MP1

Daniel Pugliese

September 30, 2014

Question 1 - WorldWithoutThief

a) With ϵ set to 0 the robot moves into a corner and then stops moving, as it tries to leave the grid. This is because with no random decisions the robot will always take the action with the best Q-value. However, because it initially has a Q-value of 0 for all actions in all states, it will never try a new action (if ties are not randomly broken) and so, once it reaches the corner, it will never try to do anything except for move off of the grid, which will never result in any movement.

b) With ϵ set to 0.1 the robot does explore, and each episode has a different reward value. Initial episodes have low rewards - negative even, as the robot moves to slipery squares and drops the packages. As time progresses the reward values of episodes goes up, and after only 23 episodes, the reward is consistently positive. It does still go quite low sometimes, and fluctuates between high values (above 200) and low ones (below 10) for the remaining episodes. This is because 10% of the time it acts randomly, so even though it may be building a good policy, it is acting off policy and not always reaping the benefits.

c) Higher values of ϵ lead to worse results in general. With $\epsilon = 0.2$, a reward over 200 is never seen. With $\epsilon = 0.3$ there are only a few episodes with rewards over 100. With $\epsilon = 0.4$ there are episodes with negative rewards mixed in with episodes with low positive rewards throughout the entire simulation. And with $\epsilon = 0.5$ there are only a few episodes that have positive reward values. This behavior is because even though the robot may know what a good policy is, it will continue to take bad actions because it acts randomly so often. The high ϵ value leads to rapid exploration, which is good, but also means that the robot does not actually benefit from anything that it learns.

Question 2 - World With Thief

a) With no knowledge of the thief, the robot never had a positive reward from an episode. This is because without knowing about the thief, the robot can't learn accurate Q-values, because two states that it thinks are the same may actually be very different (because the thief is in a different position). Thus, it will walk into the thief and never learn why this is bad - it will associate the penalty only with location, unaware of the other factor at play.

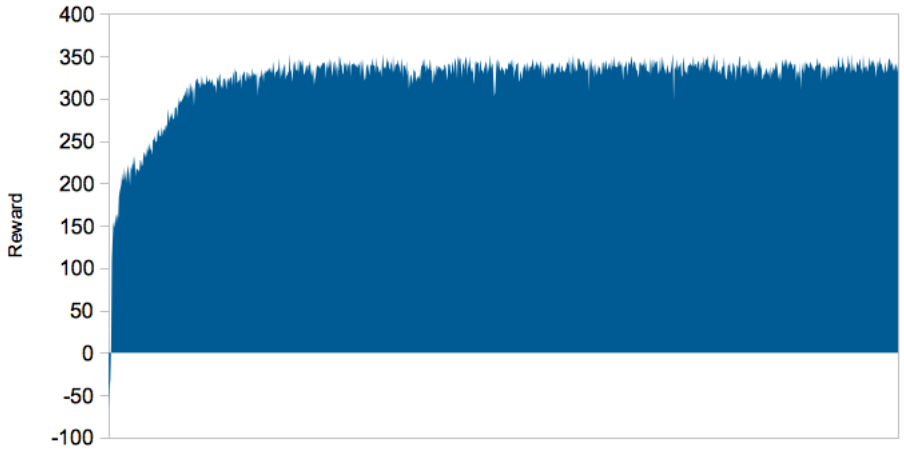
b) With knowledge of the thief the robot does considerably better, getting positive rewards in ever episode after the sixth. The drastic change is because the position of the thief now changes which state the world is in, so the robot

can acurately associate penalties and rewards with the states that they belong to. It will avoid moving into a state where it occupies the same location as the thief once learning that those states have penalties, and so it accumulates high rewards by moving around the thief (and slippery squares).

c) By performing a binary search over the posible values of ϵ and the learning rate (between 0 and 1 for both), I found the best results at $\epsilon = 0.008$ with a learning rate of 0.15. I started with both at 0.5, then halved the range and saw if increasing or decreasing lead to better results. Once the changes stopped leading to significant differences in the outcome I decided that the precision was high enough and recorded the values.

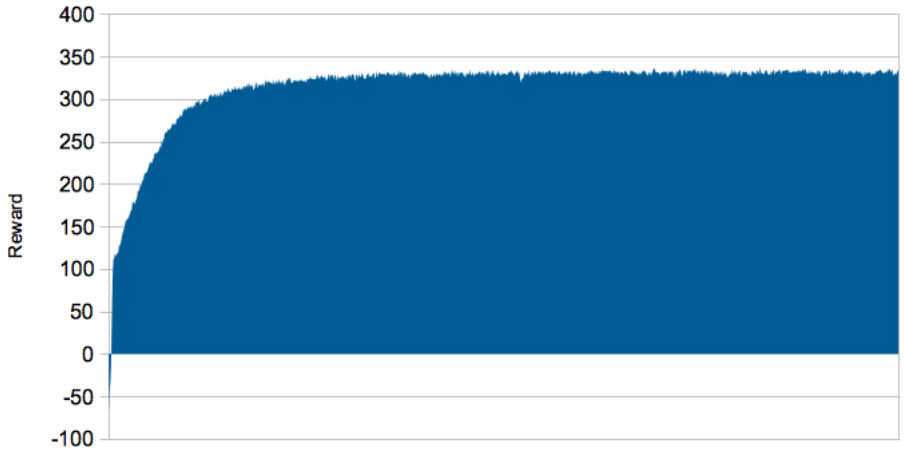
Question 3 - Graphs

a) Below is a graph of the rewards from 1000 episodes:



The reward values of later episodes is much higher than early ones, and it can be seen that a plateau is reached around the 150th episode, with only slight fluctuations after that.

Now we'll look at a graph of the averaged reward of each episode over ten simulations:



The graph is similar to that of one simluation, but the flucuations are less extreme. This is because the average reward from an episode over multiple simulations will have the effects of randomness dulled. With more simulations contributing to the average the fluctuations would continue to decrease, and eventually the graph would be flat after reaching the plateau.