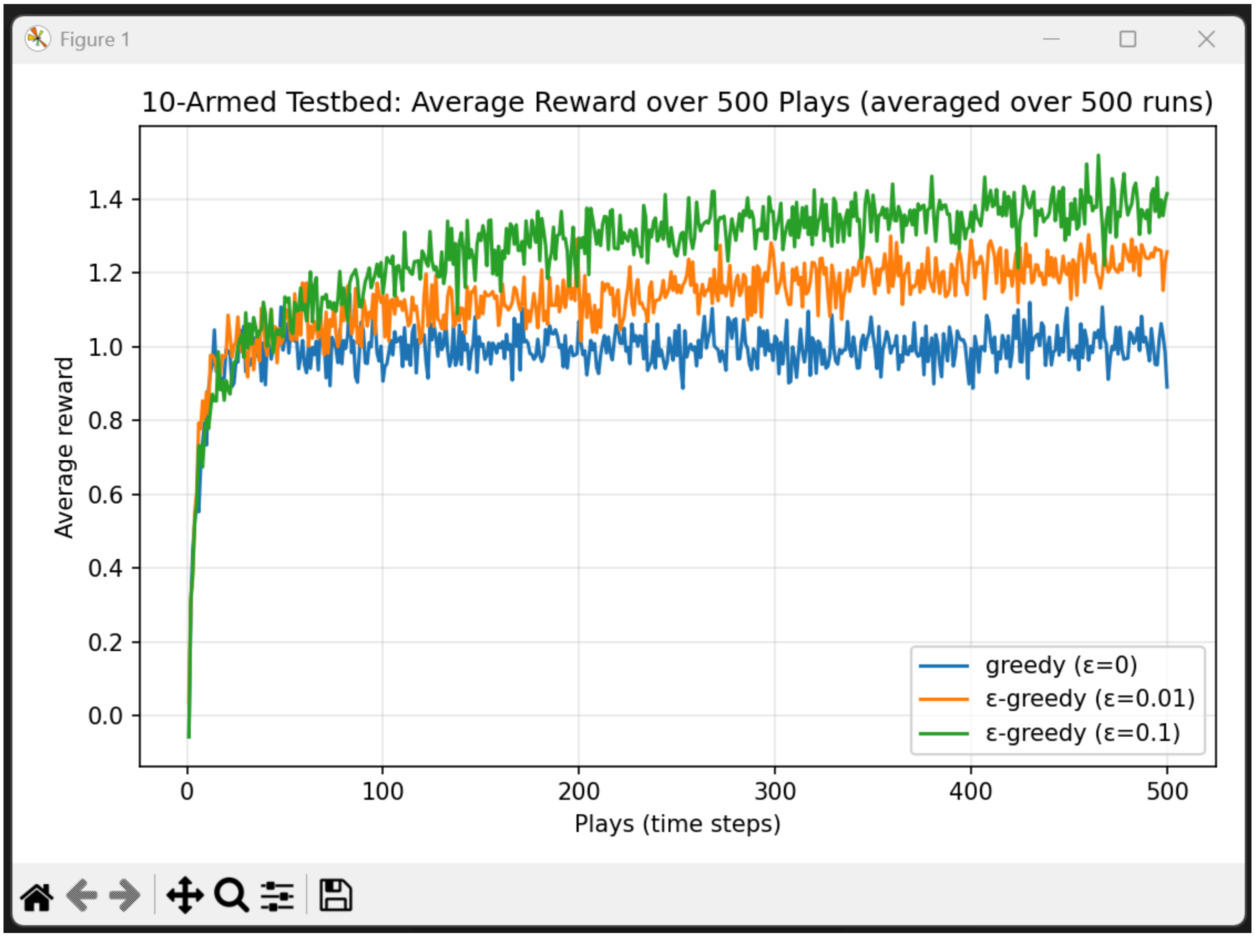
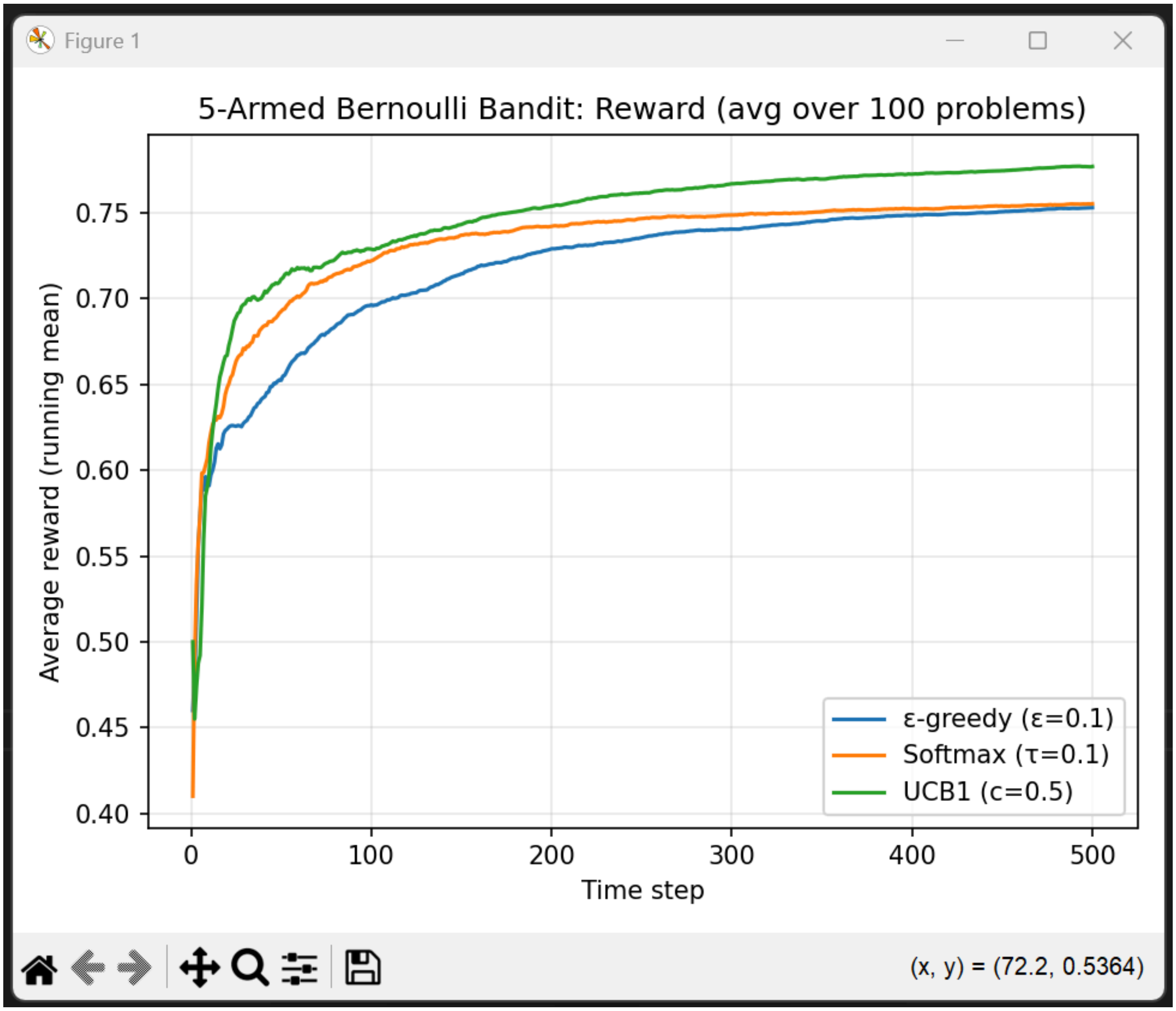
Problem 1:



I get graphs that look incredibly similar to the Sutton and Barto book. This isn't surprising because I used the exact same methodology as they did. As expected in the short term the e-greedy that takes a random action more (e-greedy 0.1) has a higher reward but we know that the e-greedy that takes a random action less (e-greedy 0.01) will in the limit have a higher reward.

Problem 2:



These results don’t really surprise me. Just like we discussed in class, UCB1 does “smart” exploration. We start by trying each action once to get a decent estimate. We then use the property that the UCB is divided by the number of times the action was taken. We basically get more and more confident as we go that the actions we are choosing are the optimal ones. I’m also not surprised by the results we got with Softwmax. Because we are using 0.1 I only expect it to be slightly better than the e-greedy approach. In last place we have e-greedy.