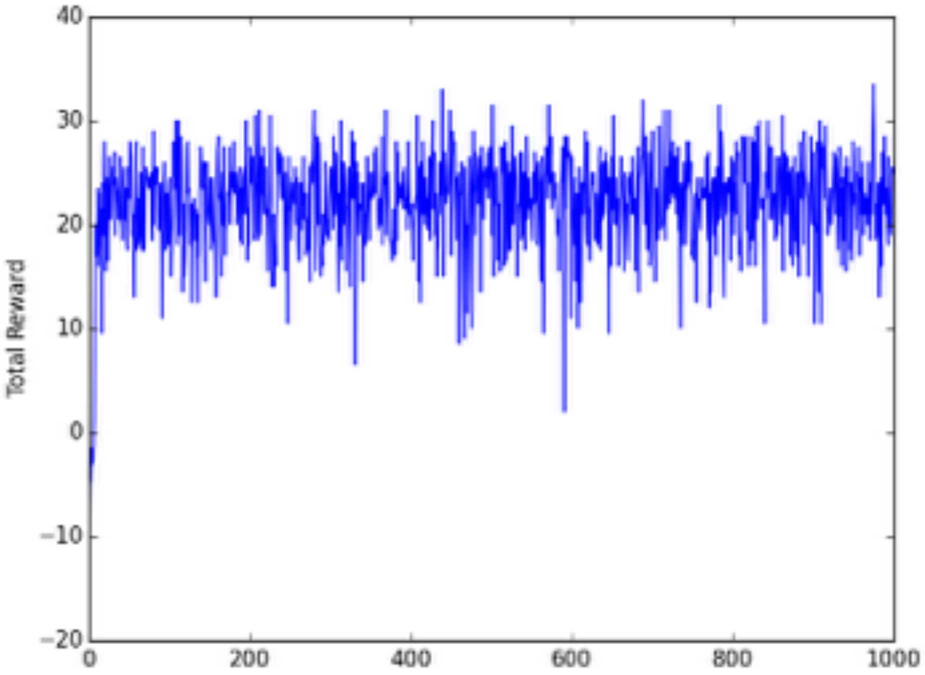
2 a) The robot will go all the way up until it hits the boundary. Then it is stuck there. Look at the episodes.txt clearly shows that the robot hasn't really learned the policy. The rewards is always negative. This is due to the fact that the thief is unknown to the robot so it cannot take into account the latter’s action. See 2 b) for more explanation on this.

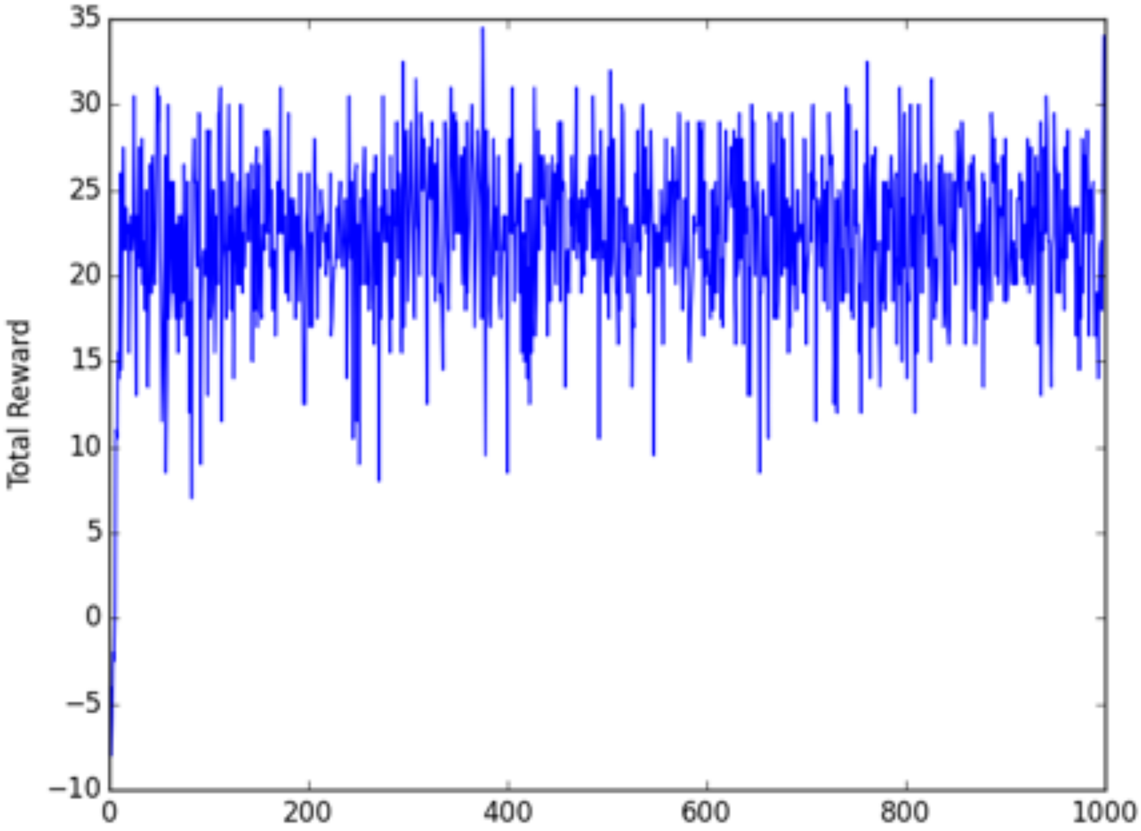
2 b) The robot is able to delivery packages to both customers and avoid thief on its way there. This is much better compared to 2 a). This is because thief is known to the robot so it can take into account of this new risk factor, better evaluate the situations and learn the policy.

2 c) I figured as long as epsilon is above 0.01 when learning rate is 0.1 it is fine (accurate and fast). I tried a combination of different numbers. In general it is better to put off rewards with smaller learning rate and explores more. And we got higher reward per episode in episodes.txt

3

The observation is the rewards value starts with a negative one, and later keep oscillating until converge to a medium one. The plot for average is on the next page.




Total Reward
35
30
25
20
15
10
5
0
−5
−10
0
200
400
600
800
1000