Programming Assignment 2: 10-Armed Testbed

Robert D. Grant

For this assignment I implemented the 10-armed testbed from Sutton and Barto, Chapter 2. In Figure 1 I show an example of the testbed; I plot sampled distributions of returns in blue and the mean values in red.

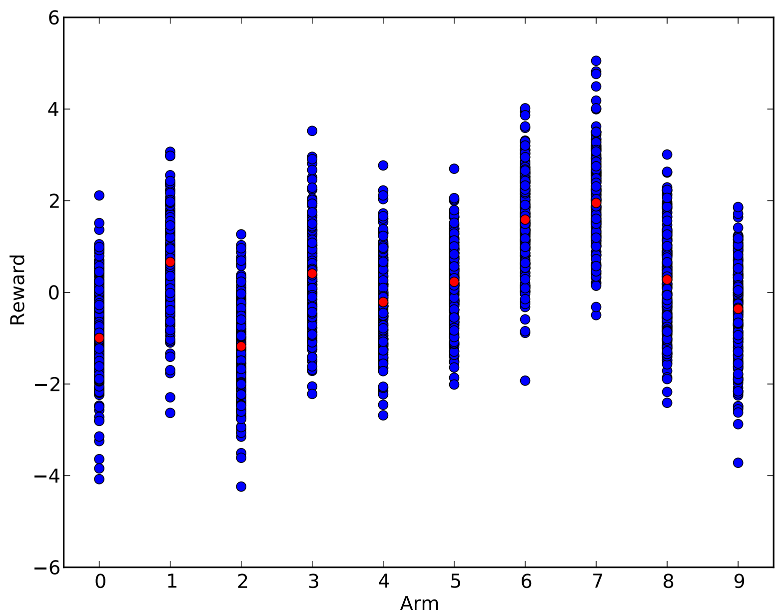

Figure 1: 10-Armed Testbed

In all experiments, averages are over 2000 tasks and mean estimates are formed using incremental sample-averages.

For my first experiment, I recreate Sutton and Barto's Figure 2.1, with the addition of $\epsilon = 0.5$ (Figure 2). My results track theirs, indicating my implementation is correct.

In my next experiment, I explore the question asked in Exercise 2.1: which method will perform best in the long run? I hypothesized it would be the lowest nonzero value of ϵ, and that the % optimal action value would converge to $100(1 - \epsilon)$. Figure 3 seems to bear this out. In this experiment I extend the play time to 5000 plays; the lowest value of ϵ, 0.01, surpassed all others in average reward and was on its way to surpassing in % optimal action. All values of ϵ seem to be converging to as expected, except for $\epsilon = 0$, which converges to a low value.

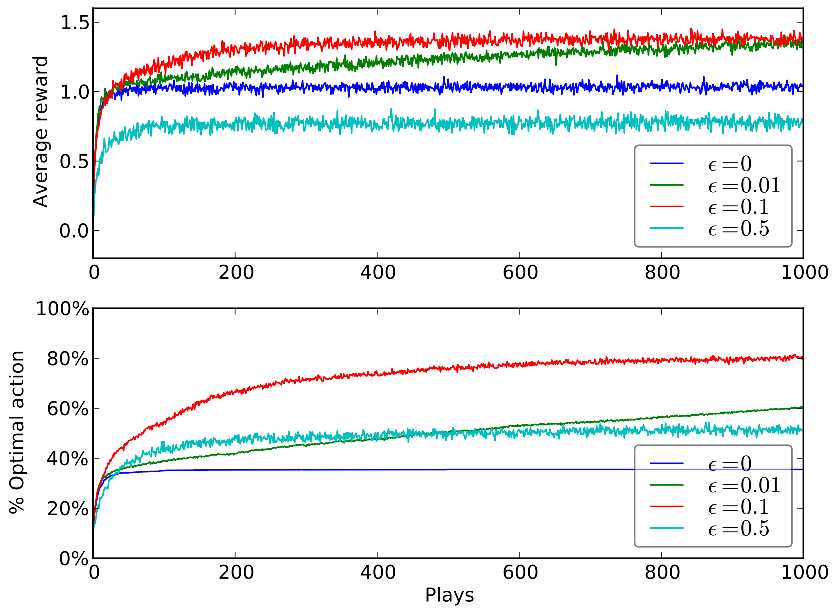

Figure 2: Sutton and Barto, Figure 2.1

Next, I explore a couple of statements about the testbed presented in Section 2.2. Namely, "suppose the reward variance had been ... 10 instead of 1 ... ϵ-greedy methods should fare even better relative to the greedy method", and "if the reward variances were zero, then the greedy method ... might actually perform best". I explore these scenarios is Figure 4 and 5.

In these experiments, the variance does not seem to change the ordering of the plots, though higher variance does make the averages noisier. In fact, the $\epsilon = 0$ value seems to do relatively better in the high variance testbed; I hypothesize that this is because it leads to more exploring.

Finally, I test softmax action selection on the 10-armed testbed, as suggested in Exercise 2.2 (Figure 6). I tested a number of different values for τ because it was difficult to determine which values would be suitable. In these runs of 1000 plays, a value of $\tau = 0.2$ seems to comes on top. The closest I get to greedy, $\tau = 0.01$, does the worst in this experiment.

