α-Fair Contextual Bandits

Siddhant Chaudhary [1] **Abhishek Sinha** [2]

Abstract

Contextual bandit algorithms are at the core of many applications, including recommender systems, clinical trials, and optimal portfolio selection. One of the most popular problems studied in the contextual bandit literature is to maximize the sum of the rewards in each round by ensuring a sublinear regret against the best-fixed context-dependent policy. However, in many applications, the cumulative reward is not the right objective - the bandit algorithm must be fair in order to avoid the echo-chamber effect and comply with the regulatory requirements. In this paper, we consider the α-FAIR CONTEXTUAL BANDITS problem, where the objective is to maximize the global α-fair utility function - a non-decreasing concave function of the cumulative rewards in the adversarial setting. The problem is challenging due to the non-separability of the objective function across rounds. We design an efficient algorithm that guarantees an approximately sublinear regret in the full-information and bandit feedback settings.

1. Introduction and related work

In applications such as personalized recommendations, greedily optimizing for the most relevant content for each user profile tends to reduce the diversity of the recommended items as it induces an unhealthy *echo-chamber* effect and propagates systematic biases (Celis et al., 2019). Recall that standard contextual bandits with a separable cumulative utility function tend to maximize the click-through rates (CTR) by recommending the most popular item for each user profile (Semenov et al., 2022). However, an over-emphasis on the CTR metric invariably leads to polarization of opinions. A similar fairness issue arises with other popular recommender systems, such as movie or song recommendations by Netflix and Spotify and various online job recommendation portals. The main objective of this paper is to design a class of *fair* contextual bandit algorithms equipped with a quantifiable fairness guarantee that holds even in the adversarial setting. Towards this goal, we propose a contextual bandit algorithm that maximizes the non-linear α-fair utility function instead of the usual time-separable utility function. Due to the diminishing return property, the optimizer of the concave α-fair utility function strikes a trade-off between the fairness and the accuracy of the recommendations through a tunable hyperparameter $\alpha \in [0, 1)$.

The α-fair metric has been widely adopted in the literature and has been considered in various dynamic resource allocation problems ((Altman et al., 2011), (Si Salem et al., 2022), (Sinha et al., 2023)). (Lan et al., 2010) gave an axiomatic characterization of fair utility functions and showed that the α-fair utility function comes out naturally. Other standard utility functions, *e.g.,* proportional fair and min-max utilities, can be shown to be a limiting form of the α-fair utility.

Fairness in bandit and online convex optimization have been extensively studied in the literature (Joseph et al., 2016; Chen et al., 2020; Agarwal et al., 2014; Patil et al., 2021; Si Salem et al., 2022; Even-Dar et al., 2009; Claure et al., 2020; Li et al., 2019). Chen et al. (2020) considered a fair contextual bandit problem with a finite number of contexts. Their online policy ensures that the probability of pulling each arm is lower-bounded by a pre-specified constant on every round. They establish a $O(\sqrt{TMN \log N})$ regret bound for the usual separable cumulative loss metric (here M is the number of contexts, and N is the number of arms). In the stochastic setting, the work by Patil et al. (2021); Claure et al. (2020), and Li et al. (2019) proposed constrained bandit policies that guarantee that the minimum *fraction* of pulls of each arm exceeds a given threshold. Our work complements this line of work where we consider an unconstrained maximization of the non-separable α-fair utility function. A detailed numerical comparison between our policy and the constrained bandit policy of Chen et al. (2020) is presented in Section 4. Badanidiyuru et al. (2014) considered a similar contextual bandit problem in

[1] Department of Computer Science, Chennai Mathematical Institute, Chennai, India [2] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India. Correspondence to: Siddhant Chaudhary <chaudhary@cmi.ac.in>, Abhishek Sinha <abhishek.sinha@tifr.res.in>.

Workshop on Foundations of Reinforcement Learning and Control at the 41^{st} International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

the stochastic setting, which was later extended to concave utility functions (Agrawal & Devanur, 2014; Agrawal et al., 2016). Agrawal et al. (2016) gave an efficient policy with $O(\sqrt{T})$ regret in the stochastic setting. However, because of the impossibility of attaining a sublinear regret bound in the full-information setting (Sinha et al., 2023, Theorem 2), their result can not be extended to the adversarial rewards, which is the main focus of this paper. Closest to this paper is the recent work by Sinha et al. (2023), which considers the problem of maximizing the α-fair utility function in the non-contextual full-information setting. In this paper, we extend their policy to the adversarial contextual bandit setting with finitely many contexts. This is accomplished by combining a recent scale-free bandit policy with non-separable rewards.

Our contributions: In this paper, we make the following contributions.

- We propose an approximately no-regret contextual bandit algorithm for the α-fair global utility function with an approximation factor at most 1.445. The non-additivity of the α-fair utility function across rounds makes this problem significantly more challenging than the classic contextual bandit problems, where the cumulative reward can be decomposed as the sum of rewards in each round. By combining seemingly unrelated recent advances in online convex optimization and scale-free bandit algorithms, we propose an efficient policy for this problem.
- As a by-product of our algorithm specialized to a single context, we give the first fair MAB algorithm with an approximately sublinear regret for the α-fair utility function in the adversarial setting.
- Because of the global non-separability of the utility function, we introduce a new analytical technique involving a novel *bootstrapping* method to bound the regret in both full-information and bandit settings.
- We perform extensive numerical simulations of our policy and compare it with the state-of-the-art benchmarks with standard datasets.

All missing proofs can be found in the accompanying supplementary material.

2. The Full-information setting

We start our discourse with the simpler full-information setting where the entire reward vector for all arms is revealed to the policy at the end of every round. The more challenging bandit feedback setting, where only the reward component corresponding to the arm that was pulled is revealed on every round (where the event $3'$ takes place), will be studied in Section 3. Specifically, we consider a fully adversarial setting with N arms [1] and a finite number of contexts M. The following sequence of events takes place on every round $t \in [T]$.

1. The adversary first 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 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 $r(t)$ is revealed to the policy. Or,

$3'$. (**Bandit-feedback 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.

For a given online algorithm, let the probability vector $\boldsymbol{x}^j(t) \in \Delta_N, j \in [M]$ denote the probability of pulling the arms when the j^{th} context is revealed to the policy on round t (here Δ_N is the set of all distribution on N items). An online policy is defined by the collection of (conditional) distributions $\big(\boldsymbol{x}^j(t), j \in [M]\big)$, where, upon observing the current context c_t, the policy samples an arm $I_t \sim \boldsymbol{x}^{c_t}(t)$ for round t. The goal of the policy is to sequentially learn the best collection of distributions $\big(\boldsymbol{x}^j(t), j \in [M]\big)$, one for each context, to maximize the α-fair utility function described next.

Note. The assumption $\delta \leq r_i(t)$ is not suitable for many applications. Please refer to section C of the appendix, where we do an analysis of the case of non-negative rewards, i.e when $0 \leq r_i(t) \leq 1$ for all $i \in [N]$.

2.0.1. Utility function and the regret metric

For each arm $i \in [N]$, the (expected) cumulative reward accrued till round t for a given policy is defined as:

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

In this paper, we consider the problem of maximizing the sum of α-fair utility functions of the arms where the utility of the i^{th} arm is defined as:

$$\phi(R_i(T)) := \frac{(R_i(T))^{1-\alpha}}{1-\alpha}, \; i \in [N], \tag{2}$$

[1] The arms could represent either distinct actions or N different candidate policies for some problem from which we want to pick the best one (Auer et al., 2002).

where $0 \leq \alpha < 1$ is some fixed constant. The parameter α strikes a trade-off between fairness and efficiency. Setting $\alpha = 0$ corresponds to the usual linear reward function. On the other hand, larger α induces fairness because of the diminishing return property, which encourages playing all arms evenly (Lan et al., 2010). Formally, our objective is to design an online policy that minimizes the c-approximate *contextual* regret, which competes with the best offline policy in hindsight (*i.e.,* a fixed mapping from contexts to arms) instead of the best arm. Formally, the contextual regret is defined as:

$$\text{Regret}_T(c) := \max_{\boldsymbol{x}_*} \sum_{i=1}^{N} \phi(R_i^*(T)) - c \sum_{i=1}^{N} \phi(R_i(T)), \quad (3)$$

where $c \geq 1$ is some small constant, and, for each arm i, $R_i^*(T)$ is the cumulative reward (1) accrued by any static policy using the *fixed* collection of distributions $\boldsymbol{x}_* \equiv (\boldsymbol{x}_*^1, ..., \boldsymbol{x}_*^M)$ used in Eq. (1). A few words on the c-regret metric (3) are in order. Clearly, $c = 1$ corresponds to the usual static regret. However, it is known from Sinha et al. (2023, Theorem 2) that even in the full-information setting, no online policy can achieve a sublinear regret for $c = 1$. The concept of c-approximate regret has been useful in other online learning problems as well (Azar et al., 2022; Emamjomeh-Zadeh et al., 2021; Paria & Sinha, 2021).

Note: 1. We initialize $R_i(0)$ to 1 so that the derivative $\phi'(R_i(t))$ remains well-defined for all $t \in [T]$.

2. In the full-information setting, we work exclusively with the expected cumulative rewards rather than the true rewards, which is stochastic due to the randomness of the policy. This allows us to carry out a simpler deterministic analysis. Using standard concentration inequalities, it can be shown that resulting bounds carry over for the true rewards as well (Sinha et al., 2023, Section 4). However, due to the limited feedback, this trick no longer works in the bandit setting, where we work with the stochastic true rewards.

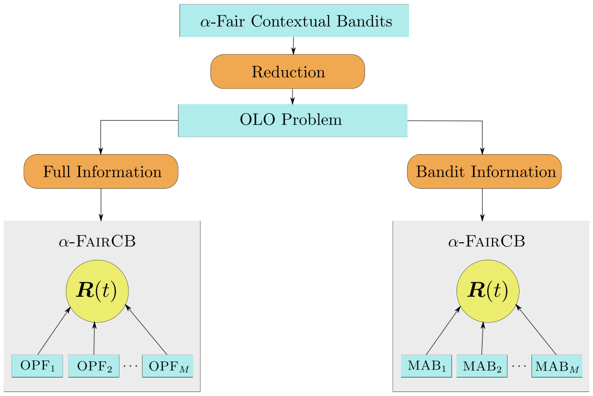

Figure 1. Diagram representing the web of reductions used in the paper. First, the contextual bandit problem with a global α-fair objective is reduced to a standard online linear optimization (OLO) problem. The reduction works the same way in both the full-information and bandit-information feedback settings. Then, in either setting, we parallelly run M instances of a non-contextual policy, and all the M policies are coupled through the *shared* vector $\boldsymbol{R}(t)$ of cumulative rewards. On a high level, after the linearization step, the j^{th} policy for $j \in [M]$ controls the regret for the jth context.

