Programming Assignment #1

One of my questions after doing the reading was whether optimistic initialization of an epsilon-greedy bandit with epilson>0 would do any better than the epsilon=0 case since the optimistic already forces it to explore quite a bit. Page 36 Fig. 2.4 from the textbook. I implemented (after taking advantage of some starter code for a bandit http://blog.yhat.com/posts/the-beer-bandit.html) the same 2000 bandit testbed with 10-armed bandits for 1000 steps that was described in the textbook.

After analyzing the behavior of the two approaches I realized that my question didn't really make sense because optimistic initialization forces full exploration until the optimal arm is found, and then only chooses the optimal value. However, if two arms have very similar values it can takes a very long time to find the optimal values, but in either case the random sampling of other arms only decreases the expected average reward.

True Values of the 10 Arms for Poor Case

0.228	**-1.205**	**0.361**	**-0.804**	**-0.469**
-0.725	**-1.603**	**0.383**	**-0.277**	**-0.421**

True Values of the 10 Arms for Perfect Case

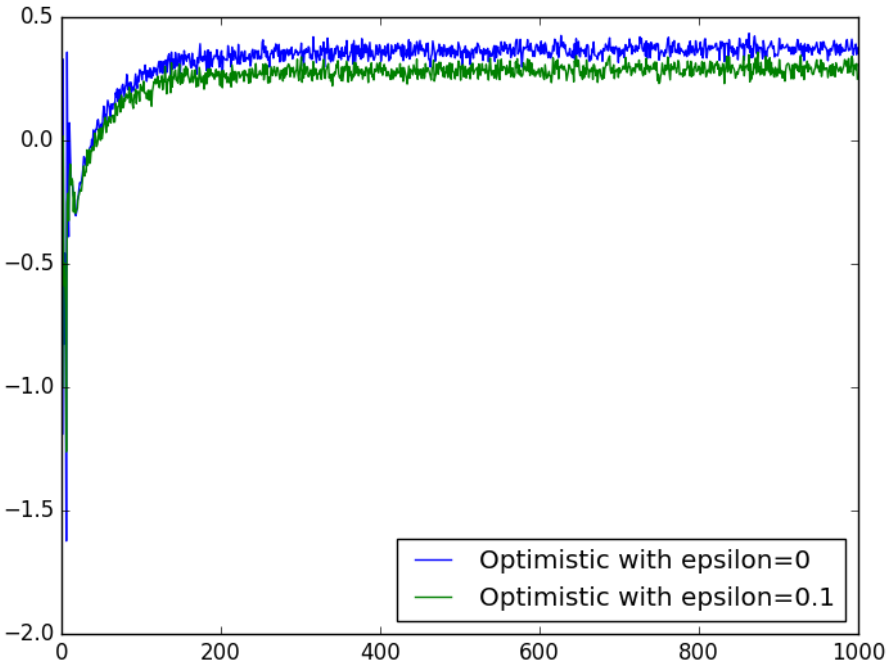
1.506	**-0.020**	**0.684**	**0.942**	**-1.481**
-0.444	**0.116**	**3.189**	**0.264**	**0.057**

We can see that in the poor case the average reward tends toward the average of the two close peak values, and the percentage of correct choice tends toward 50% because it is bouncing back and forth to between the top values. In the perfect case, the average reward nearly immediately reaches the true optimal reward and the percentage of correct choice goes to 100% after only 10s of iterations.

