Exploration and Fairness in Infinite Armed Bandit Scenarios

Akshay Channesh
Venkateshwar Ragavan

github.com/AkshayChn/modern-rl

Introduction

Introduction

1. Multi Armed Bandit Problem
2. Infinite Arm Bandit
   a. Explore known arms further
   b. Pull a new un-explored arm
3. Regret

$$R_{\mathcal{A}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t^*} - \mu_{i_t}\right],$$

1. Exploration
   a. UCB
   b. Thompson Sampling
2. Fairness Notion
3. Surplus (More on this later)

i =	1	2	3	. . .	10
p(i) =	0.1	0.7	0.4	. . .	0.32

Exploration [Ref: Chatterjee et al. UAI’17]

UCB

if $\exists j \in A_t$ *s.t.* $n_{j,t} = 0$ **then**

  Pull arm $i_t = j$.

else

  Pull arm $i_t = \text{argmax}_{i \in A_t}\left(\frac{s_{i,t}}{n_{i,t}} + \sqrt{\frac{8 \log t}{n_{i,t}}}\right)$.

end

Observe reward $r_t \sim \text{Bernoulli}(\mu_{i_t})$.

Update UCB-indices

$$s_{i_t,t+1} \leftarrow s_{i_t,t} + r_t$$

$$n_{i_t,t+1} \leftarrow n_{i_t,t} + 1$$

for $j \neq i_t$ **do**

$$s_{j,t+1} \leftarrow s_{j,t}$$

$$n_{j,t+1} \leftarrow n_{j,t}$$

end

Thompson Sampling

Sample $\theta_{i,t} \sim \text{Beta}(s_{i,t} + 1, n_{i,t} - s_{i,t} + 1)$ for each arm $i \in A_t$.

Pull arm $i_t \in \text{argmax}_{i \in A_t} \theta_{i,t}$.   (ties are broken lexicographically)

Observe reward $r_t \sim \text{Bernoulli}(\mu_{i_t})$.

Update posteriors

$$s_{i_t,t+1} \leftarrow s_{i_t,t} + r_t$$

$$n_{i_t,t+1} \leftarrow n_{i_t,t} + 1$$

for $j \neq i_t$ **do**

$$s_{j,t+1} \leftarrow s_{j,t}$$

$$n_{j,t+1} \leftarrow n_{j,t}$$

end

Fairness

- Notion: *α*-fairness
- Each Arm is guaranteed to be pulled a certain number of times with exceptions being no more than *α* fraction of times.
- Ref: Patil, V. et al. (2020). Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5379-5386. https://doi.org/10.1609/aaai.v34i04.5986

Algorithm 1: Fair-Learn

Input: $[k], (r_i)_{i\in[k]}, \alpha \geq 0$, Learn$(\cdot)$

Initialize:

$N_{i,0} = 0$ for all $i \in [k]$

$S_{i,0} = 0$ for all $i \in [k]$, where $S_{i,t}$ = total reward of arm i in t rounds

for $t = 1, 2, \ldots$ **do**

  Define : $A(t) = \left\{ i \mid r_i \cdot (t-1) - N_{i,t-1} > \alpha \right\}$

  Pull arm

$$i_t = \begin{cases} \text{argmax}_{i\in[k]} \left(r_i \cdot (t-1) - N_{i,t-1} \right) & \text{If} A(t) \neq \emptyset \\ \textsc{Learn}(N_t, S_t) & \text{Otherwise} \end{cases}$$

  Update parameters N_t and S_t

end

Surplus weighted Curiosity for New Arms

- Our Main Contribution
- Inspired by Curiosity Driven Exploration
- Infinite Arm Bandit has this tradeoff:
  - Explore known arms further
  - Pull a new un-explored arm
- Surplus: Potential excess quality that can be achieved by exploring unknown arms.
  - S = Hypothetical best - Current Best Arm
- Exploration of Unknown Arms can be done with a probability of $P = \omega.S$
- In our expts $\omega = 0.005$

Surplus weighted Curiosity Algorithm:

- Explore new arm with probability P
- Run UCB or TS over known arms with probability (1-P)

Our Work

Our Work

1 - An empirical study of various approaches to Infinite Armed Bandits

2 - Surplus based Curiosity: An algorithm which tunes exploration based on potentially available surplus rewards.