2.1. Algorithm design I: Linearization

Similar to Sinha et al. (2023), the algorithm design proceeds in two steps - (1) linearization with policy-dependent gradients and then (2) solving the linearized online optimization problem. See Figure 1 for a schematic. In the linearization step, we first reduce the problem to an instance of an online linear optimization (OLO) problem. Since the utility function $\phi(\cdot)$ is concave, we have

$$\phi(x) - \phi(y) \leq \phi'(y)(x - y) \quad (4)$$

for all $x, y > 0$. Now, let $\beta \geq 1$ be a constant, which will be fixed later. Taking $x = R_i^*(T)$ and $y = \beta R_i(T)$ in the above inequality, we get

$$\begin{aligned}
&\phi(R_i^*(T)) - \beta^{1-\alpha}\phi(R_i(T)) \\
&\overset{(a)}{=} \phi(R_i^*(T)) - \phi(\beta R_i(T)) \\
&\overset{(b)}{\leq} \phi'(\beta R_i(T))[R_i^*(T) - \beta R_i(T)] \\
&\overset{(c)}{\leq} \beta^{-\alpha}\phi'(R_i(T)) \sum_{t=1}^{T} r_i(t)[x_{*,i}^{c_t} - \beta x_i^{c_t}(t)], \quad (5)
\end{aligned}$$

where in (a), we have used the property that $\phi(\beta x) = \beta^{1-\alpha}(x)$ which holds for (2); in (b), we have used inequality (4), and in (c), we have used the definition of the cumulative rewards given in (1), the fact that $\beta \geq 1$ and the property $\phi'(\beta x) = \beta^{-\alpha}\phi'(x)$. Summing up the bound (5) over all the arms $i \in [N]$, we obtain the following bound to the $\beta^{1-\alpha}$-approximate regret of any online policy:

$$\begin{aligned}
&\text{Regret}_T(\beta^{1-\alpha}) \\
&\leq \beta^{-\alpha} \sum_{t=1}^{T} \sum_{i \in [N]} \phi'(R_i(T)) r_i(t)[x_{*,i}^{c_t} - \beta x_i^{c_t}(t)]. \quad (6)
\end{aligned}$$

Note that $R_i(T)$ is the cumulative reward accrued in the entire horizon of length T, and hence, it depends on the entire sequence of rewards and the actions of the policy. Clearly, this non-causal information is not available to the online policy at any intermediate round $t < T$. This shows that directly minimizing the upper bound (6) using online convex optimization methods is not feasible as the reward function involves the variables $\phi'(R_i(T))$'s. To get around this fundamental difficulty, we now define a *surrogate* online linear optimization problem by replacing the t^{th} coefficient $\phi'(R_i(T))$ in the RHS of the upper bound (6) with its causal surrogate $\phi'(R_i(t-1))$. With this substitution, the problem of minimizing (6) becomes an instance of the online linear optimization (OLO) problem. However, in contrast with the standard OLO problem, here the reward functions are no longer oblivious as they depend on the policy through its past actions. By bounding the regret of the surrogate problem, we show that it is possible to derive an approximate regret bound to the original regret minimization problem (3). Hence, dropping the factor $\beta^{-\alpha}$, the surrogate regret that we minimize is:

$$\begin{aligned}&\text{Surrogate Regret}_T\\ &= \max_{\boldsymbol{x}_*} \sum_{t=1}^{T} \sum_{i\in[N]} \phi'(R_i(t-1)) r_i(t)[x^{c_t}_{*,i} - x^{c_t}_i(t)] \quad (7)\end{aligned}$$

In particular, for the surrogate problem, the linear reward vector at time step t is given by $\phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)$, which implicitly depends on the past actions of the policy (through the first term). Here, $\phi'(\boldsymbol{R}(t-1)) \equiv (\phi'(R_1(t-1)), ..., \phi'(R_N(t-1)))$. Upon setting $\beta \equiv (1-\alpha)^{-1}$, the following result relates the original regret (3) with the surrogate regret (7) for any policy.

Lemma 2.1. *For any $T \geq 1$ and for any policy, we have*

$$\text{Regret}_T(c_\alpha) \leq (1-\alpha)^{\alpha} \text{Surrogate Regret}_T + c_\alpha N \quad (8)$$

where $c_\alpha = (1-\alpha)^{-(1-\alpha)} \leq e^{1/e} < 1.445$.

After accounting for M different contexts with a common cumulative reward vector $\boldsymbol{R}(t)$, the proof generalizes the arguments in Sinha et al. (2023, Lemma 1). See Section A.1 in the Appendix for the complete proof.

2.2. Algorithm design I: Solving the linearized problem with full information

In view of the regret bound (8), we now propose α-FairCB - an online policy to approximately minimize the surrogate regret (7). In brief, α-FairCB runs M instances of adaptive online gradient descent policy in parallel, where the j^{th} instance is responsible for controlling the regret for the j^{th} context. These parallel policies are coupled through the common state vector $\boldsymbol{R}(t)$ - the cumulative reward accrued up to time t, which is affected by all contexts. Technically, this strategy works because, after the linearization step above, using the Cauchy-Scwarz inequality, the regret can be upper-bounded by the sum of policy-dependent gradients over all M instances. Finally, the norm of these policy-dependent gradients are controlled using a novel *bootstrapping* technique. The following lemma gives a precise regret bound for the surrogate problem.

Algorithm 1 α-FairCB (Full Information Setting)

```
Input: Fairness parameter $0 \leq \alpha < 1$, Sequence of reward vectors $\boldsymbol{r}(1)$, ..., $\boldsymbol{r}(T)$, Sequence of contexts $c_1, ..., c_T$, Euclidean projection oracle on the simplex $\Pi_{\Delta_N}$, and an upper bound $D = \sqrt{2}$ to the Euclidean diameter of the simplex $\Delta_N$.
Output: Distributions $\boldsymbol{x}^{c_t}(t)$ for each round $t$.
Initialization:
    $R_i(0) \leftarrow 1, S_j \leftarrow 0, \boldsymbol{x}^j \leftarrow \frac{\mathbf{1}}{N}, \forall i, j.$
for $t = 1$ to $T$ do
    Receive the context $c_t$ for round $t$.
    if Context $c_t$ is seen for the first time then
        Output $\boldsymbol{x}^{c_t}(t) = \boldsymbol{x}^{c_t}$ (uniform distribution).
    else
        Let $t'$ be the last time step when context $c_t$ was seen.
        Compute gradient vector $\boldsymbol{g}$ as follows:
            $g_i = \frac{r_i(t')}{R_i^\alpha} \quad \forall i \in [N]$
        Update the cumulative gradient norm:
            $S_{c_t} \leftarrow S_{c_t} + \|\boldsymbol{g}\|_2^2$
        Carry out the online gradient ascent update:
            $\boldsymbol{x}^{c_t} \leftarrow \Pi_{\Delta_N}\left(\boldsymbol{x}^{c_t} + \frac{D}{\sqrt{2S_{c_t}}}\boldsymbol{g}\right)$
        Output $\boldsymbol{x}^{c_t}(t) = \boldsymbol{x}^{c_t}$.
    end if
    Observe reward vector $\boldsymbol{r}(t)$.
    Update $R_i(t) \leftarrow R_i(t-1) + x_i^{c_t}(t) r_i(t)$.
end for
```

Lemma 2.2. *The α-FairCB policy described in **Algorithm 1** achieves the following static regret bound for the surrogate problem (7):*

$$\text{Surrogate Regret}_T = \begin{cases} O(N^3 M T^{1/2-\alpha}), & \textit{if } 0 < \alpha < \frac{1}{2} \\ O(N^3 M \sqrt{\log T}), & \textit{if } \alpha = \frac{1}{2} \\ O(1), & \textit{if } \frac{1}{2} < \alpha < 1. \end{cases} \quad (9)$$

See Section A.2 for the proof of the result. The proof of this lemma exploits a novel *bootstrapping technique* which repeatedly boosts the estimate of the gradients, which are controlled by the policy, to obtain a better adaptive regret bound. Combining **Lemma** 2.1 and **Lemma** (2.2), we establish our main result.

Theorem 2.3. ***Algorithm 1** achieves the following approximate regret bound for the contextual bandit problem in the* *full information setting with the α-fair utility function:*

$$\text{Regret}_T(c_\alpha) = (1-\alpha)^\alpha \begin{cases} O(N^3MT^{1/2-\alpha}), \textit{if } 0 < \alpha < \frac{1}{2} \\ O(N^3M\sqrt{\log T}), \textit{if } \alpha = \frac{1}{2} \\ O(1), \textit{ if } \frac{1}{2} < \alpha < 1. \end{cases}$$

where $c_\alpha = (1-\alpha)^{-(1-\alpha)} < 1.445$.

3. The Bandit feedback setting

We now study the same problem in the more challenging bandit feedback model. In this setup, only the reward of the arm selected by the policy, *i.e.,* $r^{c_t}_{I_t}(t)$, is revealed on each round. Following standard practice, we assume that the reward vectors $\boldsymbol{r}(t)$ and the context sequence $c_t \in [M]$ for each time step t are generated by an *oblivious adversary*, *i.e.*, the sequence of rewards and contexts is fixed *a priori.*

Because of the limited feedback, an online policy cannot observe the expected cumulative rewards defined in Eqn. (1) as one needs to know the entire reward vector $\boldsymbol{r}(t)$ to compute the expected reward. Hence, instead of using the distribution $\boldsymbol{x}^{c_t}(t)$, we directly use the random *one-hot encoded vector* $\boldsymbol{X}^{c_t}(t)$ to define the true cumulative rewards [2]. Here, the I_t th component (which corresponds to the selected arm) of the vector $\boldsymbol{X}^{c_t}(t)$ is set to one, and the rest of the components are set to zero. Hence, the true cumulative reward vector, which the policy can observe under the bandit feedback setting, evolves as follows:

$$R_i(t) = R_i(t-1) + X_i^{c_t}(t)r_i(t),\ R_i(0) = 1. \tag{10}$$

As before, we will use the notation $\boldsymbol{x}^{c_t}(t) \in \Delta_N$ to denote the probability distribution of pulling the arms on step t. Hence, for all $i \in [N]$ and $t \in [1,T]$, we have

$$\mathbb{P}[X_i^{c_t}(t) = 1] = x_i^{c_t}(t). \tag{11}$$

Our objective is to design a policy which minimizes the expected c-approximate regret defined below:

$$\begin{aligned} &\text{Regret}_T(c) \\ &:= \max_{\boldsymbol{x}_* \in (\Delta_N)^M} \mathbb{E}\left[\sum_{i\in[N]} \phi(R^*_i(T)) - c\sum_{i\in[N]}\phi(R_i(T))\right]. \end{aligned} \tag{12}$$

In the above definition, $c \geq 1$ is a small constant whose value will be specified later and $\boldsymbol{R}^*(T)$ is the cumulative reward vector obtained for a stationary contextual bandit policy which pulls arms according to the fixed collection of distributions $\boldsymbol{x}_* \equiv (\boldsymbol{x}^1_*, ..., \boldsymbol{x}^M_*)$ depending on the current context. Let $(\boldsymbol{x}^1_*, ..., \boldsymbol{x}^M_*) \in (\Delta_N)^M$ be the best-fixed collection of distributions which achieves the maximum in (12). We have

$$\begin{aligned} &\text{Regret}_T(c) \\ &= \mathbb{E}\left[\sum_{i\in[N]} \phi(R^*_i(T)) - c\sum_{i\in[N]}\phi(R_i(T))\right] \\ &\overset{(a)}{=} \sum_{i\in[N]} \mathbb{E}[\phi(R^*_i(T))] - c\mathbb{E}\left[\sum_{i\in[N]}\phi(R_i(T))\right] \\ &\overset{(b)}{\leq} \sum_{i\in[N]} \phi(\mathbb{E}[R^*_i(T)]) - c\mathbb{E}\left[\sum_{i\in[N]}\phi(R_i(T))\right] \\ &\overset{(c)}{=} \sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^T r_i(t)x^{c_t}_{*,i}\right) - c\mathbb{E}\left[\sum_{i\in[N]}\phi(R_i(T))\right] \end{aligned} \tag{13}$$

Above, in (a), we have used the linearity of expectation. In (b), we have used Jensen's Inequality on the concave function ϕ. In (c), we have just expanded $\mathbb{E}[R^*_i(T)]$ using (10) and (11).

