Exploration and Fairness in Infinite Armed Bandit Problems

Akshay Channesh and **Venkateshwar Ragavan**
University of Illinois Chicago
{achann4,vragv2}@uic.edu

Abstract

1 Introduction

Patil et al. (2021)

2 Related Work and Preliminary Concepts

2.1 Infinite Multi-Armed-Bandit

2.2 Regret Bound

2.3 UCB

2.4 Thompson Sampling

2.5 Fairness

3 Proposed Method

3.1 Curiosity

3.2 Surplus

4 Experiments and Discussion

We evaluate eight agents in three environments. We record the regret accumulated by each agent as it runs and graph it. We present the regret vs. time graphs and show the relative ordering of each of the agents.

Among the eight agents the first two are trivial ones whose regret increases linearly with time. Every other agent is more intelligent and achieves sub-linear regret. We plot the regret vs. time graph of all agents except the first two since they are very trivial.

4.1 The Agents

The eight agents that we focus on are as follows.

1. **Random Agent** picks an unseen arm half the times and a random seen arm otherwise.

2. **Always New Arm Agent** always picks an unseen arm.

3. **Naive UCB Agent** picks a new unseen arm with $p = 1\%$ and every other time it picks among the known arms using UCB.

4. **Naive Thompson Sampling Agent** picks a new unseen arm with $p = 1\%$ and every other time it picks among the known arms using Thompson Sampling.

5. **Fair UCB Agent** picks a new unseen arm with $p = 1\%$ and every other time it picks among the known arms using UCB while also ensuring α-Fairness among known arms.

6. **Fair TS Agent** picks a new unseen arm with $p = 1\%$ and every other time it picks among the known arms using Thompson Sampling while also ensuring α-Fairness among known arms.

7. **Surplus Curiosity UCB Agent** picks a new unseen arm with a probability based on the surplus, and every other time it picks among known arms using UCB.

8. **Surplus Curiosity TS Agent** picks a new unseen arm with a probability based on the surplus, and every other time it picks among known arms using Thompson Sampling.

4.2 Environments

The three environments that we evaluate the agents in are as follows.

1. **Uniform Environment** is one in which every new arm has a quality μ which is sampled from a uniform distribution between 0 and 1.

2. **Increasingly Better Environment** is one in which arms with bad quality are available in the begining and progressively better and better ones become available. In this environment every new arm has a quality μ which

is sampled from a uniform distribution between 0 and T, where T starts from 0 and increases slowly till 1 by the time we reach the last round.

3. **Progressively Worse Environment** is the reverse of the previous environment. We start with good arms and later sampled arms go on getting worse and worse. Here T starts from 1 and decreases slowly till 0 by the time we reach the last round.

4.3 Results

4.3.1 Uniform Environment

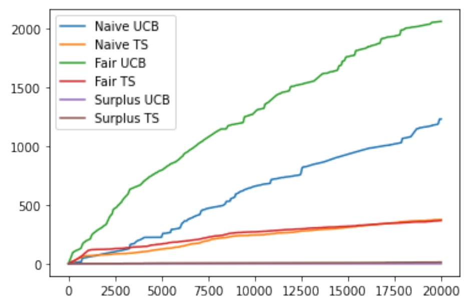

Figure 1: Uniform Environment. Linear-Linear Graph

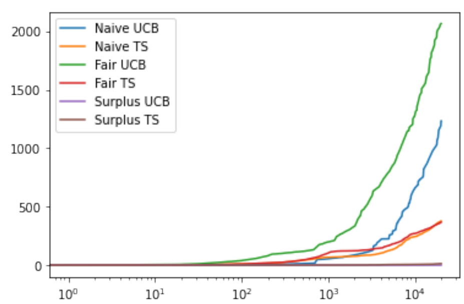

Figure 2: Uniform Environment. Log-Linear Graph

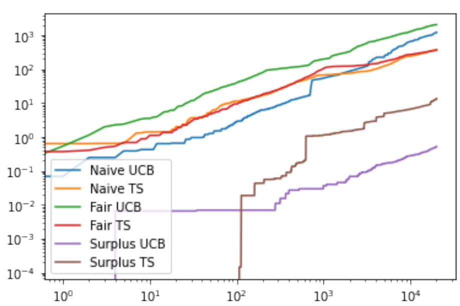

Figure 3: Uniform Environment. Log-Log Graph

Figures 1 to 3 show the same data associated with the Uniform environment in three different scales. We show it as such only to highlight this relative ordering of each of the agents, since only this is important.

