Fair Best Arm Identification with Fixed Confidence

Alessio Russo[1,⋆] and Filippo Vannella[2,⋆]

***Abstract*— In this work, we present a novel framework for Best Arm Identification (BAI) under fairness constraints, a setting that we refer to as *fair BAI*. Unlike traditional BAI, which solely focuses on identifying the optimal arm with minimal sample complexity, fair BAI also includes a set of fairness constraints. These constraints impose a lower limit on the selection rate of each arm and can be either model-agnostic or model-dependent. For this setting, we establish an instance-specific sample complexity lower bound and analyze the *price of fairness*, quantifying how fairness impacts sample complexity. Based on the sample complexity lower bound, we propose** F-BAI**, an algorithm provably matching the sample complexity lower bound, while ensuring that the fairness constraints are satisfied. Numerical results, conducted using both a synthetic model and a practical wireless scheduling application, show the efficiency of** F-BAI **in minimizing the sample complexity while achieving low fairness violations**

I. Introduction

In recent years, a large body of work has focused on making machine learning systems more fair [1]. This effort reflects a broader societal shift towards more ethical algorithms, recognizing the impact these systems have across various sectors of society.

Notably, the importance of fairness has been also recently acknowledged within the context of Multi-Armed Bandit (MABs) [2]. MABs (or simply bandits) [3] are sequential decision-making problems under uncertainty, in which a learner must strategically select arms to maximize a given objective over time.

Bandit algorithms have seen widespread adoption in various applications: online advertisement [4], recommender systems [5], [6], wireless network optimization [7], and others [8]. Due to the importance and impact of these applications, there has been an increasing understanding of the need to include fairness aspects in the decision-making process. For example, in wireless scheduling problems with multiple Quality of Service (QoS) classes (see Sec. VI), fairness is achieved by ensuring that users within each class receive appropriate performance according to their specific requirements, e.g., in terms of throughput [9].

However, traditional bandit algorithms do not inherently address such fairness aspects. Indeed, while guaranteeing fairness has received attention in the setting of regret minimization [10], [9], [11], [12], this aspect remains largely unexplored within the problem of Best Arm Identification (BAI) [13], [14]. In BAI, the objective is to find an optimal arm with a prescribed level of confidence and with minimal sample complexity.

⋆Equal contribution.
[1]Ericsson AB, Stockholm, Sweden.
[2]Ericsson Research, Stockholm, Sweden.

In this work we investigate how to include fairness constraints into the BAI problem, leading to a novel setting that we refer to as *fair BAI*. Fair BAI extends the classical BAI problem by imposing fairness constraints on arm selection rates. These constraints ensure that no arm is underrepresented in the sampling process, addressing potential biases that may arise from pure exploration algorithms. Nonetheless, the introduction of fairness into BAI raises some challenges, notably how to balance the inherent trade-off between fairness and sample complexity, a duality that reflects wider challenges in algorithmic fairness and decision-making. We summarize our contributions as follows:

1) *The fair BAI setting.* We introduce fair BAI in §III-B, a novel and general bandit setting for BAI under fairness constraints. Our approach to fairness is broad, encompassing various classical notions such as proportional fairness and individual fairness. This versatility allows our framework to be applicable in various settings.
2) *Sample complexity and price of fairness.* In §IV we derive an instance-specific lower bound on the sample complexity of any Probably Approximately Correct (PAC) algorithm that adheres to these fairness constraints. Based on this bound, we quantify the *price of fairness* in fair BAI. This price refers to the additional samples required to identify the best arm while complying with the fairness constraints, providing additional insights into the trade-off between sample complexity and fairness.
3) *Optimal and fair algorithm.* In §V we devise F-BAI, an algorithm that asymptotically matches the sample complexity lower bound as the confidence grows higher. Furthermore, the algorithm guarantees that the fairness constraints are satisfied at each round of the interaction (for pre-specified values of fairness, i.e., model-agnostic constraints) or asymptotically (for fairness constraints that depend on the model parameter, which needs to be learned by the algorithm).
4) *Numerical experiments.* Lastly, in §VI we test our algorithm on both synthetic instances and a fair scheduling problem in radio wireless networks. Numerical results show that F-BAI is not only sample efficient but also consistently achieves minimal fairness violation.

II. Related Work

Different notions of fairness have been considered in MAB problems and, more generally, in sequential decision-making problems. For an extended literature review, please refer to App. J. For comprehensive surveys on this topic, see [2], [12].

In MAB problems, fairness has been investigated in different settings including plain, combinatorial, contextual, and linear [10], [9], [15], [16] and with different notions of fairness. The majority of these notions, generally fall into the following categories: pre-specified fairness, individual fairness, counterfactual fairness and group fairness [12], [2]. Of these notions, the closest to our work are the first two, and we focus on these in the remainder of this section.

Selection with pre-specified values of fairness [9], [17], [11], [18], [19] simply demands that the rate, or probability, at which an algorithm selects an arm stays within a pre-specified range. Related works in this category mostly target finite-time fairness constraints, which require that a given number of arms pull must be satisfied in each round [11], [19], [20]. These constraints are, in general, model-agnostic.

On the other hand, *individual fairness* [21], [22] requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [10], [20], [23], [24], [25], [26]. These constraints impose a minimal rate at which an algorithm must select arms, and they are generally model-dependent. Moreover, in this setting, algorithms often provide asymptotic guarantees as the time horizon grows large. In general, it is hard to avoid asymptotic guarantees when the constraints depend on the unknown model parameter, which needs to be estimated during the learning process.

Other important works consider the α-fairness criterion [27], [28], [29] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α. This criterion includes different notions of fairness, such as *max-min* fairness, which allocates resources as equally as possible, or *proportional fairness*, which allocates resources in a proportional manner [26], [25], [30].

Notably, as explained in the next section, our fairness definition is very general and includes for example the case of individual, proportional, and pre-specified values of fairness (see Remark III.2 for details).

Last, but not least, the totality of the above-mentioned works focuses on the regret minimization setting, where the aim is to maximize the expected cumulative rewards [3]. In contrast, our work focuses on the setting of pure exploration (a.k.a. BAI) with fixed confidence [14]. To the best of our knowledge, the only work investigating fairness in BAI is [31], where the authors consider fairness constraints on sub-populations (see App. J for details). However, our setting is inherently different and assumes fairness constraints on each arm rather than sub-populations.

III. Problem Setting

In this section, we outline the bandit model considered in this paper and we present our fair BAI setting.

A. Multi-Armed Bandit Model

We consider a stochastic bandit problem with a finite set of K arms, that we denote by $[K] := \{1, \dots, K\}$. In each round $t \geq 1$, the learner selects an arm $a_t \in [K]$, and observes a Gaussian reward $r_t \sim \mathcal{N}(\theta_{a_t}, 1)$. The rewards are i.i.d. (over rounds) and $\theta = (\theta_a)_{a\in[K]}$ is the unknown parameter vector. We indicate by $a^\star = \arg\max_{a\in[K]} \theta_a$ the arm with highest average reward, that we assume to be unique, and indicate by $\Theta = \{\theta \in \mathbb{R}^K : |\arg\max_{a\in[K]} \theta_a| = 1\}$ the set of models satisfying this property. We define the sub-optimality gap for an arm $a \neq a^\star$ as $\Delta_a = \theta_{a^\star} - \theta_a$, and the maximal (resp. minimal) gap as $\Delta_{\max} = \max_{a\neq a^\star} \Delta_a$ (resp. $\Delta_{\min} = \min_{a\neq a^\star} \Delta_a$). We also indicate by $N_a(t) = \sum_{s=1}^t \mathbf{1}_{\{a_s=a\}}$ the number of times an arm a has been selected up to round t. The empirical average of $\theta = (\theta_a)_{a\in[K]}$ at round t is denoted as $\hat{\theta}(t) = (\hat{\theta}_a(t))_{a\in[K]}$, where $\hat{\theta}_a(t) = \frac{1}{N_a(t)} \sum_{s\in[t]} r_s \mathbf{1}_{\{a_s=a\}}$. We denote by $\mathcal{F}_t$ the σ-algebra generated by $(a_1, r_1, \dots, a_t, r_t)$, the history of observations. We indicate by $\mathrm{kl}(p, q)$ the Kullback–Leibler divergence between two Bernoulli distributions of mean p and q. For any two vectors we write $x, y \in [0, 1]^K$, $x \geq y$ to denote $x_a \in [y_a, 1], \forall a \in [K]$.

B. Fair Best Arm Identification

We now briefly introduce the BAI setting, then proceed to explain how to incorporate fairness constraints.

Best Arm Identification: In BAI [32], the objective is to identify the best arm $a^\star$ with probability at least $1 - \delta$, where $\delta \in (0, 1/2)$, using the least number of samples. A BAI algorithm $\mathcal{A}$ consists of a sampling rule π_t, a stopping rule, and a decision rule. The sampling rule π_t decides which arm is selected in round t based on past observations: a_t is $\mathcal{F}_{t-1}$-measurable. The stopping rule decides when to stop sampling, and is defined by τ_δ, a stopping time w.r.t. the filtration $(\mathcal{F}_t)_{t\geq 1}$. The sample complexity for an algorithm $\mathcal{A}$ is denoted by $\mathbb{E}_{\theta,\mathcal{A}}[\tau_\delta]$, and w.l.o.g. in the following we simply denote it by $\mathbb{E}_\theta[\tau_\delta]$. Lastly, the decision rule outputs a guess of the best arm $\hat{a}_{\tau_\delta}$, based on observations collected up to round τ_δ.

Fairness Constraints: In fair BAI, we seek to identify the best arm as quickly as possible while satisfying a set of fairness constraints. We consider general types of constraints which can be either *pre-specified* or *dependent on the problem parameter* θ, as detailed in the following.

1) *Pre-specified constraints*: the selection rate at the random stopping time τ_δ, needs be larger than some *pre-specified* value $p_a \in [0, 1]$:
$$\frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \geq p_a, \forall a \in [K]. \tag{1}$$
2) *θ-dependent constraints*: asymptotically, as the confidence grows larger, the minimum selection rate at the stopping time τ_δ needs to be larger than some function of θ that we denote by $p_a(\theta) : \mathbb{R}^K \to [0, 1]$:
$$\liminf_{\delta\to 0} \frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \geq p_a(\theta), \forall a \in [K], \tag{2}$$
In this case, we further assume $p_a(\theta)$ to be continuous in θ, for every arm $a \in [K]$.

To lighten the notation, when there is no ambiguity, we omit the dependence on θ for $p(\theta)$. In the remainder of the paper, we refer to the vector $p = (p_a)_{a\in[K]}$ as the *fairness*

rate or simply rate. We denote by $p_{\text{sum}} \coloneqq \sum_{a\in[K]} p_a \leq 1$ the sum of fairness rates, and by $p_{\min} = \min_{a\in[K]:p_a>0} p_a$ the minimal fairness rate. Note that, for θ-dependent constraints, the guarantees are asymptotic (as $\delta \to 0$), since the model parameter θ (from which $p(\theta)$ depends) is unknown at the start of the interaction.

Our goal is to devise a p-fair δ-PAC algorithm with minimal sample complexity τ_δ, according to the following definition.

Definition III.1. An algorithm is p-fair δ-PAC (resp. asymptotically $p(\theta)$-fair δ-PAC) if for all $\theta \in \Theta$, $\delta \in (0, 1/2)$, it satisfies (i) Eq. (1) (resp. Eq. (2)), (ii) $\mathbb{P}_\theta(\hat{a}_\tau \neq a^\star) \leq \delta$, and (iii) $\mathbb{P}_\theta(\tau < \infty) = 1$.

Remark III.2. The fairness constraints that we propose are general enough to include the classical notions of *individual fairness* or *proportional fair*. For example, one can set $p(\theta) = p_0 \cdot \texttt{softmax}(\theta)$ for some $p_0 \in [0, 1]$, or $p_a(\theta) = p_0\theta_a / \sum_b \theta_b$ when the values of θ are positive. In the latter case, with $p_0 = 1$, we recover the proportional fair constraints used in previous works [25], [26], [30] (we refer the reader to the extended related work in App. J for more details). On the other hand, by setting a constant constraint p_a we find the classical notion of *selection with pre-specified range* [2]. For example, in [17], they use the same value of p_a for all arms, while in [11], the authors select a fixed $p_a \in [0, 1/K)$, for all $a \in [K]$.

IV. Sample Complexity Lower Bound and the Price of Fairness

In this section, we first provide an instance-specific sample complexity lower bound that is valid for any p-fair δ-PAC algorithm. Next, we analyse the *price of fairness*, quantifying how fairness constraints impact sample complexity.

A. Sample complexity lower bound

The following theorem states a lower bound on the sample complexity of any p-fair δ-PAC algorithm. Notably, the sample complexity is characterized by the following constant, that we refer to as *characteristic time*

$$T_p^\star = \inf_{w\in\Sigma_p} \max_{a\neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}, \tag{3}$$

where $\Sigma_p = \{w \geq p : \sum_{a\in[K]} w_a = 1\}$ is the clipped simplex.

Theorem IV.1. *Any p-fair δ-PAC algorithm satisfies, $\forall\theta \in \Theta$, $\mathbb{E}_\theta[\tau_\delta]/\log(1/2.4\delta) \geq 2T_p^\star$. Any asymptotically $p(\theta)$-fair δ-PAC algorithm, instead, we have $\forall\theta \in \Theta, \liminf_{\delta\to 0} \mathbb{E}_\theta[\tau_\delta]/\log(1/\delta) \geq 2T_p^\star$.*

The proof (see App. C) leverages classical change-of-measure arguments [3] and is a straightforward extension of the one in the plain bandit setting by [14]. The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm for a given fairness vector $p = (p_a)_{a\in[K]}$. The *allocation vector* $w = (w_a)_{a\in[K]}$, where $w_a \coloneqq \mathbb{E}_\theta[N_a(\tau_\delta)]/\mathbb{E}_\theta[\tau_\delta]$, characterizes the asymptotic proportion of rounds in which an arm a is selected. Furthermore, an allocation $w_p^\star$ solving the optimization problem (3) is *optimal* and *fair*: an algorithm relying on a sampling strategy realizing $w_p^\star$ would yield the lowest possible sample complexity while satisfying the fairness constraints.

The main difference with the lower bound in [14] lies in the set of "clipped" allocations Σ_p, which accounts for the additional fairness constraints $w \geq p$. Notably, for $p = (0, \dots, 0)$, we recover the lower bound for the plain BAI setting without fairness constraints [14]. In this case, we refer to the characteristic constant in this setting as $T^\star \coloneqq T_0^\star$ and to the corresponding optimal allocations as $w^\star \coloneqq w_0^\star$. Additionally for any $p \neq (0, \dots, 0)$, we have $T_p^\star \geq T^\star$; hence, ensuring fairness yields increased sample complexity.

The case of unitary p: A notable set of cases involves fair BAI instances where $p_{\text{sum}} = 1$, including the important example of *proportional fairness* (see Remark III.2). In these cases, the optimal fair allocations can be simply expressed as $w_{p,a}^\star = p_a$, for all $a \in [K]$. This observation allows to derive a computationally efficient algorithm that avoids the optimization step over Σ_p, as detailed in Sec. V. This step is typically regarded as the main bottleneck on the computational complexity of BAI algorithms [14], especially when the number of arms K grows large.

B. The Price of Fairness

The next lemma states an upper bound on the ratio $T_p^\star/T^\star$. This ratio quantifies the price in sample complexity that the learner has to pay in order to guarantee fairness.

Lemma IV.2. *For a set of fairness constraints $p = (p_a)_{a\in[K]}$, and for all $\theta \in \Theta$, we have that*

$$1 \leq \frac{T_p^\star}{T^\star} \leq O\left(\min\left(\frac{1}{1-p_{\text{sum}}}, \frac{1}{Kp_{\min}}\right)\right). \tag{4}$$

The proof is reported in App. D. The lemma shows that the price typically scales either as $(1-p_{\text{sum}})^{-1}$ or $(p_{\min})^{-1}$. In the remainder of this section, we discuss two exemplary cases that shed light on the nature of this scaling. We also refer the reader to App. E for further details and examples.

Case 1, larger fairness rate for suboptimal arms: We consider a scenario where the fairness rates p for the sub-optimal arms are significantly larger than the optimal frequencies $w^\star$ prescribed in the unconstrained case.

Under the assumption that $p_{\text{sum}} < 1$, if $p_{a^\star} \leq w_{a^\star}^\star$ and $p_a \geq w_a^\star$ for all $a \neq a^\star$, using Lem. 1 (in App. B.2) we can derive the following upper bound on the ratio $T_p^\star/T^\star$:

$$\frac{T_p^\star}{T^\star} \leq \frac{\Delta_{\max}^2}{K\Delta_{\min}^2}\left(\frac{1}{1-p_{\text{sum}}} + \frac{1}{p_{\min}}\right). \tag{5}$$

See App. E for a detailed derivation. However, as explained in the following example, it is possible in some cases to characterize the behavior of $T_p^\star/T^\star$ by solely looking at one of the two terms $(1-p_{\text{sum}})^{-1}$ or $p_{\min}^{-1}$.

Example IV.3. We consider an *antagonistic* scenario with the following θ-dependent fairness rate: $p_a(\theta) = Kp_0\frac{\Delta_a}{\sum_b \Delta_b}$,

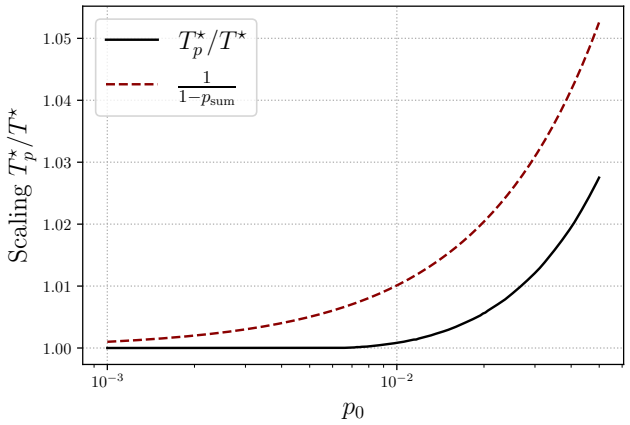

Fig. 1: **Price of fairness** for an instance with $K = 30$ arms, and higher fairness rates for sub-optimal arms (see Ex. IV.3). The price of fairness $T_p^\star/T^\star$ scales closely with $(1-p_{\text{sum}})^{-1}$

for $p_0 \in [0, 1/K]$. This scenario is termed *antagonistic* because the rates $p_a(\theta)$ are roughly proportional to Δ_a, and hence, larger for sub-optimal arms. To fix the ideas, consider a model where the rewards $\theta = (\theta_a)_{a\in[K]}$ are evenly distributed in $[0, 1]$. For small values of p_0 we have that $p_{\min}^{-1} > (1 - p_{\text{sum}})^{-1}$, with $(p_{\min})^{-1} = O(1/p_0)$. Despite the term $(p_{\min})^{-1}$ being large, we find it is not necessary to characterize $T_p^\star/T^\star$. For this specific model, as depicted in Fig. 1, the ratio $T_p^\star/T^\star$ aligns more closely with $(1-p_{\text{sum}})^{-1}$ for small values of p_0.

Case 2, the equal gap case: Another notable case involves instances with equal sub-optimality gaps, i.e., $\Delta_a = \Delta$, for all $a \neq a^\star$. In this case, we can characterize the ratio $T_p^\star/T^\star$ exactly. If $p_{a^\star} = 0$ and $p_a \geq w_a^\star$, for all $a \neq a^\star$, we have

$$\frac{T_p^\star}{T^\star} = \frac{1}{(1+\sqrt{K-1})^2}\left(\frac{1}{1-p_{\text{sum}}} + \frac{1}{p_{\min}}\right). \tag{6}$$

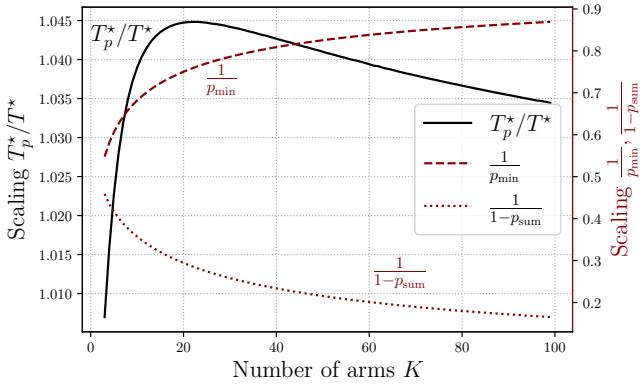

Fig. 2: **Price of fairness** for a MAB problem with equal gaps for different number of arms K. The fairness constraints are set to discourage exploration of the optimal arm by selecting $p_{a^\star} = 0$ and otherwise $p_a = (w_a^\star + 1/K)/2$ for $a \neq a^\star$. In black, on the left axis, it's depicted the ratio $T_p^\star/T^\star$. On the right axis, in dark-red, we plot the individual contributions due to $p_{\min}^{-1}$ and $(1 - p_{\text{sum}})^{-1}$.

