MP1

Gaurav Bansal

September 2014

Problem 1 In World Without Thief

Part A
Discount Factor: 0.9
Learning Rate: 0.1
Epsilon 0.0
1,000 Episodes with 10,000 Steps
I observe that a positive reward is never found. The agent uses the randomly generated policy and follows it with no randomness, thus it never finds a reward or updates the policy

Part B
Discount Factor: 0.9
Learning Rate: 0.1
Epsilon 0.1
1,000 Episodes with 10,000 Steps
I observe that a positive reward is found and the policy is updated. The randomness allowed the agent to deviate from the original policy and eventually find a reward, which in turn updates the poilcy. All rewards are positive since the agent will tend toward using the policy

Part C
Discount Factor: 0.9
Learning Rate: 0.1
Epsilon 0.5
1,000 Episodes with 10,000 Steps
I observe that both a positive and negative rewards are found. The randomness allowed the agnet to deviate from theoriginal policy and eventually find a reward, which in turn updates the poilcy. However, even though we have a policy that will point us toward positive reward, the randomness facotr is too high and the agent deviates from the policy more often.

Problem 2 In World With Thief

Part A

Discount Factor: 0.9
Learning Rate: 0.1
Epsilon 0.05
Knows Thief: n
1,000 Episodes with 10,000 Steps
I obeserve that we always find a negative reward. Since the agent doesnt know there is a thief on the board, it will move regardess of the thiefs location. This means he will keep walking into him and recieve a negative reward. The non-zero randomness allows teh agent to explore the field.

Part B

Discount Factor: 0.9
Learning Rate: 0.1
Epsilon 0.05
Knows Thief: y
1,000 Episodes with 10,000 Steps
I obeserve that we always find a positive reward. Since now the agent knows about the thief a decision based on the thiefs location can be made such that a negative reward is not recieved.The randomness is still there to help the agent explore.

Part C

Discount Factor: 0.9
Knows Thief: y
1,000 Episodes with 10,000 Steps
This references the figures 1,2,3, and 4 provided. The optimal combination I found is a Learning Rate 0.2 and an Epsilon of 0.01. I started by fixing the Rate at 0.1 and interating thorugh epsilons from 0.005, 0.01, 0.05, 0.1, and 0.2. This number array was choosen based on which made the reward go up and stopped when it started going down. I did the same Eplison array for Rates of 0.2 and 0.3. These cases seems to bound the maxima in terms of max reward and convergence time. It seemd that lower values cause the solution to converge slowly. As I raised values the solution converged faster to a similar maximum. If I continued higher, the convergence rate was the same but the maximum started to reduce. I picked the cases that converged the fastest to the highest maximum. This happen to be all the runs at Epsilon of 0.01. I then plotted the 3 learning rates a for Epsilon of 0.01. Learn rate og 0.1 and 0.2 were similiar but 0.1 seems to have large ourliers, so I settled on a rate of 0.2.

Problem 3

Discount Factor: 0.9
Learning Rate: 0.2
Epsilon 0.01
1,000 Episodes with 10,000 Steps

Figures 5 and 6 have my answer to this problem. I notice that the agent rapidly converges to a reward of about 330. The reward after the agent converges is still noisy however.This makes sense since the noise is cause by the epsilon that is there. When I average 10 runs worth of data, we can see the noise go away and the max reward settle out. This also makes sense since the noise should be random and averaging it over multiple runs should hide it.

EXTRA CREDIT - Eligibility Trace
Figure 7 shows the result of my eligibility trace experiment. I tested values of lamda of 0.95, 0.99, and 0.999. I compared these against my optimal averaged run. They all showed that they found the maximum reward, however a lamda of 0.99 converged noticibly faster than the others. I believe this is because there is now history of the reward that trickles back up the path through the Q value. This allows 'smarter' decions to be made sooner.

