Joon Young Seo (jmseo2)
CS 498 Homework 3

Problem 1

a) The reward obtained is 0 all across the episodes. It may be the case that the agent may have to get some minor penalty on the way to get larger reward (delivering successfully). Since the agent does not take any random step, it tries to choose the best action always - it will never try to go to the state where the agent got minor penalty, so the agent will never obtain any reward.

b) The agent started to get some rewards. Unlike a), it takes a random action with probability ϵ, so even if the local reward negative, it explores the world more thoroughly and eventually reach a large reward.

c) The agent gets negative rewards across the episodes. Unlike b), it takes random actions way too frequently. While it may reach a large reward eventually, it has gone through way too many negative rewards on the way, which accumulated as negative total rewards.

Problem 2

a) The agent gets negative rewards (around -10) across the episodes. The agent tries to make the best action to its knowledge, but no knowledge of thief make this attempt futile.

b) Now the agent makes high positive rewards (> 200) across the episdoes. The knowledge of the existence of thief allows the agent to try to avoid the thief.

c)

$\alpha = 0.09$

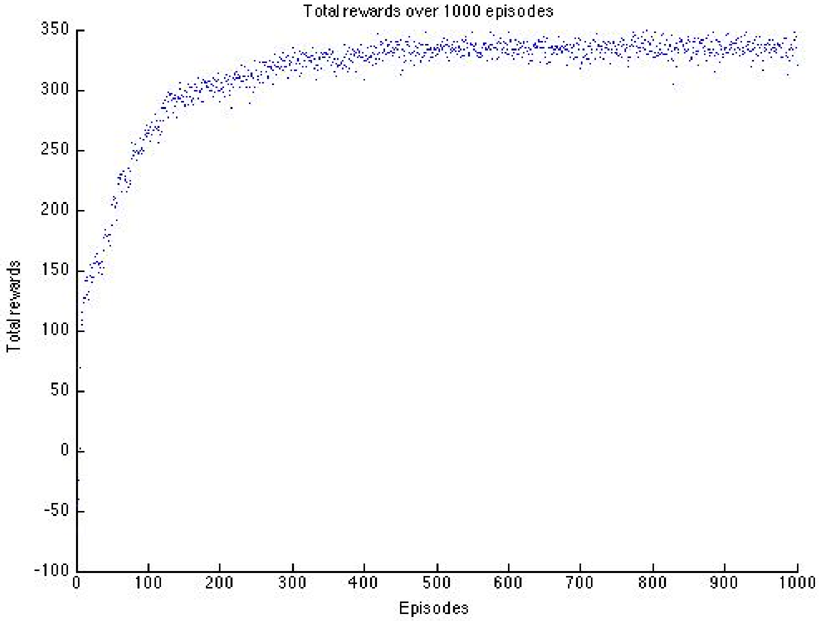
$\epsilon = 0.01$

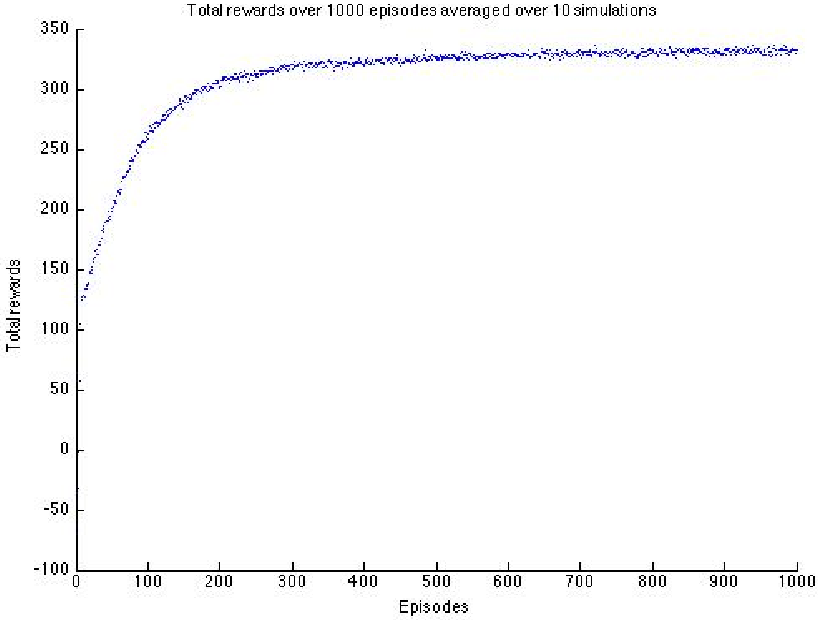
I assumed that the function $f(\alpha, \epsilon) =$ total reward is a convex function and performed some sort of procedure similar to hill climbing.

Start with random $\alpha = a$. Try to find ϵ that maximizes $f(\alpha = a, \epsilon)$. Assuming f is convex, $f(\alpha = a, \epsilon)$ should also be convex.

Increase the value of α, to say, $\alpha = b$, and find ϵ that maximizes $f(\alpha = b, \epsilon)$. If this obtained reward is smaller than the previous, decrease the α. Otherwise, increase. I repeated this procedure until I found what I thought would be suitable values of α and ϵ.

Problem 3





The reward starts out low at the beginning of the episodes, and start to rise at a rapid rate. Then it converges around to some value (around 330) and continues to stay at that value. More episodes will not result in much better rewards.