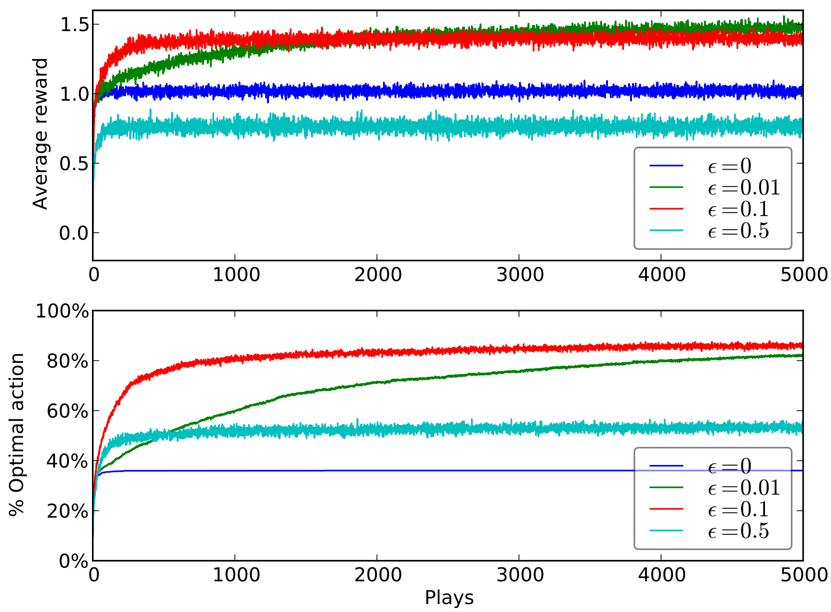

Figure 3: Exercise 2.1

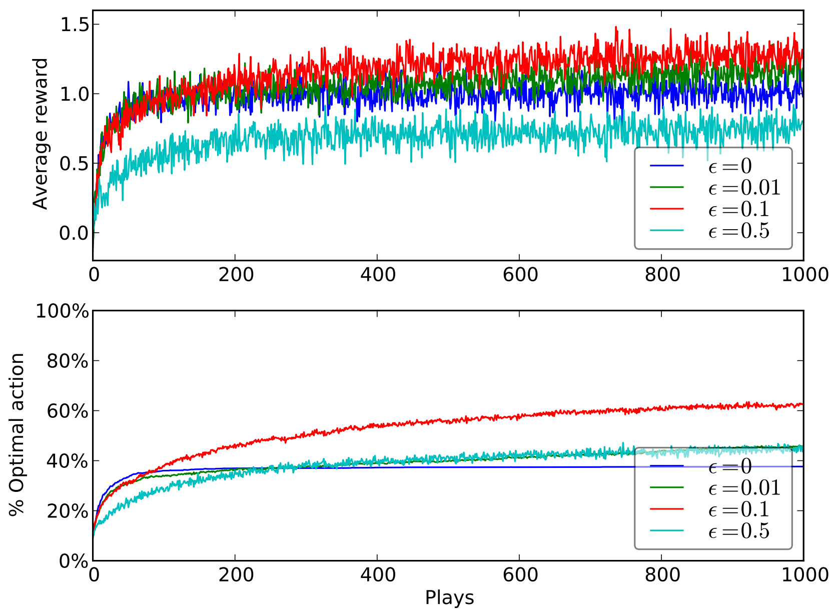

Figure 4: High Variance

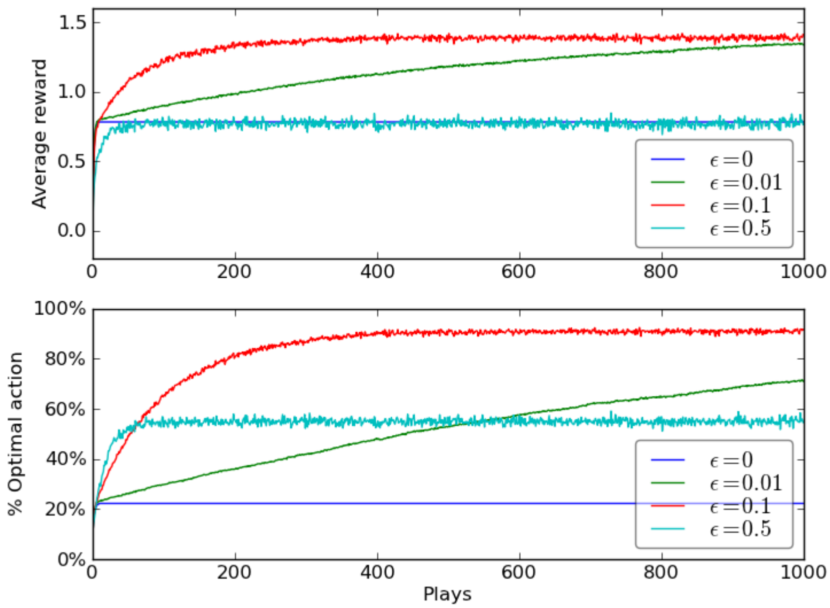

Figure 5: No Variance

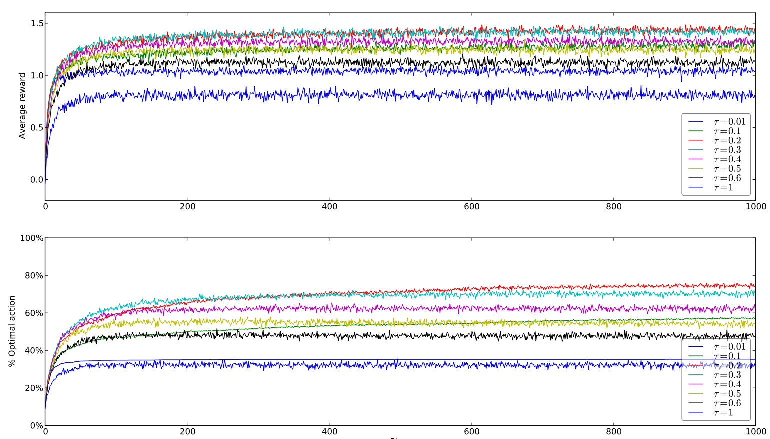

Figure 6: Exercise 2.2