Experiments

Experiments

Three environments:

1. Uniform (Static Distribution) Environment
2. Increasingly Better Environment
3. Progressively Worse Environment

Eight Agents:

1. Random Agent
2. Always New Arm Agent
3. Naive UCB Agent
4. Naive Thompson Sampling Agent
5. Fair UCB Agent
6. Fair TS Agent
7. Surplus Curiosity UCB Agent
8. Surplus Curiosity TS Agent

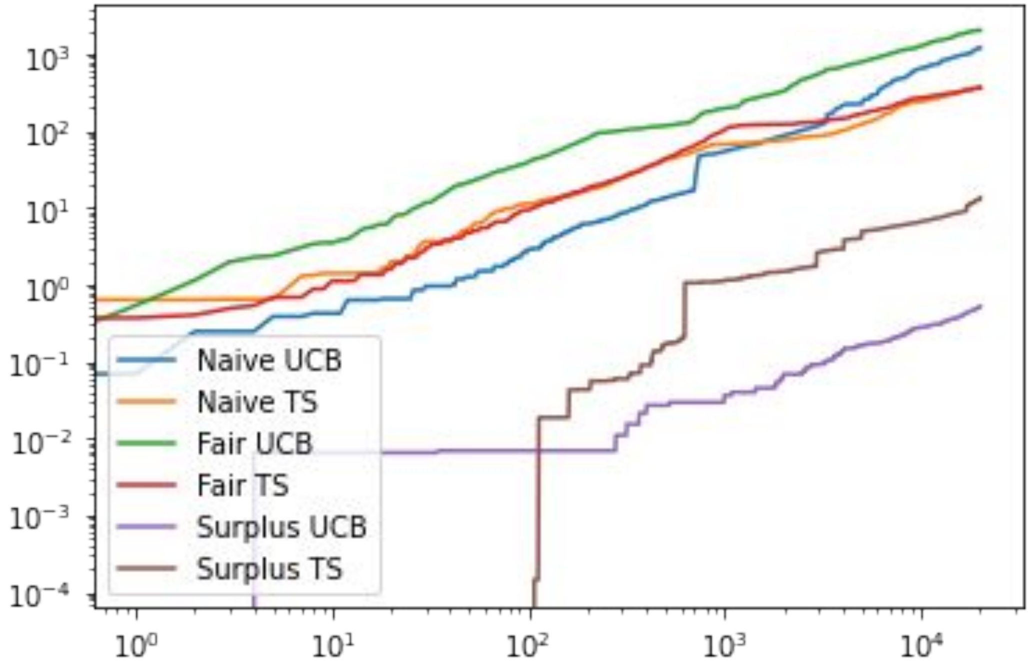
Experiments

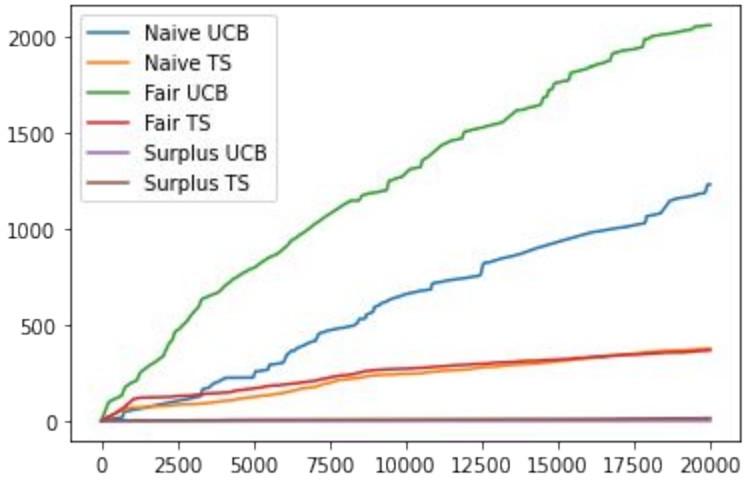
- Each agent runs **20,000 rounds**
- All eight agents take a total of **~10 min** to train per environment
- We track both kinds of regrets
  - wrt. known Arms
  - wrt. unknown Arms