3.1. Algorithm design I: Linearization

Similar to the full-information setting, we handle the non-linearity by reducing the problem to a standard bandit problem with appropriately constructed linear reward functions. Following (5), we have

$$\begin{aligned} &\phi(\mathbb{E}R^*_i(T)) - \beta^{1-\alpha}\phi(R_i(T)) \\ &\leq \beta^{-\alpha}\phi'(R_i(T))\sum_{t=1}^T r_i(t)[x^{c_t}_{*,i} - \beta X^{c_t}_i(t)] \end{aligned} \tag{14}$$

where above, $\beta \geq 1$ is some constant to be fixed later. Summing the above inequality for all $i \in [N]$ and taking expectations w.r.t the actions of the policy, we have

$$\begin{aligned} &\sum_{i\in[N]} \phi(\mathbb{E}R^*_i(T)) - \beta^{1-\alpha}\mathbb{E}\left[\sum_{i\in[N]}\phi(R_i(T))\right] \\ &\leq \beta^{-\alpha}\mathbb{E}\left[\sum_{i\in[N]}\sum_{t=1}^T \phi'(R_i(T))r_i(t)[x^{c_t}_{*,i} - \beta X^{c_t}_i(t)]\right]. \end{aligned} \tag{15}$$

Combining the last inequality with (13), we get

$$\begin{aligned} &\text{Regret}_T(\beta^{1-\alpha}) \\ &\leq \beta^{-\alpha}\mathbb{E}\left[\sum_{i\in[N]}\sum_{t=1}^T \phi'(R_i(T))r_i(t)[x^{c_t}_{*,i} - \beta X^{c_t}_i(t)]\right]. \end{aligned} \tag{16}$$

[2] With a slight abuse of notation, we use the same symbol $\boldsymbol{R}(t)$ to denote the expected cumulative rewards in the full-information setting (1) and true cumulative rewards in the bandit feedback setting (10).

Motivated by the above bound, we now consider a surrogate bandit problem by replacing the term $\phi'(R_i(T))$ with its causal counterpart $\phi'(R_i(t-1))$. We now design an online policy to minimize the surrogate regret defined as follows:

$$\begin{aligned}&\text{Surrogate Regret}_T\\ &\equiv \mathbb{E}\left[\sum_{i\in[N]}\sum_{t=1}^{T}\phi'(R_i(t-1))r_i(t)[x_{*,i}^{c_t}-X_i^{c_t}(t)]\right]\\ &=\mathbb{E}\left[\sum_{t=1}^{T}\langle\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t),\boldsymbol{x}_*^{c_t}-\boldsymbol{X}^{c_t}(t)\rangle\right].\end{aligned}\tag{17}$$

Above, $\odot$ is the elementwise product of vectors. As before, $\phi'(\boldsymbol{R}(t-1))\equiv(\phi'(R_1(t-1)),...,\phi'(R_N(t-1)))$. Analogous to **Lemma** 2.1, we have the following result, which relates the regret defined in (12) to the surrogate regret defined in (17).

Lemma 3.1. *For any $T\geq 1$, we have*

$$\text{Regret}_T(c_\alpha)\leq(1-\alpha)^\alpha\text{Surrogate Regret}_T+c_\alpha N,\tag{18}$$

where $c_\alpha=(1-\alpha)^{-(1-\alpha)}\leq e^{1/e}<1.445$.

3.2. Algorithm design II: Solving the linearized problem with bandit feedback

Lemma 3.1 motivates us to design an online policy that minimizes the regret (17) for the surrogate bandit problem. However, unlike the standard adversarial bandit problem, where the reward vectors are fixed *a priori* in an oblivious fashion, in this case, the rewards for each round t, defined as $\boldsymbol{g}_t\equiv\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}_t$, *depends on the past actions* of the policy. We can decompose the surrogate regret over different contexts as follows:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{g}_t,\boldsymbol{x}_*^{c_t}-\boldsymbol{X}^{c_t}(t)\rangle\right]\\ &=\mathbb{E}\left[\sum_{j\in[M]}\sum_{t:c_t=j}\langle\boldsymbol{g}_t,\boldsymbol{x}_*^{j}-\boldsymbol{X}^{j}(t)\rangle\right]\\ &\overset{(a)}{=}\sum_{j\in[M]}\mathbb{E}\left[\sum_{t:c_t=j}\langle\boldsymbol{g}_t,\boldsymbol{x}_*^{j}-\boldsymbol{X}^{j}(t)\rangle\right]\\ &\overset{(b)}{\leq}\sum_{j\in[M]}\underbrace{\mathbb{E}\left[\max_{\boldsymbol{y}\in\{\boldsymbol{e}_k\}_{k=1}^N}\sum_{t:c_t=j}\langle\boldsymbol{g}_t,\boldsymbol{y}-\boldsymbol{X}^{j}(t)\rangle\right]}_{\text{regret due to the } j^{\text{th}} \text{ context}}\\ &=:\widehat{\text{Regret}}_T.\end{aligned}\tag{19}$$

Above, in (a), we have used the linearity of expectation; in (b), we have used the fact that for any fixed sequence of rewards in a bandit OLO problem, the best offline benchmark is the *best fixed arm in hindsight*. The above inequality can be written as

$$\text{Surrogate Regret}_T\leq\widehat{\text{Regret}}_T\tag{20}$$

To minimize the surrogate regret, we now design a policy that minimizes $\widehat{\text{Regret}}_T$, which is the sum of the regret for each context. Note that since the cumulative reward vector is common to all contexts, the regret bounds for different contexts are coupled with each other. To solve the per-context learning problem, we use the *adaptive* and *scale-free* multi-armed bandit policy, proposed by (Putta & Agrawal, 2022), as a black box. Specifically, we run M parallel instances of this policy, one for each context where they share the global cumulative reward vector $\boldsymbol{R}(t)$. For ease of reference, we quote the regret bound achieved by the bandit policy of Putta & Agrawal (2022) in the following theorem.

Algorithm 2 α-FAIRCB (Bandit Information Setting)

```
Input: Fairness parameter $0 \leq \alpha < 1$, Sequence of reward vectors $r(1)$, ..., $r(T)$, 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 (Putta & Agrawal, 2022). Let $\mathscr{A}_j$ denote the $j$th 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 $X^{c_t}(t)$ denote the one-hot vector representing arm $I_t$.
    Feed the modified reward vector $\phi'(R(t-1)) \odot r(t)$ to policy $\mathscr{A}_{c_t}$. [3]
    Update $R_i(t) \leftarrow R_i(t-1) + X_i^{c_t}(t) r_i(t)$ for all $i \in [N]$.
end for
```

Theorem 3.2 (**Theorem 1** of (Putta & Agrawal, 2022))**.** *For any oblivious sequence of reward vectors $\boldsymbol{l}_1,...,\boldsymbol{l}_T\in\mathbb{R}^N$, the adaptive version of **Algorithm 1** of Putta & Agrawal (2022) achieves the following regret bound:*

$$\begin{aligned}&\mathbb{E}\left[\max_{\{\boldsymbol{e}_k\}_{k=1}^N}\sum_{t=1}^{T}\langle\boldsymbol{l}_t,\boldsymbol{e}_k-\boldsymbol{X}(t)\rangle\right]\\ &=O(\log T\cdot[\sqrt{NL_2}+L_\infty\sqrt{NL_1}]).\end{aligned}\tag{21}$$

In the above, $\boldsymbol{X}(t)$ is the one-hot encoded vector denoting the arm pulled on round t, $L_\infty=\max_t\|\boldsymbol{l}_t\|_\infty$, $L_2=\sum_{t=1}^T\|\boldsymbol{l}_t\|_2^2$, $L_1=\sum_{t=1}^T\|\boldsymbol{l}_t\|_1$ and the expectation is taken w.r.t. the actions of the policy.

[3] Even though we pass the full vector $\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t)$ to the bandit subroutine, it only "sees" the reward $\phi'(R_{I_t}(t-1))r_{I_t}(t)$ for the arm I_t it has just picked.

Remarks: Technically, the regret bound in Theorem 3.2 was originally established for oblivious adversaries. However, in our case, the surrogate reward vector $\boldsymbol{g}_t$ depends on the past actions of the policy up to round $t-1$. To see why we can still plug in the generic regret bound (21), note that the reward vector $\boldsymbol{g}_t$ on round t does not depend on the action $\boldsymbol{X}(t)$ taken on round t. Hence, we can use the regret bound for an *imaginary* adversary that fixes the reward vector $\boldsymbol{g}_t$ at the end of round $t-1$. Since the reward on round t does not affect the previous actions of the policy, the regret bound (21) holds. Adapting the above bound to our contextual setting, we have the following scale-free regret bound.

Lemma 3.3. *For any $t \in [1, T]$, let $\boldsymbol{g}_t := \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)$. The adaptive version of **Algorithm 1** of (Putta & Agrawal, 2022) achieves the following bound for any $j \in [M]$:*

$$\mathbb{E}\left[\max_{y\in\{\boldsymbol{e}_k\}_{k=1}^N}\sum_{t:c_t=j}\langle \boldsymbol{g}_t, \boldsymbol{y}-\boldsymbol{X}^j(t)\rangle\right] \le$$

$$\tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_2^2}+\max_{t:c_t=j}\|\boldsymbol{g}_t\|_\infty\sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_1}\right]\right), \tag{22}$$

where the $\tilde{O}(\cdot)$ notation hides the logarithmic factors. Above, the expectation is taken w.r.t the policy actions.

Please refer to Section A.5 for the proof. The following result bounds the surrogate regret (19).

Lemma 3.4. *The α-FAIRCB policy described in **Algorithm 2** achieves the following bound on the regret of the surrogate bandit OLO problem for the α-fair utility function:*

$$\widehat{\text{Regret}}_T = \tilde{O}(MN^2T^{\frac{1-\alpha}{2}}) \tag{23}$$

where the $\tilde{O}(\cdot)$ notation hides the $\log T$ factor.

Finally, combining **Lemma** 3.1, (20) and **Lemma** 3.4, we establish our main result.

Theorem 3.5. ***Algorithm** 2 achieves the following approximate regret bound for the contextual bandit problem in the bandit information feedback setting with the α-fair utility function:*

$$\text{Regret}_T(c_\alpha) = (1-\alpha)^\alpha \tilde{O}(MN^2T^{\frac{1-\alpha}{2}}) \tag{24}$$

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

4. Experiments

We evaluate the performance of the proposed algorithm on a movie genre recommendation problem using the MOVIELENS 25M dataset (Harper & Konstan, 2015). The dataset consists of 25 million data points, each consisting of a movie rating given by a user. For our experiments, we take a small sample comprising of the around $5,000$ data points. These samples are users having a frequency of 1000 within the dataset. The underlying contextual bandit problem is formulated as follows: we interpret the users as contexts and movie genres as arms. In the selected sample, the number of contexts turns out to be $M = 5$ (number of users), and the number of arms featured is $N = 19$ (number of movie genres). The dataset is sorted by the column containing the timestamps at which the ratings were reported, and this is taken to be the order of request arrivals. Since our policy requires a positive lower bound to the rewards, we take the minimum reward to be $\delta = 0.001$ if the recommended genre doesn't fit the current movie (i.e a reward very close to 0) and 1 otherwise.

Performance metrics: To measure fairness, we use the popular *Jain's Fairness Index* (Jain et al., 1998), calculated for the vector of cumulative rewards at the end of the time horizon. For any round $t \in [1, T]$, Jain's fairness index is defined as:

$$\text{Jain's Fairness Index} := \frac{(\sum_{i=1}^N R_i(t))^2}{N\sum_{i=1}^N R_i^2(t)}. \tag{25}$$

Jain's fairness index assumes a value between 0 and 1, where a value of 1 is obtained when all components of the reward vector are the same (i.e., fully fair). In particular, if each arm receives an equal share of cumulative rewards, this index will be 1. Throughout our experiments, we take $\alpha = 0.9$ (i.e. a high level of fairness). We also plot the approximate contextual regret as defined in equations (3) and (12) for the full information and bandit feedback settings, respectively.

