CS440 MP1 Part 2

ajmori2

September 2014

1 World Without Thief

a. We see reward is at best 0.0 every time. Because the robot has a high likelihood to slip in a puddle and it must cross at minimum three puddles to get a reward, it is considerably better to just stay at the company and not face a punishment rather than risk the immediate negative reward.

b. When some randomness is introduced, we see skyrocketed rewards (there seems to be a great variance but a good number of them are above 100). Because our robot now may do something that is at the time a worse reward but will later yield a great reward, it can get past the local barrier choosing 0 over a negative reward and actually yield positive ones!

c. As ϵ grow (the system becomes more random), the rewards start to deteriorate and eventually become negative, as past experience dictates our actions less and thus we cannot make a consistent string of logical decisions.

2 World With Thief

a. The robot generally receives negative rewards because it does not know where the thief is and thus has no way to avoid the one part of the simulation that would most frequently give it a negative reward.

b. Now that the robot knows where the thief is and can properly avoid it, the robot is scoring very high rewards.

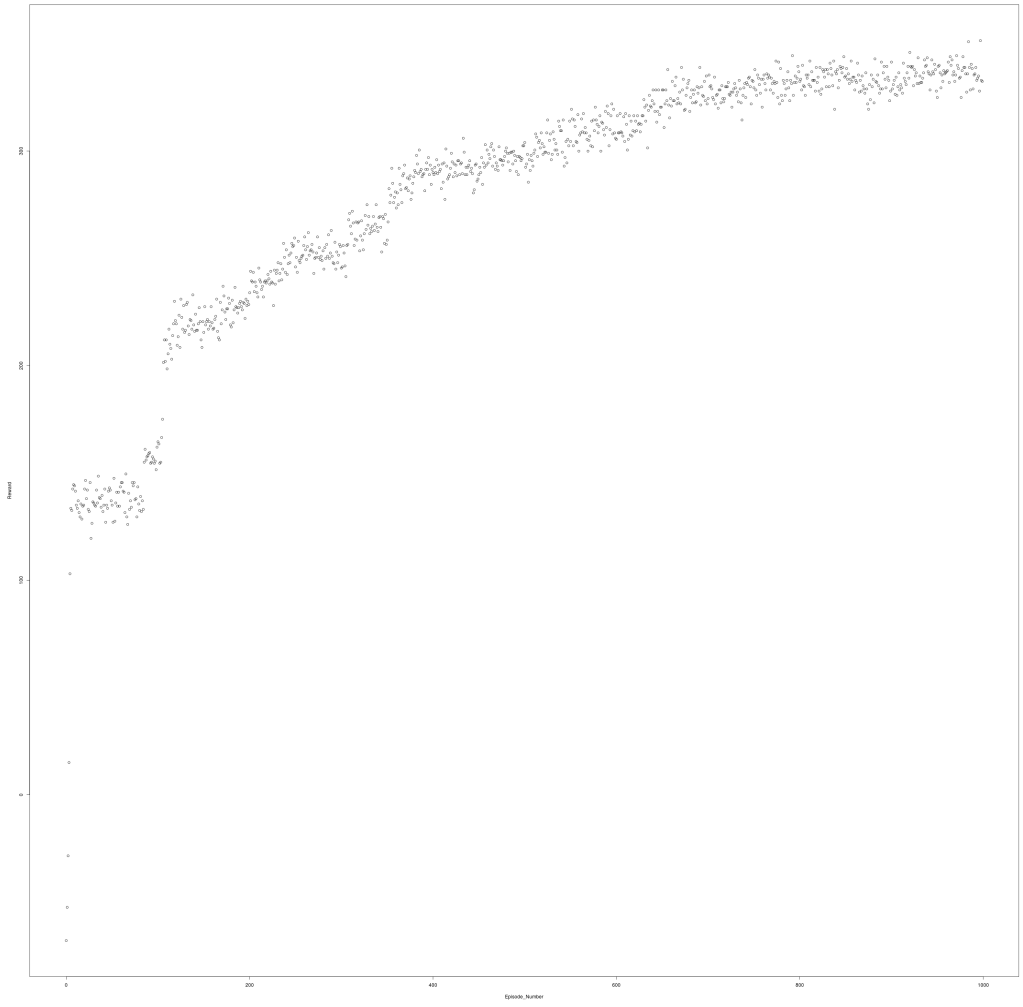
c. Although it seemed logical that a small learning rate (very incremental changes) and a small ϵ (Very little randomness) would be ideal, I decided it was best to check all combinations for rate and ϵ where $0 \leq rate \leq 1$ and $0 \leq \epsilon \leq 1$ out to two decimal points. After testing this, it seemed consistent for all rates that $0 < \epsilon < .02$ consistenly held higher rewards, and $.02 < rate < .06$ yielded the best rates. So I ran the tests again, this time to three decimal places, and found a rate of .047 and ϵ of .006 were

the best, which makes sense, as a large rate would make one drastic measurement near the end completely change our reward, and a large epsilon would yield illogical action choices.

3 Graph

In both graphs, we can see the reward go from very awful (negative) and after only a couple dozen episodes shoot up to high positives. It then slowly converges upon some value that could potentially be the maximum reward. On the average case, however, there is less variance of the reward value.

SingleEpisode



$MultipleEpisodes$