We refer the reader to App. E for a detailed derivation.

To better understand the scaling of the two terms, we note that in the unconstrained case, the allocation $w_a^\star$ for all $a \neq a^\star$, decreases as $O(1/K)$. Hence, for a large value of K, if $p_a \approx w_a^\star$ for all $a \neq a^\star$, we may expect the term $p_{\min}^{-1}$ to be larger than $(1 - p_{\text{sum}})^{-1}$ (see Fig. 2 for an example of this case). Otherwise, the term $(1 - p_{\text{sum}})^{-1}$ is expected to be larger.

V. The F-BAI Algorithm

In this section, we propose F-BAI, an (asymptotically) p-fair and δ-PAC algorithm. The algorithm belongs to the family of Track-and-Stop (TaS) algorithms [14], which track the sampling allocations $w_p^\star$ as suggested by the solution to the lower bound optimization problem (3). The algorithm mainly consists of (i) a sampling rule and (ii) a stopping rule. We detail these steps in the remainder of this section, and present the pseudo-code for F-BAI in Alg. 1.

A. Sampling rule

The main idea of the algorithm is that sampling the arms according to $w_p^\star$ is automatically optimal in terms of sample complexity, and satisfies the fairness constraints. However, as the instance parameter θ is unknown, we leverage a *certainty-equivalence* principle and use the current estimate $\hat{\theta}(t) = (\hat{\theta}_a(t))_{a\in[K]}$ in place of the true parameter.

In the algorithm, we denote by $w_p^\star(t)$ the solution to Eq. (3) with $\hat{\theta}(t)$ plugged into the expression, i.e.,

$$w_p^\star(t) = \arg \inf_{w\in\Sigma_p} \max_{a\neq a_t^\star} \frac{w_a^{-1} + w_{a_t^\star}^{-1}}{\Delta_a(t)^2},$$

where $a_t^\star = \arg\max_a \hat{\theta}_a(t)$ and $\Delta_a(t) = \hat{\theta}_{a_t^\star}(t) - \hat{\theta}_a(t)$.

To enforce that the parametric uncertainty asymptotically goes to 0 (*i.e.*, $\hat{\theta}(t) \to \theta$ a.s.), we take a convex combination of $w_p^\star(t)$ with a constant policy $\pi_c = (\pi_{c,a})_{a\in[K]}$, using a parameter ϵ_t. This can be interpreted as a form of *forced exploration* [14] and guarantees that, asymptotically, each arm is sampled infinitely often.

Algorithm 1 F-BAI

Input: Fairness vector $p = (p_a)_{a\in[K]}$, confidence δ
Set $t \leftarrow 1$
while $Z(t) < \beta(\delta, t)$ **do**
  Compute $w_p^\star(t)$ and set $\pi(t) \leftarrow (1 - \epsilon_t)w_p^\star(t) + \epsilon_t \pi_c$
  Select $a_t \sim \pi(t)$ and observe reward r_t
  Update statistics $\hat{\theta}(t), N_a(t)$ and set $t \leftarrow t + 1$
end while
Return $\hat{a}_{\tau_\delta} = \arg\max_a \hat{\theta}_a(\tau_\delta)$

The constant policy π_c, and the value of ϵ_t depend on the type of fairness constraint as follows:

- *Pre-specified constraints*: Let $K_0 = |\{a \in [K] : p_a = 0\}|$ be the number of arms for which $p_a = 0$. In the simple case that $K_0 = 0$, we set $\pi_{c,a} = p_a + (1 -$

$p_{\text{sum}})/K$. Otherwise we set $\epsilon_t = 1/2\sqrt{t}$, and define π_c as

$$\pi_{c,a} = \begin{cases} p_a & p_a > 0 \\ \frac{1-p_{\text{sum}}}{K_0} & \text{otherwise.} \end{cases}$$

- *θ-dependent constraints*: in this case, we select $\pi_{c,a} = 1/K$, i.e., a uniform policy for all $a \in [K]$, and we set $\epsilon_t = 1/2\sqrt{t}$.

The choice of π_c is justified by the fact that, in the pre-specified setting, the fairness constraint naturally induces a linear exploration rate and hence we do not require any additional forced exploration. On the other hand, if the fairness constraints depend on θ, we leverage a uniform policy π_c to ensure that each arm is sufficiently explored.

Note that our tracking procedure is probabilistic (we sample an arm from $\pi(t)$) and differs from the deterministic versions commonly employed in classical Track-and-Stop algorithms [14]. Therefore, our approach, inspired by best policy identification techniques [33], requires different arguments in order to prove its optimality and fairness guarantee. See also app. G for a detailed discussion.

B. Stopping rule

The stopping rule is defined through two components: (1) a generalized-likelihood ratio test (GLRT) $Z(t)$ and (2) a threshold function $\beta(\delta, t)$. Following [14], the GLRT can be expressed as $Z(t) \coloneqq t/(2T_p^\star(t))$, where

$$T_p^\star(t) := \max_{a \neq a_t^\star} \frac{w_a(t)^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2},\ w_a(t) \coloneqq \frac{N_a(t)}{t}.$$

Next, we consider the following threshold function from [34]

$$\beta(\delta, t) = 3 \sum_{a \in [K]} \log(1 + \log(N_a(t))) + K\mathcal{C}_{exp}\left(\frac{\log(\frac{1}{\delta})}{K}\right),$$

where $\mathcal{C}_{exp}$ is a function defined in Thm. 7 in [34].

C. Sample Complexity and Fairness Guarantees

Fairness guarantees: We obtain the following guarantees on the fairness of our algorithm.

Proposition V.1. F-BAI *is p-fair δ-PAC (resp. asymptotically $p(\theta)$-fair δ-PAC). Furthermore, for pre-specified constraints,* F-BAI *satisfies the fairness constraints for all rounds, i.e.,* $\frac{\mathbb{E}[N_a(t)]}{t} \geq p_a, \forall t \geq 1,\ \forall a \in [K]$.

Sample complexity guarantees: Next, we establish that our algorithm achieves optimal sample complexity asymptotically (as $\delta \to 0$).

Theorem V.2. *For all $\delta \in (0, 1/2)$,* F-BAI *has a finite expected sample complexity $\mathbb{E}_\theta[\tau_\delta] < \infty$, and it satisfies:*
(1) Almost sure asymptotic optimality:

$$\mathbb{P}_\theta\left(\limsup_{\delta \to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\log(1/\delta)} \leq T_p^\star\right) = 1,$$

(2) Asymptotic optimality in expectation:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\log(1/\delta)} \leq T_p^\star.$$

See App. H and App. I for a detailed derivation of Prop. V.1 and Thm V.2.

VI. Numerical Results

In this section, we numerically evaluate the performance of F-BAI. We propose two sets of experiments: we apply F-BAI to a synthetic bandit instance (in Sec. VI-A), and to an industrial use-case from the radio communication domain: wireless scheduling (in Sec. VI-B). Additional results are reported in the appendix.

Fairness criteria: For both sets of experiments, we focus on two settings: *agonistic fairness* and *antagonistic fairness*. These terms relate to how the fairness parameter p impacts exploration. In the former setting, p promotes exploration (e.g., by aligning with the optimal allocation in the unconstrained setting $w^\star$), while in the latter, it inhibits exploration. We clarify these concepts below in the context of pre-specified and θ-dependent constraints.

(i) *Pre-specified constraints*: we select the fairness vector as $p_a = p_0[\alpha w_a^\star + (1-\alpha)\bar{w}_a^\star]$, where $p_0 \in (0,1), \alpha \in (0,1)$ and $\bar{w}_a^\star = (1/w_a^\star)/\sum_{b \in [K]}(1/w_b^\star)$. The parameter p_0 regulates the "amount of fairness" in the problem. We set $\alpha = 0.9$ for the *agonistic* case and $\alpha = 0.1$ in the *antagonistic* one. Note that in the latter case, p_a is almost inversely proportional to the optimal allocation $w^\star$ in the unconstrained case.

(ii) *θ-dependent constraints:* in the *agonistic case* we select the fairness functions as $p_a(\theta) = p_0 \frac{1/\max(\Delta_a, \Delta_{\min})}{\sum_{b \in [K]} 1/\max(\Delta_b, \Delta_{\min})}$, with $p_0 \in (0,1)$. In the *antagonistic case* we select $p_a(\theta) = p_0 \frac{\Delta_a}{\sum_{b \in [K]} \Delta_b}$. In these two cases, we see how the fairness rates are proportional, or inversely proportional, to the sub-optimality gaps $\Delta_a = \theta_{a^\star} - \theta_a$.

Fairness violation: We measure the *fairness violation* at time $t \leq \tau_\delta$ as $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)_+$, where and $(x)_+ = \max(x, 0)$. We also measure the expected fairness violation at the stopping time τ_δ as

$$\text{Fairness Violation} = \mathbb{E}_\theta[\rho(\tau_\delta)]. \tag{7}$$

This metric measures the average maximum amount of fairness violation at the stopping time τ_δ. For instance, a 5% violation, suggests that an arm has been sampled 5% less frequently than the rate prescribed by p_a (at most).

Baseline algorithms: We compare our F-BAI algorithm to Track-and-Stop (TAS) from [14], a baseline that does not consider any fairness constraint, and UNIFORM FAIR, an algorithm selecting an arm a in round t with probability $p_a(\hat{\theta}(t)) + (1 - p_{\text{sum}}(\hat{\theta}(t)))/K$. Hence, UNIFORM FAIR guarantees that $\mathbb{E}_\theta[N_a(t)]/t \geq p_a, \forall t \geq 1$, or $\lim_{t \to \infty} \mathbb{E}[N_a(t)]/t \geq p_a(\theta)$, for pre-specified or θ-dependent constraints, respectively. We test these algorithms by varying the values of $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. The results are averaged over $N = 100$ independent runs. All the confidence intervals refer to 95% confidence.

A. Synthetic Experiments

Model: For the pre-specified setting we consider a bandit model where the expected rewards $(\theta_a)_{a \in [K]}$ linearly range in $[0, K/2.5]$. We consider both *agonistic* and

	Algorithm	Pre-specified constraints				θ-dependent constraints			
		Sample Complexity		Fairness Violation		Sample Complexity		Fairness Violation	
		Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic
$\delta = 0.1$	F-BAI	258.63 ± 27.56	935.25 ± 129.53	$3.21\% \pm 1.30\%$	$2.11\% \pm 0.69\%$	404.89 ± 48.64	936.52 ± 130.58	$2.48\% \pm 0.46\%$	$1.64\% \pm 0.21\%$
	TAS	258.87 ± 38.29	258.87 ± 38.57	$7.50\% \pm 1.59\%$	$18.57\% \pm 0.33\%$	488.01 ± 73.29	488.01 ± 74.66	$4.48\% \pm 0.48\%$	$6.31\% \pm 0.08\%$
	UNIFORM FAIR	313.67 ± 28.86	1552.43 ± 240.79	$3.11\% \pm 1.16\%$	$1.03\% \pm 0.57\%$	673.91 ± 72.51	4905.09 ± 697.26	$1.44\% \pm 0.48\%$	$0.17\% \pm 0.23\%$
$\delta = 0.01$	F-BAI	390.72 ± 31.42	1522.11 ± 152.75	$1.51\% \pm 0.38\%$	$1.34\% \pm 0.17\%$	658.86 ± 64.51	1559.28 ± 192.29	$1.79\% \pm 0.29\%$	$1.20\% \pm 0.12\%$
	TAS	436.01 ± 49.84	436.01 ± 49.46	$4.05\% \pm 1.24\%$	$19.33\% \pm 0.13\%$	837.94 ± 96.03	837.94 ± 95.43	$4.78\% \pm 0.37\%$	$6.63\% \pm 0.08\%$
	UNIFORM FAIR	475.54 ± 33.65	2504.16 ± 289.01	$1.78\% \pm 0.34\%$	$0.31\% \pm 0.11\%$	1052.03 ± 112.42	7666.75 ± 948.81	$0.97\% \pm 0.45\%$	$0.03\% \pm 0.08\%$
$\delta = 0.001$	F-BAI	508.11 ± 32.49	2039.71 ± 199.70	$1.22\% \pm 0.32\%$	$1.12\% \pm 0.17\%$	891.82 ± 68.97	2053.25 ± 204.73	$1.44\% \pm 0.21\%$	$1.03\% \pm 0.10\%$
	TAS	628.32 ± 56.20	628.32 ± 57.14	$2.94\% \pm 0.91\%$	$19.80\% \pm 0.12\%$	1272.28 ± 123.52	1272.28 ± 119.12	$5.12\% \pm 0.34\%$	$6.97\% \pm 0.09\%$
	UNIFORM FAIR	604.36 ± 36.91	3416.73 ± 341.82	$1.43\% \pm 0.28\%$	$0.17\% \pm 0.09\%$	1433.40 ± 115.00	10740.08 ± 1062.28	$0.51\% \pm 0.28\%$	$0.02\% \pm 0.11\%$

TABLE I: Synthetic experiments: sample complexity and fairness violation for F-BAI, TAS, and UNIFORM FAIR. The fairness violation, as defined in Eq. (7), measures the average maximum extent of fairness deviation at the stopping time τ_δ.

antagonistic fairness rates and select $K = 10$, $p_0 = 0.9$. For θ-dependent constraints we consider an instance with $K = 15$ arms with rewards linearly ranging in $[0, 5]$, and $p_0 = 0.7$.

Results: In Tab. I we summarize the main results for this experiment for the sample complexity $\mathbb{E}_\theta[\tau_\delta]$ and fairness violation at the stopping time $\mathbb{E}_\theta[\rho(\tau_\delta)]$.

In terms of sample complexity, F-BAI shows similar performances to TAS in the agonistic setting. This is expected, since in this case the fairness constraints p are closely related to $w^\star$, and thus greatly favor exploration. At the same time F-BAI is able to guarantee a lower fairness violation, twice as low than TAS.

In case the fairness constraints are antagonistic, and thus do not favour exploration, we see how the sample complexity of F-BAI increases, while still maintaining a low fairness violation. In comparison, the sample complexity of UNIFORM FAIR is almost 50% as high, while having similar violations. For θ-dependent constraints the difference in sample complexity is even higher.

In Fig. 3 we show the distribution of maximum violation over all experimental runs. These results offer a comprehensive view of the algorithms' fairness throughout the duration of observation. Furthermore, the mean of these distributions effectively represents the average violation per round for each algorithm. From the results, we see that the behavior of F-BAI is close to that of FAIR UNIFORM, while TAS has larger violations overall.

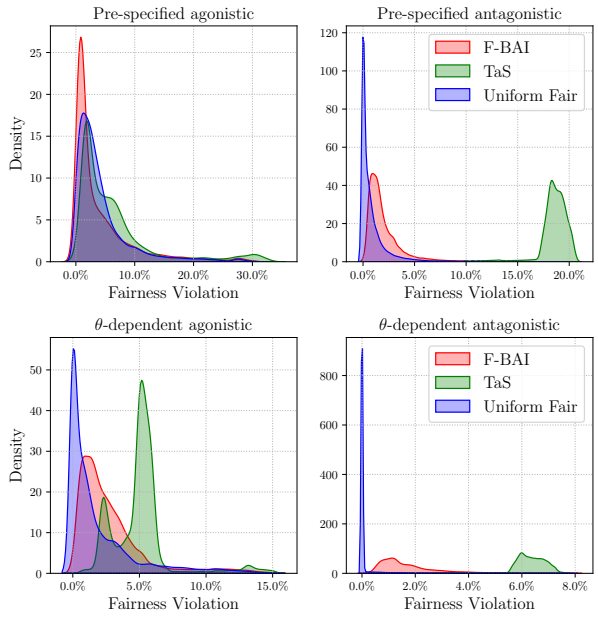

Fig. 3: Violations for the synthetic experiments with $\delta = 0.01$. Each subplot illustrates the distribution of maximum violation $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)_+$, across all rounds $t \leq \tau_\delta$ and experimental runs.

B. Wireless scheduling

Model: We consider a wireless radio environment with a Base Station (BS) and a set of K User Equipments (UEs) connected to the BS (see Fig. 4). Communication proceeds in time slots in a down-link fashion. The BS is placed at the center of a cell (or sector) of radius d [m] (measured in meters), and the K UEs are randomly distributed in the cell.

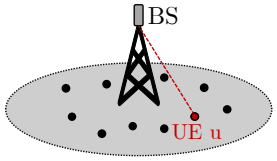

Fig. 4: Visual depiction of the scheduling environment with $K = 10$ UEs.

At each round, $t \geq 1$, the BS selects a single UE out of the K to be scheduled for transmission. Naturally, in this formulation, the BS represents the learner, and the set of UEs $[K]$ represents the various arms that can be selected by the BS at each round.

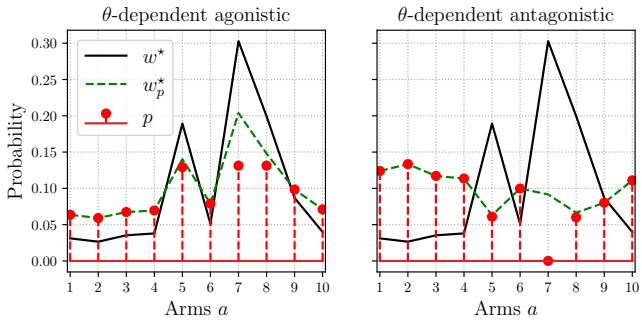

Fig. 5: Allocations for the scheduling experiments with θ-dependent constraints. Agonistic constraints favour exploration, since $w_p^\star \approx w^\star$, while antagonistic ones discourage exploration of good arms.

	Algorithm	Pre-specified constraints				θ-dependent constraints			
		Sample Complexity		Fairness Violation		Sample Complexity		Fairness Violation	
		Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic
$\delta = 0.1$	F-BAI	199.10 ± 15.96	457.90 ± 48.15	$3.03\% \pm 0.39\%$	$2.13\% \pm 0.24\%$	197.80 ± 17.05	599.79 ± 68.83	$4.60\% \pm 0.43\%$	$2.97\% \pm 0.32\%$
	TaS	136.88 ± 9.59	136.88 ± 9.78	$6.55\% \pm 0.68\%$	$10.76\% \pm 0.12\%$	136.88 ± 9.48	136.88 ± 9.86	$5.32\% \pm 0.36\%$	$8.22\% \pm 0.08\%$
	Uniform Fair	236.50 ± 16.11	726.52 ± 85.13	$2.45\% \pm 0.37\%$	$1.12\% \pm 0.25\%$	220.07 ± 18.00	1889.56 ± 287.37	$4.07\% \pm 0.35\%$	$1.94\% \pm 0.48\%$
$\delta = 0.01$	F-BAI	285.41 ± 15.74	696.11 ± 58.62	$2.35\% \pm 0.27\%$	$1.79\% \pm 0.20\%$	298.68 ± 21.88	833.55 ± 78.24	$3.96\% \pm 0.37\%$	$2.38\% \pm 0.23\%$
	TaS	207.79 ± 13.53	207.79 ± 13.64	$5.71\% \pm 0.67\%$	$11.14\% \pm 0.13\%$	207.79 ± 13.84	207.79 ± 13.28	$4.92\% \pm 0.37\%$	$8.55\% \pm 0.11\%$
	Uniform Fair	323.86 ± 19.23	1071.62 ± 91.97	$1.91\% \pm 0.29\%$	$0.68\% \pm 0.18\%$	359.49 ± 24.66	2853.99 ± 319.41	$3.00\% \pm 0.26\%$	$1.21\% \pm 0.40\%$
$\delta = 0.001$	F-BAI	358.81 ± 17.44	899.13 ± 74.28	$2.00\% \pm 0.29\%$	$1.60\% \pm 0.18\%$	398.94 ± 24.53	1048.52 ± 84.89	$3.43\% \pm 0.34\%$	$2.02\% \pm 0.18\%$
	TaS	271.05 ± 16.99	271.05 ± 16.87	$5.22\% \pm 0.62\%$	$11.51\% \pm 0.10\%$	271.05 ± 16.93	271.05 ± 17.11	$4.67\% \pm 0.33\%$	$8.90\% \pm 0.10\%$
	Uniform Fair	410.72 ± 22.63	1383.06 ± 95.08	$1.52\% \pm 0.21\%$	$0.41\% \pm 0.12\%$	476.13 ± 32.11	3703.97 ± 354.92	$2.58\% \pm 0.24\%$	$0.86\% \pm 0.37\%$

TABLE II: Sample complexity and fairness violations for the scheduling experiments.

