The Multi-armed Bandit Problem

Ben Wiley and **Micah Brown**
{bewiley,msbrown}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

This paper analyzes a well known division of reinforcement learning known as the multi-armed bandit problem. The multi-armed bandit problem attempts to determine the best policy when faced with multiple choices. We analyze two different policies for picking machines: upper confidence bound (UCB1) and $\mathcal{E}_n$-GREEDY. We conclude that the $\mathcal{E}_n$-GREEDY policy is better than UCB1 due to the optimal arm being played more frequently and a lower regret.

1 Introduction

When given multiple options in a situation with unknown outcomes, naturally one does not know whether to explore more options or stick to the best option that one knows. A multi-armed bandit problem enables us to study the exploration-exploitation tradeoff in reinforcement learning. In this problem, an agent is faced with multiple "arms" (though often the term slot-machine is used), in which each arm has a reward distribution. The agent's goal is to determine a policy that informs it which arm to pick at any given state. Should it pick the arm that it knows to be the most successful to this point, or should it explore other options in the case that it can find a better alternative? That is the central question of the exploration-exploitation trade-off that we would like to answer.

This problem is interesting to Artificial Intelligence researchers due to the number of interesting practical uses that it has been applied to. For example, the multi-armed bandit problem has been used to model clinical trial treatments, ad placement, website optimization, and computer game-playing (Bubeck and Cesa-Bianchi 2012). Furthermore, the idea of reinforcement learning can summarize most of the field of Artificial Intelligence: place an agent in an unknown environment, with unknown rules, and have the agent determine what to do to thrive (Russell and Norvig 2003).

The multi-armed bandit problem has been analyzed extensively by (Auer, Cesa-Bianchi, and Fischer 2002), and this paper attempts to reproduce a subset of their work. The remainder of this paper includes background information on reinforcement learning and the multi-armed bandit problem, the experiments we performed, the results, and finally our conclusions.

2 Background

At every state, the agent picks an arm to play depending on a policy. The two deterministic policies that were analyzed were the UCB1 and the $\mathcal{E}_n$-GREEDY policies, as defined by Auer, Cesa-Bianchi, and Fischer (2002). In the UCB1 policy, which Auer, Cesa-Bianchi, and Fischer (2002) derived from Agrawal (1995), the agent plays the machine that maximizes:

$$\bar{x_j} + \sqrt{\frac{2 \ln n}{n_j}} \tag{1}$$

where $\bar{x_j}$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the total number of plays so far.

The $\mathcal{E}_n$-GREEDY policy is defined by playing the machine with the current highest reward with probability 1-$\mathcal{E}_n$, otherwise play a random arm. $\mathcal{E}_n$ is defined by:

$$\mathcal{E}_n = \min\Big\{1, \frac{cK}{d^2 n}\Big\} \tag{2}$$

with c, d as parameters, K is the number of arms in the given distribution, and n is the total number of plays so far.

To determine the quality of a policy, we use two metrics. The first is the proportion of times that the actual best machine was played. Naturally, a good policy will play the actual best machine a higher percentage of times, especially over time once it has learned appropriately. The second metric to measure the quality of a policy is regret. Regret is the difference between how well you could have done, had you known the probability distribution in advance, compared to how well you actually did. Essentially, you take the difference for each selection you made, of the reward from the selected arm subtracted by the reward from the optimal arm. A smaller regret means that the policy picked the optimal arm more times or an arm that was very close to optimal. Therefore, smaller regret values indicate a better policy.

3 Experiments

In effort to reproduce the results found by Auer, Cesa-Bianchi, and Fischer, we run similar tests on distribution tables 1, 3, 11, and 14 as described in their paper. We run the UCB1 algorithm once per distribution table, and then the $\mathcal{E}_n$-GREEDY algorithm once for each value of c designated in the antecedent paper.

On each run, we accumulate data on the total number of plays, the number of plays for each arm, and the total reward for each arm. We are able to use this information to calculate the percentage of the time the best arm is played, as well as the total regret, and we record this data after every $10^{n}th$ play, for all n from 0 to 6.

Importantly, the way we calculate regret seems to be slightly different than Auer, et al. In their paper, the specified equation for regret includes a measure of expected number of times an arm will have been played during the first n plays. Unclear upon how that measure was calculated, we have chosen to use the actual (recorded) number of times an arm has been played in place of it. This may lead to somewhat different results, particularly for lower values of n.

4 Results

Our results are summarized in Figures 1—8 in the appendix section. For each distribution, we have graphed the % of the time that the best machine was picked, as well as the regret on a logarithmic x-axis of the number of plays.

For all experiments, the $\mathcal{E}_n$-GREEDY policy outperforms the UCB1 policy. This was measured by the fact that the $\mathcal{E}_n$-GREEDY policies converged to the best policy quicker than the UCB1 policy, and that the UCB1 policy had a much higher regret than the other policies. Furthermore, it appears that higher values of the parameter c result in worse performing policies judging by regret, however they did not differ greatly.