Calculating the offline baseline metrics: Note that the offline benchmarks in equations (3) and (12) required for computing the approximate regret involve computing the best offline collection of M distributions maximizing the cumulative α-fair utility function. Since $\phi(\cdot)$ is a concave function, this is a standard concave maximization problem over the convex domain $(\Delta_N)^M$. In our experiments, we use the CVXPY package for solving this problem (Diamond & Boyd, 2016).

Baseline Policies (Full Information Setting): We consider two baselines (1) a context-agnostic HEDGE policy (i.e. a policy that ignores contexts) and (2) the FAIRCB policy from (Chen et al., 2020). Note that inherently HEDGE is not a fair policy as its objective is to optimize the total reward. On the other hand, FAIRCB's fairness constraint is specified by a tunable parameter $\nu \in (0, \frac{1}{N})$; in particular, the constraint is that the marginal probability of each arm being pulled at any given time step is at least ν. To determine the appropriate value of ν for our experiments, we run the FAIRCB policy for 50 different values of ν evenly

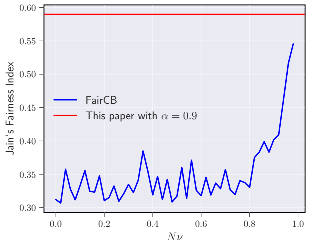

Figure 2. Fairness levels for varying values of ν against the fairness level of our policy for $\alpha = 0.9$.

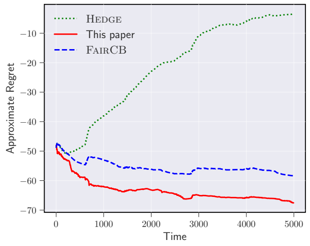

Figure 3. Approximate regret for the full information setting.

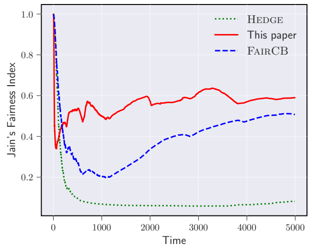

Figure 4. Jain's Fairness Index for the full information setting.

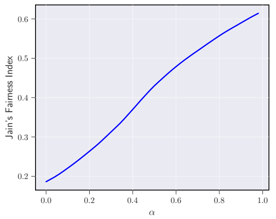

Figure 5. Fairness index of our policy for varying values of α.

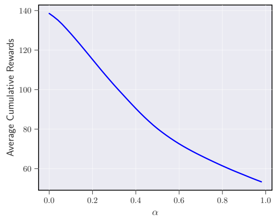

Figure 6. Averaged cumulative rewards of our policy for varying values of α.

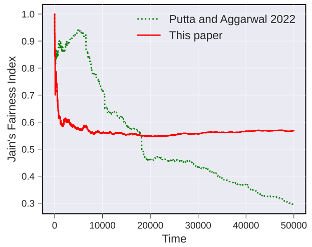

Figure 7. Jain's Fairness Index for the bandit information setting.

distributed in the interval $\left[0, \frac{1}{N}\right)$, and compare the fairness levels achieved for each of these policies with the fairness level achieved by our policy with $\alpha = 0.9$ (see Figure 2). It can be seen that for increasing ν, the fairness levels of FAIRCB show a generally increasing trend, and even for the highest value of ν in our sample ($\nu = \frac{0.98}{N}$), the fairness level is still not as high as our policy's. For further experiments, we take $\nu = 0.98/N$ (i.e the fairness level for FAIRCB achieving the largest fairness index).

Results (Full Information Setting): Figure 3 shows that the proposed α-FAIRCB policy outperforms the HEDGE and FAIRCB policies in terms of the approximate contextual regret (3). As expected, the context-agnostic HEDGE policy with no inherent notion of fairness performs the worst among the three policies under consideration. Finally, in terms of Jain's Fairness Index (25), we observe that the proposed α-FAIRCB outperforms both the non-contextual HEDGE and FAIRCB policies even for a moderately large time horizon (Figure 4).

Varying values of α: We also study the effects of different values of α in the performance of α-FAIRCB. It is expected that as α increases, a higher level of fairness is achieved. Figure 5 shows this trend. Also, Figure 6 shows that as α increases, the averaged total cumulative reward of our policy decreases, which is expected with increasing levels of fairness.

Baseline Policies (Bandit Information Setting): For this setting, we do the computations on about $50k$ data points. As a baseline policy, we run the context-agnostic *adaptive* multi-armed bandit policy proposed by Putta & Agrawal (2022), which is also used by our contextual bandit policy as a subroutine.

Results (Bandit Information Setting): In terms of Jain's Fairness Index, it is observed from Figure 7 that although for the first few rounds, (Putta & Agrawal, 2022)'s policy outperforms α-FAIRCB, but over the entire time horizon, α-FAIRCB achieves a significantly better fairness index. The behaviour for the first few time steps can be explained by the fact that Putta & Agrawal (2022)'s policy has an exploration component, which makes the policy choose each arm with an approximately equal probability in the initial stages. However, since their policy maximizes the cumulative rewards, it achieves a worse fairness index over a longer horizon. In terms of the approximate regret, we observe results similar to that of the full information setting. Please refer to section B in the Appendix for the plot and additional experimental results.

5. Conclusion and Future Work

In this paper, we considered the problem of learning adversarial unstructured context-to-reward mapping and proposed an approximately regret-optimal policy in the full-information and bandit feedback setting. 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.

6. Acknowledgement

The work was supported in part by a Google India faculty Research award and partly by an NSF-DST collaborative grant coordinated by IDEAS-Technology Innovation Hub (TIH) at the Indian Statistical Institute, Kolkata.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Altman, E., Avrachenkov, K., and Ramanath, S. Multiscale fairness and its application to resource allocation in wireless networks. In Domingo-Pascual, J., Manzoni, P., Palazzo, S., Pont, A., and Scoglio, C. (eds.), *NETWORKING 2011*, pp. 225–237, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-20798-3.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Azar, Y., Fiat, A., and Fusco, F. An $alpha$-regret analysis of adversarial bilateral trade. *Advances in Neural Information Processing Systems*, 35:1685–1697, 2022.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

Celis, L. E., Kapoor, S., Salehi, F., and Vishnoi, N. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 160–169, 2019.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., and Nikolaidis, S. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020.

Claure, H., Chen, Y., Modi, J., Jung, M., and Nikolaidis, S. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 299–308, 2020.

Diamond, S. and Boyd, S. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

Emamjomeh-Zadeh, E., Wei, C.-Y., Luo, H., and Kempe, D. Adversarial online learning with changing action sets: Efficient algorithms with approximate regret bounds. In *Algorithmic Learning Theory*, pp. 599–618. PMLR, 2021.

Even-Dar, E., Kleinberg, R., Mannor, S., and Mansour, Y. Online learning with global cost functions. In *22nd Annual Conference on Learning Theory, COLT*, 2009. URL http://www.cs.mcgill.ca/~colt2009/papers/005.pdf#page=1.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5 (4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Jain, R., Chiu, D., and Hawe, W. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems, 1998.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Lan, T., Kao, D., Chiang, M., and Sabharwal, A. *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

Li, F., Liu, J., and Ji, B. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Paria, D. and Sinha, A. `LeadCache` : Regret-optimal caching in networks. *Advances in Neural Information Processing Systems*, 34:4435–4447, 2021.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research*, 22(1):7885–7915, 2021.

Putta, S. R. and Agrawal, S. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 910–930. PMLR, 2022.

Semenov, A., Rysz, M., Pandey, G., and Xu, G. Diversity in news recommendations using contextual bandits. *Expert Systems with Applications*, 195:116478, 2022.

Si Salem, T., Iosifidis, G., and Neglia, G. Enabling long-term fairness in dynamic resource allocation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(3):1–36, 2022.

Sinha, A., Joshi, A., Bhattacharjee, R., Musco, C., and Hajiesmaili, M. No-regret algorithms for fair resource allocation. *arXiv preprint arXiv:2303.06396*, 2023.

Supplementary Material for α-Fair Contextual Bandits

July 4, 2024

A. Appendix

A.1. Proof of Lemma 2.1

Before proving the claim, we establish an auxiliary result that will be useful later.

Lemma A.1. *Under any policy which updates the cumulative rewards of the i^{th} user $R_i(\cdot)$ as in (1) $\forall i \in [N]$, the following inequality holds:*

$$\phi'(R_i(t-1))[R_i(t) - R_i(t-1)] \leq \int_{R_i(t-1)-1}^{R_i(t)-1} \phi'(R)\mathrm{d}R. \tag{26}$$

Proof. Since $0 \leq \alpha < 1$, observe that the utility function $\phi(\cdot)$ given by Eq. (2) is well-defined on $[0, \infty)$ and is differentiable in $(0, \infty)$. Also, because $R_i(\cdot)$ is monotonically non-decreasing and $R_i(0) = 1$, we note that $R_i(t-1) - 1 \geq 0$ for all $t \in [1, T]$. By the fundamental theorem of calculus combined with the mean value theorem, we have

$$\int_{R_i(t-1)-1}^{R_i(t)-1} \phi'(R)\mathrm{d}R = \phi'(c_0)[R_i(t) - R_i(t-1)] \tag{27}$$

for some $c_0 \in (R_i(t-1) - 1, R_i(t) - 1)$; in particular, we have $c_0 < R_i(t) - 1$. Now, from the defintion (1) observe that $R_i(t) - R_i(t-1) = x_i^{c_t}(t) r_i(t) \leq 1$, where we have used the fact that $x_i^{c_t}(t), r_i(t) \leq 1$. This implies that $R_i(t) - 1 \leq R_i(t-1)$, and hence, $c_0 < R_i(t-1)$.

Finally, since $\phi(\cdot)$ is concave, $\phi'(\cdot)$ is non-increasing; this implies that $\phi'(c_0) \geq \phi'(R_i(t-1))$. Combining this with (27), the claim follows. □

We now establish Lemma 2.1.

Proof. The upper bound for $\text{Regret}_T(\beta^{1-\alpha})$ from Eq. (6) can be split into the difference of two terms A and B as defined below:

$$\text{Regret}_T(\beta^{1-\alpha}) \leq \beta^{-\alpha}[A - \beta B], \tag{28}$$

where

$$A = \sum_{i \in [N]} \phi'(R_i(T)) \sum_{t=1}^{T} r_i(t) x_{*,i}^{c_t}, \tag{29}$$

$$B = \sum_{i \in [N]} \phi'(R_i(T)) \sum_{t=1}^{T} r_i(t) x_i^{c_t}(t). \tag{30}$$

Also, let A' and B' denote the corresponding terms in the regret expression (7) for the surrogate OLO problem. We will now bound the terms A and B in terms of A' and B', respectively.

