CS 440 MP1

Victoria Vanderbach, netID: vndrbch2

Part 2: Experimentation

1) In WorldWithoutThief

a) When the agent has no chance to explore, it also has no chance to take risks. It finds the best option is to not take on negative penalties and thus it doesn't gain any rewards. Thus, it finds that the best policy to is simply avoid stepping on a slippery spaces. However, do to the structure of the WorldWithoutThief, the agent finds it impossible to drop off packages and receive positive rewards because it fears crossing the slippery spaces in the center of the map because of the negative penalty associated with them. Without the exploration factor, it will never learn what is on the other side of the slippery spots and will never achieve a reward.

b) Now that the agent has some ability to act randomly, it is able to complete its task and drop off its packages to receive a positive reward. In a 10000 step episode, it can accumulate around 200 reward points now. It also minimizes the amount of time it is on slippery spots while still realizing that it must cross occasionally to receive a reward. The random actions have allowed the robot to take on negative penalties to cross an obstacle and see that in doing so it can achieve a net positive reward in the end. Thus, the agent policy now accounts for risks needing to be taken in order to fulfill the task and achieve the best reward in the end. The agent learns better by taking random risks occasionally to perfect its policy.

c) There is a down side to being too random: the agent takes too many risks now. Being overly random results in large amounts of negative rewards being accumulated that are not compensated for in the net positive rewards the agent gets by delivering packages. At larger epsilon values, the agent is acting more like a random agent rather than a learning agent. The best epsilon, or exploration rate, will result in maximizing the net rewards, both positive and negative.

2) In WorldWithThief

a) The world is more stochastic now with the introduction of the unknown thief. The thief becomes an element that negatively affects the utility of the middle column for the agent. Due to large amounts of testing, the agent has encountered the thief on all the rows of this column, thus decreasing the utility of entering the middle column drastically. The agent now sees the middle column as undesirable, and will not cross it like it did with the slippery spots in the WorldWithoutThief. Also, the exploration rate is too low for the agent to want cross over this column enough times to find the positive rewards on the other side. The agent chooses to not take on the negative penalties associated with those states and will do nothing to achieve positive rewards.

b) Now that the agent knows where the thief is, it understands that the middle column has two states per entry in it, one where there is a thief and one where there is not a thief. It can now distinguish between these two states and treat them with different utilities unlike it was doing before. The agent