5 Conclusions

In this section, briefly summarize your paper — what problem did you start out to study, and what did you find? What is the key result / take-away message? It's also traditional to suggest one or two avenues for further work, but this is optional.

6 Appendix

References

Agrawal, R. 1995. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability* 27(4):pp. 1054–1078.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47(2-3):235–256.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR* abs/1204.5721.

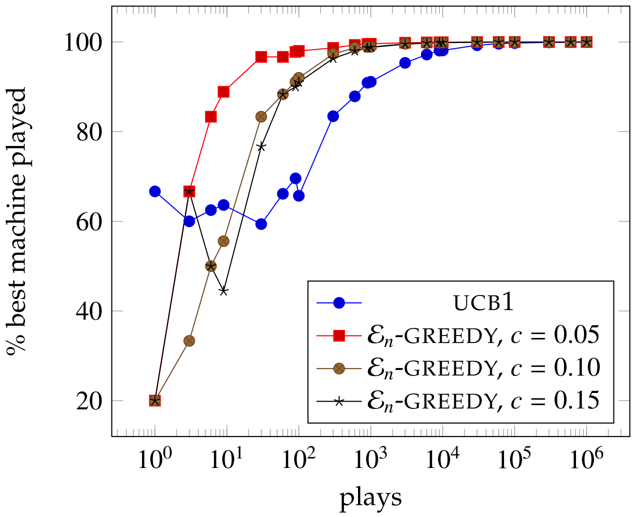

Figure 1: Comparison of policies on distribution 1.

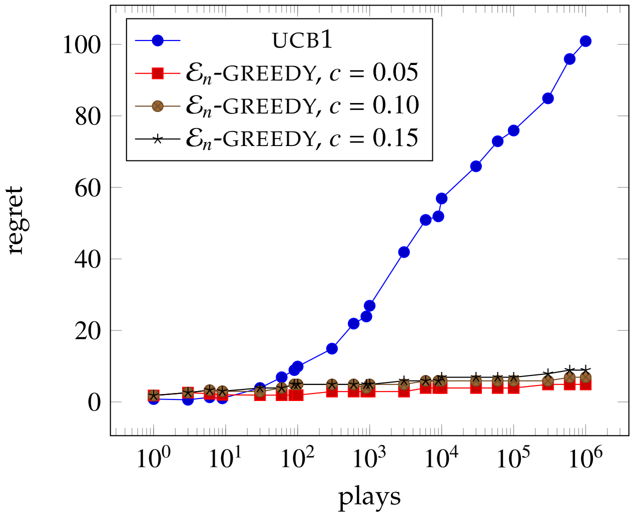

Figure 2: Comparison of policies on distribution 1.

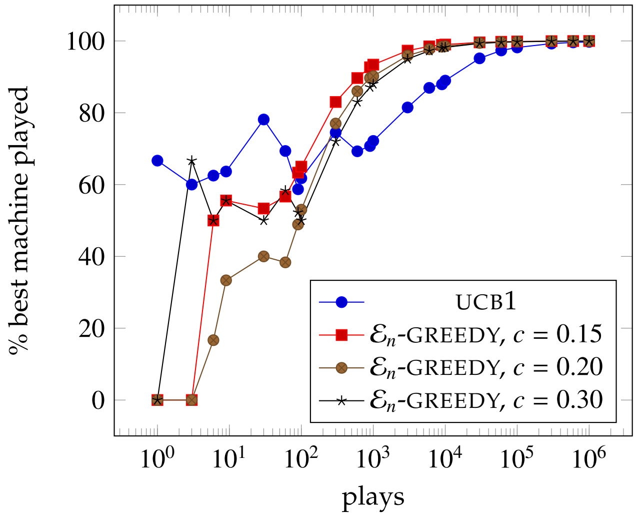

Figure 3: Comparison of policies on distribution 3.

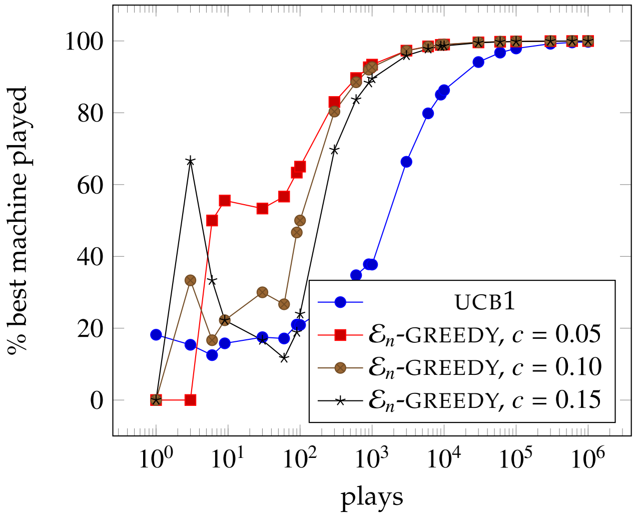

Figure 5: Comparison of policies on distribution 11.