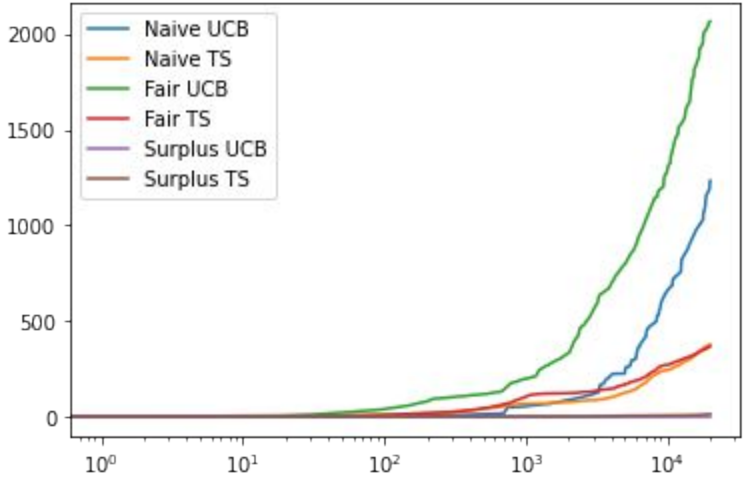
Plots (from next slide):

- X Axis: Time Horizon in Rounds
- Y Axis: Cumulative Unseen Regret
- Each slide is about one environment.
- Three graphs is each slide.
  - linear - linear
  - log - linear
  - log - log
- Highlights the order. Lower regret is better.

