α-Fair Contextual Bandits

Siddhant Chaudhary and Abhishek Sinha

Chennai Mathematical Institute and Tata Institute of Fundamental Research, Mumbai, India

Contextual Bandits with α-fair utility

Consider the usual contextual bandits problem in the (oblivious) adversarial setting, where the number of arms is N and the number of contexts is M. The following sequence of events takes place on every round $t \in [T]$:

1. The adversary decides a context-reward pair (c_t, r_t), where $c_t \in [M]$ and $\delta \leq r_i(t) \leq 1, \forall i \in [N]$. Here $\delta > 0$ is a fixed positive constant.
2. The context c_t is then revealed to the online policy, which then uses this information to choose an arm (possibly randomly) $I_t \in [N]$.
3. **(Full Information Setting)** The policy obtains a reward of $r_{I_t}(t)$ and the entire reward vector $\boldsymbol{r}(t)$ is revealed to the policy, or,
4. **(Bandit Information Setting)** The policy obtains a reward of $r_{I_t}(t)$ and only the value of $r_{I_t}(t)$ is revealed to the policy.

If $j \in [M]$ is the context at time t, let $\boldsymbol{x}^j(t) \in \Delta_N$ denote the probability vector of each armed being pulled, i.e upon observing the current context c_t, the policy samples an arm $I_t \sim \boldsymbol{x}^j(t)$. Also, let $\boldsymbol{X}^j(t)$ denote the random *one-hot encoded vector* representing the (random) arm picked by the policy.

Cumulative Rewards: For each arm $i \in [N]$, the total cumulative reward accrued till round t for a full information policy is defined as:

$$R_i(t) = R_i(t-1) + x_i^{c_t}(t) r_i(t), \quad R_i(0) = 1 \tag{1}$$

where the same definition holds for a bandit information policy as well, with $\boldsymbol{x}^{c_t}(t)$ replaced with $\boldsymbol{X}^{c_t}(t)$.

Objective: The objective is to minimize the c-approximate *contextual* regret (for $c \geq 1$), defined as follows:

$$\text{Regret}_T(c) := \max_{\boldsymbol{x}^*} \sum_{i=1}^{N} \phi(R_i^*(T)) - c \sum_{i=1}^{N} \phi(R_i(T)) \tag{2}$$

where $\phi(R) = \frac{R^{1-\alpha}}{1-\alpha}$, $\alpha \in [0, 1)$ controls the amount of fairness, and for the bandit information setting the above definition also includes an expectation over the summation.

Linearization

Our algorithm design is based on two steps - (1) linearization of the problem with policy-dependent gradients and (2) solving the linearized OLO problem. By the concavity of the utility function $\phi(\cdot)$, we know that $\phi(x) - \phi(y) \leq \phi'(y)(x - y)$ for all $x, y > 0$. So, by taking $x = R_i^*(T)$ and $y = \beta R_i(T)$ with $\beta = \frac{1}{1-\alpha}$ in this inequality and summing over all the arms $i \in [N]$, we can get:

$$\text{Regret}_T((1-\alpha)^{-(1-\alpha)}) \leq \beta^{-\alpha} \sum_{t=1}^{T} \sum_{i \in [N]} \phi'(R_i(T)) r_i(t) [x_{*,i}^{c_t} - \beta x_i^{c_t}(t)] \tag{3}$$

Dropping the pre-factor β completely from the RHS, we arrive at our OLO problem with policy-dependent gradients:

$$\text{Surrogate Regret}_T := \max_{\boldsymbol{x}^*} \sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(t-1) \odot \boldsymbol{r}(t)), \boldsymbol{x}_*^{c_t} - \boldsymbol{x}^{c_t}(t) \rangle \tag{4}$$

where, in the bandit setting, we also have an expectation outside the summation, and where $\boldsymbol{x}^{c_t}(t)$ is replaced by $\boldsymbol{X}^{c_t}(t)$.

α-FairCB

We propose the full information and bandit information versions of α-FairCB, a policy which efficiently solves the OLO problem in (4). In brief, α-FairCB runs M instances of an adaptive OLO algorithm, one for each context, and couples the M instances via the cumulative reward vector $\boldsymbol{R}(t)$, which is affected by all contexts. For the full information problem, we use the OPF policy from [Sinha *et al.*, 2023], and for the bandit information version we use the adaptive, scale-free MAB algorithm from [Putta et Agrawal, 2021]. This strategy works because, after the linearization step, using the Cauchy-Schwarz inequality, the total regret can be upped bounded by the sum of regrets over all M instances. Finally, the norms of the policy-dependent gradients are controlled by a novel *bootstrapping* technique.

Theoretical Guarantees

The α-FairCB policy, outlined in **Algorithm** 1, achieves the following approximate regret bound for the contextual regret in the bandit information feedback setting with the α-fair utility function:

$$\text{Regret}_T(c_\alpha) = (1-\alpha)^\alpha \tilde{O}(MN^2 T^{\frac{1-\alpha}{2}}) \tag{5}$$

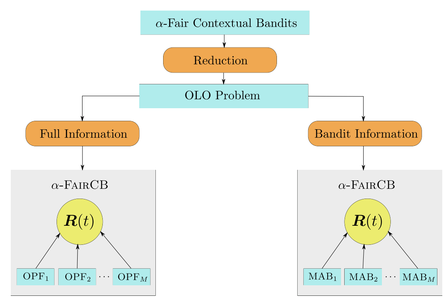
where $c_\alpha = (1-\alpha)^{-(1-\alpha)} < 1.445$, and the $\tilde{O}$ notation hides factors logarithmic in T.