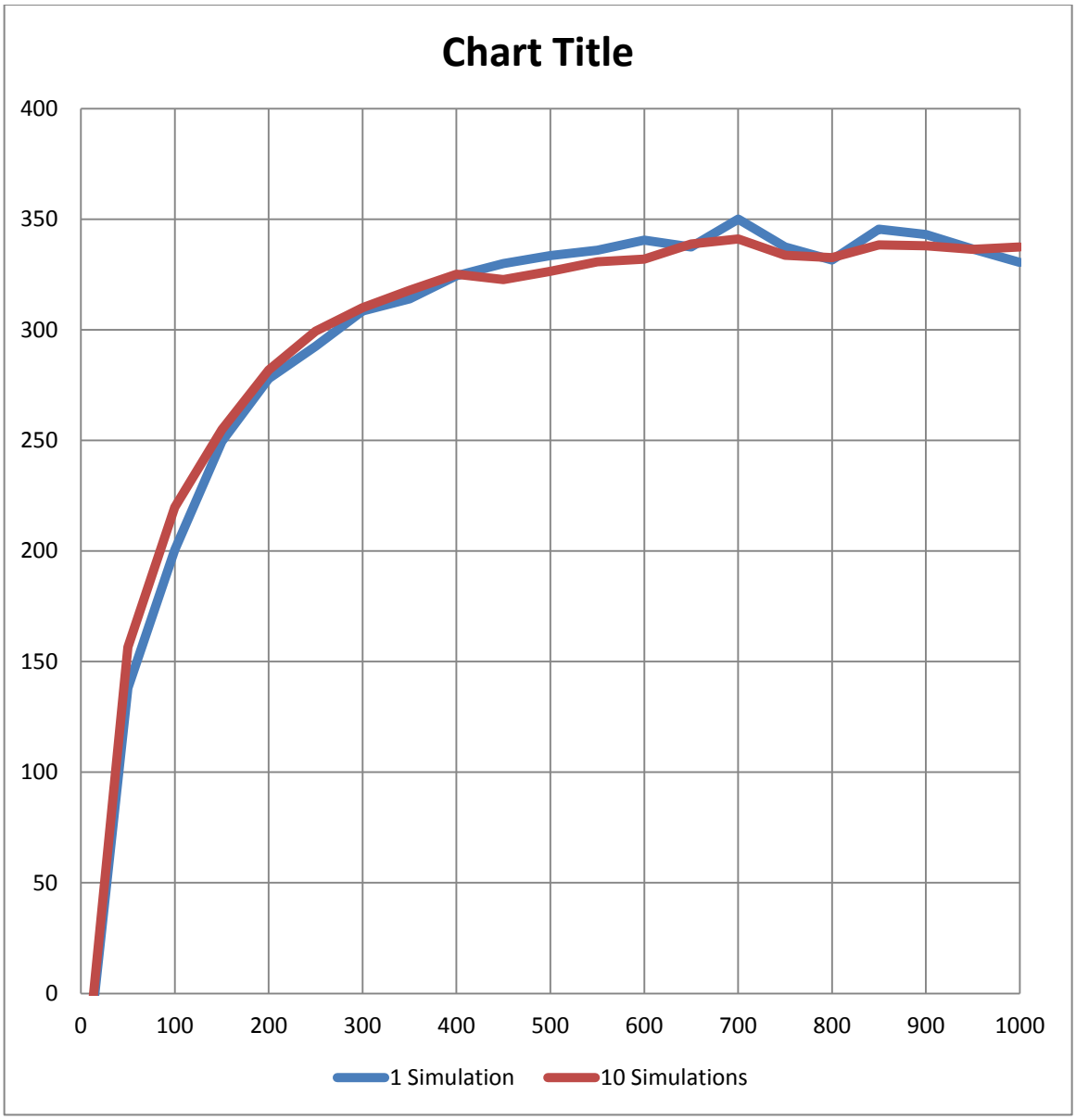
will now take measures to avoid entering a state where a thief is, but will also see that there is a path around the thief where it can reach its positive rewards. Because the world has now opened up and there is always a way to reach a goal without taking on a negative penalty, the agent can act accordingly to achieve the max positive reward.

c) Results: epsilon = 0.005, learning rate = 0.1

| Epsilon | Learning Rate | Rewards |
|---|---|---|
| 0.0 | 0.1 | 130 |
| 0.05 | 0.1 | 280 |
| 0.1 | 0.1 | 200 |
| 0.075 | 0.1 | 250 |
| 0.03 | 0.1 | 310 |
| 0.04 | 0.1 | 300 |
| 0.02 | 0.1 | 320 |
| 0.01 | 0.1 | 340 |
| 0.005 | 0.1 | 345 |
| 0.0025 | 0.1 | 340 |
| 0.0075 | 0.1 | 340 |
| 0.005 | 0.2 | 330 |
| 0.005 | 0.05 | 330 |
| 0.005 | 0.15 | 340 |
| 0.005 | 0.4 | 330 |

First, I tried maximizing rewards by varying epsilon and found the max rewards were achieved at 0.005 with a learning rate of 0.1. The max benefit of exploration seems to be when the exploration rate is close to zero, but the max point is hard to pinpoint. It is like gingerly approaching the edge of a steep cliff. The highest point is at the very edge, but go too far and you fall off. Epsilon at 0.005 seems to be close to that point where rewards are maximized.

Next I tried playing with the learning rate. The learning rate curve seems to be flat for a good range of values, but has a max around 0.1. It is less drastic of an impact on overall performance to alter the learning rate by a little than it is for altering epsilon. However, too little ability to learn or too much ability to learn will cause performance to suffer.




3) The agent drastically learns the better and better policies early on, but struggles to improve the policy further over time. It reaches a horizontal asymptote and fluctuates between making a policy better and making it worse. The rate of growth in reward acquisition with more episodes is extremely small and requires significantly more episodes to improve only a small amount. The Q learning algorithm probably gets stuck at a local minimum, and requires a lot of hill climbing to even try and explore more regions. More randomness would improve this hill climbing effort, but at the same time reduce the likeliness that the agent is actually learning instead of just guessing.