Objective: The objective is to maximize the sum throughput across all UEs. The throughput $T_{u,t}$ of UE u at round t represents the rate at which information is delivered to the UE. This quantity depends on channel conditions (or *fading*) between the BS antenna and the user. These conditions rapidly evolve over time around their mean. The fadings between pairs of (antenna, user) are typically stochastically independent across users and antennas [35], and we assume that can be modeled as independent Gaussian r.v. The reward at round t is defined as the sum throughput across UEs in the cell, i.e., $r_t = \sum_{u \in [K]} T_{u,t} \mathbf{1}_{\{a_t = u\}}$.

Fairness constraints: In wireless scheduling, the fairness constraints represent the minimal fraction of rounds in which each UE is scheduled for transmission. This constraint naturally captures UE guarantees in terms of throughput: the higher the number of slots in which a UE is scheduled, the higher will be the throughput experienced.

Experimental setup: We test F-BAI using mob-env, an open-source simulation environment [36] based on the gymnasium interface. As for the synthetic setting, we consider two sets of experiments with *pre-specified* and *θ-dependent* fairness. The fairness parameter p and the optimal allocations $w^\star$ and $w_p^\star$ for the θ-dependent setting are shown in Fig. 5. We set the number of UEs to $K = 10$ and $p_0 = 0.9$. We refer the reader to the appendix for more details on the model and experimental setup.

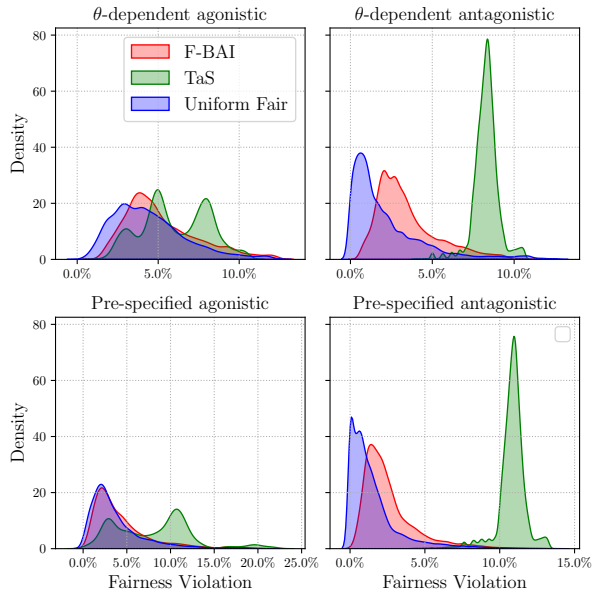

Fig. 6: Distribution of the fairness violations density for the scheduling experiments.

Results: The sample complexity and fairness violation results are presented in Tab. II, while Fig. 6 shows the distribution of the fairness violation metric. The results are generally in line with the experimental findings of the previous section: F-BAI achieves lower violation w.r.t. the non-fair baseline (TaS) while outperforming the fair baseline (Uniform-Fair) in terms of sample complexity.

VII. Conclusions

In this paper, we introduced Fair Best Arm Identification (Fair BAI), a novel setting that integrates the classical BAI framework with fairness constraints, which are either model-agnostic or model-dependent.

For both scenarios, we derived a sample complexity lower bound and quantified the price of fairness in terms of sample complexity. Leveraging this lower bound, we devised F-BAI, an algorithm that provably matches this bound while complying with the fairness constraints.

Our experimental results, obtained from both synthetic and wireless scheduling scenarios, demonstrate that F-BAI effectively achieves low sample complexity while minimizing fairness violations.

The limitations of our work include: (i) the asymptotic nature of our fairness constraints in the θ-dependent constraint; (ii) the sample complexity analysis operates in the asymptotic regime; (iii) quantifying the variance of our method is technically challenging. Future research directions involve extending the fair BAI concept to bandits with additional structures, such as linear, Lipschitz, and unimodal bandits. Furthermore, incorporating regret minimization into our framework represents another exciting area for exploration.

References

[1] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

[2] Pratik Gajane, Akrati Saxena, Maryam Tavakol, George Fletcher, and Mykola Pechenizkiy. Survey on fair reinforcement learning: Theory and practice. *arXiv preprint arXiv:2205.10032*, 2022.

[3] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 1985.

[4] Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Proc. of NeurIPS*, 2013.

[5] Kaito Ariu, Narae Ryu, Se-Young Yun, and Alexandre Proutière. Regret in online recommendation systems. In *Proc. of NeurIPS*, 2020.

[6] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proc. of AISTATS*, 2011.

[7] Thi Thuy Nga Nguyen, Urtzi Ayesta, and Balakrishna Prabhu. Scheduling users in drive-thru internet: a multi-armed bandit approach. In *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019.

[8] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *IEEE Congress on Evolutionary Computation (CEC)*, 2020.

[9] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 2019.

[10] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proc. of NeurIPS*, 2016.

[11] Vitshakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *JMLR*, 2021.

[12] Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 2021.

[13] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proc. of COLT*, 2010.

[14] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proc. of COLT*. PMLR, 2016.

[15] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *ICML*, 2017.

[16] Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil. Group meritocratic fairness in linear contextual bandits. In *Proc. of NeurIPS*, 2022.

[17] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[18] Qingsong Liu, Weihang Xu, Siwei Wang, and Zhixuan Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. In *Proc. of NeurIPS*, 2022.

[19] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

[20] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proc. of the conference on fairness, accountability, and transparency*, 2019.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. of the 3rd innovations in theoretical computer science conference*, 2012.

[22] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016.

[23] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[24] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Proc. of NeurIPS*, 2018.

[25] Tianyu Wang and Cynthia Rudin. Bandit learning for proportionally fair allocations. `https://wangt1anyu.github.io/papers/prop-fair-bandit.pdf`, 2021.

[26] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *ICML*, 2021.

[27] Anthony B Atkinson et al. On the measurement of inequality. *Journal of economic theory*, 1970.

[28] Bozidar Radunovic and Jean-Yves Le Boudec. A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Transactions on networking*, 2007.

[29] Tareq Si Salem, Georgios Iosifidis, and Giovanni Neglia. Enabling long-term fairness in dynamic resource allocation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2022.

[30] Mohammad Sadegh Talebi and Alexandre Proutiere. Learning proportionally fair allocations with low regret. *Proc. of the ACM on Measurement and Analysis of Computing Systems*, 2018.

[31] Yuhang Wu, Zeyu Zheng, and Tingyu Zhu. Best arm identification with fairness constraints on subpopulations. *arXiv preprint arXiv:2304.04091*, 2023.

[32] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. In *JMLR*, 2016.

[33] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. In *Proc. of NeurIPS*, 2021.

[34] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. In *JMLR*, 2021.

[35] David N. C. Tse and Pramod Viswanath. Fundamentals of wireless communication. *IEEE Trans. Inf. Theory*, 2009.

[36] Stefan Schneider, Stefan Werner, Ramin Khalili, Artur Hecker, and Holger Karl. mobile-env: An open platform for reinforcement learning in wireless mobile networks. In *NOMS IEEE/IFIP Network Operations and Management Symposium*, 2022.

[37] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[38] Alessio Russo and Alexandre Proutiere. On the sample complexity of representation learning in multi-task bandits with global and local structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[39] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[40] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*, 2019.

[41] Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan. Achieving user-side fairness in contextual bandits. *Human-Centric Intelligent Systems*, 2022.

[42] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proc. of NeurIPS*, 2017.

[43] Jackie Baek and Vivek Farias. Fair exploration via axiomatic bargaining. In *Proc. of NeurIPS*, 2021.

[44] Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. In *Proc. of NeurIPS*, 2021.

[45] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. Offline contextual bandits with high probability fairness guarantees. In *Proc. of NeurIPS*, 2019.

APPENDIX

CONTENTS

In this section we present additional numerical results. We briefly summarize some technical information regarding the experiments. Then, in App. A we present additional results on the synthetic model, with pre-specified constraints and θ-dependent constraints. Later, in App. B.1, we present additional experiments on the scheduling problem as well as present some technical details regarding the environment.

As previously mentioned, we tested all the algorithms for different values of $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. The results are averaged over $N = 100$ independent runs. All the confidence intervals refer to 95% confidence. Moreover, in our implementation, we tested both the exploration threshold in Sec. V-B $\beta(\delta, t) = 3\sum_{a\in[K]} \log(1+\log(N_a(t))) + K\mathcal{C}_{exp}\left(\frac{\log(\frac{1}{\delta})}{K}\right)$ [34], and $\beta(t,\delta) = \log((\log(t)+1)/\delta)$ introduced in [14]. We report the results using the latter threshold for simplicity. Lastly, the instructions to run the code can be found in the `README.md` file in the supplementary material.

Fairness criteria.: For all experiments, we focus on two settings: *agonistic fairness* and *antagonistic fairness*. These terms relate to how the fairness parameter p impacts exploration. In the former setting, p promotes exploration (e.g., by aligning with the optimal allocation in the unconstrained setting $w^\star$), while in the latter, it inhibits exploration. We clarify these concepts below in the context of pre-specified and θ-dependent rates.

(i) *Pre-specified constraints*: we select the fairness vector as $p_a = p_0[\alpha w_a^\star + (1-\alpha)\bar{w}_a^\star]$, where $p_0 \in (0,1), \alpha \in (0,1)$ and $\bar{w}_a^\star = (1/w_a^\star)/\sum_{b\in[K]}(1/w_b^\star)$. The parameter p_0 regulates the "amount of fairness" in the problem. We set $\alpha = 0.9$ for the *agonistic* case and $\alpha = 0.1$ in the *antagonistic* one. Note that in the latter case, p_a is almost inversely proportional to the optimal allocation in the unconstrained case.

(ii) *θ-dependent constraints:* in the *agonistic case* we select the fairness functions as $p_a(\theta) = p_0 \frac{1/\max(\Delta_a, \Delta_{\min})}{\sum_{b\in[K]} 1/\max(\Delta_b, \Delta_{\min})}$, with $p_0 \in (0,1)$. In the *antagonistic case* we select $p_a(\theta) = p_0 \frac{\Delta_a}{\sum_{b\in[K]} \Delta_b}$. In these two cases, we see how the fairness rates are proportional, or inversely proportional, to the sub-optimality gaps $\Delta_a = \theta_{a^\star} - \theta_a$.

Fairness violation.: We measure the *fairness violation* at time $t \le \tau_\delta$ as $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)_+$, where and $(x)_+ = \max(x, 0)$. We also measure the expected fairness violation at the stopping time τ_δ as

$$\text{Fairness Violation} = \mathbb{E}_\theta[\rho(\tau_\delta)].$$

A. Synthetic experiments

1) Numerical Results: Synthetic Model with Pre-specified Constraints:

Model.: We considered two bandit models. First, a model where the expected rewards $(\theta_a)_{a\in[K]}$ linearly range in $[0, K/2.5]$, with $K = 10$ and $p_0 = 0.9$. Secondly, a model where all the suboptimal gaps Δ_a have the same value $\Delta = K/5$, and we set $p_0 = 0.99$.

Allocations.: In Fig. 7 are depicted the optimal unconstrained allocation $w^\star$, the constrained one $w_p^\star$, and the fairness constraints $(p_a)_{a\in[K]}$. In the agonistic case we see how the fairness rates are closely related to the optimal exploration, while in the antagonistic one are inversely proportional.

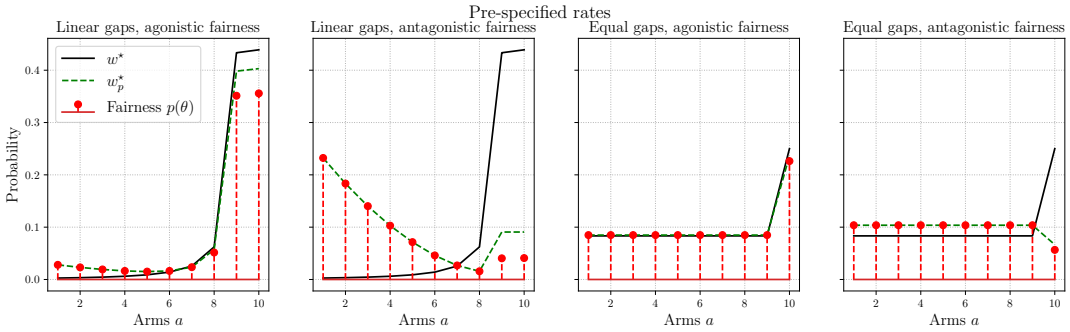

Fig. 7: Synthetic experiments with pre-specified constraints. We show the the optimal unconstrained allocation $w^\star$, the constrained one $w_p^\star$, and the fairness constraints $(p_a)_{a\in[K]}$ for both the model with rewards linearly ranging in $[0, K/2.5]$ and the model with equal-gaps.

Sample complexity.: In Fig. 8 we show the sample complexity results for each case, as well as the unconstrained sample complexity lower bound, and the constrained one.

Fairness violation.: In Fig. 9 we depict an aggregate distribution of the fairness violation $\rho(t)$ over all rounds. These plots offer a comprehensive understanding of the behavior of the algorithm.

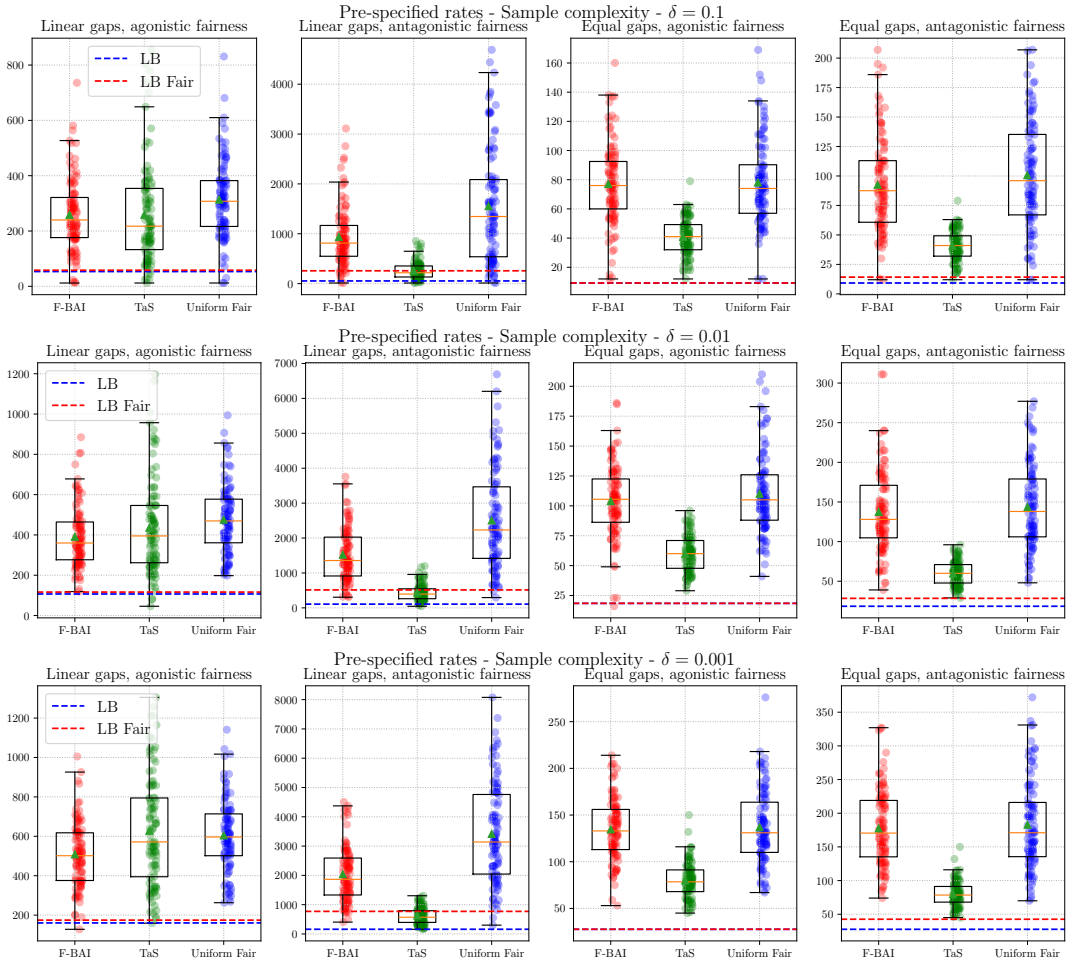

Fig. 8: Synthetic model with pre-specified constraints. Sample complexity results for different values of δ are shown in each row.

2) Numerical Results: Synthetic Model with θ-dependent Constraints:

Model.: We considered a single bandit model, with $K = 15$ arms and the reward linearly ranging in $[0, 5]$. We used a value of $p_0 = 0.7$ in the fairness constraints.

Allocations.: In Fig. 10 are depicted the optimal unconstrained allocation $w^\star$, the constrained one $w_p^\star$, and the fairness constraints $(p_a)_{a\in[K]}$. In the agonistic case we see how the fairness rates are closely related to the optimal exploration, while in the antagonistic one are inversely proportional.

Sample complexity.: In Fig. 11 we show the sample complexity results for each case, as well as the unconstrained sample complexity lower bound, and the constrained one.

Fairness violation.: In Fig. 12 we depict an aggregate distribution of the fairness violation $\rho(t)$ over all rounds. These plots offer a comprehensive understanding of the behavior of the algorithm.

B. Wireless Scheduling

This appendix is organized as follows. In App. B.1 we report additional experimental results on the wireless scheduling use-case and in App. B.2 we present more details on the experimental setup.

1) Wireless Scheduling: Additional Numerical Results: In this appendix we provide additional numerical results on the wireless scheduling experiments. More precisely we report extended results on (i) *optimal allocations*, (ii) *sample complexity*, and (iii) fairness violations in all the experimental setup described at the beginning of this appendix.

Allocations.: In Fig. 13 are depicted (i) the optimal unconstrained allocation $w^\star$, (ii) the optimal fair allocations $w_p^\star$, and the fairness constraints $(p_a)_{a\in[K]}$. In the agonistic case, we see how the fairness rates are closely related to the optimal exploration, while in the antagonistic one are inversely proportional.

Sample complexity.: In Fig. 14 we show the boxplots for the sample complexity results. The points in each figure shows the realization of the sample complexity for each run.

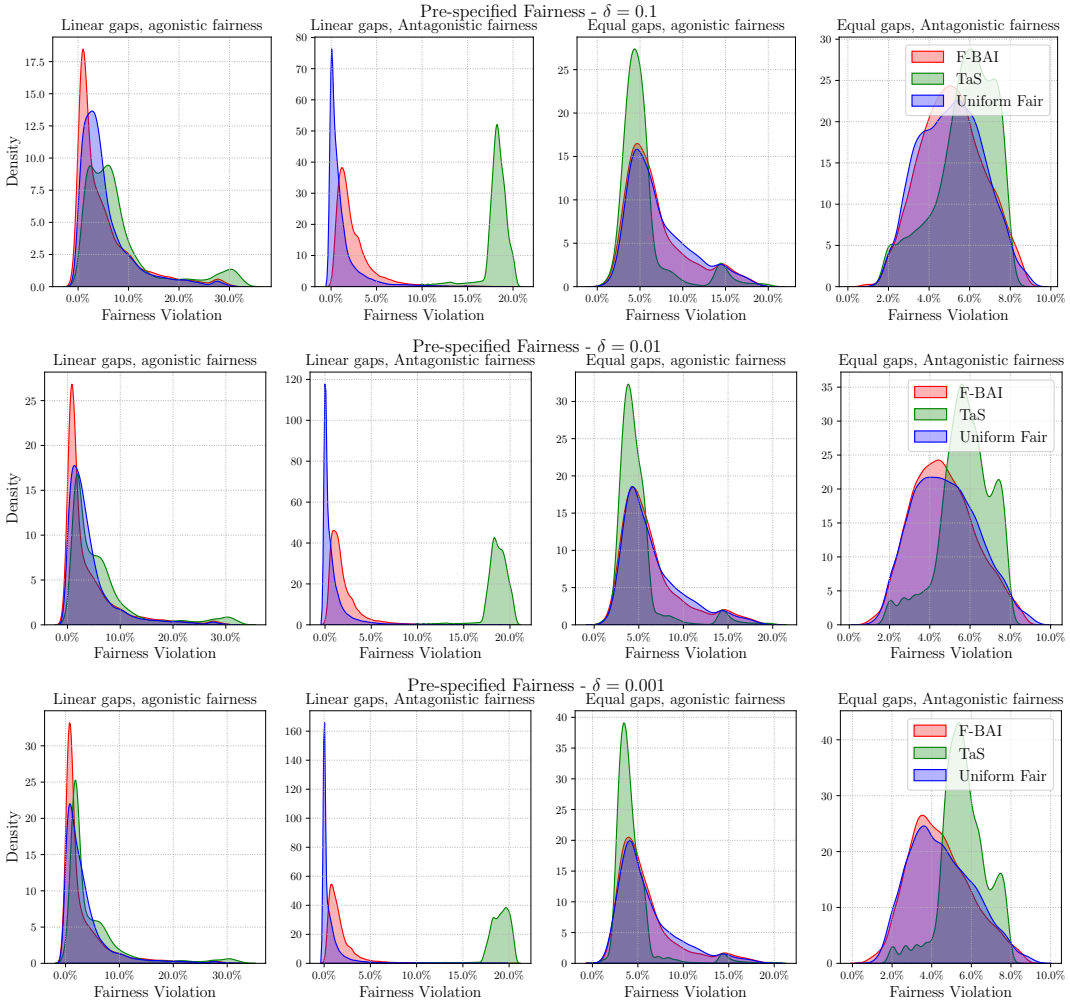

Fig. 9: Violations for the synthetic experiments with pre-specified constraints for different values of δ in each row. Each subplot illustrates the distribution of maximum violation $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)+$, across all rounds $t \leq \tau_\delta$ and experimental runs.