α-FairCB (Bandit Information Setting)

Algorithm 1 α-FairCB (Bandit Information Setting)

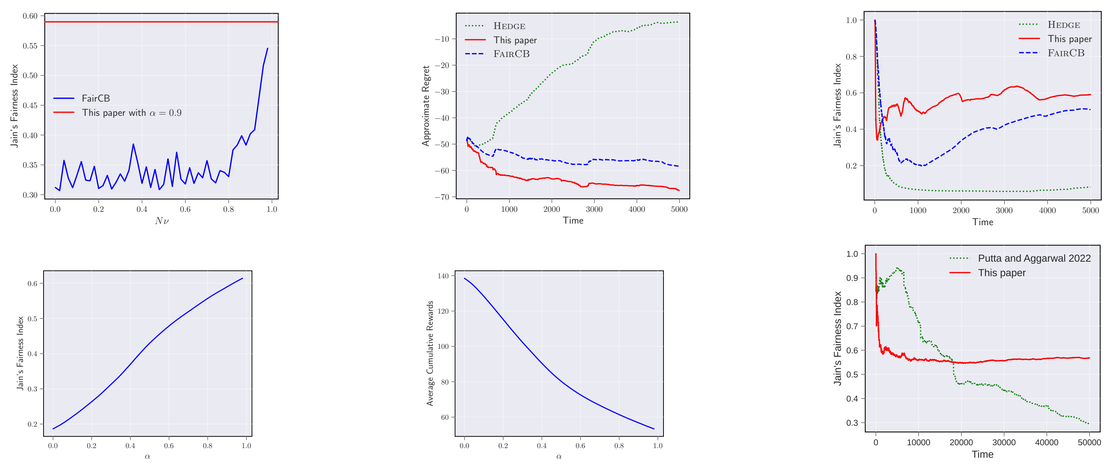
Input: Fairness parameter $0 \leq \alpha < 1$, Sequence of reward vectors $\boldsymbol{r}(1)$, ..., $\boldsymbol{r}(T)$, Sequence of contexts $c_1, ..., c_t$.
Output: Arm $I_t \in [N]$ to be played at round t, for $t \in [1, T]$.
Initialize $R_i(0) \leftarrow 1$ for all $i \in [N]$.
Initialize M adaptive, scale-free MAB policies from [**?**]. Let $\mathscr{A}_j$ denote the jth instance of the policy, for $j \in [M]$.
for $t = 1$ to T **do**
  Observe context c_t.
  Play an arm I_t picked by policy $\mathscr{A}_{c_t}$. Let $\boldsymbol{X}^{c_t}(t)$ denote the one-hot vector representing arm I_t.
  Feed the modified reward vector $\phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)$ to policy $\mathscr{A}_{c_t}$.
  Update $R_i(t) \leftarrow R_i(t-1) + X_i^{c_t}(t) r_i(t)$ for all $i \in [N]$.
end for

Schematic Diagram



Experiments

We evaluate the performance of α-FairCB on a movie genre recommendation problem using the Movie-Lens 25M dataset [Harper et Konstan, 2015]. We take a sample of about 5K datapoints from the dataset, with each user (context) having a frequency of 1000. We interpret users as contexts and genres as arms; thus, for our dataset, we have $M = 5$ and $N = 19$. We pick $\delta = 0.001$. For the full information setting, we compare our policy with a standard non-contextual Hedge policy, and the FairCB policy from [Chen *et al.*, 2020]. For the bandit setting, we compare our policy to a non-contextual scale free MAB instance from [Putta et Agrawal, 2021]. We use *Jain's Fairness Index* [Jain *et al.*, 1998] and the approximate contextual regret as our comparison metrics.



As we can see from the plots, our policy beats all the other baselines in terms of the mentioned metrics in both the full-information and the bandit settings.

Open Problems

In this paper, we considered the problem of learning adversarial unstructured context-to-reward mapping and proposed an approximately regret-optimal policy in both the full-information and bandit-information settings. In the future, it will be interesting to design efficient algorithms for the case of structured contexts. Finally, similar to [Chen *et al.*, 2020], designing α-fair bandit algorithms that guarantee a fixed fraction of pulls to each arm would also be interesting to investigate.

References

[Chen *et al.*, 2020] Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H. et Nikolaidis, S. (2020). Fair contextual multi-armed bandits: Theory and experiments. *In* Peters, J. et Sontag, D., éditeurs : *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 de *Proceedings of Machine Learning Research*, pages 181–190. PMLR.

[Harper et Konstan, 2015] Harper, F. M. et Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

[Jain *et al.*, 1998] Jain, R., Chiu, D. et Hawe, W. (1998). A quantitative measure of fairness and discrimination for resource allocation in shared computer systems.

[Putta et Agrawal, 2021] Putta, S. R. et Agrawal, S. (2021). Scale free adversarial multi armed bandits.

[Sinha *et al.*, 2023] Sinha, A., Joshi, A., Bhattacharjee, R., Musco, C. et Hajiesmaili, M. (2023). No-regret algorithms for fair resource allocation.