Steven Hill
CS 533
Homework 3

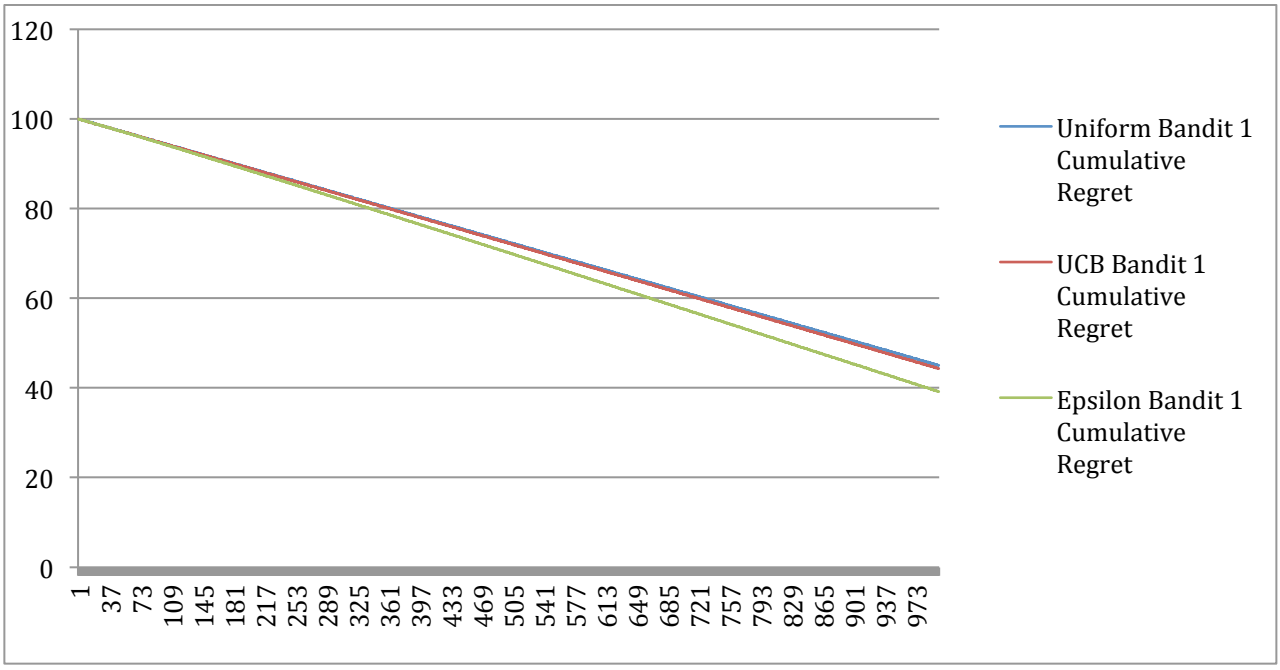
Multi-armed Bandit Algorithms

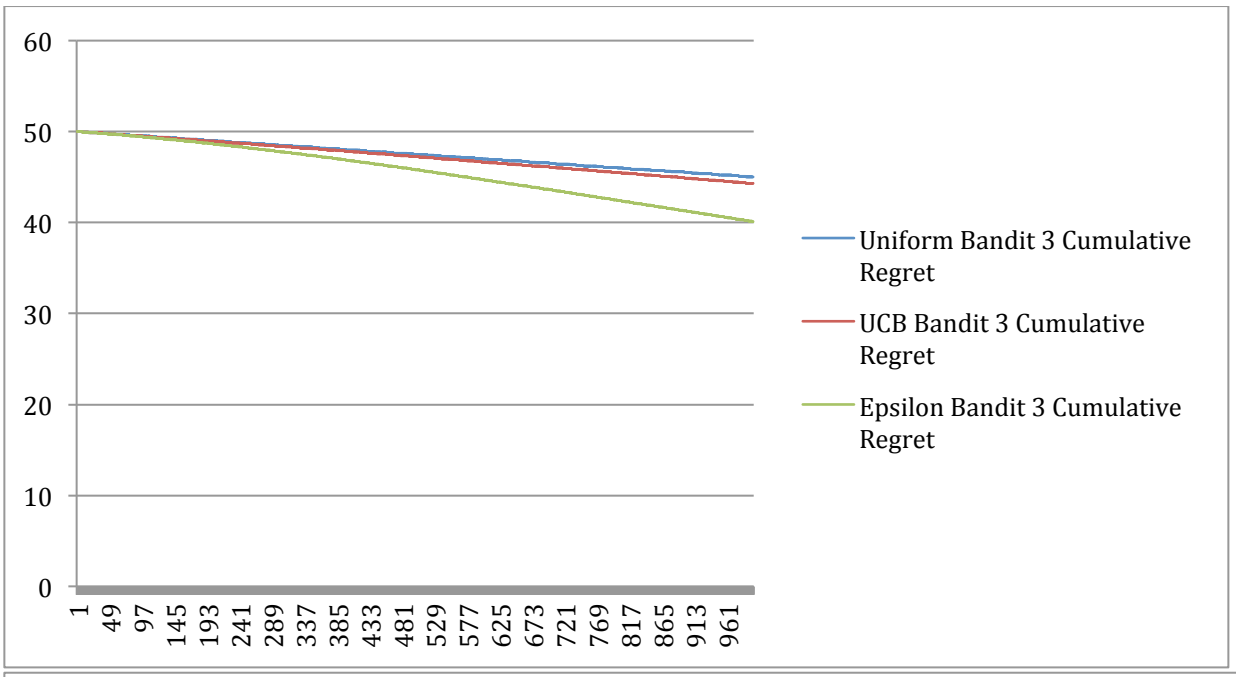
In this assignment we evaluated various multi-armed bandit problems. The experiments performed were based on some passed bandit, n-arm pulls, and t-trials to remove variation. Some interesting things to keep in mind, for both simple and cumulative regret, we "cheated" when calculating the true estimated value of the arm under evaluation. Since we know the true probabilities, we were able to calculate these values to completely evaluate our algorithms.