4.3.2 Increasingly Better Environment

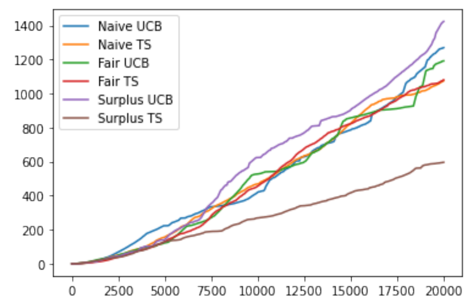

Figure 4: Increasingly Better Environment. Linear-Linear Graph

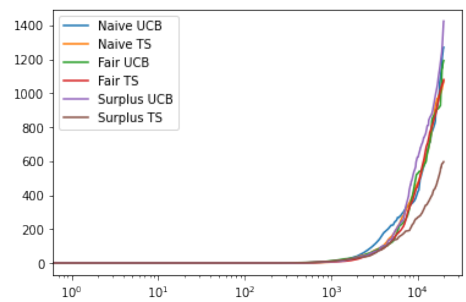

Figure 5: Increasingly Better Environment. Log-Linear Graph

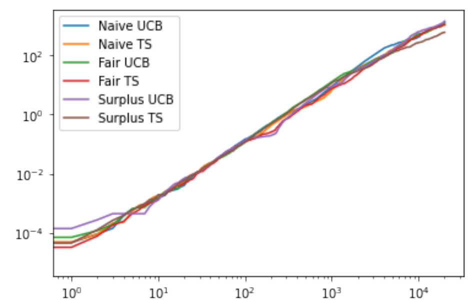

Figure 6: Increasingly Better Environment. Log-Log Graph

Similarly Figures 4 to 6 show regret vs. time graphs associated with the Increasingly Better Environment.

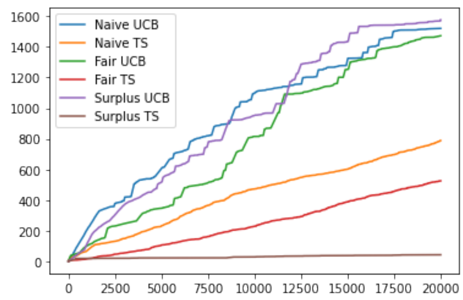

Figure 7: Progressively Worse Environment. Linear-Linear Graph

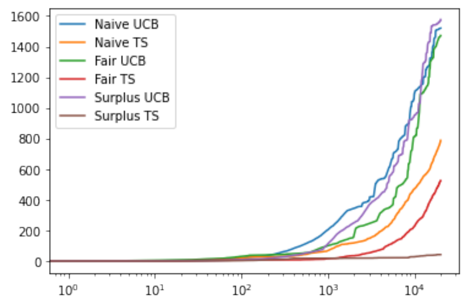

Figure 8: Progressively Worse Environment. Log-Linear Graph

4.3.3 Progressively Worse Environment

Figures 7 to 9 show the regret vs. time graphs in the Progressively Worse Environment.

4.4 Discussion

We observe the following trends from these graphs.

1. Agents using Thompson Sampling perform better than those using UCB. Chatterjee et al. (2017) make the claim that Thompson Sampling adapts better to the environment and performs better. We observe this in our experiments as well.

2. There is a cost to ensuring fairness. Algorithms that incorporate fairness fare worse than their naive counterparts, although not by much.

3. Algorithms using the surplus weighted curiosity perform better than those that do not. Of these the Surplus Curiosity Thompson Sampling agent performs the best.

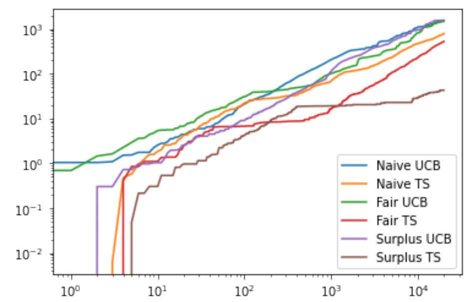

Figure 9: Progressively Worse Environment. Log-Log Graph

5 Conclusion

References

Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish, and Y. Narahari. 2017. Analysis of thompson sampling for stochastic sleeping bandits. In *UAI*.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174):1–31.