Figure 1: Fixed Rate of 0.1

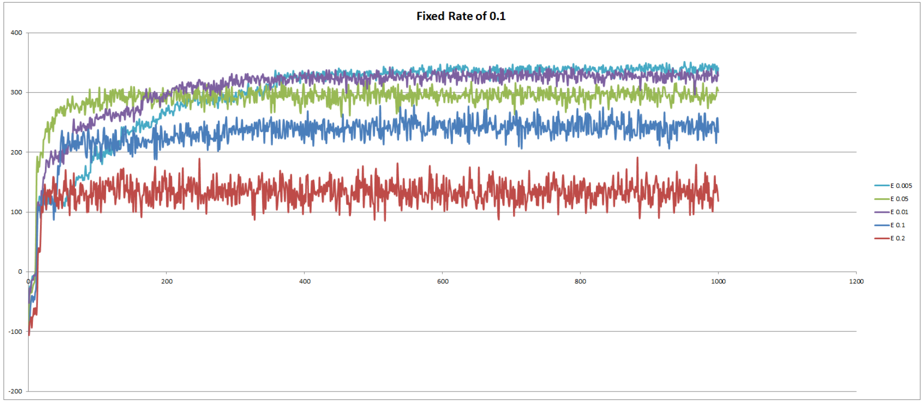


Figure 2: Fixed Rate of 0.2

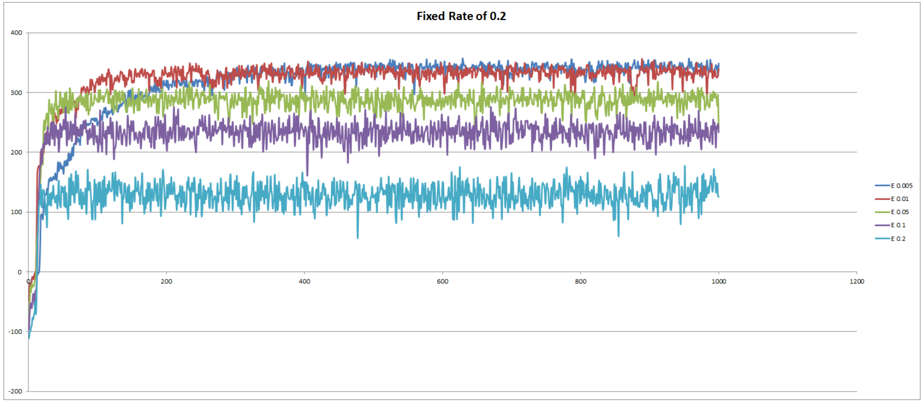


Figure 3: Fixed Rate of 0.3

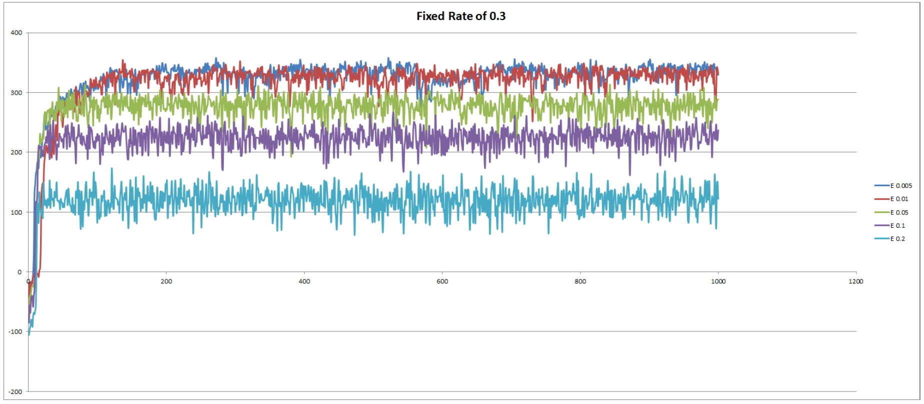


Figure 4: Fixed Epsilon of 0.01

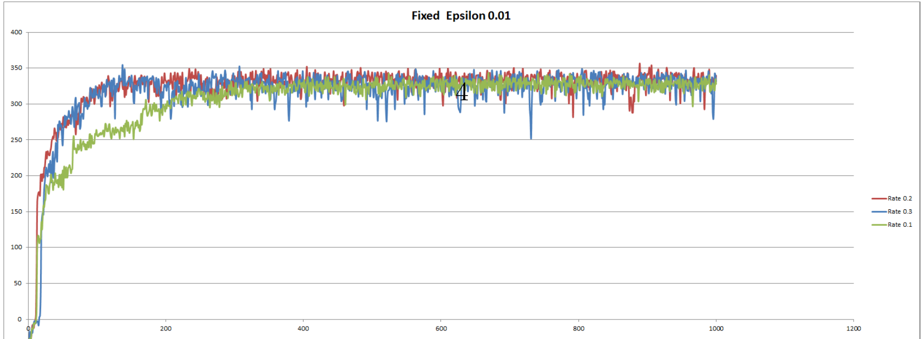