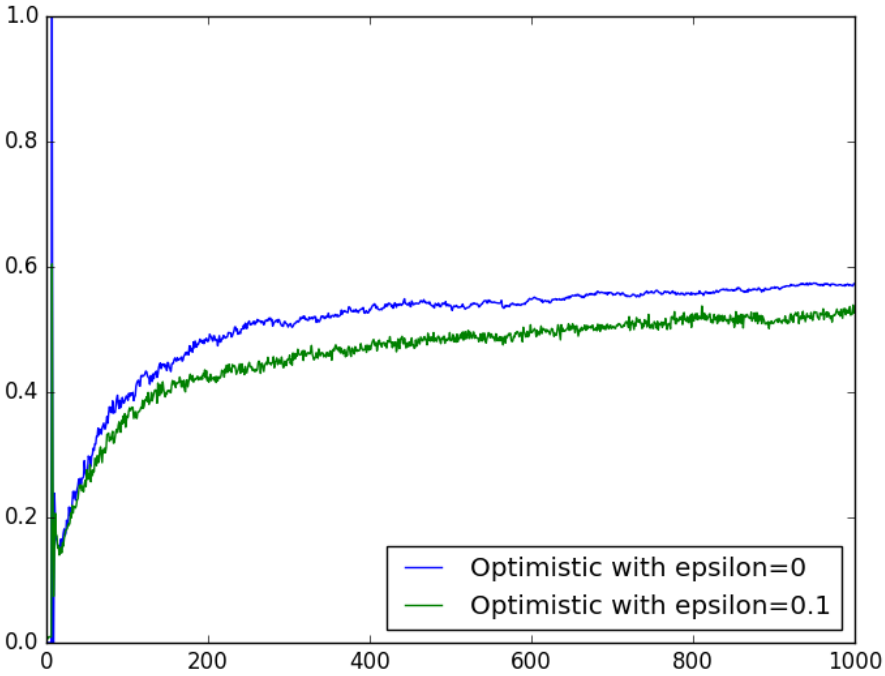
In both cases, having epsilon>0 just degrades performance because it already knows where the optimal or two best choices after a few 10s of choices. Doing further exploration is unnecessary when the rewards are stationary.

To conclude, epsilon>0 provides no benefit to optimistic initialization and actually hurts performance under stationary rewards. It could be that under highly non-stationary rewards that the extra exploration helps, but that would require further investigation.

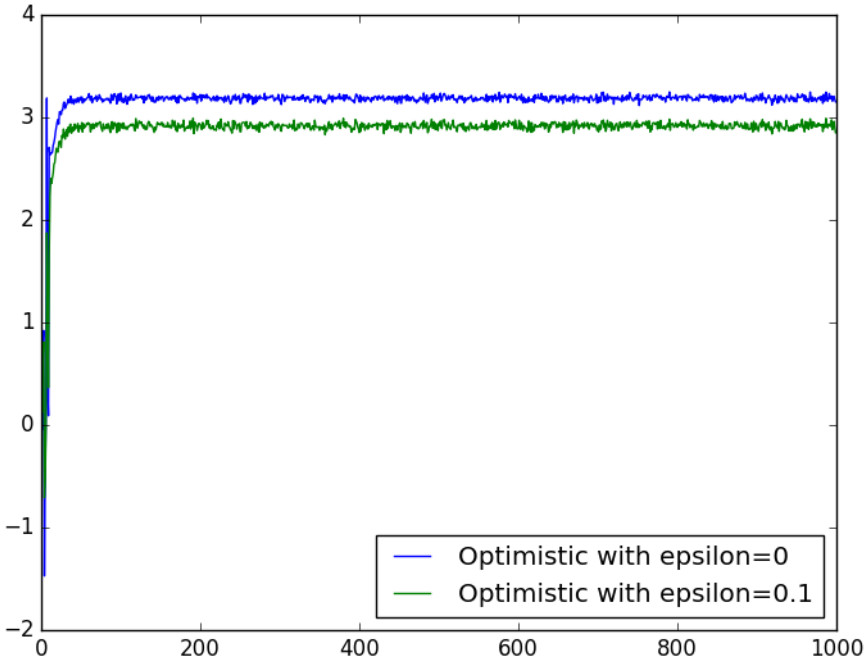
Average Reward over 2000 Bandits – Poor Case



Percentage Choosing the Optimal Value – Poor Case



Average Reward over 2000 Bandits – Perfect Case



Percentage Choosing the Optimal Value – Perfect Case