Haneul Kim
kim705
ECE448/MP1.2

1. In WorldWithoutThief

(a) The reward was negative for the first episode, and 0.0 from second to 1000th episode. The ϵ value denotes the the percentage of random action. Since the agent does not perform any random actions, it learns to avoid negative reinforcement but doesn't explore for the possible higher reward. So the agent moves back and forth where there is no slippery slope with possible negative reinceforcement.

(b) The reward increased to positive values. The agent performs random actions so it explores for the possible higher reward, and it eventually learns the way to get higher rewards. However, the randomness is limited to once out of ten times, so the agent's action doesn't primarily depend on the randomness. This allows the agent to learn to achieve positive reward once the randomness leads it to the goal.

(c) The reward decreased significantly to negative values. Setting ϵ too high leads the agent to move randomly too frequently instead of taking the learned actions. This leads the agent to perform worse than setting the ϵ to zero.

2. In WorldWithThief

(a) Throughout the 1000 episodes of 10,000 steps, the average reward is -12.9. Since the thief's move is not taken into account when learning, there is a very high possibility of encountering the thief and getting a negative reward. The first reward was -28, which is significantly higher than the average reward, so the agent learns to avoid the slippery slopes but not thief. After the first reward the reward is consistent.

(b) The reward is positive value in [200, 300], which is increased rapidly from episode 0 to episode 20. The reward from the first episode was -82 but it rapidly improved and then stabilized after episode 20. The negative reward from the thief influences the next action and the agent learns to avoid the action that might lead to the possible thief state.

(c) The best learning rate is 0.1 and the best ϵ value is 0.018. The difference is minimal within [0.04 0.5] for the learning rate, and within [0.01 0.03] for ϵ. As the learning rate decreases, the reward goes to negative values, and as it increases the reward decreases slowly. This is because the rate determines how much weight is put on the new information to reflect on the next action. When ϵ reaches to zero, the expected reward decreases but doesn't go negative. However, when it increases, the expected reward decreases to negative values. From these observations, we can conclude that there is optimal values of ϵ and learning rate, with which the agent performs the highest given other parameters constant.

3. After 1000 episodes of 10,000 steps with ϵ of 0.018 and learning rate of 0.1, we can see that the expected discounted reward increases rapidly at first 50 episodes and gradually increases in a slower rate. It stabilizes at around episode 300 with the reward value of approximately 330. Although it has some invariance from episode to episode, the general trend seems to stabilize. After running this nine more times and averaging it, the trend seems more clear that the reward gets stabilized around episode 300.

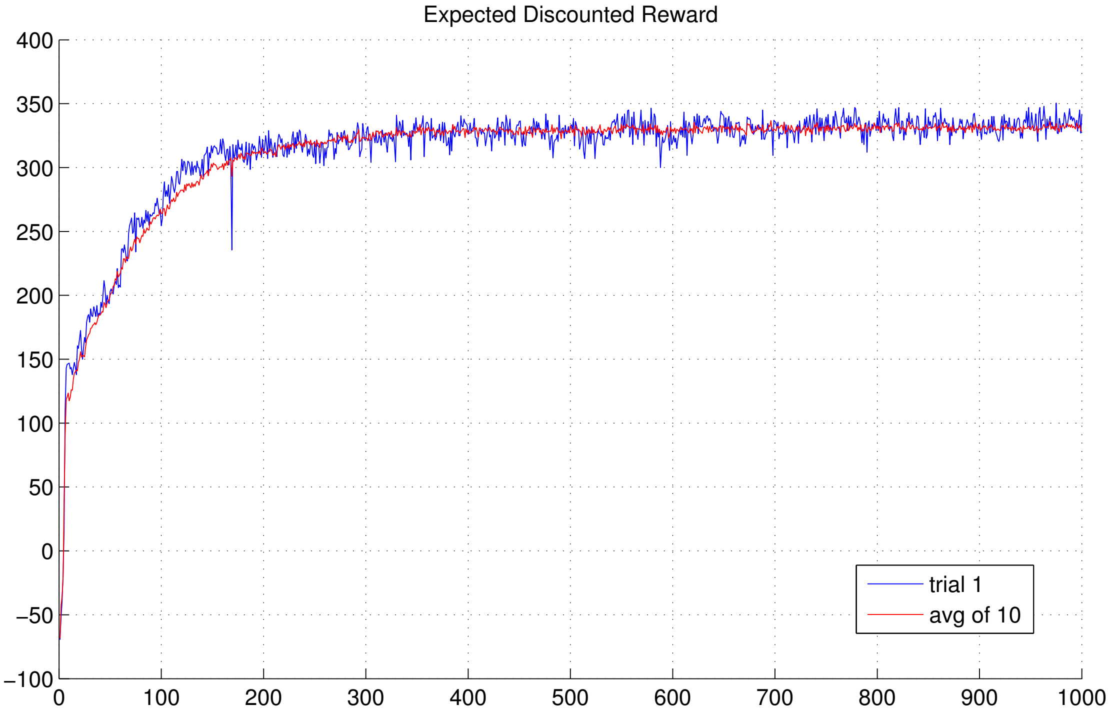

Figure 1: Expected discounted reward over 1,000 episodes of 10,000 steps