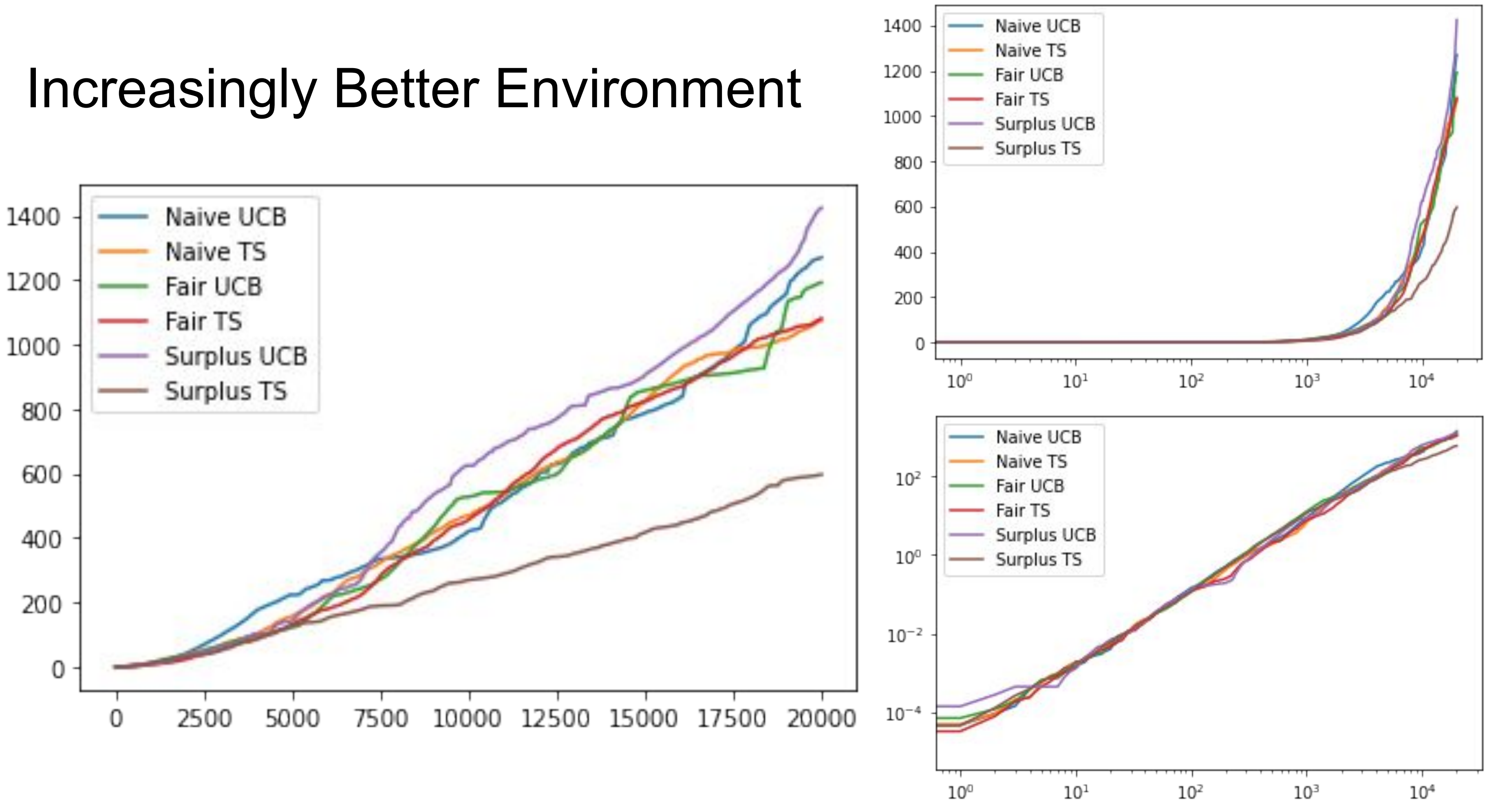
Uniform Environment



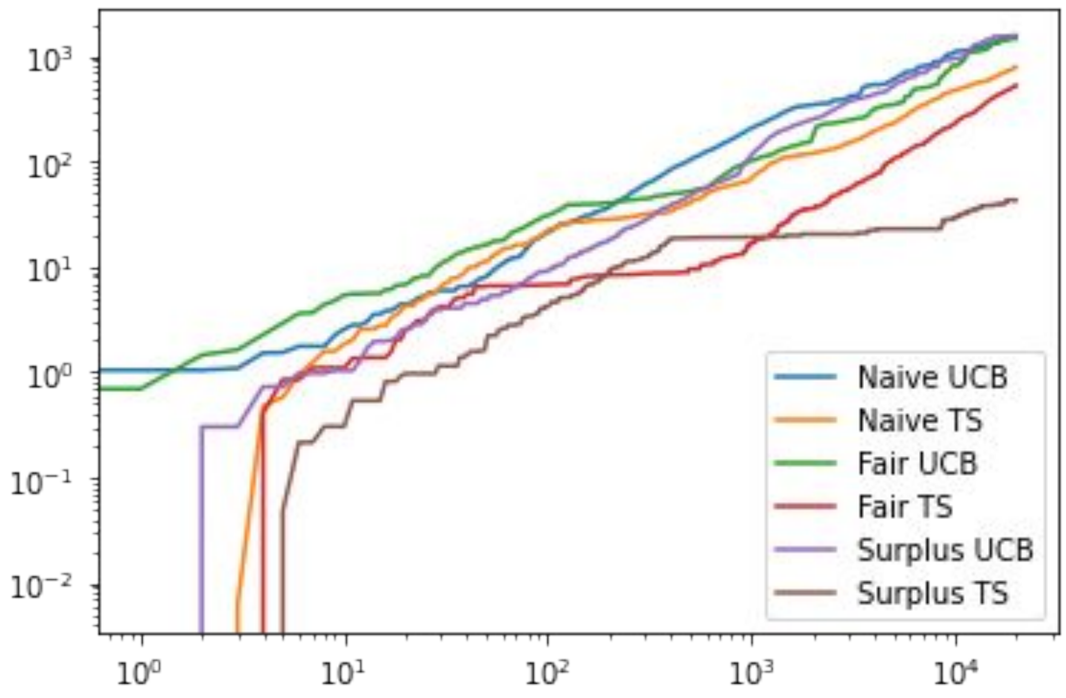


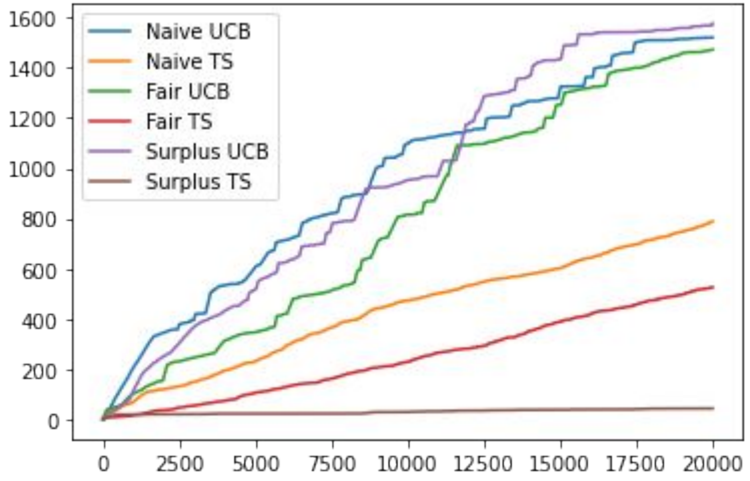


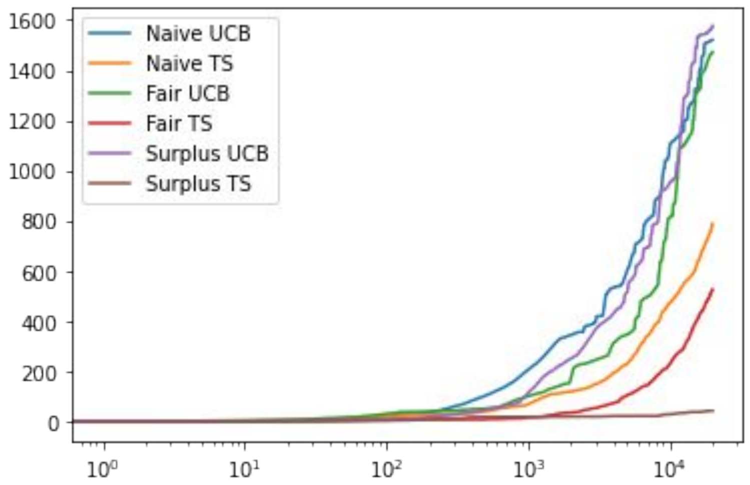
Increasingly Better Environment



Progressively Worse Environment







Conclusion from Results

Conclusions from Results

1. UCB is worse than TS Thompson Sampling outperforms UCB and is less prone to getting thrown off by newly explored arms. [Chatterjee et al. UAI 2017]
2. Fair is worse than Naive Fairness has a cost associated with it.
3. TS Surplus is better than others. Empirically seen from the results.

Further Conclusions and Future Work

Conclusions

- UCB is asymptotically optimal. Worse off in real settings.
- Weighting curiosity about new states on “surplus” gives good results.

Future Work

- Fair algorithms that use Surplus
- Theoretical bounds
- Non-bandit settings.