Proving $A \leq A'$: Note that the utility function $\phi(\cdot)$ is concave, and hence its derivative is non-increasing. Also, from the recurrence equation for the cumulative rewards (1), it is clear that under any policy, $R_i(\cdot)$ is non-decreasing for any $i \in [N]$. Hence, we see that $\phi'(R_i(t-1)) \geq \phi'(R_i(T))$ for all $t \in [1, T]$ and $i \in [N]$. This implies that

$$\begin{aligned} A &= \sum_{i \in [N]} \phi'(R_i(T)) \sum_{t=1}^{T} r_i(t) x_{*,i}^{c_t} \\ &\leq \sum_{i \in [N]} \phi'(R_i(t-1)) \sum_{t=1}^{T} r_i(t) x_{*,i}^{c_t} \\ &= A' \end{aligned} \tag{31}$$

Proving $B' \leq (1-\alpha)^{-1}(B+N)$: We now argue that the following set of inequalities holds:

$$\begin{aligned} B' &= \sum_i \sum_{t=1}^{T} \phi'(R_i(t-1)) r_i(t) x_i^{c_t}(t) \\ &\overset{(a)}{=} \sum_i \sum_{t=1}^{T} \phi'(R_i(t-1))[R_i(t) - R_i(t-1)]] \\ &\overset{(b)}{\leq} \sum_i \sum_{t=1}^{T} \int_{R_i(t-1)-1}^{R_i(t)-1} \phi'(R) \mathrm{d}R \\ &\overset{(c)}{\leq} \sum_i \int_0^{R_i(T)} \phi'(R) \mathrm{d}R \\ &\overset{(d)}{=} \sum_i \phi(R_i(T)) \\ &\overset{(e)}{=} (1-\alpha)^{-1} \sum_i \phi'(R_i(T)) R_i(T) \\ &\overset{(f)}{=} (1-\alpha)^{-1} \sum_{i \in [N]} \phi'(R_i(T)) \left(1 + \sum_{t=1}^{T} x_i^{c_t}(t) r_i(t)\right) \\ &\overset{(h)}{\leq} (1-\alpha)^{-1}(B+N) \end{aligned} \tag{32}$$

where in (a), we have used the recurrence for $R_i(\cdot)$ as given in (1). In (b), we have used (26). In (c), we have simply used the fact that $R_i(0) - 1 = 0$ and $R_i(T) - 1 \leq R_i(T)$. In (d), we have used the fundamental theorem of calculus and the fact that $\phi(0) = 0$. In (e), we have used the fact that $x\phi'(x) = (1-\alpha)\phi(x)$ which holds for the α-fair utility function $\phi(\cdot)$. In (f), we have used the definition of the cumulative rewards as in (1). In (h), we have used the definition of B and the fact that $\phi'(x) = x^{-\alpha} \leq 1$ for all $x \geq 1$.

Now, the inequality $B' \leq (1-\alpha)^{-1}(B+N)$ implies that $(1-\alpha)B' - N \leq B$. Since $\beta > 0$, we have $\beta B \geq \beta(1-\alpha)B' - \beta N$. Combining this with $A \leq A'$, we have that

$$A - \beta B \leq A' - \beta(1-\alpha)B' + \beta N. \tag{33}$$

Now, pick $\beta = (1-\alpha)^{-1}$ (which ensures that $\beta \geq 1$), and hence we obtain

$$A - \beta B \leq A' - B' + (1-\alpha)^{-1} N, \tag{34}$$

and from Eq. (28), we see that

$$\begin{aligned} \text{Regret}_T(c_\alpha) &\leq (1-\alpha)^{\alpha}(A' - B') + c_\alpha N \\ &= (1-\alpha)^{\alpha} \text{Surrogate Regret}_T + c_\alpha N, \end{aligned} \tag{35}$$

which completes the proof of the lemma. □

A.2. Proof of Lemma 2.2

For ease of notation, let $(\boldsymbol{x}_*^1, ..., \boldsymbol{x}_*^M) \in (\Delta_N)^M$ be the collection of distributions achieving the maximum in equation (7). Now, observe that Surrogate Regret$_T$ defined in (7) for the surrogate problem can be split into the sum of regrets over each of the contexts as follows:

$$\begin{aligned}
\text{Surrogate Regret}_T &= \sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t), \boldsymbol{x}_*^{c_t} - \boldsymbol{x}^{c_t}(t) \rangle \\
&= \sum_{j\in[M]} \sum_{t:c_t=j} \langle \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t), \boldsymbol{x}_*^{j} - \boldsymbol{x}^{j}(t) \rangle \\
&\overset{(a)}{\leq} \sum_{j\in[M]} \underbrace{\max_{\boldsymbol{x}_\circ^j \in \Delta_N} \sum_{t:c_t=j} \langle \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t), \boldsymbol{x}_\circ^{j} - \boldsymbol{x}^{j}(t) \rangle}_{\text{Regret for the } j^{\text{th}} \text{ context}}
\end{aligned} \tag{36}$$

where above in (a), we have simply used that the regret w.r.t $\boldsymbol{x}_*^j$ for context j is upper bounded by the regret associated to the *best* offline benchmark $\boldsymbol{x}_\circ^j$ for context j.

Next, from the pseudocode of α-FAIRCB (Full Information Version, Algorithm 1), note that a Projected Online Gradient Ascent (OGA) policy with adaptive step sizes (**Theorem 4.14** of (Orabona, 2019)) controls the regret for each context $j \in [M]$. For the sake of completeness, we mention the complete statement of the regret guarantee of the OGA policy.

Theorem A.2 (Theorem 4.14 of (Orabona, 2019))**.** *Let $\Delta \subset \mathbb{R}^d$ be a convex set with diameter D. Let us consider a sequence of linear reward functions with gradients $\{\boldsymbol{g}_t\}_{t\geq 1}$. Run the Online Gradient Ascent policy with step sizes $\eta_t = \dfrac{D}{\sqrt{2}\sum_{\tau=1}^{T}\|\boldsymbol{g}_\tau\|^2}$, $1 \leq t \leq T$. Then, the standard regret under the OGA policy can be upper bounded as follows:*

$$\textit{Regret}_T \leq D\sqrt{2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2}. \tag{37}$$

Note that, for our case we have $D = \sqrt{2}$. So, by the regret bound of the OGA policy (37), for any $j \in [M]$ we have

$$\begin{aligned}
&\max_{\boldsymbol{x}_\circ^j \in \Delta_N} \sum_{t:c_t=j} \langle \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t), \boldsymbol{x}_\circ^{j} - \boldsymbol{x}^{j}(t) \rangle \\
&\leq D\sqrt{2\sum_{t:c_t=j} \|\phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)\|_2^2} \\
&\overset{(a)}{\leq} D\sqrt{2\sum_{t:c_t=j} \|\phi'(\boldsymbol{R}(t-1))\|_2^2} \\
&\overset{(b)}{=} D\sqrt{2\sum_{t:c_t=j} \sum_{i\in[N]} \frac{1}{R_i^{2\alpha}(t-1)}}
\end{aligned} \tag{38}$$

where above in (a), we have used the fact that $\boldsymbol{r}(t) \leq \mathbf{1}$ for all t, and in (b) we have used the fact that $\phi'(x) = x^{-\alpha}$. Now, summing (38) over all the contexts $j \in [M]$ and combining this with (36), we see that

$$\begin{aligned}
\text{Surrogate Regret}_T &\leq \sum_{j\in[M]} D\sqrt{2\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{2\alpha}(t-1)}} \\
&= M\sum_{j\in[M]}\frac{1}{M}D\sqrt{2\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{2\alpha}(t-1)}} \\
&\overset{(a)}{\leq} DM\sqrt{\frac{2}{M}\sum_{j\in[M]}\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{2\alpha}(t-1)}} \\
&= D\sqrt{M}\sqrt{2\sum_{t=1}^{T}\sum_{i\in[N]}\frac{1}{R_i^{2\alpha}(t-1)}}, \qquad (39)
\end{aligned}$$

where above in (a), we have used Jensen's Inequality for the square root function. Using the fact that $R_i(t-1) \geq 1$ for all t, bound (39) implies that

$$\text{Surrogate Regret}_T \leq O(\sqrt{MNT}). \qquad (40)$$

In the following, we show that the above $O(\sqrt{T})$ regret bound can be substantially improved using a novel *bootstrapping* technique described below.

Bootstrapping: Note that the adaptive regret bound depends on the sum of the norm of gradients of the reward vectors, which are controlled by the policy itself. This is in sharp contrast with the usual OCO setting where the policy does not explicitly control the gradients, and the final regret bound is given in terms of the sum of the norm of gradients as given in (37). The *bootstrapping* technique starts with a trivial upper bound on the gradient norms and then uses the regret bound itself to improve the upper bounds on the gradient norms. This, in turn, improves the regret bound through the adaptive regret bound (37). The process is repeated a few times to get the best possible bound.

We now apply the general bootstrapping method to our problem. Note that by the definition of Surrogate Regret$_T$ in (7), we have the following inequality for any *fixed* collection $(\boldsymbol{x}_0^1, ..., \boldsymbol{x}_0^M) \in (\Delta_N)^M$ of distributions:

$$\begin{aligned}
&\sum_{t=1}^{T}\langle\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t), \boldsymbol{x}^{c_t}(t)\rangle \\
&\geq \sum_{t=1}^{T}\langle\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t), \boldsymbol{x}_0^{c_t}\rangle - \text{Surrogate Regret}_T. \qquad (41)
\end{aligned}$$

Also, using the fact that $x_i^{c_t}(t)r_i(t) = R_i(t) - R_i(t-1)$ and following the same calculations up to step (d) of (32), we see that

$$\sum_{t=1}^{T}\langle\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t), \boldsymbol{x}^{c_t}(t)\rangle \leq \sum_{i\in[N]}\phi(R_i(T)). \qquad (42)$$

Combining the above inequality with (41), we have

$$\begin{aligned}
&\sum_{i\in[N]}\phi(R_i(T)) \\
&\geq \sum_{t=1}^{T}\langle\phi'(\boldsymbol{R}(t-1))\odot\boldsymbol{r}(t), \boldsymbol{x}_0^{c_t}\rangle - \text{Surrogate Regret}_T. \qquad (43)
\end{aligned}$$

Next, we lower bound $\phi'(\boldsymbol{R}(t-1))$ by $\phi'(\boldsymbol{R}(T))$ and pick $\boldsymbol{x}_0^j = \frac{1}{N}\mathbf{1}$ for all $j \in [M]$ (i.e., we pick the uniform distribution as an offline benchmark for each context). Doing so, and using the fact that $\boldsymbol{r}(t) \geq \delta \mathbf{1}$ for all t, we have

$$\begin{aligned}\sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t), \boldsymbol{x}_0^{c_t} \rangle &\geq \sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(T)) \odot \boldsymbol{r}(t), \boldsymbol{x}_0^{c_t} \rangle \\ &= \sum_{i \in [N]} \sum_{t=1}^{T} \phi'(R_i(T)) r_i(t) \frac{1}{N} \\ &\geq T \sum_{i \in [N]} \phi'(R_i(T)) \frac{\delta}{N}. \qquad (44)\end{aligned}$$

Plugging the last inequality in (43), we conclude that

$$\sum_{i \in [N]} \phi(R_i(T)) \geq T \sum_{i \in [N]} \phi'(R_i(T)) \frac{\delta}{N} - \text{Surrogate Regret}_T. \qquad (45)$$

Now, noting that $0 < R_i(T) \leq T$ for all i, and that $\phi(\cdot)$ is monotone non-decreasing, we see that for any $i \in [N]$ the above inequality implies

$$\frac{NT^{1-\alpha}}{1-\alpha} \geq T\phi'(R_i(T)) \frac{\delta}{N} - \text{Surrogate Regret}_T, \qquad (46)$$

which implies the following inequality after dividing throughout by T and replacing $\phi'(R_i(T))$ by $\frac{1}{R_i^\alpha(T)}$:

$$\frac{N}{(1-\alpha)T^\alpha} \geq \frac{1}{R_i^\alpha(T)} \frac{\delta}{N} - \frac{\text{Surrogate Regret}_T}{T}, \qquad (47)$$

which is equivalent to

$$\frac{1}{R_i^\alpha(T)} \leq \frac{N}{\delta} \left[\frac{N}{(1-\alpha)T^\alpha} + \frac{\text{Surrogate Regret}_T}{T} \right]. \qquad (48)$$

Now, from Eq. (40), we have the following preliminary bound Surrogate Regret$_T \leq O(\sqrt{MNT})$. We use the *bootstrapping* technique by plugging this in (48) to derive the following improved bound on the cumulative reward accrued by the i^{th} arm.

$$\frac{1}{R_i^\alpha(T)} \leq O\left(\frac{N^2\sqrt{MN}}{T^{\min(1/2,\alpha)}} \right), \ \forall i \in [N]. \qquad (49)$$

Now, we consider the following two cases:

Case 1: $0 \leq \alpha \leq 1/2$**.** In this case, from (49) we see that $\frac{1}{R_i^\alpha(T)} \leq O(\frac{N^2\sqrt{MN}}{T^\alpha})$, and hence $\frac{1}{R_i^{2\alpha}(T)} \leq O(\frac{N^5 M}{T^{2\alpha}})$. Note that this bound holds for all T. Hence, plugging this in (39), we get

$$\text{Surrogate Regret}_T \leq O\left(DN^{\frac{5}{2}} M \sqrt{2 \sum_{t=2}^{T} \sum_{i \in [N]} \frac{1}{(t-1)^{2\alpha}}} \right) \qquad (50)$$

If $0 \leq \alpha < 1/2$, the above bound becomes Surrogate Regret$_T \leq O\left(DN^3 M T^{\frac{1}{2}-\alpha}\right)$. If $\alpha = \frac{1}{2}$, the above bound becomes Surrogate Regret$_T \leq O\left(DN^3 M \sqrt{\log T}\right)$.

