# CS 440 HW1

mhalm2

September 7, 2014

## 1

(a) I consistently see a reward of 0.0. When you have an epsilon of 0, you have no chance of randomly exploring new options; the learning algorithm always chooses what it believes to be the best option. Therefore, reward-giving paths that are missed in the first episode will always be missed, as the algorithm will not take the chance at randomly exploring them.

(b) The episodes quickly converge on a reward range of 100 to 200, with a small fraction of low reward episodes. It reaches higher values than $\epsilon = 0$ because it explores more options. The small fraction of low reward is due to the low epsilon value. The algorithm will choose random paths few times, and the episodes during which it chooses suboptimal paths result in low rewards.

(c) For $\epsilon = 0.5$, every episode has a negative reward. This happens because the epsilon value is so high that choosing enough high utility states in a row to get a reward that takes many steeps is extremely unlikely.

## 2

(a) The episode rewards are all negative. Has the bot is unaware of the location of the Thief, to the algorithm, seemingly random locations have negative reward, so the Q-Learning model doesn't account for it well and is unable to learn how to avoid it.

(b) The episode reward consistently converges on a 275-300 range. With the knowledge of the Thief, the agent can properly learn how to avoid it.

(c) My initial investigation involved checking the last 20 results of each $\epsilon$ in $0.0, 0.1, 0.2, ...1.0$ and found that 0.1 and 0.2 were the only viable options. Further investigations on the 0.01 scale around these values revealed 0.01 as the yeilding the highest and most reliable data.

A similar search scheme and measurement of viability was used for the learning rate, which revealed that 0.2 provided the best average scores.

## 3

As shown in the figures on the following page, the algorithm seems to converge on an expected value of around 320 at an exponential rate. When averaged over 10 trials, the variance amongst trials close to 320 decreased dramatically do to random distribution due to $\epsilon$ being concealed by averaging.

The algorithm learns at an exponential rate due to its greedy nature. It very quickly "decides" which paths are the best and takes them almost all the time.

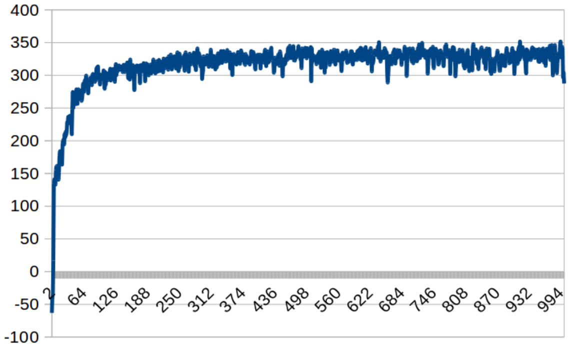

Figure 1: One Trial

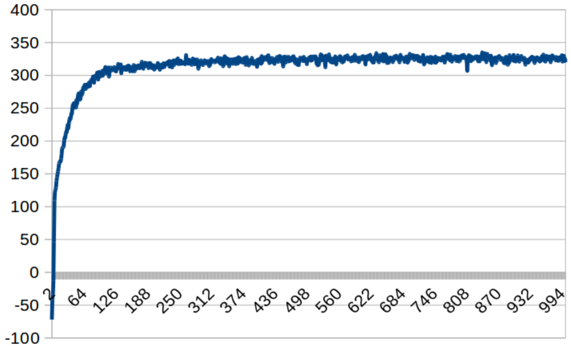

Figure 2: Ten Trials