## MP1 Part 2

1. In WorldWithoutThief:

   **(a)** Every episode has a reward of zero or less when the discount is set to 0.9, the learning rate to 0.1, and epsilon to 0.0. It faces the risk of slipping at fewest three puddles to collect a reward, therefore, it'd make sense to just stay put.

   **(b)** The reward varies greatly, and is largely positive. The magnitude of the negative values is far less than the average magnitude of the positive values. Now that the robot makes decisions that leads to a potentially worse immediate result at the cost of greater payoff later.

   **(c)** For epsilon equals 0.5, the reward tends more negative and the overall magnitudes are less than for epsilon 0.1, as expected if taking random movements more often.
   For epsilon equals 0.75, the rewards are overwhelmingly negative, again as expected when taking random movements more often.
   Generally as epsilon increases, reward deteriorates.

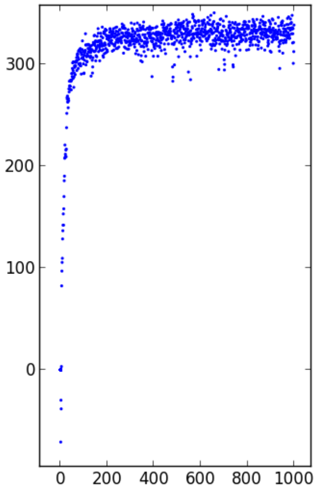
2. In WorldWithThief:

   **(a)** The rewards are overwhelmingly negative with and extremely few small positive rewards. This is consistent with what we'd assume, as the robot cannot avoid the thief if it doesn't know where he is.

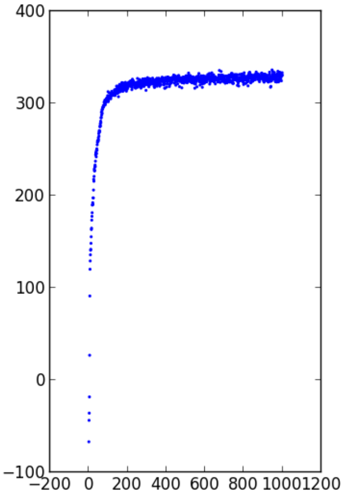
   **(b)** The rewards are extremely and overwhelmingly positive, as expected when the robot can now evade the theif.

   **(c)** The reward tends to peak at low values of epsilon (i.e. $\epsilon \leq 0.1$), and learning rate less than 0.1.

3. The reward grows logarithmically, and seems to converge between 300 and 350.
   Reward vs episode, one run



Reward vs episode, ten runs averaged


400
300
200
100
0
−100
−200
0
200
400
600
800
1000
1200