Case 2: $1/2 < \alpha < 1$**.** In this case, bound (49) implies that $\frac{1}{R_i^\alpha(T)} \leq \left(\frac{N^2\sqrt{MN}}{T^{1/2}} \right)$, and hence $\frac{1}{R_i^{2\alpha}(T)} \leq O\left(\frac{N^5 M}{T}\right)$. Again, this is true for all T. So, plugging this in (39), we get

$$\begin{aligned}\text{Surrogate Regret}_T &\leq O\left(DN^{\frac{5}{2}} M \sqrt{2 \sum_{t=2}^{T} \sum_{i \in [N]} \frac{1}{(t-1)}} \right) \\ &= O(DN^3 M \sqrt{\log T}) \qquad (51)\end{aligned}$$

Plugging this back in (48), we get that $\frac{1}{R_i^\alpha(T)} \leq O(\frac{N^5 M}{T^\alpha})$, and hence $\frac{1}{R_i^{2\alpha}(T)} \leq O(\frac{N^{10} M^2}{T^{2\alpha}})$. Again, note that this holds for all T. Hence, plugging this in (39), we see that

$$\begin{aligned}\text{Surrogate Regret}_T &\leq O\left(DN^5 M^{\frac{3}{2}} \sqrt{2\sum_{t=2}^{T}\sum_{i\in[N]} \frac{1}{(t-1)^{2\alpha}}}\right) \\ &= O(1) \end{aligned} \tag{52}$$

where above, we have used the fact that $2\alpha > 1$.

A.3. Proof of Lemma 3.1

The proof of **Lemma** 2.1 works here with minor modifications. Again, the upper bound in (16) for $\text{Regret}_T(\beta^{1-\alpha})$ can be split into the difference of two terms A and B as follows:

$$\text{Regret}_T(\beta^{1-\alpha}) \leq \beta^{-\alpha}\mathbb{E}[A - \beta B] \tag{53}$$

where

$$A = \sum_{i\in[N]} \phi'(R_i(T)) \sum_{t=1}^{T} r_i(t) x_{*,i}^{c_t} \tag{54}$$

$$B = \sum_{i\in[N]} \phi'(R_i(T)) \sum_{t=1}^{T} r_i(t) X_i^{c_t}(t) \tag{55}$$

Also, let A' and B' denote the corresponding terms in the surrogate regret for the OLO problem defined in (17). Following the same argument as in the proof of **Lemma** 2.1, we can obtain $A \leq A'$ and $B' \leq (1-\alpha)^{-1}(B+N)$.

As before, the inequality $B' \leq (1-\alpha)^{-1}(B+N)$ implies that $(1-\alpha)B' - N \leq B$. Since $\beta > 0$, we have $\beta B \geq \beta(1-\alpha)B' - \beta N$. Combining this with $A \leq A'$, we see that

$$A - \beta B \leq A' - \beta(1-\alpha)B' + \beta N. \tag{56}$$

Now, pick $\beta = (1-\alpha)^{-1}$ (ensuring that $\beta \geq 1$), and hence, we obtain

$$A - \beta B \leq A' - B' + (1-\alpha)^{-1}N. \tag{57}$$

Taking expectations w.r.t the policy actions, we get

$$\mathbb{E}[A - \beta B] \leq \mathbb{E}[A' - B'] + (1-\alpha)^{-1}N. \tag{58}$$

Finally, from (53), we get

$$\begin{aligned}\text{Regret}(c_\alpha) &\leq \beta^{-\alpha}\mathbb{E}[A' - B'] + \beta^{-\alpha}(1-\alpha)^{-1}N \\ &= (1-\alpha)^\alpha \text{Surrogate Regret}_T + c_\alpha N, \end{aligned} \tag{59}$$

completing the proof of the lemma.

A.4. Proof of Lemma 3.4

Consider some context $j \in [M]$. As before, for any $t \in [1, T]$ let $\boldsymbol{g}_t := \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)$. Then, we have the following set of inequalities considering the adaptive regret bound of the MAB policy handling context j:

$$\begin{aligned}
&\tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_2^2} + \max_{t:c_t=j}\|\boldsymbol{g}_t\|_\infty \sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_1}\right]\right) \\
&\overset{(a)}{\leq} \tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_2^2} + \sqrt{N\sum_{t:c_t=j}\|\boldsymbol{g}_t\|_1}\right]\right) \\
&\overset{(b)}{\leq} \tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\|\phi'(\boldsymbol{R}(t-1))\|_2^2} + \sqrt{N\sum_{t:c_t=j}\|\phi'(\boldsymbol{R}(t-1))\|_1}\right]\right) \\
&\overset{(c)}{=} \tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{2\alpha}(t-1)}} + \sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{\alpha}(t-1)}}\right]\right) \\
&\overset{(d)}{\leq} \tilde{O}\left(\mathbb{E}\left[\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\frac{1}{R_i^{\alpha}(t-1)}}\right]\right) \\
&\overset{(e)}{\leq} \tilde{O}\left(\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right).
\end{aligned} \tag{60}$$

Above, in (a) we have used the fact that $\max_{t:c_t=j}\|\boldsymbol{g}_t\|_\infty \leq 1$, which follows because $\boldsymbol{r}(t) \leq \boldsymbol{1}$ and $\phi'(\boldsymbol{R}(t-1)) \leq \boldsymbol{1}$. In (b), we have used the fact that $\boldsymbol{r}(t) \leq \boldsymbol{1}$. In (c), we have used $\phi'(x) = x^{-\alpha}$. In (d), we have used the fact that for each $i \in [N]$, $R_i(t-1) \geq 1$. Finally, in (e), we have applied Jensen's Inequality to the concave square root function. So, from the last inequality and **Lemma** 3.3, we get

$$\begin{aligned}
&\mathbb{E}\left[\max_{y\in\{\boldsymbol{e}_k\}_{k=1}^N}\sum_{t:c_t=j}\langle\phi'(\boldsymbol{R}(t-1)\odot\boldsymbol{r}(t)), \boldsymbol{y}-\boldsymbol{X}^j(t)\rangle\right] \\
&\leq \tilde{O}\left(2\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right)
\end{aligned} \tag{61}$$

Summing the above inquality over all contexts $j \in [M]$, we get the following inequality on $\hat{\text{Regret}}_T$ defined in (19):

$$\begin{aligned}
\hat{\text{Regret}}_T &\leq \sum_{j\in[M]}\tilde{O}\left(\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right) \\
&= M\sum_{j\in[M]}\frac{1}{M}\tilde{O}\left(\sqrt{N\sum_{t:c_t=j}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right) \\
&\overset{(a)}{\leq} M\tilde{O}\left(\sqrt{\sum_{j\in[M]}\frac{N}{M}\sum_{t:c_t=j}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right) \\
&= \sqrt{MN}\tilde{O}\left(\sqrt{\sum_{t=1}^{T}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^{\alpha}(t-1)}}\right),
\end{aligned} \tag{62}$$

where in (a) above, we have used Jensen's Inequality on the concave square root function. Note that this bound is similar to the bound in (39) for the full information feedback setting, with the only difference being in the exponent of the cumulative reward sequence (2α versus α).

Next, we will derive a bound similar to (48). Note that by the definition of $\hat{\text{Regret}}_T$ in (19), we have the following inequality for any fixed collection of distributions $(\boldsymbol{x}_0^1, ..., \boldsymbol{x}_0^M)$:

$$\begin{aligned}&\sum_{j\in[M]} \mathbb{E}\left[\sum_{t:c_t=j} \langle \boldsymbol{g}_t, \boldsymbol{X}^j(t)\rangle\right]\\ &\geq \sum_{j\in[M]} \mathbb{E}\left[\sum_{t:c_t=j} \langle \boldsymbol{g}_t, \boldsymbol{x}_0^j\rangle\right] - \hat{\text{Regret}}_T.\end{aligned} \tag{63}$$

Using the linearity of expectation, the above inequality can be written as

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{X}^{c_t}(t)\rangle\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_0^{c_t}\rangle\right] - \hat{\text{Regret}}_T. \tag{64}$$

Next, observing that $X_i^{c_t}(t)r_i(t) = R_i(t) - R_i(t-1)$ and following the same calculations up to step (d) of (32) and taking expectations w.r.t the policy actions, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{X}^{c_t}(t)\rangle\right] \leq \mathbb{E}\left[\sum_{i\in[N]} \phi(R_i(T))\right] \tag{65}$$

Lower bounding $\phi'(R_i(t-1))$ by $\phi'(R_i(T))$ yields

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_0^{c_t}\rangle \geq \sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(T)) \odot \boldsymbol{r}(t), \boldsymbol{x}_0^{c_t}\rangle. \tag{66}$$

Finally, taking expectations w.r.t. the policy actions, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_0^{c_t}\rangle\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \phi'(\boldsymbol{R}(T)) \odot \boldsymbol{r}(t), \boldsymbol{x}_0^{c_t}\rangle\right]. \tag{67}$$

Now, let us take the offline benchmark policy to be the uniform distribution for all contexts, i.e., $\boldsymbol{x}_0^j = \frac{1}{N}\mathbf{1}$ for all $j \in [M]$, which will imply that $\sum_{t=1}^{T} r_i(t)x_{0,i}^{c_t} \geq \frac{\delta T}{N}$ for all $i \in [N]$. So, the RHS in the last equation can be lower bounded by $\sum_{i\in[N]} \mathbb{E}[\phi'(R_i(T))] \cdot \frac{\delta T}{N}$. Hence, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_0^{c_t}\rangle\right] \geq \sum_{i\in[N]} \mathbb{E}[\phi'(R_i(T))] \cdot \frac{\delta T}{N} \tag{68}$$

So, from the last equation and equations (64) and (65), we get

$$\mathbb{E}\left[\sum_{i\in[N]} \phi(R_i(T))\right] \geq \sum_{i\in[N]} \mathbb{E}[\phi'(R_i(T))] \cdot \frac{\delta T}{N} - \hat{\text{Regret}}_T \tag{69}$$

So from here, following the same steps as in the full information feedback setting, we obtain

$$\mathbb{E}\left[\frac{1}{R_i^{\alpha}(T)}\right] \leq \frac{N}{\delta}\left[\frac{N}{(1-\alpha)T^{\alpha}} + \frac{\hat{\text{Regret}}_T}{T}\right] \tag{70}$$

Note the similarity between the above inequality and inequality (48) for the full information setting.

