Jason Troutner
troutne2
9/28/14

CS 440 MP1 Part 2

1) WorldWithoutThief
   a) With ε set to 0.0, the agent consistently returns policies with a reward of 0. The robot will always run straight north to the corner and then stay there. This phenomenon occurs because the robot will never take the risk of moving into a slippery square to the east. By avoiding all squares with penalties, and never making it to the squares with rewards, the robot consistently receives a reward of 0.
   b) By keeping all settings from part a, but changing ε from 0.0 to 0.1, the agent quickly discovers a policy with a positive reward. Initially the agent returns policies with negative rewards due to the randomness ε. After 15-20 episodes, however, the agent has learned enough to choose a policy with a positive reward. The element of randomness initially produces very poor policies, but uses the feedback on these policies to later decide on a policy with a positive reward
   c) If ε is increased to a much larger value, such as 0.5, the agent often does not find a policy with a positive reward within 1000 episodes. This occurs because the agent is not given enough opportunity to make decisions based on the information it has learned. Instead, the randomness dominates and the resulting policy does not produce a helpful result.

2) WorldWithThief
   a) With knows_thief set as n, the agent is not able to produce a policy with a positive reward. This happens because the unknown thief effectively introduces an element of randomness that the agent is not capable of learning about. For instance, the agent may decide that the best course of action is to walk right into the thief, because in a previous episode reward was received for the same action with the thief in a different location.
   b) With knows_thief set to y, and all other parameters left the same, the agent quickly converges on an effective policy. This happens because the agent has enough information to completely describe the state. With this information, the agent can know that an action a, taken in state s, will be helpful, because state s describes the location of the thief.
   c) I was able to achieve the best results using a learning rate of 0.22 and an epsilon value of 0.04. My process to find these values involved keeping one parameter constant, and making a small change to the second parameter. I ran multiple tests for a slightly higher parameter value, for a slightly lower parameter value, and for an unchanged parameter value, and then selected the one that converged most quickly. I alternated parameters to vary until I reached a point where my chosen values were better than the slightly higher and lower values. This point would therefor correspond to an optimum combination of learning rate and epsilon.

3) Plotting expected discounted reward for each episode in one trial (Figure 1) shows a rapidly increasing reward at the beginning, that quickly converges to high values around 300. Due to the effects of randomness, the reward still varies between 250 and 300, but remains high. When averaged over ten trials (Figure 2), we see more consistent asymptotic behavior. The reward still quickly increases to about 300, then stays relatively flat.

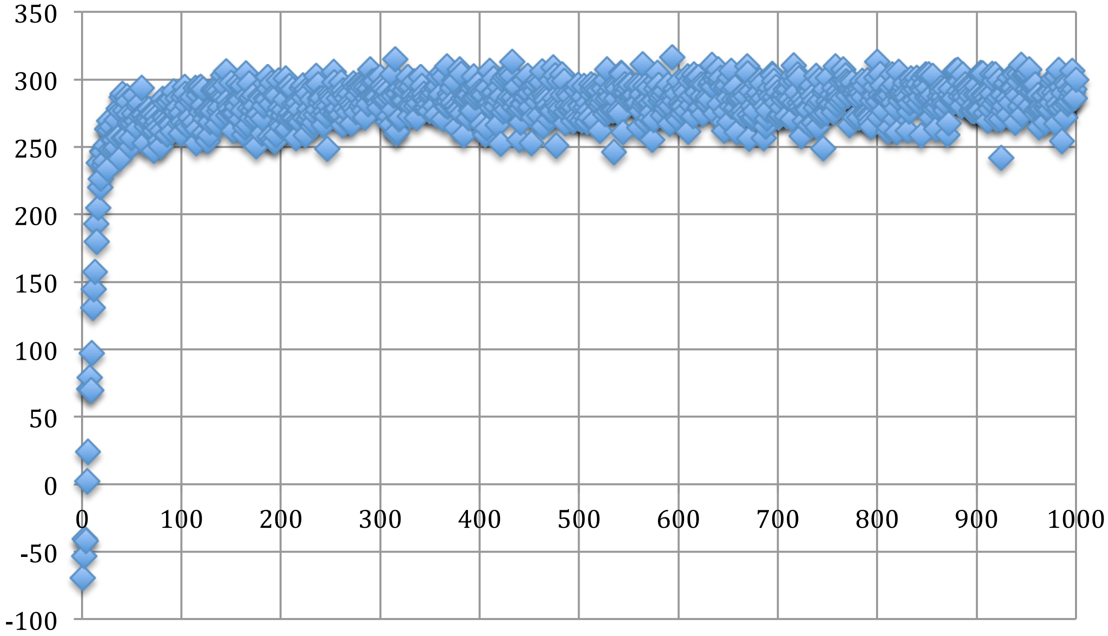

Figure 1. Expected discounted reward vs. episode for one trial

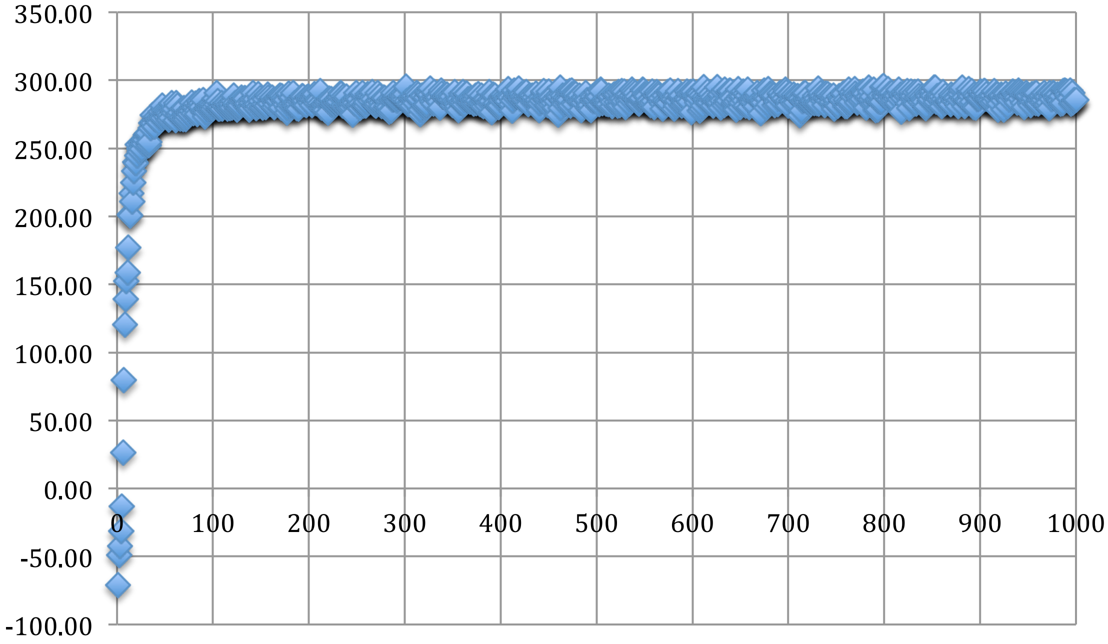

Figure 2. Expected reward vs. episode averaged over ten trials