Fairness violation.: In Fig. 15 we depict an aggregate distribution of the fairness violation $\rho(t)$ over all rounds. These plots offer a comprehensive understanding of the behavior of the algorithm.

2) Wireless Scheduling: Detailed Experimental Setting: In this appendix, we present additional details on the scheduling experiments.

Simulator. We test our F-BAI algorithm using mob-env, an open-source simulation environment [36] based on the gymnasium interface. The simulator environment consist of a mobile network with a set of K UEs and a BS. The BS is equipped with an antenna placed at a high of h m. The antenna operates at a carrier frequency of f MHz and the channel bandwidth is set at W Hz. The size of the sector considered in the experiments is set at L m^2. We report the configuration used in our experiments in Tab. III.

TABLE III: Simulator parameters.

PARAMETER	SYMBOL	VALUE
Number of UEs	K	10
Bandwidth	W	9 MHz
Carrier frequency	f	2500 MHz
Antenna height	h	50 m
Sector size	L	0.4 km^2

Once the user positions and network parameters are provided, the simulator computes the path loss in the urban environment

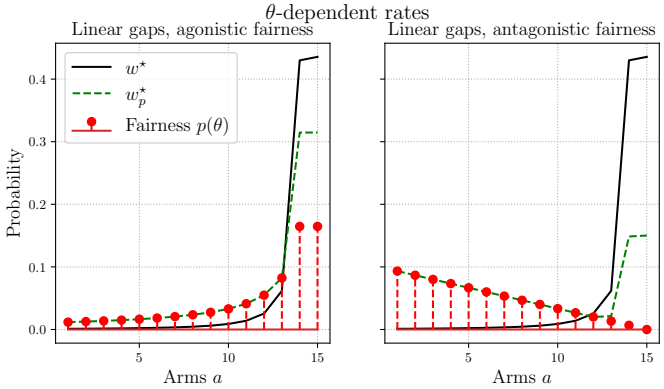

Fig. 10: Synthetic experiments with θ-dependent constraints. We show the the optimal unconstrained allocation $w^\star$, the constrained one $w_p^\star$, and the fairness constraints $(p_a(\theta))_{a\in[K]}$.

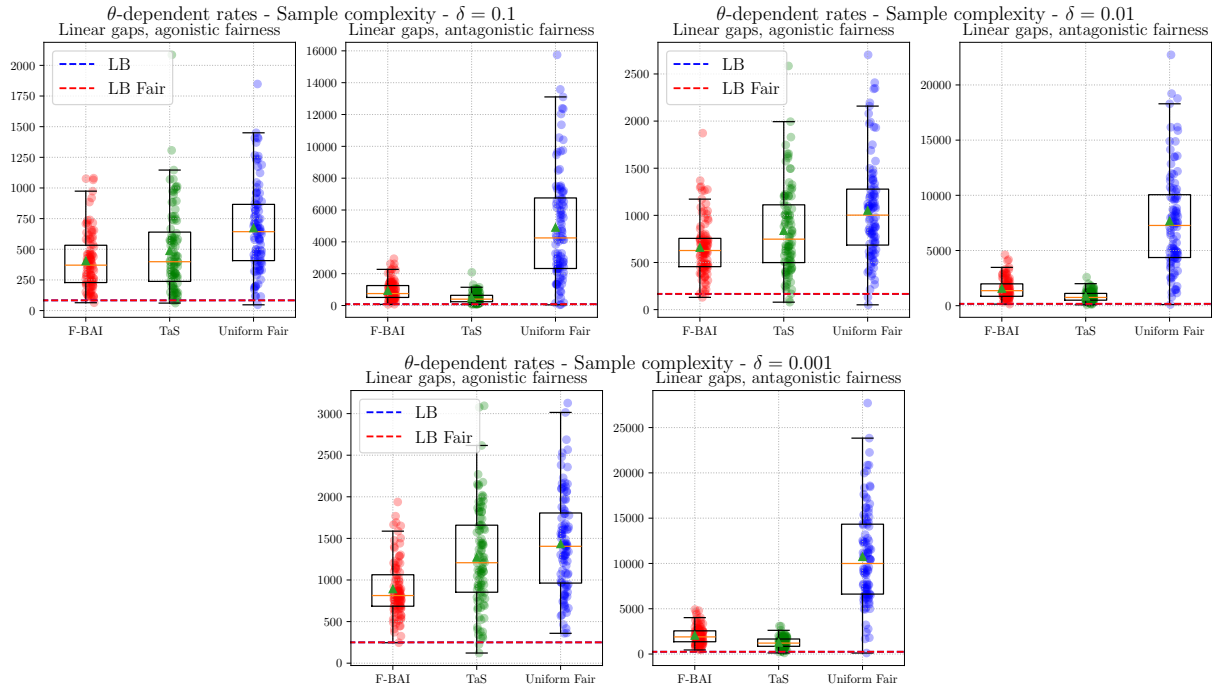

Fig. 11: Synthetic model with θ-dependent constraints. Sample complexity results for different values of δ are shown in each row.

using the Okomura-Hata propagation model [22], and computes a set of performance indicators. In our experiments we base the definition of our reward function on the sum-throughput, an important metric detailed in the following.

Throughput. The throughput $T_{u,t}$ of a UE $u \in \mathcal{U}$ at round $t \geq 1$, is formally defined in terms of the Signal-to-Noise Ratio (SNR), a metric that measures the quality of a signal in the presence of noise. Specifically, denote the SNR of a UE $u \in \mathcal{U}$ at time $t \geq 1$ is defined as $\text{SNR}_{u,t} = \frac{P_{\text{TX}}}{P_{\text{S}}}$, where P_{S} and P_{N} are the transmitted signal and noise power, respectively. Then the throughput (or rate) is expressed as

$$T_{u,t} = W \log_2(1 + \text{SNR}),$$

where W is the channel bandwidth [Hz].

Additional details. Although different works in the literature assume that a single UE can be scheduled at each time [7], we mention that more complex formulations allow for the BS to schedule a subset of the UEs at each round (see e.g., [9]). This extension yields an interesting combinatorial structure in the action selection. However, analyzing our fair bandit framework for such combinatorial bandit structures is out of the scope of this paper and is left as a future work.

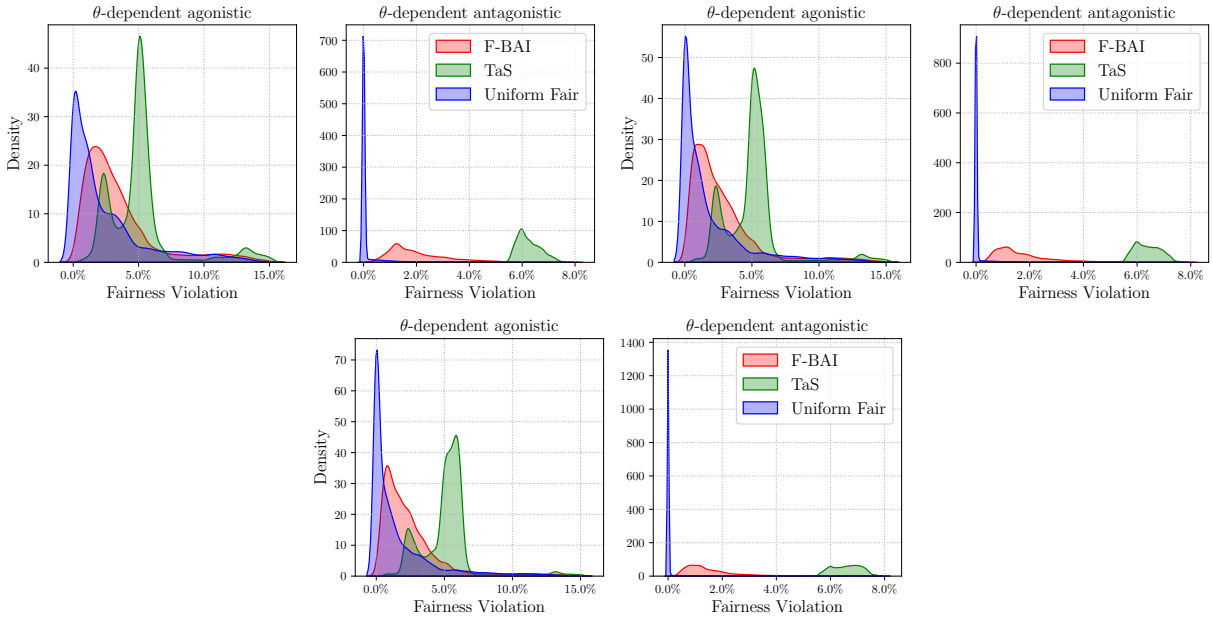

Fig. 12: Violations for the synthetic experiments with θ-dependent constraints for different values of δ in each row. Each subplot illustrates the distribution of maximum violation $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)+$, across all rounds $t \leq \tau_\delta$ and experimental runs.

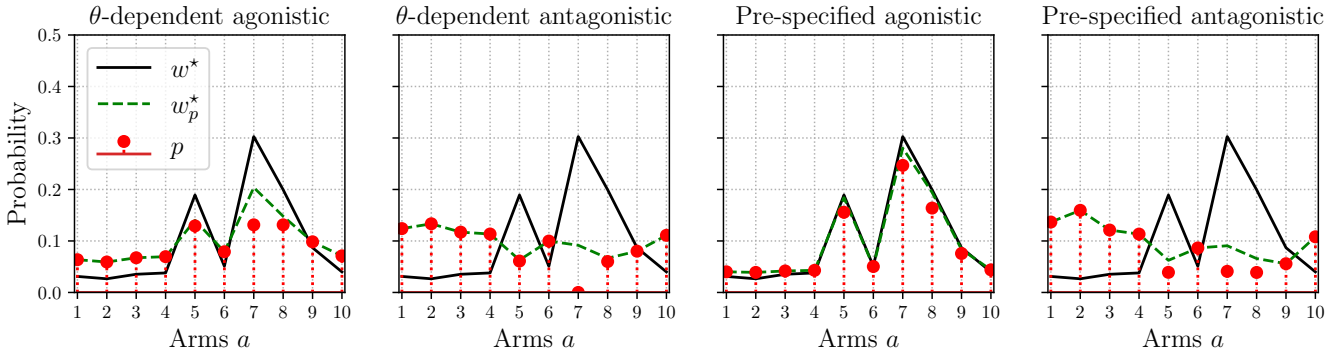

Fig. 13: Wireless scheduling experiments: optimal unconstrained allocation $w^\star$, optimal fair allocations $w_p^\star$, and fairness rates $(p_a)_{a\in[K]}$ for both pre-specified θ-dependent constraints in the agonistic and antagonistic setting.

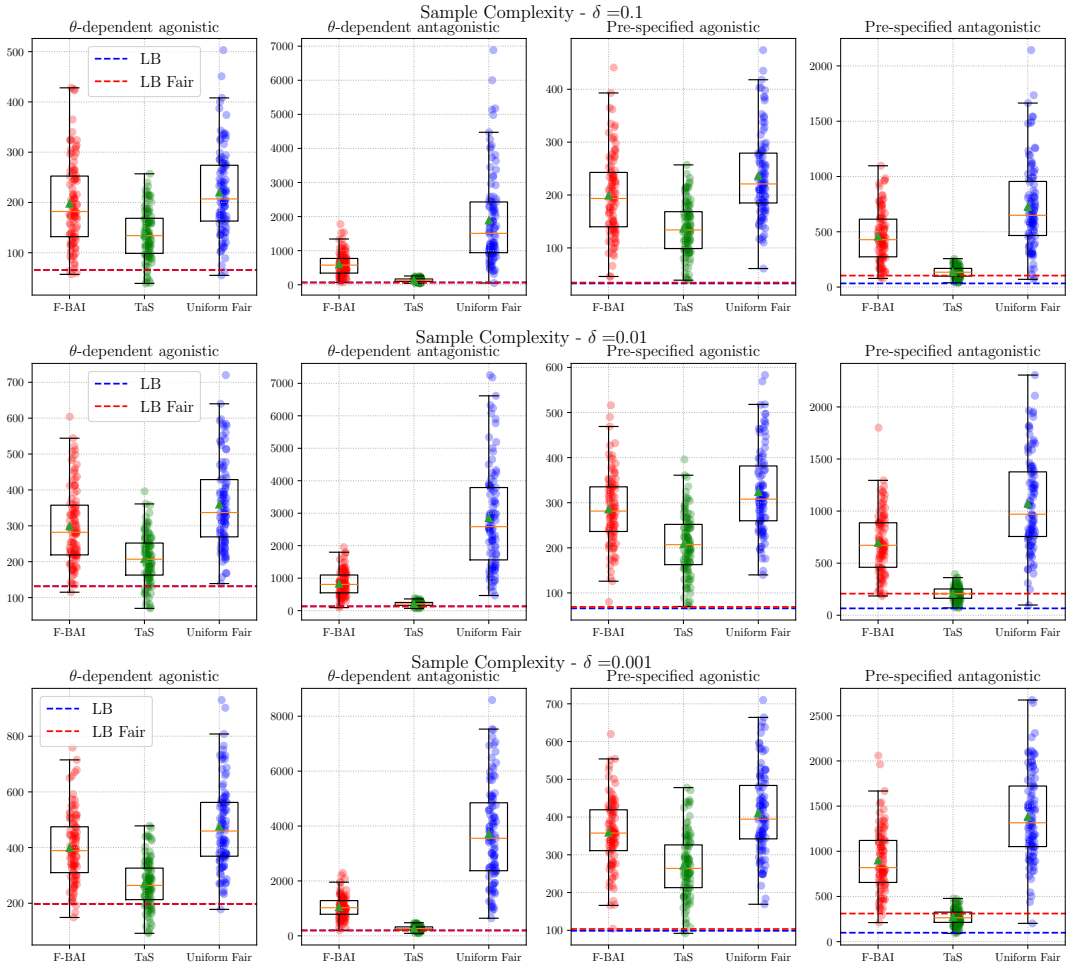

Fig. 14: Wireless scheduling experiments: sample complexity results for different values of δ.

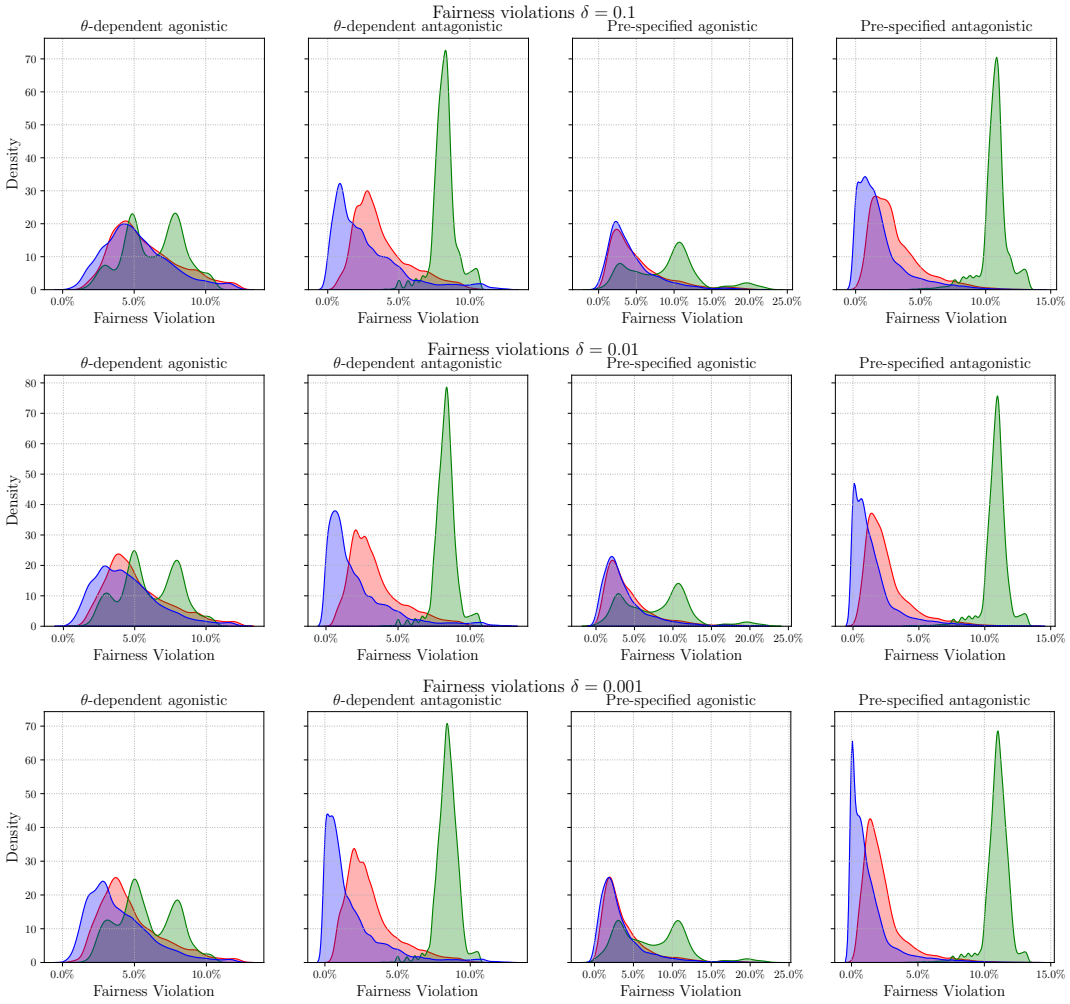

Fig. 15: Violations for the scheduling experiments different values of δ in each row. Each subplot illustrates the distribution of maximum violation $\rho(t) = (\max_a p_a(\theta) - N_a(t)/t)+$, across all rounds $t \leq \tau_\delta$ and experimental runs.

In this appendix, we prove the sample complexity lower bound (Theorem IV.1), in App. C, and the upper bound on the price of fairness $T_p^\star/T^\star$ (Lemma IV.2), in App. D. We also discuss the price of fairness for various specific bandit instances (App. E and App. F) and provide an additional result on the characterization of the optimal allocations in fair BAI (App. G).

C. Lower Bound: proof of Theorem IV.1

The proof is a straightforward extension of the one in the plain bandit setting in [14]. The main difference is that, due to the p-fair δ-PAC definition, the allocations must satisfy the conditions $w_a \geq p_a(\theta)$ for all $a \in [K]$. We sketch the main steps in the following.

Proof:

Consider a p-fair δ-PAC algorithm $(a_t, \tau, \hat{a}_{\tau_\delta})$, and define the set of confusing parameters $B(\theta) = \{\lambda \in \Theta : a_\theta^\star \neq a_\lambda^\star\}$, where $a_\theta^\star = \arg\max_a \theta_a$. Let $\mathcal{E}_\theta = \{\hat{a}_{\tau_\delta} = a_\theta^\star\}$, and note that for all θ,

$$\mathbb{P}_\theta(\mathcal{E}_\theta) \geq 1 - \delta,$$

while for all $\lambda \in B(\theta)$ we have

$$\mathbb{P}_\lambda(\mathcal{E}_\theta) \leq \delta.$$

By Lemma 1 [14], for any a.s. finite stopping time τ_δ, we have that

$$\sum_{a \in [K]} \mathbb{E}_\theta[N_a(\tau_\delta)] \frac{(\theta(a) - \lambda(a))^2}{2} \geq \mathrm{kl}(1-\delta, \delta).$$

By letting $w_a = \frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{E_\theta[\tau_\delta]}$, we can rewrite the previous equation as

$$\mathbb{E}_\theta[\tau_\delta] \sum_{a \in [K]} w_a \frac{(\theta(a) - \lambda(a))^2}{2} \geq \mathrm{kl}(1-\delta, \delta).$$

By optimizing over the set of confusing parameters we get

$$\mathbb{E}_\theta[\tau_\delta] \min_{a \neq a_\theta^\star} \frac{\Delta(a)^2}{2(w_{a_\theta^\star}^{-1} + w_a^{-1})} \geq \mathrm{kl}(1-\delta, \delta).$$

Pre-specified fairness. As the lower bound holds for any p-fair δ-PAC algorithm, we must have that (1) holds, and hence $\mathbb{E}[N_a(\tau_\delta)]/\mathbb{E}[\tau_\delta] \geq p_a$. The result is finally obtained by optimizing the allocations over the set of clipped allocations $\Sigma_p = \{w \geq p : \sum_{a \in \mathcal{A}} w_a = 1\}$.