Now, we know that $R_i^\alpha(t-1) \geq 1$ for all $i \in [N]$ and t. Plugging this in (62), we get our first bound, which is $\hat{\text{Regret}}_T \leq \tilde{O}(N\sqrt{MT})$.

As before, we do a tighter analysis to get a better regret bound. So, let $\alpha_0 \in [0,1)$ be any number. As the result of **Lemma** 3.4 claims, we want to show that $\hat{\text{Regret}}_T = \tilde{O}(T^{\frac{1-\alpha_0}{2}})$. Since $\alpha_0 \in [0,1)$, there is some positive integer $N_0 \geq 0$ such that

$$\frac{2^{N_0}-1}{2^{N_0}} \leq \alpha_0 < \frac{2^{N_0+1}-1}{2^{N_0+1}} \tag{71}$$

Now, let $\epsilon_0 > 0$ be a very small number which satisfies the following inequalities for all $0 \leq n \leq N_0$:

$$\frac{2^n-1}{2^n} < \frac{2^{n+1}-1}{2^{n+1}} - \left(\frac{2^{n+1}-1}{2^n}\right)\epsilon_0 \tag{72}$$

$$\alpha_0 \leq \frac{2^{N_0+1}-1}{2^{N_0+1}} - \left(\frac{2^{N_0+1}-1}{2^{N_0}}\right)\epsilon_0 \tag{73}$$

Note that, the above two conditions are equivalent to the following two conditions for all $0 \leq n \leq N_0$:

$$\epsilon_0 < \frac{2^n}{2^{n+1}-1}\left[\frac{2^{n+1}-1}{2^{n+1}} - \frac{2^n-1}{2^n}\right] \tag{74}$$

$$\epsilon_0 < \frac{2^{N_0}}{2^{N_0+1}-1}\left[\frac{2^{N_0+1}-1}{2^{N_0+1}} - \alpha_0\right] \tag{75}$$

Since all the quantities on the RHS in the two equations above are positive, ϵ_0 can be taken to be something smaller than the minimum of all the above quantities. Now, we have obtained $\hat{\text{Regret}}_T \leq \tilde{O}(N\sqrt{MT}) = O(\log T \cdot N\sqrt{MT})$. Plugging this in (70), we get the following:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{R_i^\alpha(T)}\right] &\leq O\left(N^2\left[\frac{1}{T^\alpha} + \frac{\hat{\text{Regret}}_T}{T}\right]\right) \\
&= O\left(N^3\sqrt{M}\left[\frac{1}{T^\alpha} + \frac{\log T}{\sqrt{T}}\right]\right) \\
&\overset{(a)}{=} O\left(N^3\sqrt{M}\left(\frac{1}{T^\alpha} + \frac{T^{\epsilon_0}}{\sqrt{T}}\right)\right) \\
&= O\left(\frac{N^3\sqrt{M}}{T^{\min(\alpha, \frac{1}{2}-\epsilon_0)}}\right)
\end{aligned} \tag{76}$$

where above in (a), we have used the simple fact that $\log T = O(T^{\epsilon_0})$. Plugging the above bound in (62), we get the following bound for any $0 \leq \alpha \leq \frac{1}{2} - \epsilon_0$:

$$\begin{aligned}
\hat{\text{Regret}}_T &\leq \tilde{O}\left(\sqrt{MN}\sqrt{\sum_{t=1}^{T}\sum_{i\in[N]}\mathbb{E}\frac{1}{R_i^\alpha(t-1)}}\right) \\
&\leq \tilde{O}\left(M^{\frac{1}{2}+\frac{1}{4}}N^{\frac{1}{2}+\frac{3}{2}}\sqrt{N\sum_{t=1}^{T}\frac{1}{(t-1)^\alpha}}\right) \\
&= \tilde{O}(M^{\frac{1}{2}+\frac{1}{4}}N^{\frac{1}{2}+\frac{3}{2}+\frac{1}{2}}T^{\frac{1-\alpha}{2}}) \\
&= \tilde{O}(M^{\frac{3}{4}}N^{\frac{5}{2}}T^{\frac{1-\alpha}{2}})
\end{aligned} \tag{77}$$

By the same inequalities as above, for any $\frac{1}{2} - \epsilon_0 \leq \alpha < 1$ we will have the bound:

$$\begin{aligned}
\hat{\text{Regret}}_T &\leq \tilde{O}\left(M^{\frac{3}{4}}N^{\frac{5}{2}}T^{\frac{1-(\frac{1}{2}-\epsilon_0)}{2}}\right) \\
&= \tilde{O}(M^{\frac{3}{4}}N^{\frac{5}{2}}T^{\frac{1}{4}+\frac{\epsilon_0}{2}})
\end{aligned} \tag{78}$$

More generally, suppose for some $0 \leq n < N_0$, we have

$$\hat{\text{Regret}}_T \leq \tilde{O}(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}} T^{\frac{1-\alpha}{2}}) \tag{79}$$

$$\hat{\text{Regret}}_T \leq \tilde{O}(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}} T^{\frac{1}{2^{n+2}}+\frac{2^{n+1}-1}{2^{n+1}}\epsilon_0}) \tag{80}$$

where (79) holds for all $\alpha \in \left[\frac{2^n-1}{2^n}, \frac{2^{n+1}-1}{2^{n+1}} - \left(\frac{2^{n+1}-1}{2^n}\right)\epsilon_0\right]$, and (80) holds for all $\alpha \in \left[\frac{2^{n+1}-1}{2^{n+1}} - \left(\frac{2^{n+1}-1}{2^n}\right)\epsilon_0, 1\right)$. Note that, by our choice of ϵ_0, both these intervals have non-negative measure (recall (72)). Also, note that we have shown the base case for $n = 0$ via inequalities (77) and (78). We will now show that (79) and (80) continue to hold for $n + 1$.

Now, since we know that (80) holds for all $\alpha \in \left[\frac{2^{n+1}-1}{2^{n+1}} - \left(\frac{2^{n+1}-1}{2^n}\right)\epsilon_0, 1\right)$, we plug the bound (80) in (70) and get the following for such α:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{R_i^\alpha(t)}\right] \\
&\leq O\left(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}+2}\left[\frac{1}{T^\alpha} + \frac{\log T \cdot T^{\frac{1}{2^{n+2}}+\frac{2^{n+1}-1}{2^{n+1}}\epsilon_0}}{T}\right]\right) \\
&= O\left(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}+2}\left[\frac{1}{T^\alpha} + \frac{\log T}{T^{1-\frac{1}{2^{n+2}}-\left(\frac{2^{n+1}-1}{2^{n+1}}\right)\epsilon_0}}\right]\right) \\
&\overset{(a)}{=} O\left(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}+2}\left[\frac{1}{T^\alpha} + \frac{T^{\epsilon_0}}{T^{1-\frac{1}{2^{n+2}}-\left(\frac{2^{n+1}-1}{2^{n+1}}\right)\epsilon_0}}\right]\right) \\
&= O\left(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}+2}\left[\frac{1}{T^\alpha} + \frac{1}{T^{\frac{2^{n+2}-1}{2^{n+2}}-\left(\frac{2^{n+2}-1}{2^{n+1}}\right)\epsilon_0}}\right]\right) \\
&= O\left(M^{\frac{2^{n+2}-1}{2^{n+2}}} N^{\frac{2^{n+3}-3}{2^{n+1}}+2}\left[\frac{1}{T^{\min\left(\alpha, \frac{2^{n+2}-1}{2^{n+2}}-\left(\frac{2^{n+2}-1}{2^{n+1}}\right)\epsilon_0\right)}}\right]\right)
\end{aligned} \tag{81}$$

where in (a) above, we have simply used the fact that $\log T = O(T^{\epsilon_0})$. Note that by our choice of ϵ_0 (recall inequality (72)), we have

$$\frac{2^{n+1}-1}{2^{n+1}} \leq \frac{2^{n+2}-1}{2^{n+2}} - \left(\frac{2^{n+2}-1}{2^{n+1}}\right)\epsilon_0 \tag{82}$$

So, plugging the bound of (81) in (62), we can obtain

$$\hat{\text{Regret}}_T \leq \tilde{O}(M^{\frac{2^{n+3}-1}{2^{n+3}}} N^{\frac{2^{n+4}-3}{2^{n+2}}} T^{\frac{1-\alpha}{2}}) \tag{83}$$

$$\hat{\text{Regret}}_T \leq \tilde{O}(M^{\frac{2^{n+3}-1}{2^{n+3}}} N^{\frac{2^{n+4}-3}{2^{n+2}}} T^{\frac{1}{2^{n+3}}+\frac{2^{n+2}-1}{2^{n+2}}\epsilon_0}) \tag{84}$$

where (83) holds for all $\alpha \in \left[\frac{2^{n+1}-1}{2^{n+1}}, \frac{2^{n+2}-1}{2^{n+2}} - \left(\frac{2^{n+2}-1}{2^{n+1}}\right)\epsilon_0\right]$ and (84) holds for all $\alpha \in \left[\frac{2^{n+2}-1}{2^{n+2}} - \left(\frac{2^{n+2}-1}{2^{n+1}}\right)\epsilon_0, 1\right)$. Hence, by induction, we see that (79) and (80) hold for all $0 \leq n \leq N_0$ in the respective intervals.

Finally, (71) and (73) imply that $\alpha_0 \in \left[\frac{2^{N_0}-1}{2^{N_0}}, \frac{2^{N_0+1}-1}{2^{N_0+1}} - \left(\frac{2^{N_0+1}-1}{2^{N_0}}\right)\epsilon_0\right]$. So, by what we've shown above, we conclude that

$$\hat{\text{Regret}}_T \leq \tilde{O}(M^{\frac{2^{N_0+2}-1}{2^{N_0+2}}} N^{\frac{2^{N_0+3}-3}{2^{N_0+1}}} T^{\frac{1-\alpha_0}{2}}) \tag{85}$$

and this completes the proof of the claim.

A.5. Proof of Lemma 3.3

Fix some context $j \in [M]$ and a time horizon T. Consider the sequence $(\boldsymbol{g}_t)_{t:c_t=j}$ of all reward vector that context j sees. Recall that $\boldsymbol{g}_t = \phi'(\boldsymbol{R}(t-1)) \odot \boldsymbol{r}(t)$. By our assumption, the reward vectors $\boldsymbol{r}(t)$ are generated by an *oblivious adversary*, i.e they are fixed beforehand. However, the cumulative reward vectors $\boldsymbol{R}(t-1)$ for $t \in [1, T]$ are *policy-dependent*, i.e they are *random*. Also, note that at every time step t, our policy picks some arm $I_t \in [N]$ to be played; from equation (10) that defines how cumulative rewards are updated, we see that there are only finitely many sequences $(g_t)_{t:c_t=j}$ that our policy can see over the time horizon T.

