MP1 Part 2

Timothy J. Rogers
tjroger2

September 28, 2014

1 In WorldWithoutThief

1.1 a)

With epsilon = 0.0 the reward value varies very little from episode to episode. This is because as soon as the agent finds a rewarding path it never bothers to try other paths. This is similar to the example from lecture where you go one way and find a dollar and then never explore the path to one hundred dollars.

1.2 b)

With epsilon = 0.1 there is a general upward trend in the reward in subsequent episodes. There is only enough random choice to break out of a greedy loop like in part a but no so much random choice that it causes the agent to consistently ignore its policy.

1.3 c)

There is a general negative trend in reward with higher epsilon, like 0.5. The agent has too high of a chance to choose randomly instead of adhering to the policy and so it frequently makes poor choices.

2 In WorldWithThief

2.1 a)

The agent achieves a fairly consistent negative reward in each episode. This is because without knowledge of the thief's location the utility of each state cannot be accurately determined. State transitions that might lead to the thief, a negative reward, could compute as a positive reward because the thief's location is unknown.

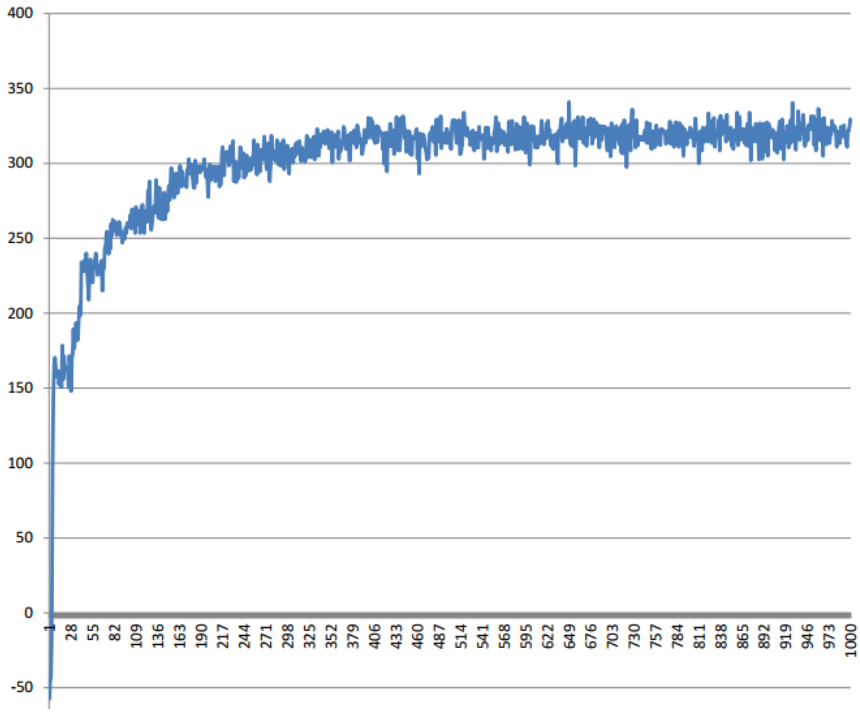
2.2 b)

With knows_thief = y the agent achieves a positive reward trend similar to that in WorldWithoutThief with a good epsilon value. This is because when the thief's position is known each states utility can be accurately calculated just like when there is no thief.

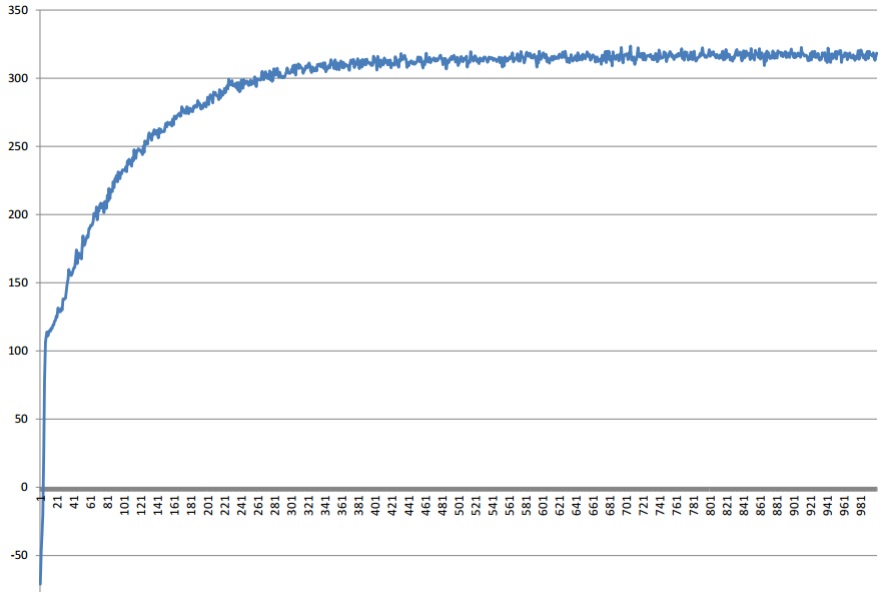
2.3 c)

The optimal value for learning rate is 0.045 and the optimal epsilon is 0.02. I determined these values by checking combinations of learning rate and epsilon values in increments of 0.005.

3 Graphs



In the graph of 1000 episodes of 10000 steps there is a general positive trend and then the reward begins oscillating around a consistent value. This is because the agent has found the best policy it can and the random actions from the epsilon value cause it to vary from the optimal reward.



The graph of the 10 simulation average shows the same general trend and oscillatory behavior but the amplitude of the oscillations are much less extreme. This is because the average mitigates the effects of some of the random epsilon based behavior.