Q1.a.

Reward for all episodes are zero. First one may have negative. All episodes have reward zero as agent always follows greedy approach so it does not dare to follow the slippery path. It makes the policy of an agent almost random.

Q1.b.

Most of the episodes get positive reward as the agent follows random path when probability is less than 0.1. So the agent may dare to explore the slippery path and choose greedily the best path to deliver among the available ones when probability is greater than 0.1. The path followed is mostly optimal always.

Q1.c.

Negative rewards are observed. This may be due to the high value of epsilon which was set to 0.5.The approach of the agent for learning is random or greedy with equal probability leading agent take bizarre steps and may follow costly path. The behavior is unpredictable at one simulation agent may deliver packets with a reasonable reward and immediately after second simulation it may learn the worst path getting trapped in slippery areas.

Q.2.a.

Epsilon = 0.05, Knows Thief = n

The reward is negative in most of the learning episodes. Which means in addition to the ignorance about the thief, the agent is greedy which stops him from travelling to slippery areas trapping him in safe path in which thief cannot enter. Agent is not travelling across a path of thief.

Q.2.b.

Epsilon = 0.05, Knows Thief = y

Agent is almost acting greedy. The reward is positive for learning episodes and agent is avoiding the thief properly for getting maximum reward. The agent is not getting trapped in the safe path. As agent knows about the thief and acting greedily, it can consider a path close to optimal one avoiding the thief.

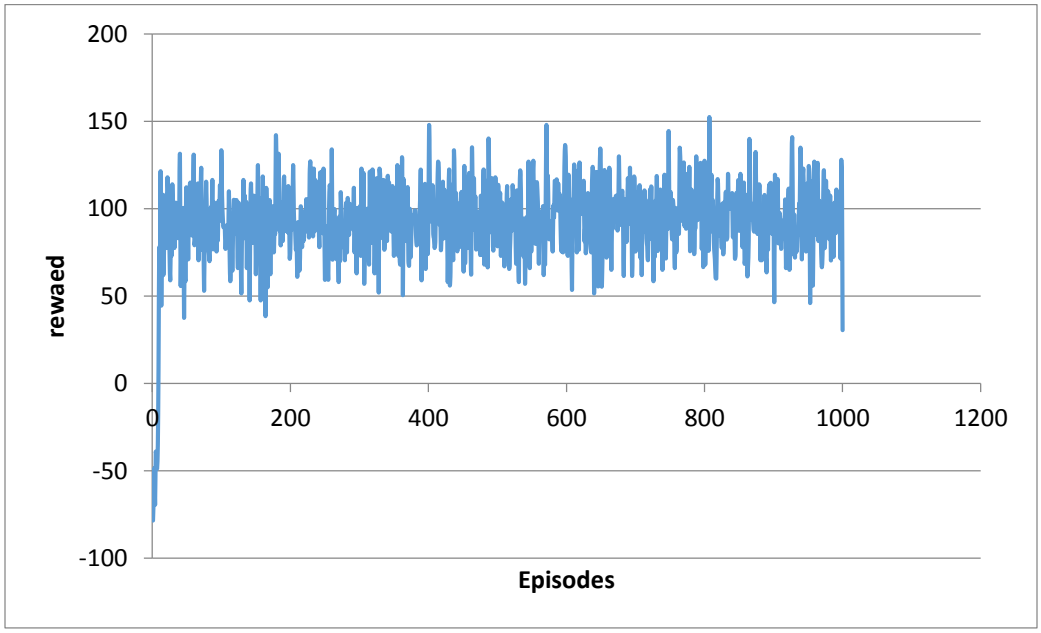
Q.2.c.

Epsilon = 0.2, Knows Thief = y

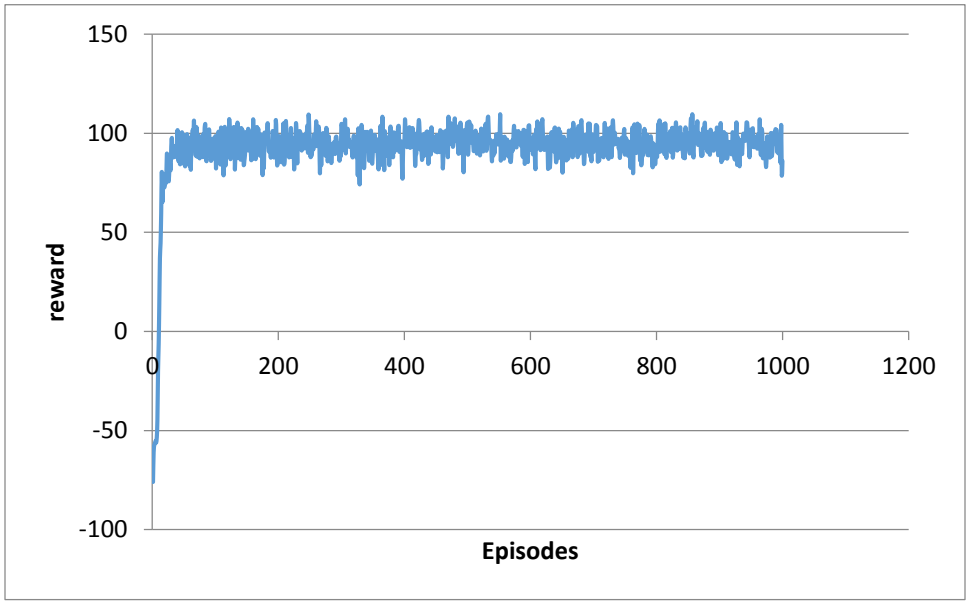
For epsilon =0.2 the agent behaves wiser, he follows optimal path. As the e = 0.2, the random choice is taken not to the great extent only once in five times probably. Which makes agent follow the optimal Q value for learning. Too much greedy (less epsilon value) makes agent restricted to the safe path and get trapped. The high epsilon value makes the agent follow bizarre path ignoring the optimal one. Also for many subsequent simulations the agent followed almost optimal path for values of 0.2.

Q3.

First Simulation:



Average of 10 Simulations:



From the first graph we can observe that the agent does not get the same reward all the time, also the reward is also not increasing after every episode. This must be due to some random decisions when probability is less than 0.2. The most of the high reward can be considered the outcome of being greedy and choosing optimum Q value for deciding actions.

The second graph which is the average of 10 simulation shows that the reward fluctuation is bit lower than first one. The average reward over 10 simulation is bit concentrated around 100. It can be due to the average effect which may be diminishing the random decision factor as epsilon is set to 0.2 and most of the time greedy decisions take place.

Extra Qestion: For the eligibility traces,

I got the negative reward values for lambda = 0.9 and 0.95 (with World with Thief and epsilon = 0.2 learning rate = 0.1, discount =0.9). The agent gets trapped in non-optimal path after delivering packet 2.

The expected observation should be the reward should get increment after subsequent episodes. As the last 36 actions (which I am keeping track in a LinkedList to implement a queue) get penalized for a wrong move or get rewarded for correct goal completion.

However in case of World with thief, the Eligibility traces may harm the learning as, it last wrong action may result in collision with thief even though the previous actions were part of optimal path. This will only slow down the learning and will need more steps and episodes.