MP1 Part 2

Jonathan Chao

September 29, 2014

1

(a) The rewards for the first two episodes are negative, then the rest are zero.

(b) The rewards for early episodes are negative, then become large at the end. This is different from the previous experiment because we made ϵ non-zero; we allowed the agent to explore rather than be purely greedy.

(c) The vast majority of the episodes result in negative rewards. This is because we have allowed too much randomness in our exploration.

2

(a) The rewards are mostly negative.

(b) The rewards start out negative, then quickly grow to about the 200-300 range. This is because the agent knows the position of the thief, and can thus factor that into its decision of the best policy.

(c) I found the best α to be about 0.2 and ϵ to be 0.01. I found that a learning rate that is too high will result in the rewards converging extremely quickly, but they are not really optimal. I also found that having an ϵ that is too high will result in negative rewards.

3

The rewards start out negative in the first few episodes, then quickly grows to about the optimal value after about 100 episdoes. It hovers around the same range of values for the rest of the simulation. When plotting ten experiments, they all seem to follow the same pattern of growth and plateau.