So, let S be the set of all sequences $(\boldsymbol{g}_t)_{t:c_t=j}$ that our policy can see. Let $\boldsymbol{q} \in \Delta_S$ be the probability distribution that the policy induces over the set S of possible reward sequences. For a fixed reward sequence $(\boldsymbol{l}_t)_{t:c_t=t} \in S$, we have the following by **Theorem** 3.2:

$$\begin{aligned}&\mathbb{E}\left[\max_{\{\boldsymbol{e}_k\}_{k=1}^N} \sum_{t:c_t=j} \langle \boldsymbol{l}_t, \boldsymbol{e}_k - \boldsymbol{X}^j(t)\rangle\right] \\ &\leq \tilde{O}\left(\sqrt{N \sum_{t:c_t=j} \|\boldsymbol{l}_t\|_2^2} + \max_{t:c_t=j} \|\boldsymbol{l}_t\|_\infty \sqrt{N \sum_{t:c_t=j} \|\boldsymbol{l}_t\|_1}\right) \end{aligned} \tag{86}$$

Above, the expectation in the first time is taken w.r.t the policy actions. Now, taking expectations in the above inequality w.r.t the distribution $\boldsymbol{q}$ over S, we get

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\max_{\{\boldsymbol{e}_k\}_{k=1}^N} \sum_{t:c_t=j} \langle \boldsymbol{l}_t, \boldsymbol{e}_k - \boldsymbol{X}^j(t)\rangle\right]\right] \\ &\leq \mathbb{E}\left[\tilde{O}\left(\sqrt{N \sum_{t:c_t=j} \|\boldsymbol{l}_t\|_2^2} + \max_{t:c_t=j} \|\boldsymbol{l}_t\|_\infty \sqrt{N \sum_{t:c_t=j} \|\boldsymbol{l}_t\|_1}\right)\right] \end{aligned} \tag{87}$$

By the tower property of conditional expectations, the first term on the LHS in the above inequality is just

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\max_{\{\boldsymbol{e}_k\}_{k=1}^N} \sum_{t:c_t=j} \langle \boldsymbol{l}_t, \boldsymbol{e}_k - \boldsymbol{X}^j(t)\rangle\right]\right] \\ &= \mathbb{E}\left[\max_{\{\boldsymbol{e}_k\}_{k=1}^N} \sum_{t:c_t=j} \langle \boldsymbol{g}_t, \boldsymbol{e}_k - \boldsymbol{X}^j(t)\rangle\right] \end{aligned} \tag{88}$$

where the expectation on the RHS above is taken w.r.t the policy actions. Combining the above with (87), the claim follows.

B. Additional Experiments

Figure 8 shows that the α-FAIRCB policy achieves a better approximate contextual regret (12) even for a moderately large time horizon in the bandit setting. Figures 9 and 10 show plots of the standard regret of all the policies in the full information and the bandit information settings respectively. As before, it is clearly seen that the α-FAIRCB policy beats all the other policies in terms of the standard regret in both the settings.

C. Extension to non-negative rewards

In this section, we will study the case when the reward vectors $\boldsymbol{r}(t)$ don't necessarily satisfy the inequality $\delta\boldsymbol{1} \leq \boldsymbol{r}(t)$, and we instead assume that $\boldsymbol{0} \leq \boldsymbol{r}(t) \leq \boldsymbol{1}$ holds for all t. Our trick will be to rescale all the reward vectors to the vector $\boldsymbol{r}'(t)$ defined by

$$r_i'(t) = \frac{\epsilon + r_i(t)}{1+\epsilon} \tag{89}$$

where $\epsilon > 0$ is to be decided later. Note that we still assume that $\boldsymbol{r}(t) \leq \boldsymbol{1}$. It is easily seen that

$$\frac{\epsilon}{1+\epsilon} \leq r_i'(t) = \frac{\epsilon + r_i(t)}{1+\epsilon} \leq 1 \tag{90}$$

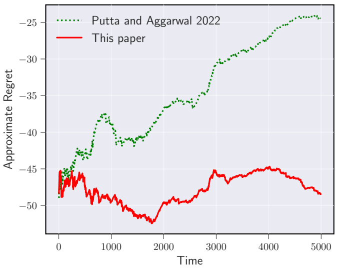

Figure 8. Approximate regret for the bandit information setting.

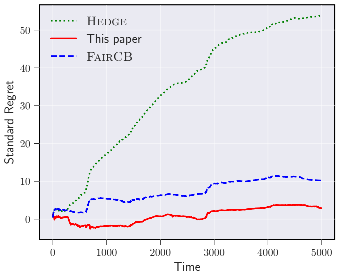

Figure 9. Standard regret for the full information setting.

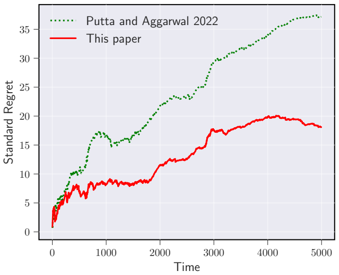

Figure 10. Standard regret for the bandit information setting.

Now, let $\boldsymbol{R}'(t)$ and $\boldsymbol{R}(t)$ be the usual cumulative reward vectors (with rewards $\boldsymbol{r'}$, $\boldsymbol{r}$ respectively) and $\boldsymbol{R}'(0) = \boldsymbol{R}(0) = \boldsymbol{1}$. Suppose we run our bandit policy for the rewards given by the vectors $\boldsymbol{r}'(0)$. In that case, we have

$$\sum_{i\in[N]} \phi(R'_{i,*}(t)) - \sum_{i\in[N]} \phi(R'_i(t)) = O(MNT^{\frac{1-\alpha}{\alpha}}) \tag{91}$$

where as usual, the first term is computed w.r.t a fixed collection $\boldsymbol{x}_* \equiv (\boldsymbol{x}^1_*, ..., \boldsymbol{x}^M_*)$ of distributions. Now, we have the following set of inequalities:

$$\begin{aligned}\text{Regret}' &:= \sum_{i\in[N]} \phi(R'_{*,i}(T)) - \sum_{i\in[N]} \phi(R'_i(T)) \\ &= \sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^{T} x^{c_t}_{i,*}(t)\frac{r_i(t)+\epsilon}{1+\epsilon}\right) - \sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^{T} x^{c_t}_{i}(t)\frac{r_i(t)+\epsilon}{1+\epsilon}\right) \end{aligned} \tag{92}$$

Now, consider the first of the two terms in (92). We have

$$\begin{aligned}\sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^{T} x^{c_t}_{i,*}(t)\frac{r_i(t)+\epsilon}{1+\epsilon}\right) &\geq \sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^{T} x^{c_t}_{i,*}(t)\frac{r_i(t)}{1+\epsilon}\right) && (\phi \text{ is monotone}) \\ &\geq \sum_{i\in[N]} \phi\left(\frac{1}{1+\epsilon} + \sum_{t=1}^{T} x^{c_t}_{i,*}(t)\frac{r_i(t)}{1+\epsilon}\right) && (\phi \text{ is monotone}) \\ &= \frac{1}{(1+\epsilon)^{1-\alpha}} \sum_{i\in[N]} \phi\left(1 + \sum_{t=1}^{T} x^{c_t}_{i,*}(t) r_i(t)\right) \\ &= \frac{1}{(1+\epsilon)^{1-\alpha}} \sum_{i\in[N]} \phi(R_{i,*}(t)) \end{aligned} \tag{93}$$

Now, consider the second term in (92). We have the following set of inequalities:

$$
\begin{aligned}
&\sum_{i\in[N]} \phi\left(1+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)+\epsilon}{1+\epsilon}\right) \\
&= \sum_{i\in[N]} \phi\left(1+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)}{1+\epsilon}+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{\epsilon}{1+\epsilon}\right) \\
&\leq \sum_{i\in[N]} \phi\left(1+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)}{1+\epsilon}+T\frac{\epsilon}{1+\epsilon}\right) && (\phi \text{ is monotone}) \\
&\leq \sum_{i\in[N]} \phi\left(1+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)}{1+\epsilon}\right)+N\phi\left(\frac{T\epsilon}{1+\epsilon}\right) && (\phi(a+b)\leq\phi(a)+\phi(b)) \\
&= \sum_{i\in[N]} \phi\left(\frac{\epsilon}{1+\epsilon}+\frac{1}{1+\epsilon}+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)}{1+\epsilon}\right)+N\phi\left(\frac{T\epsilon}{1+\epsilon}\right) \\
&\leq N\phi\left(\frac{\epsilon}{1+\epsilon}\right)+\sum_{i\in[N]} \phi\left(\frac{1}{1+\epsilon}+\sum_{t=1}^{T} x_i^{c_t}(t)\frac{r_i(t)}{1+\epsilon}\right)+N\phi\left(\frac{T\epsilon}{1+\epsilon}\right) && (\phi(a+b)\leq\phi(a)+\phi(b)) \\
&= N\phi\left(\frac{\epsilon}{1+\epsilon}\right)+\frac{1}{(1+\epsilon)^{1-\alpha}}\sum_{i\in[N]} \phi\left(R_i(t)\right)+N\phi\left(\frac{T\epsilon}{1+\epsilon}\right) && (94)
\end{aligned}
$$

So, combining (93) and (94), we get the following:

$$
\begin{aligned}
&\sum_{i\in[N]} \phi(R'_{*,i}(T))-\sum_{i\in[N]} \phi(R'_i(T)) \\
&\geq \frac{1}{(1+\epsilon)^{1-\alpha}}\left(\sum_{i\in[N]} \phi(R_{i,*}(T))-\sum_{i\in[N]} \phi(R_i(t))\right)-N\phi\left(\frac{\epsilon}{1+\epsilon}\right)-N\phi\left(\frac{T\epsilon}{1+\epsilon}\right) && (95)
\end{aligned}
$$

For convenience, this inequality can be written as

$$
\text{Regret}' \geq \frac{1}{(1+\epsilon)^{1-\alpha}}\text{Regret}-\frac{N}{(1-\alpha)}\left(\frac{\epsilon}{1+\epsilon}\right)^{1-\alpha}-\frac{NT^{1-\alpha}}{1-\alpha}\left(\frac{\epsilon}{1+\epsilon}\right)^{1-\alpha} \tag{96}
$$

Multiplying throughout by $(1+\epsilon)^{1-\alpha}$ and rearranging, we get

$$
\text{Regret} \leq (1+\epsilon)^{1-\alpha}\text{Regret}'+\frac{N\epsilon^{1-\alpha}}{(1-\alpha)}+\frac{NT^{1-\alpha}\epsilon^{1-\alpha}}{1-\alpha} \tag{97}
$$

The trick now is to fine-tune the parameter ϵ optimally, which we do by equating the first and the last terms in the RHS of (97), since these two are the dominating terms. Note that Regret$'$ (in equation (97)) is inversely proportional to δ (see (48)), and for our case we have $\delta=\frac{\epsilon}{1+\epsilon}$; hence, we see that Regret$'$ is inversely proportional to ϵ. Note that, for both the full information and bandit information settings, we have a bound of $\tilde{O}(T^{\frac{1-\alpha}{2}})$ on Regret$'$ (infact, the bound for the full information setting is better). So, the optimal value of ϵ is computed by equating

$$
\tilde{O}\left(\frac{T^{\frac{1-\alpha}{2}}}{\epsilon}\right)=(\epsilon T)^{1-\alpha} \tag{98}
$$

and solving for ϵ, this gives us

$$
\epsilon = O\left(T^{-\frac{1-\alpha}{2(2-\alpha)}}\right) \tag{99}
$$

Substituting this bound back in (97), we obtain the bound

$$\text{Regret} \leq O\left(T^{\frac{(3-\alpha)(1-\alpha)}{2(2-\alpha)}}\right) \tag{100}$$

Finally, recall that we have $\alpha < 1$. Hence, it's easy to see that

$$\frac{(3-\alpha)(1-\alpha)}{2(2-\alpha)} < 1 - \alpha \tag{101}$$

So, although the bound is worse than the ones given in Theorem 2.3 and Theorem 3.5, it is still *sublinear* w.r.t $T^{1-\alpha}$, and hence the bound is still non-trivial and meaningful.