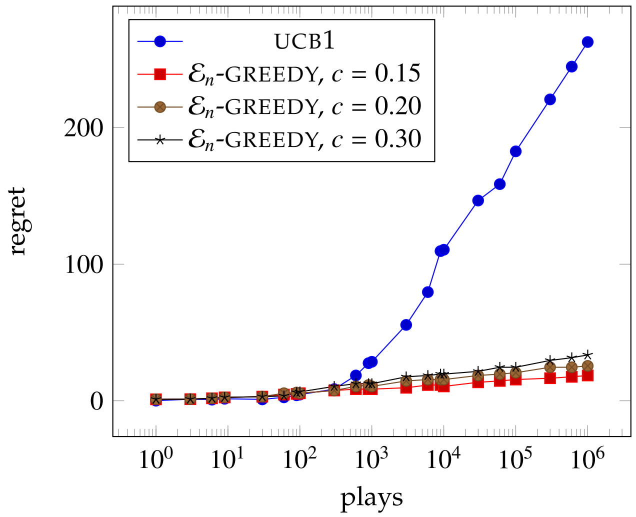

Figure 4: Comparison of regret for policies on distribution 3.

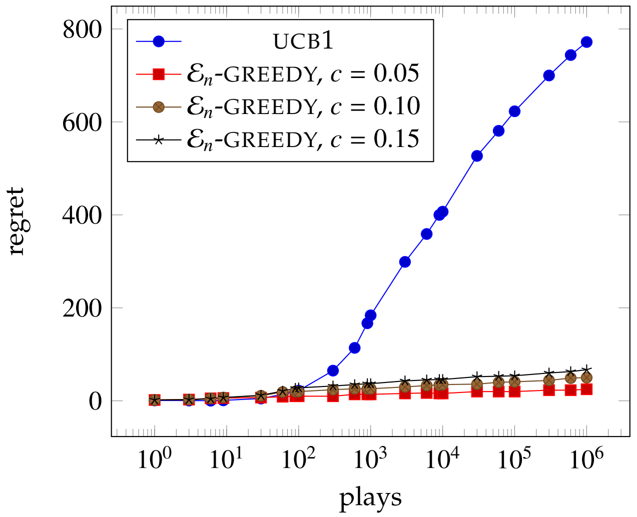

Figure 6: Comparison of regret for policies on distribution 11.

Russell, S. J., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Pearson Education.

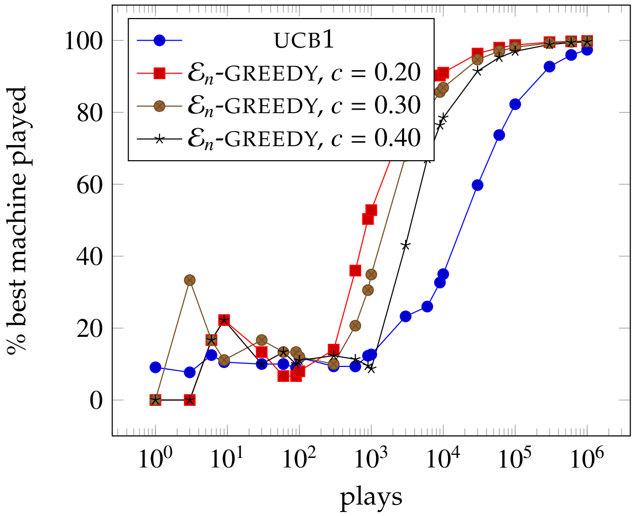

Figure 7: Comparison of policies on distribution 14.

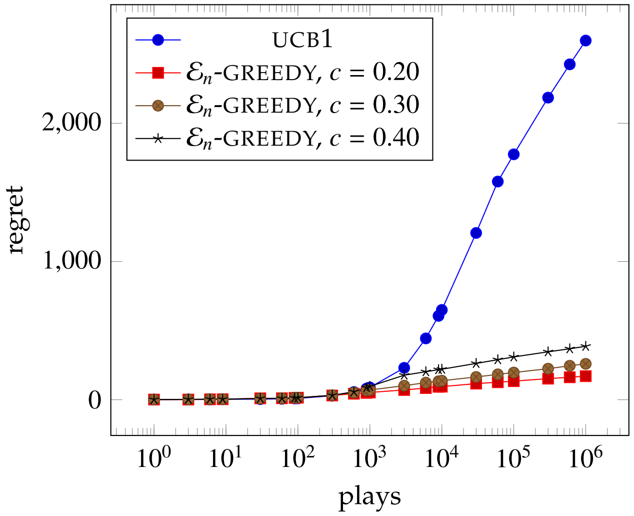

Figure 8: Comparison of regret for policies on distribution 14.