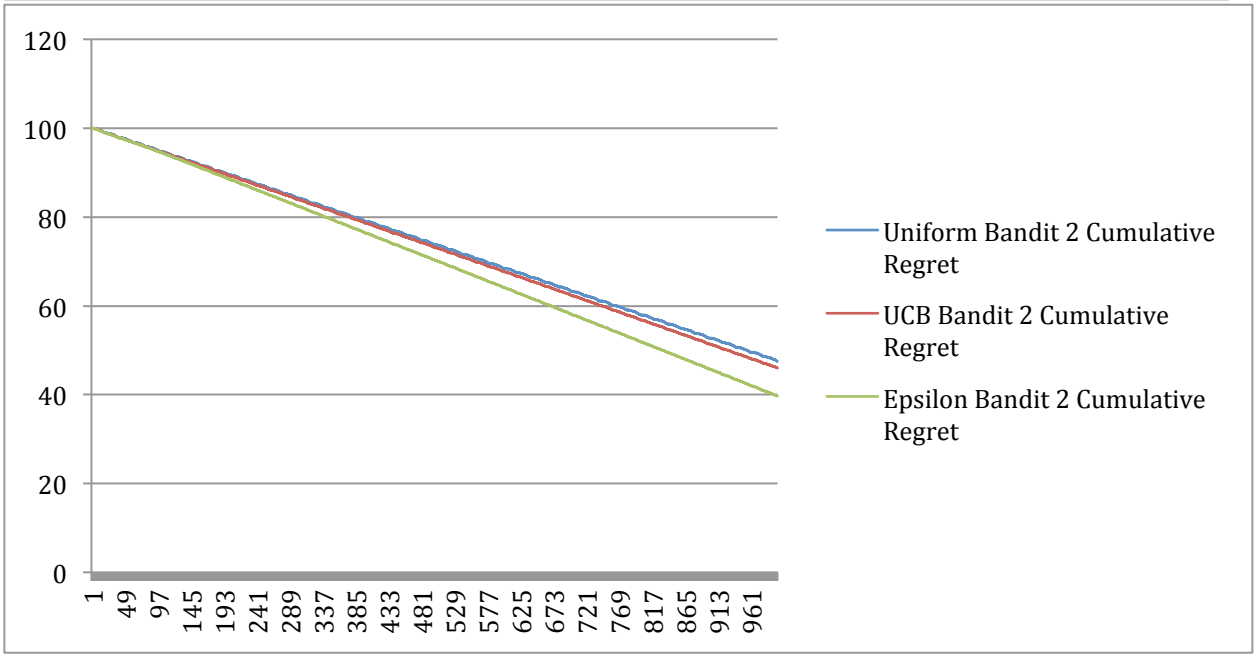
In addition to the bandits provided (bandit 1 and bandit 2), I created a bandit with 10 arms, 9 of these arms always return 0 reward, and one of these arms returns a reward of 1 with .05 probability. This bandit isn't particularly interesting in regard to the other bandits, however the two interesting cases, being "normal", and extreme, were already covered by the two provided bandits. I thought it would be somewhat interesting to observe a needle in the haystack type bandit.