θ-dependent fairness. As the lower bound holds for any asymptotically $p(\theta)$-fair δ-PAC algorithm we must have that $\liminf_{\delta \to 0} \mathbb{E}[N_a(\tau_\delta)]/\mathbb{E}[\tau_\delta] \geq p_a(\theta)$. Hence, the result is finally obtained by optimizing the allocations over the set $\Sigma_p = \{w \geq p(\theta) : \sum_a w_a = 1\}$ and letting $\delta \to 0$.

D. Price of Fairness: proof of Lemma IV.2

Proof: Consider the following feasible allocation $\tilde{w} \in \Sigma_p$: $\tilde{w}_a = p_a + \frac{1 - p_{\mathrm{sum}}}{K}$, where $p_a = p_a(\theta)$ and $p_{\mathrm{sum}} = p_{\mathrm{sum}}(\theta)$ for the sake of simplicity.

We can write

$$T_p^\star = \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2} \leq \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_{\min}^2} \leq \max_{a \neq a^\star} \frac{\tilde{w}_a^{-1} + \tilde{w}_{a^\star}^{-1}}{\Delta_{\min}^2} = \frac{1}{\Delta_{\min}^2}(\tilde{w}_{a^\star}^{-1} + \max_{a \neq a^\star} \tilde{w}_a^{-1}).$$

Let $p_{\min} = \min_{a \in [K] : p_a > 0} p_a$, then

$$\tilde{w}_{a^\star}^{-1} + \max_{a \neq a^\star} \tilde{w}_a^{-1} = \frac{K}{K p_{a^\star} + (1 - p_{\mathrm{sum}})} + \frac{K}{K p_{\min} + (1 - p_{\mathrm{sum}})} \leq \frac{2K}{K p_{\min} + (1 - p_{\mathrm{sum}})}$$

On the other hand, by App. A.4 in [14], we have that

$$T^\star \geq \sum_{a \in [K]} \frac{1}{\Delta_a^2} \geq \frac{K}{\Delta_{\max}^2}.$$

Hence, we find

$$\frac{T_p^\star}{T^\star} \leq \frac{\Delta_{\max}^2}{\Delta_{\min}^2} \frac{2}{K p_{\min} + 1 - p_{\mathrm{sum}}}.$$

Now, we consider two separate cases. If $K p_{\min} \geq 1 - p_{\mathrm{sum}}$ we get

$$\frac{2}{K p_{\min} + 1 - p_{\mathrm{sum}}} \leq \frac{1}{1 - p_{\mathrm{sum}}},$$

Otherwise, if $1 - p_{\text{sum}} > K p_{\min}$ we have

$$\frac{2}{K p_{\min} + 1 - p_{\text{sum}}} < \frac{1}{K p_{\min}},$$

and hence we can conclude

$$\frac{T_p^\star}{T^\star} \leq \frac{\Delta_{\max}^2}{\Delta_{\min}^2} \min\left(\frac{1}{1-p_{\text{sum}}}, \frac{1}{K p_{\min}}\right). \tag{8}$$

E. Price of Fairness: examples and specific instances

For specific bandit instances, we can obtain tighter bounds. We consider the following cases.

1. The 2-armed bandits case.: In the 2-armed bandit case it is known that the optimal allocation in the non-fair setting satisfies $w^\star = (1/2, 1/2)$ [14]. In the fair BAI setting, we have that $w_p^\star \neq w^\star$, when either p_1 or p_2 are greater than $1/2$ (naturally the case where both p_1 and p_2 are greater than $1/2$ is unfeasible). Without loss of generality let $p_1 > 1/2$. Then the optimal allocations satisfy $w_p^\star = (p_1, 1-p_1)$, and we have that

$$\frac{T_p^\star}{T^\star} = \frac{1}{4p_1(1-p_1)}. \tag{9}$$

This ratio quantifies the price of fairness in MAB and naturally, it is minimized for $p_1 = 1/2$. Note that when particularizing our bound in (8) for $K = 2$, with $p_1 > 1/2$, we have

$$\frac{T_p^\star}{T^\star} \leq \frac{2}{(1-p_1)}.$$

2. The case of unitary p.: Note that if p is such that $\sum_{a\in[K]} p_a = 1$, the optimal solution satisfies $w_{p,a}^\star = p_a$, for all $a \in [K]$, and hence

$$T_p^\star = \max_{a \neq a^\star} \frac{p_a^{-1} + p_{a^\star}^{-1}}{\Delta_a^2}.$$

In this case, we have $T_p^\star \leq \frac{2}{p_{\min}\Delta_{\max}^2}$, and hence

$$\frac{T_p^\star}{T^\star} \leq \frac{1}{K p_{\min}} \left(\frac{\Delta_{\max}}{\Delta_{\min}}\right)^2, \tag{10}$$

where $p_{\min} = \min_{a:p_a>0} p_a$.

3. The case $p_{a^\star} = 0$ and $p_a \geq w_a^\star$ for $a \neq a^\star$.: This scenarios is important since it displays both dependencies on p_{sum} and $p_{\min} = \min_{a:p_a>0} p_a$.

Assuming $p_{\text{sum}} < 1$, if $p_{a^\star} = 0$, and $p_a \geq w_a^\star$ for $a \neq a^\star$, by Lem. .1 we immediately have that $w_{p,a^\star}^\star = 1 - p_{\text{sum}}$ and $w_{p,a}^\star = p_a$ for $a \neq a^\star$.

We immediately conclude that $T_p^\star \leq \frac{1}{\Delta_{\min}^2}(\frac{1}{1-p_{\text{sum}}} + \frac{1}{p_{\min}})$, and thus

$$\frac{T_p^\star}{T^\star} \leq \frac{1}{K} \frac{\Delta_{\max}^2}{\Delta_{\min}^2} \left(\frac{1}{1-p_{\text{sum}}} + \frac{1}{p_{\min}}\right). \tag{11}$$

We now discuss the example in Case 1 of §VI-A in more details. We consider the case where the values of $\theta = (\theta_a)_{a\in[K]}$ linearly range in $[0,1]$. In this case the minimum gap scales as $1/(K-1)$ and $\sum_a \Delta_a = K/2$, henceforth $p_{\min} = K p_0 \frac{1/(K-1)}{K/2} = \frac{2p_0}{K-1}$, while $p_{\text{sum}} = K p_0$. Therefore $(1-p_{\text{sum}})^{-1}$ is the leading term for $p_0 \to 1/K$ (see also Fig. 1), while $p_{\min}^{-1}$ is decreasing in p_0.

4. The case $p_{a^\star} = 0$ and $p_a \geq w_a^\star$ for $a \neq a^\star$ with equal gaps.: The previous scenario can be extended to the case of equal gaps, i.e., $\Delta_a = \Delta$ for all $a \neq a^\star$. In this particular case, we are able to exactly compute $T_p^\star/T^\star$. In fact, the solution to the lower bound problem $T^\star$ satisfies

$$w_a^\star = \begin{cases} \dfrac{\sqrt{K-1}}{\sqrt{K-1}+K-1} & a = a^\star, \\ \dfrac{1}{\sqrt{K-1}+K-1} & a \neq a^\star. \end{cases} \tag{12}$$

Therefore

$$\begin{aligned} T^\star &= \frac{1}{\Delta^2} \max_{a\neq a^\star} \frac{1}{w_a^\star} + \frac{1}{w_{a^\star}^\star} = \frac{\sqrt{K-1}+K-1}{\Delta^2} \cdot \frac{1+\sqrt{K-1}}{\sqrt{K-1}}, \\ &= \frac{1+\sqrt{K-1}}{\Delta^2}(1+\sqrt{K-1}), \\ &= \frac{(1+\sqrt{K-1})^2}{\Delta^2}. \end{aligned}$$

Hence, using the fact that $T_p^\star = \frac{1}{\Delta^2}(\frac{1}{1-p_{\text{sum}}} + \frac{1}{p_{\min}})$, the ratio $T_p^\star/T^\star$ becomes

$$\frac{T_p^\star}{T^\star} = \frac{1 - p_{\text{sum}} + p_{\min}}{p_{\min}(1-p_{\text{sum}})(1+\sqrt{K-1})^2}. \tag{13}$$

Lastly, we also recover the form of the previous upper bound $\frac{T_p^\star}{T^\star} \leq \frac{1}{K-1}\left(\frac{1}{p_{\min}} + \frac{1}{1-p_{\text{sum}}}\right)$.

5. The case $p_{a^\star} \geq w_{a^\star}^\star$ and $p_a \leq (1-p_{a^\star})/(K-1)$ for $a \neq a^\star$ with equal gaps.: For the equal gap case, we can also study the scenario where p_a, for $a \neq a^\star$, is significantly small, and $p_{a^\star} \geq w_{a^\star}^\star$.

Similarly as before, if $p_a \in [0, \frac{1-p_{a^\star}}{K-1}]$, then we have $w_{p,a^\star}^\star = p_{a^\star}$ and $w_{p,a}^\star = c$ with $(K-1)c + p_{a^\star} = 1 \Rightarrow w_{p,a}^\star = (1-p_{a^\star})/(K-1)$ for all $a \neq a^\star$. Therefore $T_p^\star = \frac{1}{\Delta^2}(\frac{K-1}{1-p_{a^\star}} + \frac{1}{p_{a^\star}})$ and the ratio $T_p^\star/T^\star$ becomes

$$\frac{T_p^\star}{T^\star} = \frac{p_{a^\star}(K-2)+1}{p_{a^\star}(1-p_{a^\star})(1+\sqrt{K-1})^2}. \tag{14}$$

Note that in this case p_{sum} and $p_{\min}$ are equivalent. Furthermore, in this case the price of fairness is due to the additional sampling of the optimal arm at the cost of not sampling enough all the other arms.

F. Additional results on Price of Fairness

In this appendix, we provide additional numerical results with the goal of quantifying numerically the price of fairness and the tightness of our bound on $T_p^\star/T^\star$ presented in IV.2.

1) The two regimes in the price of fairness: In this section, we illustrate the existence of two regimes of our upper bound on $T_p^\star/T^\star$. Indeed, each of the terms appearing in the bound ($1/Kp_{\min}$ and $1/(1-p_{\text{sum}})$) is tighter in different scenarios.

To illustrate this phenomenon, we consider a set of instances where the sub-optimality gaps are fixed, i.e., $\Delta_a = \Delta$, for all $a \neq a^\star$. The fairness parameters are selected as $p = \lambda w^\star$, where $w^\star$ are the optimal allocations in the plain bandit setting (without fairness constraints), i.e., $w^\star$ are the allocations optimizing $T^\star$, and $\lambda \in [0,1]$ is a scaling parameter. Fig. 16 shows a plot of the value of $T_p^\star/T^\star$, and the terms appearing in our bound, i.e., $1/Kp_{\min}$ and $1/(1-p_{\text{sum}})$ when varying the parameter λ.

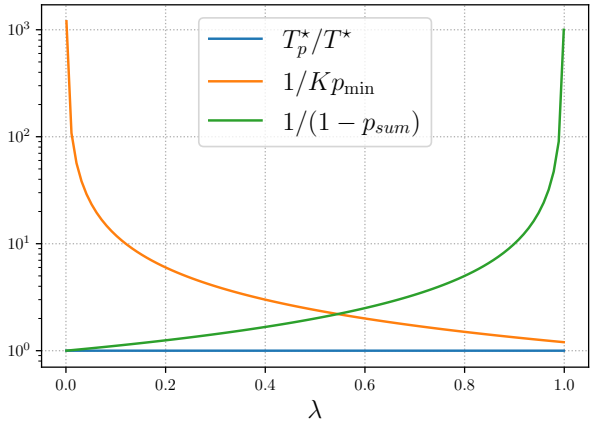

Fig. 16: Illustration of the two regimes in the upper bound on $T_p^\star/T^\star$.

Note that $T_p^\star/T^\star = 1$ as for all λ, we have $\lambda w_a^\star \leq w_a^\star$, which implies $w_a^\star = w_{p,a}^\star$. As it can be observed from the figure, generally we have two regimes: for low values of λ the term $(1-p_{\text{sum}})^{-1}$ is tighter (as $p_{\min}$ will also have low value); as λ increases, the term $(1-p_{\text{sum}})^{-1}$ (as p_{sum} approaches 1) and the term $1/Kp_{\min}$ provides a tighter bound. Our bound $T_p^\star/T^\star \leq O(\min\{(1-p_{\text{sum}})^{-1}, (Kp_{\min})^{-1}\})$ captures both these scenarios and provides a tighter bound.

2) Price of Fairness: scaling of $(1-p_{\text{sum}})^{-1}$-dependent Constraints: In this subsection we investigate the scaling of $T_p^\star/T^\star$, and how it is related to $1/(1-p_{\text{sum}})$. We study 4 different cases, each for a different number of arms $K \in \{10, 20, 30\}$.

1) In the first setting, we study a model where the rewards linearly ranging in $[0,1]$. The fairness constraints are constant, set to some value $p \in (0, 1/K)$ for all arms $a \in [K]$. Results are depicted in Fig. 17.
2) In the second setting, we study a model where all the arms have equal gap $\Delta_a = 0.1$. The fairness constraints are constant, set to some value $p \in (0, 1/K)$ for all arms $a \in [K]$. Results are depicted in Fig. 17.
3) In the third setting we study a model with rewards linearly ranging in $[0,1]$. The fairness constraints are θ-dependent. In particular, we compute it as follows

$$p_a(\theta) = pK\frac{f_a(\theta)}{\sum_b f_b(\theta)}, \text{ where } f_a(\theta) = 1 + \frac{1}{\Delta_a}, \tag{15}$$

with $f_{a^\star} = 1$. This makes the rates inversely proportional to the gap. However, since $f_{a^\star} = 1$, the optimal arm will have a very low fairness constraint, resulting in the algorithm sampling "good" arms at a very large rate. Results are depicted in Fig. 18.

4) The last setting uses the same fairness constraint as the previous one, but the arms now have the same gap $\Delta_a = 0.1$, $\forall a \in [K]$. Results are depicted in Fig. 18.

In all cases, we evaluate $T_p^\star/T^\star$ over different values of $p \in (0, 1/K)$. From the results, we see a clear scaling in $(1-p_{\text{sum}})^{-1}$ when the rewards are linearly ranging, while the equal gap case is less affected by this parameter.

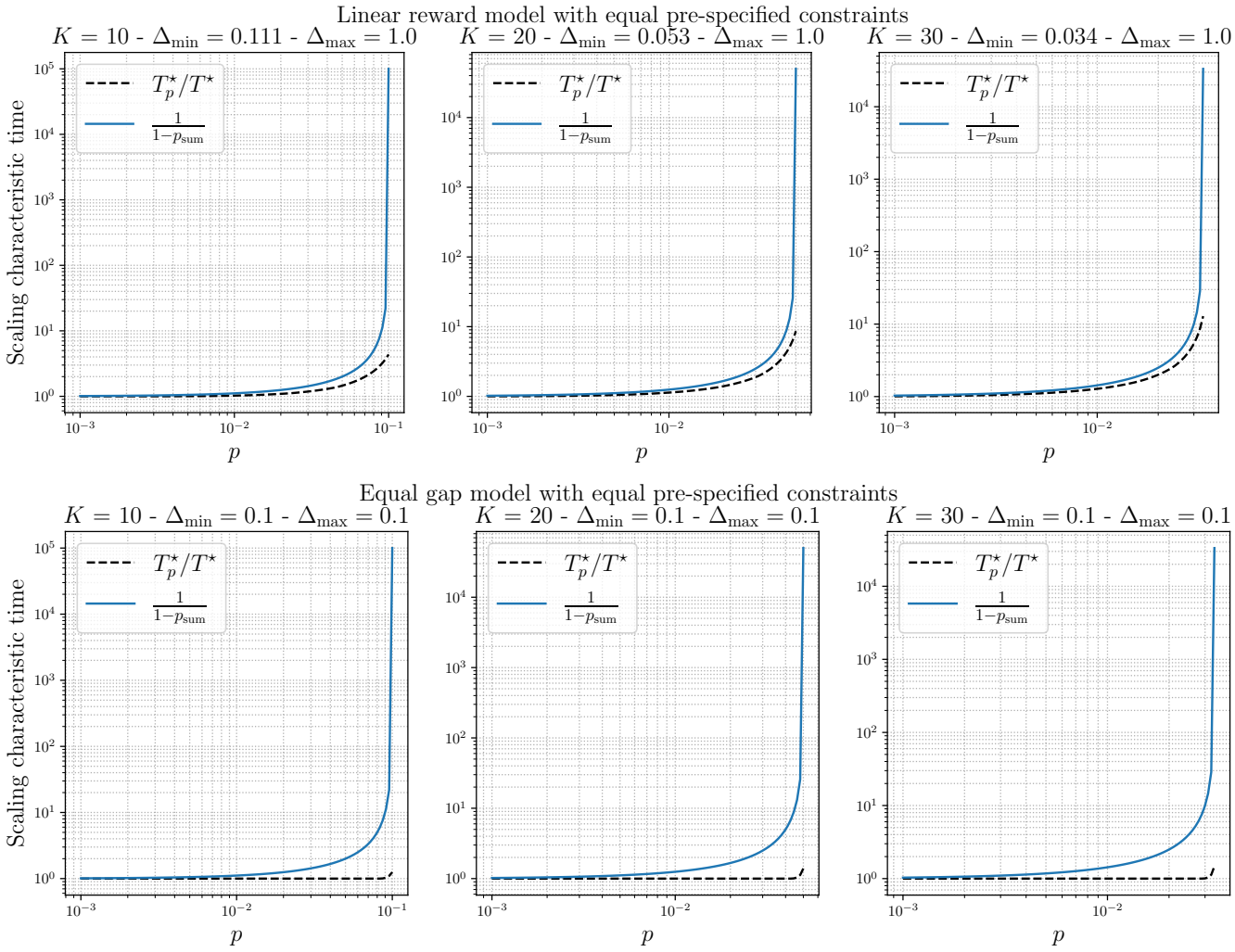

Fig. 17: Scaling of $T_p^\star/T^\star$ for two different models with equal pre-specified constraints. All the arms have the same fairness constraint $p_a \geq p$. On the top we show results for the model with rewards linearly ranging in $[0, 1]$. On the bottom we show results for the model with equal gaps $\Delta_a = 0.1$. From left to right we depict results for different number of arms $K \in \{10, 20, 30\}$.

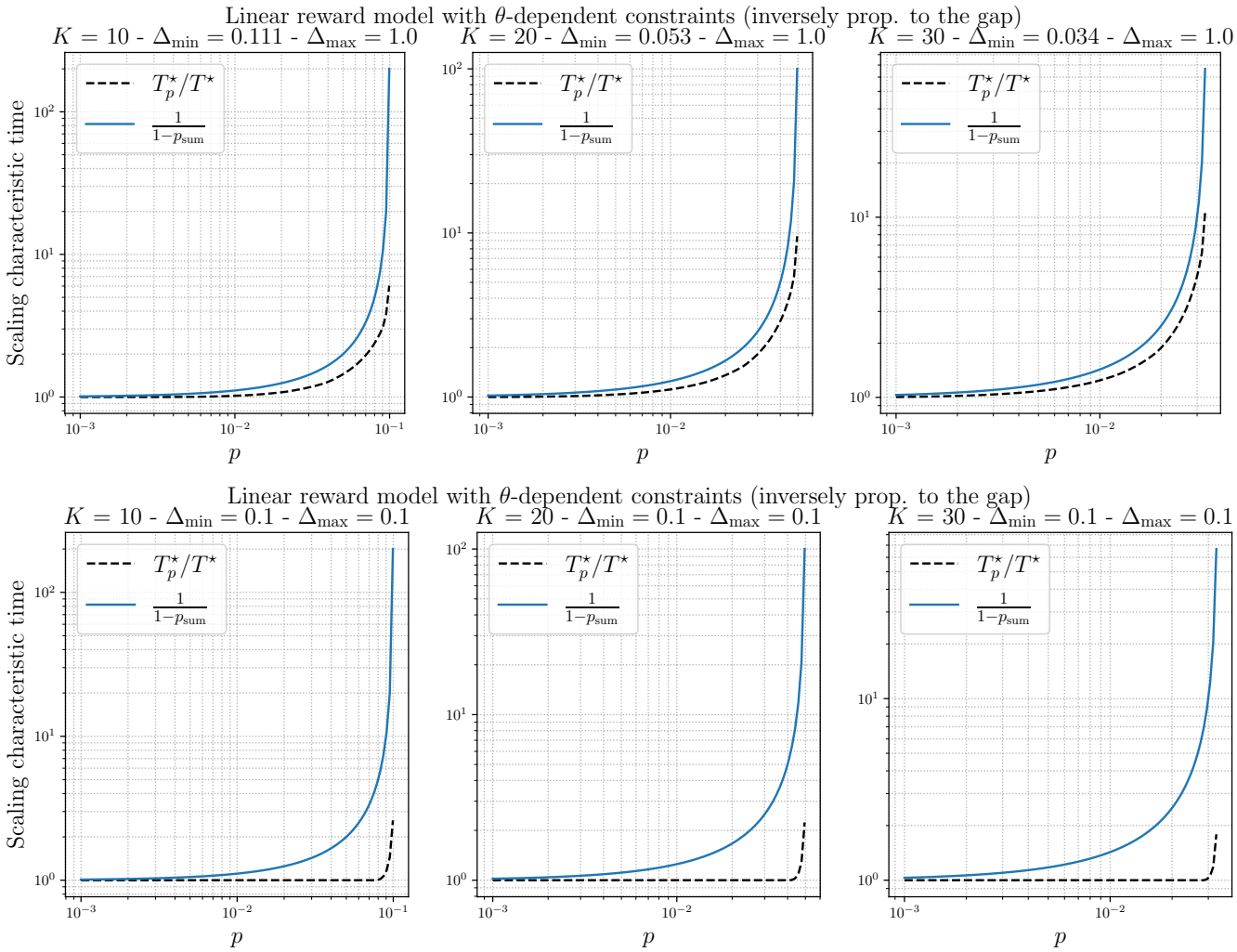

Fig. 18: Scaling of $T_p^\star/T^\star$ for two different models with equal θ-dependent constraints (see also the $p(\theta)$ function in Eq. (15)). On the top we show results for the model with rewards linearly ranging in $[0, 1]$. On the bottom we show results for the model with equal gaps $\Delta_a = 0.1$. From left to right we depict results for different number of arms $K \in \{10, 20, 30\}$.

