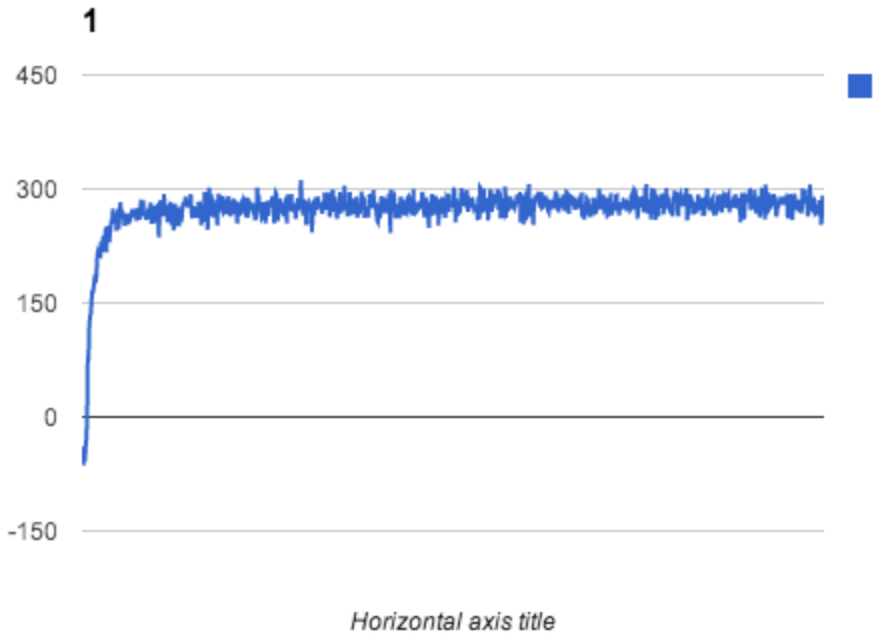
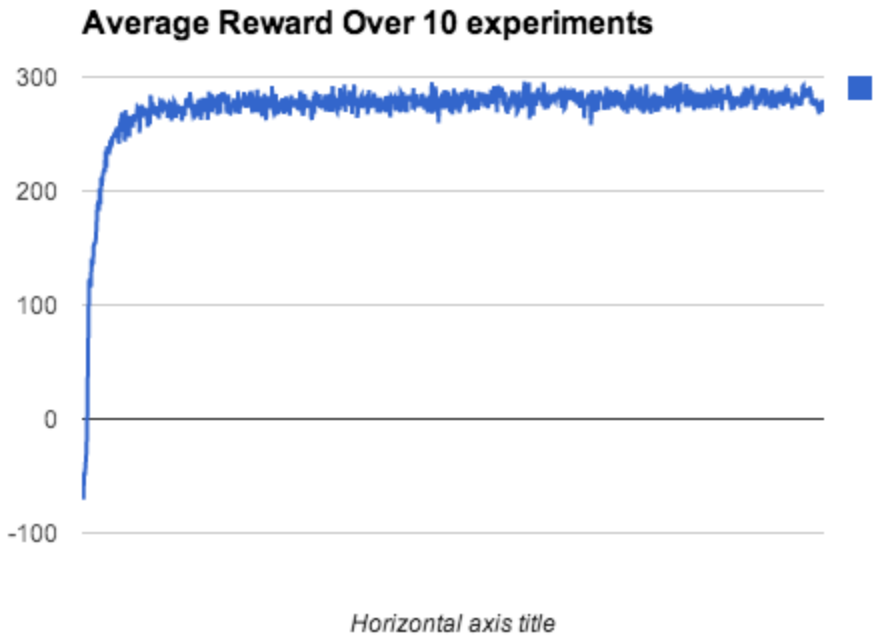
1.
   a) When $\varepsilon = 0$, the agent doesn't pass any slipperiness because the agent is following a pure greedy policy. Therefore, the agent does not go to slipperiness positions with high possibility because in most cases the expected award at the slipperiness positions is negative. The agent thus always wanders around the region where expected reward is zero.
   b) When $\varepsilon = 0.1$, the agent can successfully deliver the packages to destination, avoid more slippery positions and go back to the company again with high possibility. Because the randomness makes the agent explore different positions without following the greedy policy, it can have the chance to step into the positions with negative reward but with positive real utility, such as the slipperiness positions, which lead to the destination with positive reward.
   c) As ε gets larger, the agent can sometimes succeed to deliver packages because it has more chance to explore a random position, but after ε gets too large, the reward of each single episode decreases because the behavior of the agent is closer to a random agent and does not follow and update the policy in an optimal way given the known information within each episode.
2.
   a) The agent does not get close to the thief in the world with thief but unknown thief's position because without knowing the position of the thief, the agent does not have extra states to store the information about the thief's position, and thus cannot have different actions at a same position to respond to the different positions of the thief. Therefore, the agent doesn't have the ability to deliver the package while avoiding the thief. The policy thus constrains the agent stay around the company to have 0 reward instead of meeting the thief to get negative reward.
   b) When the position of the thief is known, the agent can learn the policy to avoid the thief and deliver the packages for most of the time. Because the agent now have extra states corresponding to the different positions of the thief, the agent has the ability to avoid the thief by choosing different actions at a position.
   c) The optimal ε is around 0.01 and the optimal learning rate is around 0.05. For searching the optimal parameters, I fixed one and search for the optimal of the other by using searching through 0 - 1 with decreasing step size to find out the value that makes the reward of each episode to converge fastest to a certain value.
3.





We can observe that the average reward of each episode increase rapidly at the beginning few episodes and converge afterwards.