Cumulative Regret

For every bandit the cumulative regret curves look interesting, it is clear the epsilon algorithm performs better than any of the others, however they all still scale linearly relative to each other. This matches the bounds provided, e^-n, e^-cn, and e^-dn. The extreme case, bandit 3, you start to see a more parabolic curve for the epsilon algorithm.





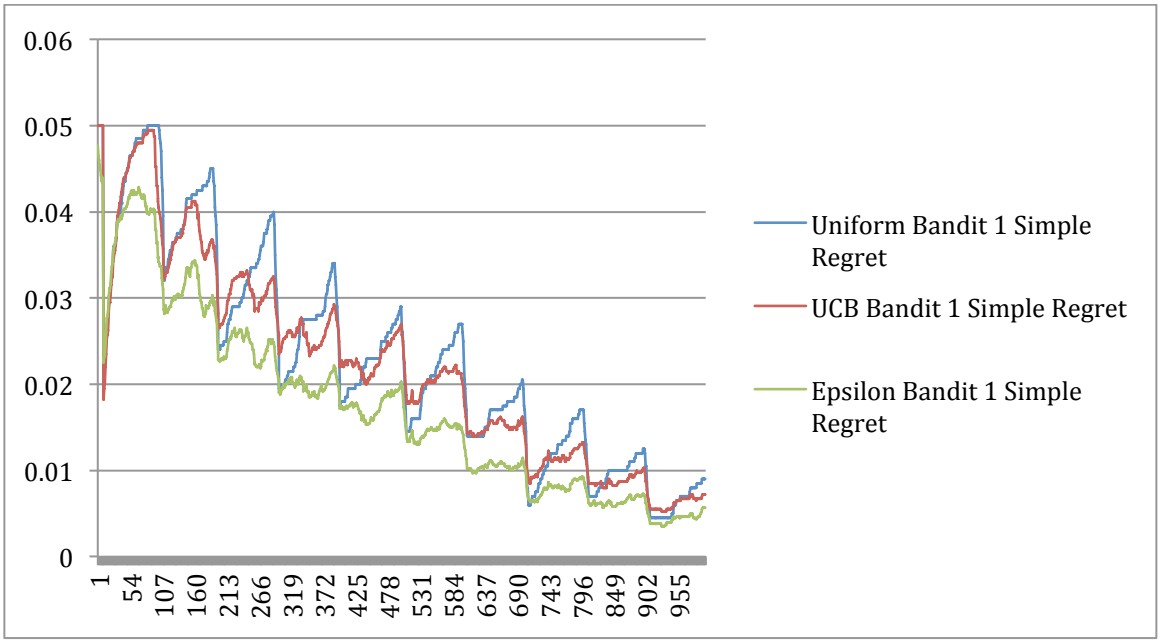


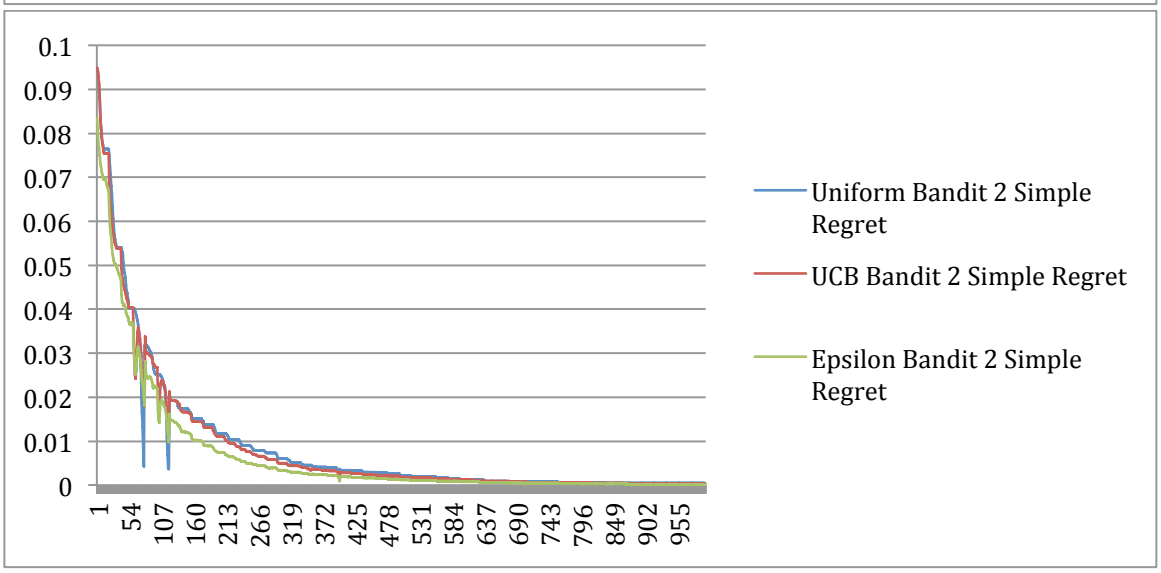
Simple Regret

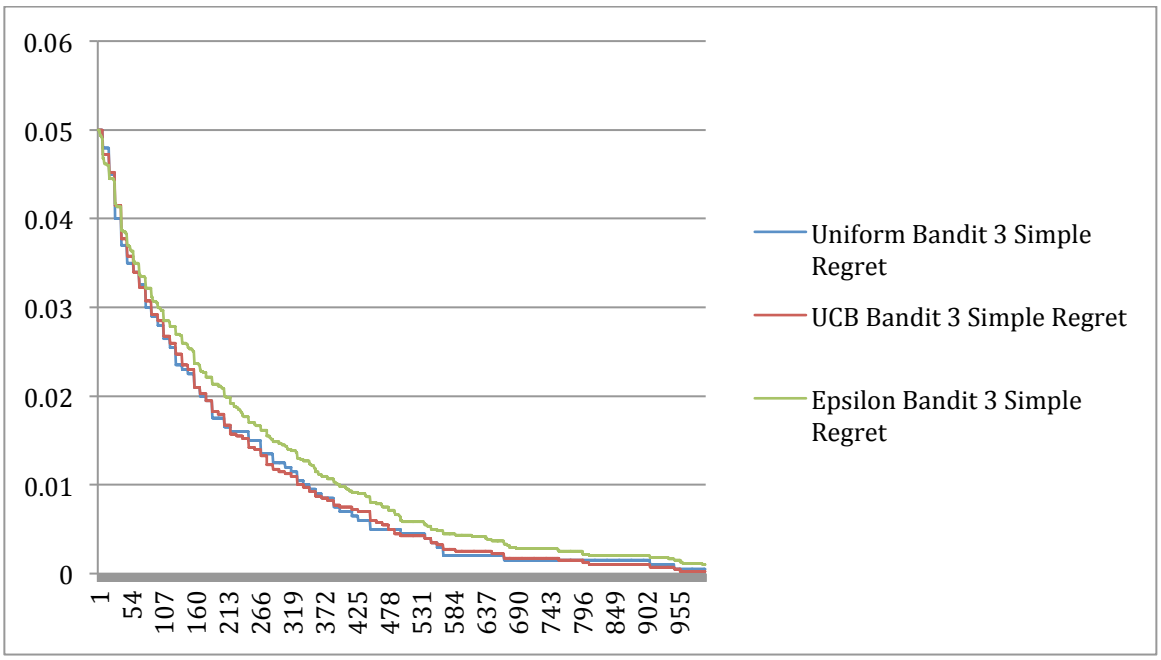
Simple regret preforms as as expected. Uniform and epsilon bandit reach the optimal node more quickly while UCB takes a bit longer to get there. The interesting thing to note about the charts is the consistently non-zero value of simple regret. This is due to the huge number of trials run. Non-zero values just indicate that in at least one trial the correct arm was not identified., however, for the purpose of this assignment it is easy to assume that very small values are in-fact 0.

The first bandit produced a very interesting looking curve. And this can be attributed to the average time it takes the algorithm to find the "better" arm. You get this interesting looking dip in regret which should closely correlate to the cumulative distribution of the arm. Again, the curve doesn't reach zero due to the high number of trials produced.

The final bandit didn't really produce anything interesting. You see the same quick convergence to the optimal arm for all bandits. It may have been more interesting to create very large problem and evaluate that, however that would have probably just led to an upward shift of the same curve.






0.06
0.05
0.04
0.03
0.02
0.01
0
1
54
107
160
213
266
319
372
425
478
531
584
637
690
743
796
849
902
955
Uniform Bandit 3 Simple Regret
UCB Bandit 3 Simple Regret
Epsilon Bandit 3 Simple Regret