Figure 5: Rate of 0.2 and Epsilon of 0.01

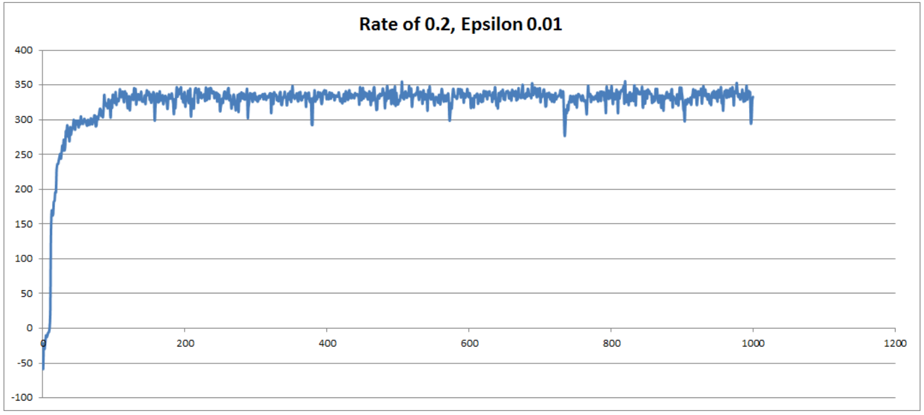


Figure 6: 10 Run Average of Rate 0.2 and Epsilon 0.01

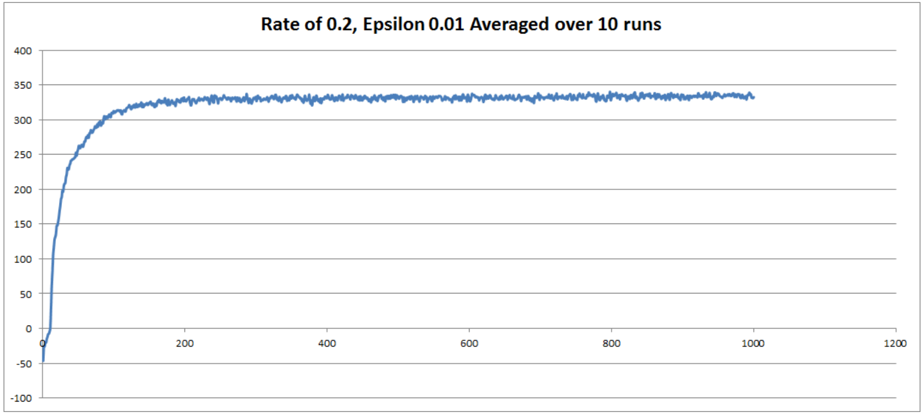
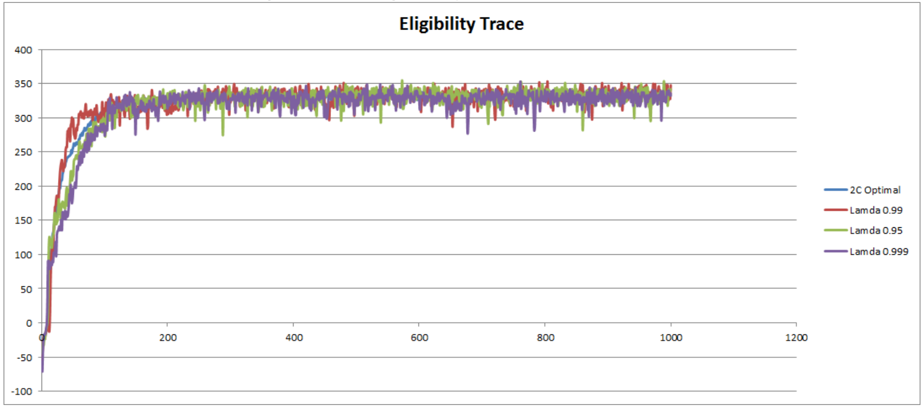


Figure 7: Eligibility Trace Experiment


Eligibility Trace
2C Optimal
Lamda 0.99
Lamda 0.95
Lamda 0.999