MP1 Part 2

Submitted by

Vineet Agarwal

1) In WorldWithoutThief
   a) Setting discount factor to 0.9, learning rate to 0.1 and ε to 0.0, the agent was simulated for 1000 episodes of 10,000 steps each. It was observed that all rewards are zero. As there is no randomness in the system, the robot is not exploring and hence not learning.
   b) As soon as the ε to 0.1, randomness is introduced into our system and robot starts exploring new states. It some cases it will reach the goal of delivering, in some cases it will slip or get robbed, and it will learn from these mistakes.
      Upon running the simulation multiple times, it was observed that for few initial steps, the reward is negative, which soon becomes positive (20~30 episodes) and then remains positive for the remaining episodes. Average of rewards for the simulation was 136.
   c) Setting ε to 0.5, we observe that learning drops and rewards after each episode is negative. The robot is taking too many random actions which is not resulting in good reward. Average reward in this case is about -28, which is much lower than earlier case.
2) In WorldWithThief
   a) Setting ε to 0.5 with the robot not having knowledge of the thief, we observe that reward in most episodes is negative. Average reward for a case was observed to be -12. This makes sense as the robot doesn't have knowledge of thief, it is unable to learn how to avoid negative rewards associated with thief.
   b) With knowledge of thief, robot does much better with reward values converging to about 275. It makes sense as now the robot can learn how to avoid negative rewards associated with thief.
   c) To look for best parameters, I modified the code to output the average reward achieved in last 100 episodes of 1000 episode simulation.
      Keeping the Learning rate constant, I varied Epsilon, I observed that for much larger values, rewards are lower. Optimal value of Epsilon was 0.01.

Epsilon	Learning Rate	Reward
0.1	0.1	213
0.05	0.1	274
0.075	0.1	244
0.04	0.1	289
0.025	0.1	307
0.01	0.1	325
0	0.1	98
0.15	0.1	154

Keeping epsilon fixed at 0.01, I varied learning rate and tried to achieve maximum average reward –

Epsilon	Learning Rate	Reward
0.01	0.05	320
0.01	0.075	332
0.01	0.1	330
0.01	0.07	325
0.01	0.06	331

Rewards are very similar for Learning value in the range 0.6-0.75. A value of 0.07 was picked as optimal.

3) Figure 1 shows rewards for each episode for one run and figure shows rewards for each episode averaged over 10 runs. In both cases, the final reward converges to a value close to 325, but in figure 2 we see that variation is much lesser due to averaging.

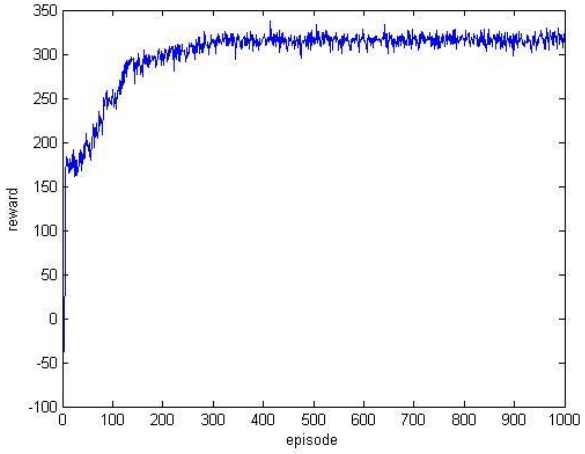

Figure 1: Reward for each episode for single run

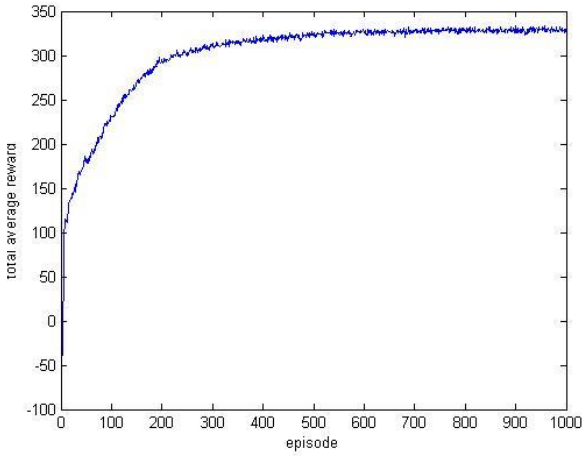

Figure 2: Reward for each episode for ten runs