G. Characterizing the optimal allocations

Lemma .1. *The optimal allocations to* (3) *satisfies* $w_{p,a}^\star = p_a$*, for all* $a \in [K]$ *such that* $p_a \geq w_a^\star$.

Proof: The problem (3) can be equivalently rewritten in epigraph form as

$$\begin{aligned} T_p^\star := \min_{w,z} \quad & z \\ \text{s.t.} \quad & z \geq \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}, \quad \forall a \neq a^\star \\ & w \geq p \\ & \sum_{a \in [K]} w_a = 1 \end{aligned}$$

Let $w^\star$ be an optimal allocation when $p = 0$. By Lemma 4 in [14], we have that such an optimal allocation satisfies

$$\frac{(w_a^\star)^{-1} + (w_{a^\star}^\star)^{-1}}{\Delta_a^2} = \frac{(w_b^\star)^{-1} + (w_{a^\star}^\star)^{-1}}{\Delta_b^2}, \quad \forall a, b \neq a^\star.$$

Now, consider a p such that $p_a \geq w_a^\star > 0$ for some $a \neq a^\star$. Naturally, for such a p, we must have that $w_{p,a}^\star \geq p_a$, and hence

$$\frac{(w_{p,a}^\star)^{-1} + (w_{p,a^\star}^\star)^{-1}}{\Delta_a^2} \leq \frac{(w_{p,b}^\star)^{-1} + (w_{p,a^\star}^\star)^{-1}}{\Delta_b^2}, \quad \forall b \neq a^\star.$$

Note that increasing $w^\star_{p,a^\star}$ beyond p_a, would only make the RHS larger as we have that $\sum_{b\neq a} w^\star_{p,b} = 1 - p_a$. This would yield to a higher $T^\star_p$, as $T^\star_p = \max_{b\neq a^\star} \frac{(w^\star_{p,b})^{-1}+(w^\star_{p,a^\star})^{-1}}{\Delta_b^2}$. Hence we conclude that the optimal allocation must satisfy $w^\star_{p,a} = p_a$.

In this appendix, we provide an analysis of the F-BAI algorithm. First, we analyze the fairness in App. H, and later the sample complexity in App. I. The main sample complexity results are given in 2 forms: almost sure sample-complexity optimality and optimality in expectation. For conciseness, we provide unified results for both *pre-specified rates* and θ*-dependent rates*.

H. Fairness and δ-PAC Results

We provide first a generic result on the guarantees for the case of *pre-specified rates*, and later provide the proof of Proposition V.1.

Proposition .2. *For $t \geq 1$,* F-BAI *with* pre-specified rates *guarantees that $\mathbb{E}_\theta[N_a(t)] \geq tp_a$ if $p_a > 0$, and $\mathbb{E}_\theta[N_a(t)] \geq (1-p_{\text{sum}})(\sqrt{t+1}-1)/K_0$ otherwise, where $K_0 = |\{a \in [K] : p_a = 0\}|$. For* F-BAI *with θ-dependent rates the arms are being selected at a rater greater than $\sqrt{t}$, i.e., $\mathbb{E}_\theta[N_a(t)] \geq O(\sqrt{t})$.*

Proof: Due to the properties of $w^\star$, for $t > 1$ and an arm a this choice guarantees that

$$\begin{aligned}
\mathbb{E}_\theta[N_a(t)] &\geq 1 + \sum_{s=1}^{t} \left(1 - \frac{1}{2\sqrt{s}}\right) p_a + \frac{\pi_{c,a}}{2\sqrt{s}}, \\
&\geq tp_a + \frac{(\pi_{c,a} - p_a)}{2} \sum_{s=1}^{t} \frac{1}{\sqrt{s}}, \\
&\geq tp_a + \frac{(\pi_{c,a} - p_a)}{2} \int_1^{t+1} \frac{1}{\sqrt{s}} ds, \\
&\geq tp_a + (\pi_{c,a} - p_a)(\sqrt{t+1} - 1).
\end{aligned}$$

Hence, for an arm a such that $p_a > 0$ we find $\mathbb{E}[N_a(t)] \geq tp_a$ since $\pi_{c,a} \geq p_a$, and $\mathbb{E}[N_a(t)] \geq (1-p_{\text{sum}})(\sqrt{t+1}-1)/K_0$ otherwise.

For θ-dependent rates the result follows immediately by noticing that $\mathbb{E}[N_a(t)] \geq (\sqrt{t+1}-1)/K$.

Then, for F-BAI we are able to give the following fairness guarantees.

Proof: [Proof of Proposition V.1] The proof for $\mathbb{P}_\theta(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^\star_\theta) \leq \delta$ comes directly from Thm. 7 in [34], and we omit it for simplicity.

Fairness with pre-specified rates.: From Proposition .2 for $p_a > 0$ we have $\mathbb{E}_\theta[N_a(\tau_\delta)] = \sum_t \mathbb{E}_\theta[N_a(t)|\tau_\delta = t]\mathbb{P}_\theta(\tau_\delta = t) \geq p_a \sum_t t\mathbb{P}_\theta(\tau_\delta = t) = p_a \mathbb{E}_\theta[\tau_\delta]$, which yields the first fairness guarantee.

Asymptotic fairness with θ-dependent rates.: The proof for the asymptotic result is more convoluted and relies on different tools that we present later in the appendix.

For $T \geq 1, \varepsilon > 0$, consider the concentration events

$$C_{T,0}(\varepsilon) = \cap_{t=h_0(T)}^{T} (\|\hat{\theta}(t) - \theta\|_\infty \leq \varepsilon) \text{ and } C_{T,1}(\varepsilon) = \cap_{t=h_1(T)}^{T} (\|N(t)/t - w_p^\star\|_\infty \leq K(\varepsilon)),$$

where $h_0(t) = t^{1/4}$, $h_1(t) = t^{1/2}$ and $K(\varepsilon)$ is a bounded function, vanishing in 0 (see also Proposition .8). Define then the value of the problem

$$T^\star_{\theta,p}(\varepsilon) = \sup_{\substack{\tilde{\theta}:\|\theta-\tilde{\theta}\|_\infty \leq \varepsilon, \\ \tilde{w}:\|\tilde{w}-w_p^\star\|_\infty \leq K(\varepsilon)}} T^\star_{\tilde{\theta},p}(\tilde{w}), \text{ with } T^\star_{\tilde{\theta},p}(\tilde{w}) \coloneqq \max_{a \neq a^\star_{\tilde{\theta}}} \frac{\tilde{w}_a^{-1} + \tilde{w}_{a^\star_{\tilde{\theta}}}^{-1}}{\tilde{\Delta}_a^2},$$

where $T^\star_{\tilde{\theta},p}(\tilde{w})$ is the value of the problem for model $\tilde{\theta}$ with allocation $\tilde{w}$ and $a^\star_{\tilde{\theta}} = \arg\max_a \tilde{\theta}_a$, $\tilde{\Delta}_a = \max_b \tilde{\theta}_b - \tilde{\theta}_a$.

Due to Proposition .8, we have that for $T_1(\varepsilon) \geq 1/\varepsilon^4$ and $T \geq T_1(\varepsilon)$, conditionally on $C_{T,0}(\varepsilon)$, the event $C_{T,1}(\varepsilon)$ occurs with high probability. Then, for every $t \in [\sqrt{T}, T]$ under $C_{T,0}(\varepsilon) \cap C_{T,1}(\varepsilon)$ we have $Z(t) \geq t/T^\star_{\theta,p}(\varepsilon)$. In the following, since we let $\delta \to 0$, we choose $\delta < \varepsilon$, and let $T_1(\varepsilon) = 1/\delta^4 \geq 1/\varepsilon^4$.

Next, as in Thm. .5 let $T_2(\varepsilon) = \inf\{t : t/T^\star_{\theta,p}(\varepsilon) \geq \ln(Bt/\delta)\}$, where $B > 0$ is a constant chosen as in Thm. .5 s.t. $\beta(\delta, t) \leq \ln(Bt/\delta)$. Then, for all $t \geq \max(T_1(\varepsilon), T_2(\varepsilon))$, under $C_{T,0}(\varepsilon) \cap C_{T,1}(\varepsilon)$, we have that

$$Z(t) \geq t/T^\star_{\theta,p}(\varepsilon) \geq \ln(Bt/\delta) \geq \beta(\delta, t).$$

Let then $T_\varepsilon = \max(T_1(\varepsilon), T_2(\varepsilon))$. Note that $T_\varepsilon \geq 1/\delta^4$. Hence $(\tau_\delta \leq T_\varepsilon) \supset C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon)$.

Next, note

$$\begin{aligned}
\mathbb{E}_\theta[N_a(\tau_\delta)] &= \mathbb{E}_\theta[N_a(\tau_\delta)|C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon)]\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon)) \\
&\quad + \mathbb{E}_\theta[N_a(\tau)|\overline{C_{T_\varepsilon,0}(\varepsilon)} \cup \overline{C_{T_\varepsilon,1}(\varepsilon)}]\mathbb{P}_\theta(\overline{C_{T_\varepsilon,0}(\varepsilon)} \cup \overline{C_{T_\varepsilon,1}(\varepsilon)}), \\
&\geq \mathbb{E}_\theta[N_a(\tau_\delta)|C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon)]\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon)), \\
&\geq (w^\star_{p,a} - K(\varepsilon))\mathbb{E}_\theta[\tau_\delta]\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon) \cap C_{T_\varepsilon,1}(\varepsilon))
\end{aligned}$$

Now we prove that $\liminf_{\delta\to 0}\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon)\cap C_{T_\varepsilon,1}(\varepsilon)) = 1$. Observe $\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon)\cap C_{T_\varepsilon,1}(\varepsilon)) = \mathbb{P}_\theta(C_{T_\varepsilon,1}(\varepsilon)|C_{T_\varepsilon,0}(\varepsilon))\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon))$.

From Proposition .8 we have

$$\mathbb{P}_\theta(C_{T_\varepsilon,1}(\varepsilon)|C_{T_\varepsilon,0}(\varepsilon)) \geq 1 - 2K\frac{\exp(-\sqrt{T_\varepsilon}\varepsilon^2/2)}{1-\exp(-\varepsilon^2/2)}$$

Since $\sqrt{T_\varepsilon}\geq 1/\delta^2$ we get $\liminf_{\delta\to 0}\mathbb{P}(C_{T_\varepsilon,1}(\varepsilon)|C_{T_\varepsilon,0}(\varepsilon))\geq 1$. Then, from Proposition .7 we have

$$\mathbb{P}_\theta(C_{T_\varepsilon,0}(\varepsilon)) \geq 1 - \frac{1}{T_\varepsilon^\alpha} - 2B_\varepsilon K T_\varepsilon \exp\left(-2\left\lfloor \frac{p_0^{1/4}T_\varepsilon^{1/16}}{(\log^2(1+K'T_\varepsilon^\alpha))^{1/4}}\right\rfloor \varepsilon^2\right),$$

where $\alpha>1$. Asymptotically, as $\delta\to 0$, we have that also this term converges to 1 due to the exponential converging to 0 and $1/T_\varepsilon^\alpha\to 0$. Hence

$$\liminf_{\delta\to 0}\frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \geq (w^\star_{p,a} - K(\varepsilon)).$$

Letting $\varepsilon\to 0$ concludes the proof since $K(0)=0$.

I. Sample complexity guarantees

1) Almost-sure Sample Complexity Upper Bound: In this section, we prove an almost sure sample complexity bound of F-BAI. To derive this result, first, we prove that each arm is sampled infinitely often. Later, we show that, asymptotically, the average number of times we select an arm a converges to $w^\star_{p,a}$ almost surely.

Proposition .3. *Each arm is sampled infinitely often in* F-BAI*, i.e.,* $\mathbb{P}_\theta(\lim_{t\to\infty}N_a(t)=\infty)=1$ *for all* a*.*

Proof: **Case: pre-specified rates.** The policy in F-BAI for a such that $p_a>0$ ensures that $\mathbb{P}_\theta(a_t=a)\geq p_a$ Consequently, we have that

$$\sum_{t=1}^\infty \mathbb{P}_\theta(a_t=a) \geq \sum_{t=1}^\infty p_a = \infty.$$

By the Borel-Cantelli lemma it follows that arm a is chosen infinitely often asymptotically.

Now consider an arm a such that $p_a=0$. Then F-BAI guarantees that $\mathbb{P}_\theta(a_t=a)\geq \epsilon_t\frac{1-p_{\text{sum}}}{K_0}$, hence

$$\sum_{t=1}^\infty \mathbb{P}_\theta(a_t=a) \geq \sum_{t=1}^\infty \frac{1-p_{\text{sum}}}{2K_0\sqrt{t}} = \infty.$$

Hence, each arm is sampled infinitely often.

Case: θ-dependent rates. In this case F-BAI guarantees that $\mathbb{P}_\theta(a_t=a)\geq\epsilon_t\frac{1}{K}$, thus

$$\sum_{t=1}^\infty \mathbb{P}_\theta(a_t=a) \geq \sum_{t=1}^\infty \frac{1}{2K\sqrt{t}} = \infty.$$

Hence, each arm is sampled infinitely often.

We now show that F-BAI asymptotically samples arms according to $w^\star_p$.

Proposition .4. *For every arm* $a\in[K]$*,* F-BAI *satisfies*

$$\mathbb{P}_\theta\left(\lim_{t\to\infty}\frac{N_a(t)}{t} = w^\star_{p,a}\right) = 1. \tag{16}$$

Proof: Proposition .3 guarantees that, by the law of large numbers, $(\hat\theta_t\to\theta)$ almost surely as $t\to\infty$. By continuity, we also have that (P1) $w^\star_p(t)\to w^\star_p$ almost surely (by an application of Berge's Theorem). Then, consider

$$\frac{1}{t}N_t(a) - w^\star_{p,a} = \frac{1}{t}\sum_{k=1}^t[\mathbf{1}_{(a_k=a)} - w^\star_{p,a}] = \underbrace{\frac{1}{t}\sum_{k=1}^t[\mathbf{1}_{(a_k=a)} - w^\star_{p,a}(k)]}_{(\circ)} + \underbrace{\frac{1}{t}\sum_{k=1}^t[w^\star_{p,q}(k) - w^\star_{p,q}]}_{(\square)}.$$

The second term $(\square)$ clearly tends to 0 almost surely from property (P1). To prove that the first term $(\circ)$ converges to 0 rewrite it as

$$(\circ) = \underbrace{\frac{1}{t}\sum_{k=1}^t[\mathbf{1}_{(a_k=a)} - \pi_a(k)]}_{M_t} + \underbrace{\frac{1}{t}\sum_{k=1}^t[\pi_a(k) - w^\star_{p,a}(k)]}_{(*)}.$$

For the first term let $S_t = tM_t$, and note that S_t is a martingale since $\mathbb{E}[tM_t|\mathcal{F}_{t-1}] = (t-1)M_{t-1}+\mathbb{E}[\mathbf{1}_{(a_t=a)}-\pi_a(t)|\mathcal{F}_{t-1}] = (t-1)M_t$. To show that $S_t/t \to 0$ we use Lemma 2.18 [37]. To that aim, it is sufficient to show that $\sum_k X_i^2/k^2 < \infty$, where $X_i = S_i - S_{i-1}$. Since $|X_i| \leq 1$ then the series convergences, from which follows that $\lim_{t\to\infty} \frac{S_t}{t} = 0$. Hence $M_t \to 0$ almost surely.

For the second term $(*)$ we find $\pi_a(k) - w^\star_{p,a}(k) = \epsilon_k(\pi_{c,a} - w^\star_{p,a}(k))$, hence $|\pi_a(k) - w^\star_{p,a}(k)| \leq \epsilon_k$. Therefore we conclude the proof by observing the convergence of $\frac{1}{t}\sum_{k=1}^t |\pi_a(k) - w^\star_{p,a}(k)|$ to 0:

$$\frac{1}{t}\sum_{k=1}^{t} |\pi_a(k) - w^\star_{p,a}(k)| \leq \frac{1}{2t}\sum_{k=1}^{t} \frac{1}{\sqrt{k}} \leq \frac{2\sqrt{t}-1}{2t} \to 0 \text{ as } t \to \infty.$$

We can now prove an almost sure upper bound of the sample complexity of F-BAI.

Theorem .5 (Sample complexity almost sure upper bound of F-BAI.)**.** F-BAI*, both for pre-specified rates and θ-dependent rates, guarantees that*

$$\mathbb{P}_\theta\left(\limsup_{\delta\to 0} \frac{\tau_\delta}{\ln(1/\delta)} \leq T^\star_p\right) = 1. \tag{17}$$

Proof: The proof follows similarly as in [14], [38], and we provide it for completeness. Denote by $T^\star_p(t) = \inf_{w\in\Sigma_p} \max_{a\neq a^\star_t} \frac{w_a^{-1}+w_{a^\star_t}^{-1}}{\Delta_a(t)^2}$ the optimal characteristic time for a model $\hat{\theta}(t) \in \Theta$. Consider the event $\mathcal{E} = \left(\forall a, \lim_{t\to\infty} \frac{N_a(t)}{t} = w^\star_{p,a}, \lim_{t\to\infty} \hat{\theta}(t) = \theta\right)$. From proposition .4 we have $\mathbb{P}_\theta(\mathcal{E}) = 1$. Then, there exists t_0 s.t. for all $t \geq t_0$ we have $\hat{\theta}(t) \in \Theta$.

Furthermore, due to the continuity of $T^\star_p(t)$, for every $\eta \in (0,1)$ there exists $t_1 \geq t_0$ such that for $t \geq t_1$ we have $T^\star_p \geq (1-\eta)T^\star_p(t) \geq (1-\eta)t/Z(t)$, thus $Z(t) \geq (1-\eta)t/T^\star_p$.

Now, recall that the stopping time is defined through $\beta(\delta,t) = 3\sum_a \ln(1+\ln(N_a(t))) + K\mathcal{C}_{exp}\left(\frac{\ln(1/\delta)}{K}\right)$. Since at infinity $\mathcal{C}_{exp}(x) \sim x + O(\ln(x))$ then there exists $C > 0$ s.t. $K\mathcal{C}_{exp}\left(\frac{\ln(1/\delta)}{K}\right) \leq \ln(C/\delta)$. Moreover, $3\sum_a \ln(1+\ln(N_a(t))) \leq 3K\ln(1+t)$. Hence, there exists a constant $B > 0$ such that $\beta(\delta,t) \leq \ln(Bt/\delta)$.

Combining all the observations, we find

$$\begin{aligned}
\tau_\delta &= \inf\{t \geq t_0, Z(t) \geq \beta(\delta,t)\},\\
&\leq t_1 \vee \inf\{t \geq t_0, (1-\eta)t/T^\star_p \geq \beta(\delta,t)\},\\
&\leq t_1 \vee \inf\{t \geq t_0, (1-\eta)t/T^\star_p \geq \ln(Bt/\delta)\}.
\end{aligned}$$

Applying Lemma 8 in the appendix of [38] with $\beta = B/\delta$ and $\gamma = T^\star_p/(1-\eta)$ gives that

$$\tau_\delta \leq \max\left(t_1, \frac{T^\star_p}{1-\eta}\left[\ln\left(\frac{BT^\star_p}{\delta(1-\eta)}\right) + \sqrt{2\left(\ln\left(\frac{BT^\star_p}{\delta(1-\eta)}\right)-1\right)}\right]\right).$$

Therefore $\limsup_{\delta\to 0} \frac{\tau_\delta}{\ln(1/\delta)} \leq \frac{T^\star_p}{1-\eta}$ almost surely. We conclude by letting $\eta \to 0$.

2) Expected Sample Complexity Bound: In order to prove the sample-complexity of F-BAI we must guarantee the forced exploration property with high probability. To this end, Proposition .9 is instrumental in the derivation of the sample complexity of F-BAI.

Using this result we can bound the probability of the event that $\hat{\theta}(t)$ is not within $\varepsilon > 0$ of the true value of θ, and derive the asymptotic sample complexity bound. Using Proposition .8 and Proposition .7 in the next theorem we provide an upper bound on the expected sample complexity of F-BAI.

Theorem .6 (Upper bound in expectation of F-BAI)**.** *For all $\delta \in (0,1/2)$* F-BAI *satisfies $\mathbb{E}_\theta[\tau_\delta] < \infty$ and $\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\ln(1/\delta)} \leq T^\star_p$.*

Proof: For $T \geq 1, \varepsilon > 0$ consider the concentration events

$$C_{T,0}(\varepsilon) = \cap_{t=h_0(T)}^T (\|\hat{\theta}(t) - \theta\|_\infty \leq \varepsilon) \text{ and } C_{T,1}(\varepsilon) = \cap_{t=h_1(T)}^T (\|N(t)/t - w^\star_p\|_\infty \leq K(\varepsilon)),$$

where $h_0(t) = t^{1/4}$, $h_1(t) = t^{1/2}$ and $K(\varepsilon)$ is a bounded function, vanishing in 0 (see also Proposition .8). Define then the value of the problem

$$T^\star_{\theta,p}(\varepsilon) = \sup_{\substack{\tilde{\theta}:\|\theta-\tilde{\theta}\|_\infty \leq \varepsilon,\\ \tilde{w}:\|\tilde{w}-w^\star_p\|_\infty \leq K(\varepsilon)}} T^\star_{\tilde{\theta},p}(\tilde{w}), \text{ with } T^\star_{\tilde{\theta},p}(\tilde{w}) \coloneqq \max_{a\neq a^\star_{\tilde{\theta}}} \frac{\tilde{w}_a^{-1} + \tilde{w}_{a^\star_{\tilde{\theta}}}^{-1}}{\tilde{\Delta}_a^2},$$

where $T^\star_{\tilde\theta,p}(\tilde w)$ is the value of the problem for model $\tilde\theta$ with allocation $\tilde w$ and $a^\star_{\tilde\theta} = \arg\max_a \tilde\theta_a$, $\tilde\Delta_a = \max_b \tilde\theta_b - \tilde\theta_a$.

Due to Proposition .8, there exists $T_1(\varepsilon)$ s.t. for all $T \geq T_1(\varepsilon)$, conditionally on $C_{T,0}(\varepsilon)$, the event $C_{T,1}(\varepsilon)$ occurs with high probability. Moreover, for every $t \in [\sqrt{T}, T]$ under $C_{T,0}(\varepsilon) \cap C_{T,1}(\varepsilon)$ we have $Z(t) \geq t/T^\star_{\theta,p}(\varepsilon)$.

Next, as in Thm. .5 let $T_2(\varepsilon) = \inf\{t : t/T^\star_{\theta,p}(\varepsilon) \geq \ln(Bt/\delta)\}$, where $B > 0$ is a constant chosen as in Thm. .5 s.t. $\beta(\delta, t) \leq \ln(Bt/\delta)$. Then, for all $t \geq \max(T_1(\varepsilon), T_2(\varepsilon))$, under $C_{T,0}(\varepsilon) \cap C_{T,1}(\varepsilon)$, we have that

$$Z(t) \geq t/T^\star_{\theta,p}(\varepsilon) \geq \ln(Bt/\delta) \geq \beta(\delta, t).$$

Hence $(\tau_\delta \leq T) \supset C_{T,0}(\varepsilon) \cap C_{T,1}(\varepsilon)$. Therefore

$$\begin{aligned}
\mathbb{E}[\tau_\delta] &= \sum_{T=1}^{\infty} \mathbb{P}_\theta(\tau_\delta > T), \\
&\leq \max(T_1(\varepsilon), T_2(\varepsilon)) + \sum_{T=\max(T_1(\varepsilon),T_2(\varepsilon))+1}^{\infty} \mathbb{P}_\theta(\tau_\delta > T), \\
&\leq \max(T_1(\varepsilon), T_2(\varepsilon)) + \sum_{T=\max(T_1(\varepsilon),T_2(\varepsilon))+1}^{\infty} \mathbb{P}_\theta(\overline{C_{T,0}(\varepsilon)} \cup \overline{C_{T,1}(\varepsilon)}), \\
&\leq \max(T_1(\varepsilon), T_2(\varepsilon)) + \sum_{T=\max(T_1(\varepsilon),T_2(\varepsilon))+1}^{\infty} \mathbb{P}_\theta(\overline{C_{T,1}(\varepsilon)} | \overline{C_{T,0}(\varepsilon)}) + \mathbb{P}_\theta(\overline{C_{T,0}(\varepsilon)}), \\
&\leq \max(T_1(\varepsilon), T_2(\varepsilon)) + \sum_{T=\max(T_1(\varepsilon),T_2(\varepsilon))+1}^{\infty} \mathbb{P}_\theta(\overline{C_{T,1}(\varepsilon)} | \overline{C_{T,0}(\varepsilon)}) + \mathbb{P}_\theta(\overline{C_{T,0}(\varepsilon)}),
\end{aligned}$$

The last sum by Proposition .7 (with $\alpha > 1$) and Proposition .8 is clearly bounded. Hence, also the expected value of τ_δ is bounded for all values of $\epsilon \in (0, 1)$. Therefore, as in Thm. .5, we get $\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\ln(1/\delta)} \leq \limsup_{\delta\to 0} \frac{\max(T_1(\varepsilon),T_2(\varepsilon))}{\ln(1/\delta)} \leq T^\star_p(\varepsilon)$. We conclude by letting $\varepsilon \to 0$.

Proposition .7 (Concentration of the estimate $\hat\theta_t$)**.** *Let $\alpha, \varepsilon > 0$, $h(t) = t^{1/4}$ and $C_T(\varepsilon) = \cap_{t=h(T)}^{T}(\|\hat\theta(t) - \theta\|_\infty \leq \varepsilon)$. Then, there exists a constant $B_\varepsilon > 0$ that depends on ε such that*

$$\forall T \geq 1, \mathbb{P}_\theta(\overline{C_T(\varepsilon)}) \leq \frac{1}{T^\alpha} + 2B_\varepsilon KT \exp\left(-2\left\lfloor \frac{p_0^{\frac{1}{4}} T^{1/16}}{\sqrt{\ln(1 + K'T^\alpha)}} \right\rfloor \varepsilon^2\right), \tag{18}$$

where, for pre-specified rates $K' = 2\max(K_0, K - K_0)$ and $p_0 = \min((1 - p_{\text{sum}})/2K_0, \min_{a:p_a>0} p_a)$. For θ-dependent rates instead $p_0 = 1/2K$ and $K' = K$.

Proof: Consider the forced exploration event of Proposition .9 with $\gamma = 1/T^\alpha$ and $\alpha > 0$

$$\mathcal{E}_T = \left(\forall a \in [K], \forall t \geq 1, N_a(t) \geq \left\lfloor \left(\frac{p_0 t}{\ln^2(1 + K'T^\alpha)}\right)^{1/4} \right\rfloor\right).$$

Then, using the same proposition we obtain

$$\begin{aligned}
\mathbb{P}_\theta(\overline{C_T(\varepsilon)}) &= \mathbb{P}_\theta((\overline{C_T(\varepsilon)} \cap \mathcal{E}_T) \cup (\overline{C_T(\varepsilon)} \cap \overline{\mathcal{E}_T})), \\
&\leq \mathbb{P}_\theta((\overline{C_T(\varepsilon)} \cap \mathcal{E}_T)) + \mathbb{P}_\theta(\overline{\mathcal{E}_T}), \\
&\leq \mathbb{P}_\theta((\overline{C_T(\varepsilon)} \cap \mathcal{E}_T)) + \frac{1}{T^\alpha}.
\end{aligned}$$

Let $\lambda(T) = \frac{\ln^2(1+K'T^\alpha)}{p_0}$ and consider now the first term: expand it using some union bounds and conclude with an application

of Hoeffding inequality:

$$\begin{aligned}
\mathbb{P}_\theta(\overline{C_T(\varepsilon)} \cap \mathcal{E}_T) &\leq \sum_{t=h(T)}^{T} \mathbb{P}_\theta(\|\hat{\theta}(t) - \theta\|_\infty > \varepsilon \cap \mathcal{E}_T), \\
&\leq \sum_{t=h(T)}^{T} \sum_{a} \mathbb{P}_\theta(|\hat{\theta}_a(t) - \theta_a| > \varepsilon \cap \mathcal{E}_T), \\
&\leq \sum_{t=h(T)}^{T} \sum_{a} \sum_{k=\lfloor (t/\lambda(T))^{1/4} \rfloor}^{t} \mathbb{P}_\theta(|\hat{\theta}_a(t) - \theta_a| > \varepsilon, N_a(t) = k), \\
&\leq \sum_{t=h(T)}^{T} \sum_{a} \sum_{k=\lfloor (t/\lambda(T))^{1/4} \rfloor}^{t} 2\exp(-2k\varepsilon^2), \\
&\leq 2 \sum_{t=h(T)}^{T} \sum_{a} \sum_{k=0}^{t-\lfloor (t/\lambda(T))^{1/4} \rfloor} \exp\left(-2(k + \lfloor (t/\lambda(T))^{1/4} \rfloor)\varepsilon^2\right), \\
&\leq 2 \sum_{t=h(T)}^{T} \sum_{a} \exp\left(-2(\lfloor (t/\lambda(T))^{1/4} \rfloor)\varepsilon^2\right) \sum_{k=0}^{t-\lfloor t/\lambda(T) \rfloor^{1/4}} \exp\left(-2k\varepsilon^2\right), \\
&\leq \frac{2K}{1 - e^{-2\varepsilon^2}} \sum_{t=h(T)}^{T} \exp\left(-2(\lfloor (t/\lambda(T))^{1/4} \rfloor)\varepsilon^2\right), \\
&\leq \frac{2KT}{1 - e^{-2\varepsilon^2}} \exp\left(-2(\lfloor (h(T)/\lambda(T))^{1/4} \rfloor)\varepsilon^2\right), \\
&\leq 2B_\varepsilon KT \exp\left(-2 \left\lfloor \frac{p_0^{\frac{1}{4}} T^{1/16}}{\sqrt{\ln(1 + K'T^\alpha)}} \right\rfloor \varepsilon^2\right).
\end{aligned}$$

Proposition .8 (Concentration of the average sampling $N_a(t)/t$)**.** *Let $h_0(t) = t^{1/4}, h_1(t) = t^{1/2}$ and $\varepsilon > 0$. Further let $T \geq 1/\varepsilon^4$, $C_{T,0}(\varepsilon) = \cap_{t=h_0(T)}^{T}(\|\hat{\theta}(t) - \theta\|_\infty \leq \varepsilon)$ and $C_{T,1}(\varepsilon) = \cap_{t=h_1(T)}^{T}(\|N(t)/t - w_p^\star\|_\infty \leq K(\varepsilon))$, where $K : [0,1] \to [0,1]$ is the modulus of continuity of $w_p^\star$ on $\|\theta - \theta'\|_\infty \leq \theta$, a function continuous in a neighbourhood of 0 satisfying $\lim_{\varepsilon \to 0} K(\varepsilon) = 0$. Then, we have*

$$\mathbb{P}_\theta(\overline{C_{T,1}(\varepsilon)} | C_{T,0}(\varepsilon)) \leq 2K \frac{\exp(-\sqrt{T}\varepsilon^2/2)}{1 - \exp(-\varepsilon^2/2)}. \tag{19}$$

Proof: We first prove that for $t \in [h_1(T), T]$ and $T \geq 1/\varepsilon^4$ we have

$$\mathbb{P}_\theta(\exists a : |N_a(t)/t - w_{p,a}^\star| > K(\varepsilon) | C_{T,0}(\varepsilon)) \leq 2K \exp(-t\varepsilon^2/2), \tag{20}$$

where $K(\varepsilon) < \infty$ for all $\varepsilon > 0$, and $\lim_{\varepsilon \to 0} K(\varepsilon) = 0$. If the last inequality holds, then, the proposition' statement follows by a union bound since

$$\mathbb{P}_\theta(\overline{C_{T,1}(\varepsilon)} | C_{T,0}(\varepsilon)) \leq \sum_{t=h_1(T)}^{T} 2K \exp(-t\varepsilon^2/2) \leq 2K \frac{\exp(-\sqrt{T}\varepsilon^2/2)}{1 - \exp(-\varepsilon^2/2)}.$$

Then, consider

$$N_a(t)/t - w_p^\star(a) = \underbrace{\frac{1}{t} \sum_{k=1}^{h_0(T)} [\mathbf{1}_{(a_t = a)} - w_{p,a}^\star]}_{(\circ)} + \underbrace{\frac{1}{t} \sum_{k=h_0(T)+1}^{t} [\mathbf{1}_{(a_t = a)} - w_{p,a}^\star(t)]}_{(\square)} + \underbrace{\frac{1}{t} \sum_{k=h_0(T)+1}^{t} [w_{p,a}^\star(t) - w_{p,a}^\star]}_{(*)}.$$

- For the first term $(\circ)$, for $t \geq h_1(T)$ and $T \geq 1/\varepsilon^4$ we have $(\circ) \leq h_0(T)/h_1(T) = 1/T^{1/4} \leq \varepsilon$.
- For the middle term $(\square)$ we have

$$(\square) = \underbrace{\frac{1}{t} \sum_{k=h_0(T)+1}^{t} [\mathbf{1}_{(a_t = a)} - \pi_a(t)]}_{M_t} + \underbrace{\frac{1}{t} \sum_{k=h_0(T)+1}^{t} [\pi_a(t) - w_{p,a}^\star(t)]}_{(\triangle)}$$

Let $S_t = tM_t$, and observe that S_t is a martingale since $\mathbb{E}[S_t|\mathcal{F}_{t-1}] = (t-1)M_{t-1}$. Then, using Azuma-Hoeffding inequality we have $\mathbb{P}_\theta(|M_t| \geq \varepsilon) = \mathbb{P}_\theta(|S_t| \geq t\varepsilon) \leq 2\exp(-\varepsilon^2 t/2)$.
Instead, for $(\triangle)$, for $t \geq h_1(T)$ we have

$$(\triangle) \leq \frac{1}{t}\sum_{k=h_0(T)+1}^{t} \frac{1}{2\sqrt{k}} \leq \frac{1}{2t}\int_{h_0(T)}^{t} \frac{1}{\sqrt{x}}dx \leq 1/\sqrt{t} \leq 1/T^{1/4}.$$

Hence for $T \geq 1/\varepsilon^4$ we have $(\triangle) \leq \varepsilon$.

- For the last term $(*)$ under $C_{T,0}(\varepsilon)$ by continuity (Berge's theorem) there exists $K'(\varepsilon)$ s.t. $\|w^\star_{p,a}(t) - w^\star_{p,a}\|_\infty \leq K'(\varepsilon) \leq 1$, hence $(*) \leq K'(\varepsilon)$.

In conclusion, by letting $K(\varepsilon) = (3\varepsilon + K'(\varepsilon))$ we obtain

$$\mathbb{P}_\theta(\exists a : |N_a(t)/t - w^\star_{p,a}| > K(\varepsilon)|C_{T,0}(\varepsilon)) \leq 2K\exp(-t\varepsilon^2/2).$$

J. Fair-BAI: Other Results

Here we provide the following result that is instrumental to prove the sample-complexity of FAIR-BAI .

Proposition .9 (Forced exploration of FAIR-BAI)**.** *In* FAIR-BAI *we have for* $\gamma \in (0,1)$

$$\mathbb{P}_\theta\left(\forall a \in [K], \forall t \geq 1, N_a(t) \geq \left\lfloor \left(\frac{p_0 t}{\log^2(1+\frac{K'}{\gamma})}\right)^{1/4} \right\rfloor\right) \geq 1-\gamma,$$

where, for pre-specified rates $K' = 2\max(K_0, K-K_0)$ *and* $p_0 = \min((1-p_{\rm sum})/2K_0, \min_{a:p_a>0} p_a)$*. For* θ*-dependent rates instead* $p_0 = 1/2K$ *and* $K' = K$*.*

Proof: The proof is inspired by the forced exploration property of best-policy identification techniques for MDPs [33]. We provide the proof for *pre-specified rates* and later extend it to the case with θ*-dependent rates*.

Define the event $\mathcal{E} = \{\forall a \in [K], \forall k \geq 1, \tau_a(k) \leq g(k)\}$, where $\tau_a(k)$ is the time arm a is sampled the k-th time, and let $g(k)$ be an increasing function of k with $g(k) = 0$. Later we specialize to $g(k) = \lambda k^4$ for some $\lambda > 0$. To prove the claim, we can instead prove

$$\mathbb{P}_\theta\left(\forall a \in [K], \forall k \geq 1, \tau_a(k) \leq \frac{\log^2(1+\frac{K'}{\gamma})}{p_0}k^4\right) \geq 1-\gamma.$$

A strategy is to bound $\mathbb{P}_\theta(\bar{\mathcal{E}})$, where $\bar{\mathcal{E}} = \{\exists a \in [K], \exists k \geq 1, \tau_a(k) > g(k) \wedge \forall n = 1, \dots, k-1, \tau_a(n) \leq g(n)\}$.

Decomposition of $\mathbb{P}_\theta(\bar{\mathcal{E}})$*.:* We begin by using a union bound and rewriting the terms that appear.

$$\begin{aligned}
\mathbb{P}_\theta(\bar{\mathcal{E}}) &\leq \sum_a \Big[\mathbb{P}_\theta(\tau_a(1) > g(1)) + \sum_{k\geq 2}\mathbb{P}_\theta(\tau_a(k) > g(k), \tau_a(k-1) \leq g(k-1))\Big],\\
&\leq \sum_a \Big[\mathbb{P}_\theta(\tau_a(1) > g(1)) + \sum_{k\geq 2}\mathbb{P}_\theta(\tau_a(k) - \tau_a(k-1) > g(k) - g(k-1), \tau_a(k-1) \leq g(k-1))\Big],\\
&\leq \sum_a \Big[\mathbb{P}_\theta(\tau_a(1) > g(1)) + \sum_{k\geq 2}\sum_{n=1}^{g(k-1)}\mathbb{P}_\theta(\tau_a(k) - \tau_a(k-1) > g(k) - g(k-1)|\tau_a(k-1) = n)\mathbb{P}_\theta(\tau_a(k-1) = n)\Big].
\end{aligned}$$

Upper bound of the main two terms.: Now we bound the two terms that appear in the last sentence. Regarding the first term, observe that $\mathbb{P}_\theta(\tau_a(1) > g(1))$ is the probability the first time arm a is picked after $g(1)$ trials, then $\mathbb{P}_\theta(\tau_a(1) > g(1)) \leq \prod_{i=1}^{g(1)}(1-\pi_{i,a})$. For an arm s.t. $p_a > 0$ we find that $\mathbb{P}_\theta(\tau_a(1) > g(1)) \leq (1-p_a)^{g(1)}$ and $\mathbb{P}_\theta(\tau_a(1) > g(1)) \leq \prod_{i=1}^{g(1)}(1-\epsilon_i(1-p_{\rm sum})/K_0) \leq (1-\epsilon_{g(1)}(1-p_{\rm sum})/K_0)^{g(1)}$ otherwise (since ϵ_i is a decreasing sequence).

For the second term we find

$$\mathbb{P}_\theta(\tau_a(k) - \tau_a(k-1) > N|\tau_a(k-1) = n) \leq \prod_{i=n+1}^{N+n}(1-\pi_{i,a}) \leq \begin{cases}(1-p_a)^N & \text{if } p_a > 0,\\ \prod_{i=n+1}^{N+n}(1-\epsilon_i(1-p_{\rm sum})/K_0) & \text{otherwise.}\end{cases}$$

Hence

$$\mathbb{P}_\theta(\tau_a(k) - \tau_a(k-1) > g(k) - g(k-1)|\tau_a(k-1) = n) \leq \begin{cases}(1-p_a)^{g(k)-g(k-1)} & \text{if } p_a > 0,\\ \prod_{i=n+1}^{g(k)-g(k-1)+n}(1-\epsilon_i(1-p_{\rm sum})/K_0) & \text{otherwise.}\end{cases}$$

In the last term, perform the change of variable $\prod_{i=n+1}^{g(k)-g(k-1)+n}(1-\epsilon_i(1-p_{\rm sum})/K_0) = \prod_{i=0}^{g(k)-g(k-1)-1}(1-\epsilon_{i+n+1}(1-p_{\rm sum})/K_0)$. Next, use the fact that $n \le g(k-1)$ and that ϵ_t is decreasing in t, to obtain

$$\prod_{i=0}^{g(k)-g(k-1)-1}(1-\epsilon_{i+n+1}(1-p)/K') \le \prod_{i=0}^{g(k)-g(k-1)-1}(1-\epsilon_{i+g(k-1)+1}(1-p_{\rm sum})/K_0) \le (1-\epsilon_{g(k)}(1-p_{\rm sum})/K_0)^{g(k)-g(k-1)}.$$

Let $b_{k,a} = (1-p_a)^{g(k)-g(k-1)}$ for a s.t. $p_a > 0$. Then for $g(k) = \lambda k^\alpha$ we have $g(k)-g(k-1) = \lambda(k^\alpha-(k-1)^\alpha) \ge \lambda k^{\alpha-1}$, implying $b_{k,a} \le (1-p_a)^{\lambda k^{\alpha-1}}$. Applying the inequality $1-x \le \exp(-x)$ we find

$$b_{k,a} \le \exp(-p_a\lambda k^{\alpha-1}) \le \exp(-p_a\lambda k) \Rightarrow \sum_{k\ge 1} b_{k,a} \le \sum_{k\ge .1}\exp(-p_a\lambda k) = \frac{\exp(-p_a\lambda)}{1-\exp(-p_a\lambda)}.$$

Now, let $b'_{k,a} = (1-\epsilon_{g(k)}(1-p_{\rm sum})/K_0)^{g(k)-g(k-1)}$. As before, we find $b'_{k,a} \le (1-\epsilon_{g(k)}(1-p_{\rm sum})/K_0)^{\lambda k^{\alpha-1}}$. Now, use $\epsilon_{g(k)} = 1/2\sqrt{\lambda k^\alpha}$, thus

$$b'_{k,a} \le \left(1-\frac{1-p_{\rm sum}}{2K_0\sqrt{\lambda k^\alpha}}\right)^{\lambda k^{\alpha-1}} \le \exp\left(-\frac{\lambda k^{\alpha-1}(1-p_{\rm sum})}{2K_0\sqrt{\lambda k^\alpha}}\right) = \exp\left(-\frac{\lambda k^{\alpha-1}(1-p_{\rm sum})}{2K_0\sqrt{\lambda k^\alpha}}\right) \le \exp\left(-\frac{\sqrt{\lambda}k^{\alpha/2-1}(1-p_{\rm sum})}{2K_0}\right).$$

letting $\alpha = 4$, we have $b'_{k,a} \le \exp\left(-\frac{\sqrt{\lambda}k(1-p_{\rm sum})}{2K_0}\right)$, hence $\sum_{k\ge1} b'_{k,a} \le \frac{\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}{1-\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}$.

Final step.: In conclusion, letting $p_{\min} = \min_{a:p_a>0} p_a$ and using the fact that $e^{-x}/(1-e^{-x})$ is a decreasing function for $x > 0$:

$$\begin{aligned}
\mathbb{P}_\theta(\bar{\mathcal{E}}) &\le \sum_{a:p_a>0}\frac{\exp(-p_a\lambda)}{1-\exp(-p_a\lambda)} + K_0\frac{\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}{1-\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)},\\
&\le (K-K_0)\frac{\exp(-p_{\min}\lambda)}{1-\exp(-p_{\min}\lambda)} + K_0\frac{\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}{1-\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)},\\
&\le K'\left[\frac{\exp(-p_{\min}\lambda)}{1-\exp(-p_{\min}\lambda)} + \frac{\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}{1-\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}\right],\\
&\le K'\left[\frac{\exp(-p_{\min}\lambda)}{1-\exp(-p_{\min}\lambda)} + \frac{\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}{1-\exp\left(-\frac{\sqrt{\lambda}(1-p_{\rm sum})}{2K_0}\right)}\right],\\
&\le K'\left[\underbrace{\frac{\exp(-p_0\lambda)}{1-\exp(-p_0\lambda)}}_{(\circ)} + \underbrace{\frac{\exp\left(-p_0\sqrt{\lambda}\right)}{1-\exp\left(-p_0\sqrt{\lambda}\right)}}_{(\square)}\right].
\end{aligned}$$

where $K' = 2\max(K-K_0, K_0)$ and $p_0 = \min(p_{\min}, (1-p_{\rm sum})/2K_0)$. We want to verify for what value of λ the last inequality is smaller than δ. In case the first term $(\circ)$ dominates, we can upper bound the last expression by two times $(\circ)$ and obtain that

$$K'\frac{\exp(-p_0\lambda)}{1-\exp(-p_0\lambda)} = \delta \Rightarrow \lambda = \frac{\log(1+\frac{K'}{\delta})}{p_0}$$

and otherwise if the second term $(\square)$ dominates we find

$$K'\frac{\exp\left(-p_0\sqrt{\lambda}\right)}{1-\exp\left(-p_0\sqrt{\lambda}\right)} = \delta \Rightarrow \lambda = \frac{\log^2(1+\frac{K'}{\delta})}{p_0}$$

In both cases, since $K'/\delta > 2$, we have $\ln(1+\frac{K'}{\delta}) < \log^2(1+\frac{K'}{\delta})$. Hence $\lambda = \frac{\log^2(1+K'/\delta)}{p_0}$ guarantees $\mathbb{P}_\theta(\bar{\mathcal{E}}) \le \delta$.

Adaptation with θ-dependent rates.: The adaptation in this setting is straightforward by noting now that we only have the contribution due to $b'_{k,a}$. In fact, for all arms a we can bound $\mathbb{P}_\theta(\tau_a(1) > g(1)) \leq \prod_{i=1}^{g(1)}(1-\epsilon_i/K) \leq (1-\epsilon_{g(1)}/K)^{g(1)}$ and

$\mathbb{P}_\theta(\tau_a(k) - \tau_a(k-1) > g(k) - g(k-1) | \tau_a(k-1) = n) \leq (1-\epsilon_{g(k)}/K)^{g(k)-g(k-1)}$. Choosing $g(k) = \lambda k^4$ leads to

$$b'_{k,a} \leq \left(1 - \frac{\epsilon_{g(k)}}{K}\right)^{\lambda k^\alpha - 1} \leq \exp\left(-\frac{k\sqrt{\lambda}}{2K}\right).$$

Therefore $\mathbb{P}_\theta(\bar{\mathcal{E}}) \leq K \frac{\exp(-\sqrt{\lambda}/2K)}{1-\exp(-\sqrt{\lambda}/2K)}$. Choosing λ s.t. this latter probability is bounded by δ yields the result.

Fairness in machine learning has been extensively studied [1] for various fairness criteria. Similarly, different notions of fairness have been considered in the bandit literature [12], [2], which have predominantly focused on the framework of regret minimization rather than pure exploration (*a.k.a.*, Best-Arm Identification [39]).

The majority of these notions deal with the problem that arms should not be neglected, emphasizing the importance of selecting each arm adequately. This selection could be based on an average or a specific probability, and may or may not depend on the arm's reward distribution. Such an approach essentially places a constraint on the algorithm to guarantee balanced arm selection.

In particular, fairness concepts in the literature generally fall into the following categories: *individual fairness*, *selection with pre-specified range*, *counterfactual fairness* and *group fairness* [12], [2].

- *Individual fairness* [21], [10] requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [25], [23].
- *Selection with pre-specified range* [20], [11], [9] simply demands that the rate, or probability, at which an algorithm selects an arm stays within a pre-specified range.
- *Group fairness* imposes constraints based on some statistical parity across subgroups [2]. For example, in [40] divide arms into several subgroups $G_1, \dots, G_n$ and ensure that the probability of pulling an arm is constant given the group membership. In contextual bandit problems, one can ensure fairness among different contexts, as in [41] or between groups similarly to the non-contextual setting [16].
- In [42] the authors study the concept of *counterfactual fairness*. Their definition captures the idea that a decision is fair towards an individual if it is fair also in an alternative situation where the individual belong to a different group while keeping all the other important variables unchanged.

In the following, we focus on the first two notions of fairness for single-agent systems in the online setting (we refer the reader to [2] for more information about the other cases). For the multi-agent case some recent works in the bandit setting are [43], where they study the notion of *Nash bargaining solution*, and [44], where they use the *Nash social welfare* as notion of fairness. For the offline case, in [45] the authors present ROBINHOOD, an offline contextual bandit method designed to satisfy, in probability, a generic fairness criterion defined through a constraint objective.

Selection with pre-specified range.: This type of fairness constraint demands the rate at which an arm is pulled to stay within a pre-specified range, and thus does not depend on θ.

In [20] the authors use a notion of fairness that constrains the probability that an arm a is selected to stay within a pre-specified constant interval $[l_a, u_a]$. This type of constraint yields a polytope $\mathcal{C}$ on the possible set of policies, and they compare the performance of their algorithm to the best-performing policy in $\mathcal{C}$. They propose CONSTRAINED-ε-GREEDY, an algorithm that achieves a regret of $O(K \ln T/\eta\Delta^2_{\min})$, where η is some small constant.

The authors in [9] provide a notion of *asymptotic fairness* in a combinatorial sleeping bandit setting

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{N_a(T)}{T}\right] \geq p_a \quad \forall a \in [K],$$

where $(p_a)_a \in [0,1]^K$ are known fixed values. This constraint effectively limits the rate at which arms are selected asymptotically and does not depend on the value of θ.

In [17] the authors propose two algorithms: Strictly-rate-constrainted UCB and Stochastic-rate-constrained UCB. The former algorithm guarantees that at any time the pulling rate for any arm is at least $p - 1/t$, with a regret upper bound of $O(\sum_{a\neq a^\star} \ln T/\Delta_a)$. The latter algorithm, Stochastic-rate-constrained UCB, guarantees that at each time t each arm has to be pulled with a probability greater than p. The regret is computed by comparing to a policy that pulls the best-estimated arm with probability $(1 - Kp)$, and uniformly otherwise, leading to a regret upper bounded of $O(\sum_{a\neq a^\star} \ln T/\Delta_a)$.

In [11] the authors propose a constraint similar to Strictly-rate-constrained UCB that holds uniformly over time. They say an algorithm to be η-fair if

$$\lfloor p_a t \rfloor - N_a(t) \leq \eta, \forall t \in [T], \forall a \in \mathcal{A}.$$

	Setting	Fairness definition	Lower Bound	Upper Bound
[20]	Pre-specified range	$l_a \leq \pi_a(t) \leq u_a, \forall a \in [K], \forall t \in [T]$		$O\left(K \log T/\Delta^2_{\min}\right)$
[17]	Pre-specified range	$\mathbb{E}_\theta[N_a(T)/T] \geq p, \forall a \in [K]$		$O\left(\sum_a \log(T)/\Delta_a\right)$
[11]	Pre-specified range	$\lfloor p_a t \rfloor - \eta \leq N_a(t), \forall t \in [T], \forall a \in [K]$		$O\left(\sum_a \Delta_a\right)$
[9]	Pre-specified range	$\liminf_{T\to\infty} \mathbb{E}[N_a(T)/T] \geq p_a, \forall a \in [K]$		$O\left(\sqrt{T \log(T)}\right)$
[19]	Pre-specified range	$\mathbb{E}_{x\sim p_\mathcal{X}}[\pi_{x,a}(t)] \geq p, \forall a \in [K], \forall t \in [T]$		$O(\sqrt{TMK \ln(K)})$
[10]	Individual fairness	$\mathbb{P}_\theta(\pi_a(t) > \pi_b(t)$ only if $\theta_a > \theta_b | \mathcal{H}_t) \geq 1 - \delta$	$\Omega(K^3 \log(1/\delta))$	$O(\sqrt{K^3 T \log(TK/\delta)})$
[23]	Individual fairness	$D_1(\pi_a(t), \pi_a(t)) \leq \varepsilon_1 D_2(\theta_a, \theta_b) + \varepsilon_2$ w.p. $1 - \delta, \forall t \in [T]$		$O((KT)^{2/3})$
[24]	Individual fairness	Oracle feedback		$O(\sqrt{T})$
[26]	Individual fairness	Proportional fair: $\pi^\star_a/\pi^\star_b = p(\theta_a)/p(\theta_b), \forall a, b \in [K]$.		$\tilde{O}\left(\sqrt{TK}\right)$ w.p. $1 - \delta$.
[25]	Individual fairness	Proportional fair: $\pi^\star_a/\pi^\star_b = \theta_a/\theta_b, \forall a, b \in [K]$.		

TABLE IV: Summary of bandit fairness settings.

where $\eta \geq 0$ and $p_a \in [0, 1/K)$ for every arm a. The parameter η quantifies the *unfairness tolerance* allowed in the system. Furthermore, the parameter p_a is constrained in $[0, 1/K)$. Lastly, to account for the fact that now any fair algorithm must incur a linear regret, they define a new notion of regret that does not account for the regret accumulated due to the fairness constraint:

$$R_F(T) = \sum_{a \in \mathcal{A}} \Delta_a \left(\mathbb{E}[N_a(T)] - \max(0, \lfloor p_a T \rfloor - \eta) \right).$$

They provide an instance-specific upper bound to their FAIR-UCB algorithm that for large T becomes $R_F(T) \leq (1 + \pi^2/3) \sum_{a \neq a^\star} \Delta_a$. The fact that the regret does not scale in time is due to the fact that for large values of T the algorithm will pull sub-optimal arms only to satisfy the fairness constraints. In [18] the authors show a more generic UCB-LP algorithm that is able to deal with this type of fairness constraint, and other types of combinatorial constraints.

The authors of [19] study an adversarial contextual bandit setting with M contexts and K arms. In this work fairness is defined as a minimum rate that a task is assigned to a user, and the constraint is formalized as:

$$\mathbb{E}_{x \sim p_{\mathcal{X}}}[\pi_{x,a}(t)] \geq p, \forall a \in [K], \forall t \in [T],$$

where $p_{\mathcal{X}}(x)$ is the probability of observing context x, and $\pi_{x,a}(t)$ is the conditional probability of selecting an arm a for a given context x at time t. They propose a variant of Follow-the-Regularized-Leader (FRTL) which yields $O(\sqrt{TMK \ln(K)})$ regret.

Individual fairness.: These types of fairness constraints demand making similar decisions for similar arms. One of the first works to consider this type of fairness in stochastic MABs and Contextual Bandits (CBs) is [10]. Their notion of fairness considers an history of observations $\mathcal{H}_t$ up to time t, and define an algorithm to be δ-fair if with probability at least $1 - \delta$, over the realization history $\mathcal{H}_t$, for all rounds $t \in [T]$ and all pairs of arms $a, b \in [K]$ we have

$$\pi_a(t) > \pi_b(t) \text{ only if } \theta_a > \theta_b,$$

where $\pi_a(t)$ is the probability that at time t the algorithms chooses arm a. A similar fairness criterion is considered also in [15] for Markov decision processes (by using the Q-values of the optimal policy). In other words, the condition above ensures that a better arm is always selected with a higher probability than a worse arm. They propose an adaptation of the UCB algorithm that satisfies this fairness condition and whose pseudo-regret satisfies $R(T) = O(\sqrt{K^3 T \ln(TK/\delta)})$. They also state a lower bound $\Omega(K^3 \ln(1/\delta))$ that suggests that this cubic rate in K may be hard to improve.

In [23] they consider stochastic and dueling bandits, and the authors impose two specific fairness constraints: *smooth fairness* and *calibrated fairness*. Smooth fairness indicates that two arms with similar reward distributions should be selected with comparable probabilities. Technically, for all t with probability $1 - \delta$ we have $D_1(\pi_t(a), \pi_t(b)) \leq \varepsilon_1 D_2(\theta_a, \theta_b) + \varepsilon_2$, where D_1, D_2 are suitable divergence functions with $\varepsilon_1, \varepsilon_2 \geq 0$ are suitable constant. They develop a Thompson-Sampling method that achieves a fairness regret of $O((KT)^{2/3})$. Calibrated fairness, on the other hand, requires that each arm be sampled with a probability proportional to the likelihood of its reward being the greatest.

In [24] the authors study fairness in a linear contextual bandit setting. They highlight the difficulty of defining a precise fairness metric over individuals. To avoid this issue, they assume the algorithm has access to an oracle that understands fairness but cannot define it explicitly. The algorithm learns about fairness through feedback on its decisions from the oracle, adjusting accordingly to meet the fairness constraint, and achieve a regret of $O(\sqrt{T})$.

α-fairness criterion.: Another important body of work considers the α-fairness criterion [27], [28], [29] for fair resource allocation, which yields different fairness criteria based on the value of α. Generally speaking, the aim is to find a policy maximizing the α-criterion

$$f_\alpha(\theta) = \begin{cases} \frac{\theta^{1-\alpha}}{1-\alpha} & \alpha \in [0, 1) \cup (1, \infty), \\ \log(\theta) & \alpha = 1. \end{cases}$$

For $\alpha \to \infty$ we obtain the notion of *max-min* fairness, which is used when we want to allocate as equal resources as possible to users/items However, sometimes it is unwise to allocate resources to users that are much more expensive than others. For $\alpha = 0$ we obtain the classical greedy solution, while the case $\alpha = 1$ is also known as *proportional fair*. This latter case tries to allocate resources to users/items in a proportional manner. Therefore, we believe that the case $\alpha = 1$ is part of the more general notion of *individual fairness*, which is considered in [26], [25], [30].

The authors of [26] study a MAB setting in which the goal is to devise a fair allocation according to some merit function p that is strictly positive. Their aim is to devise a policy $\pi \in \{\pi \in [0, 1]^K : \sum_a \pi_a = 1\}$ that ensures each arm has a selection rate proportional to its merit, that is

$$\frac{\pi_a}{\pi_b} = \frac{p(\theta_a)}{p(\theta_b)}, \quad \forall a, b \in [K].$$

This constraint yields an optimal fair policy of the type (Th. 3.1.1 in [26])

$$\pi_a^\star = \frac{p(\theta_a)}{\sum_{b \in [K]} p(\theta_b)}, \quad \forall a \in [K],$$

and they measure the fairness of a policy based on the so-called *fair regret* up to time T, which is defined as

$$R_F(T) = \sum_{t\in[T]} \sum_a |\pi_a^\star - \pi_a(t)|,$$

where $\pi_a(t)$ is the policy selected by the agent at time t. The regret analysis relies on two conditions: (1) that the merit of each arm is positive, lower bounded by some known constant γ; (2) that the merit function is L-Lipschitz continuous. Without one of these conditions, they show that the minimax regret lower bound is of order $O(T)$. Their UCB-type algorithm satisfies a fairness regret of $\tilde{O}(L\sqrt{KT}/\gamma)$ and a classical reward regret of $\tilde{O}(\sqrt{KT})$ with probability $1-\delta$.

In [25] the authors aim to devise a purely proportional fair allocation (with no consideration for regret). In this setting, an allocation is defined as a vector over actions $\pi = (\pi_a)_{a\in[K]}$ such that $\pi_a \geq 0$, and $\sum_a \pi_a = T$. Given the arm utilities vector $\theta = (\theta_a)_{a\in\mathcal{A}}$, an allocation is *proportionally fair* if it solves the following optimization problem:

$$\max_\pi \sum_a \theta_a \log(\pi_a) \text{ s.t. } \sum_a \pi_a = T.$$

from which one can find the optimal fair allocation as follows

$$\pi_a^\star = \frac{T\theta_a}{\sum_{b\in[K]} \theta_b}.$$

Therefore the optimal solution satisfies $\pi_a^\star/\pi_b^\star = \theta_a/\theta_b$, similarly to [26]. They formulate PROPORTIONAL CATCH-UP, an algorithm that tries to play arms as to guarantee that $N_a(t)/N_b(t) \approx \hat{\theta}_a(t)/\hat{\theta}_b(t)$, where $\hat{\theta}(t) = (\hat{\theta}_a(t))_{a\in[K]}$ is the utility estimator at time t.

Best arm identification (BAI).: The primary objective in standard best arm identification problems is to identify an arm that yields the highest expected reward, To the best of our knowledge, the setting of BAI with fairness constraint has been studied only in [31], where the authors consider fairness constraints of subpopulations. This type of constraint requires that the chosen arm must be fair across various subpopulations (such as different ethnic groups, age brackets, etc.). This is achieved by ensuring that the expected reward for each subpopulation exceeds certain predefined thresholds.