Factored-Reward Bandits with Intermediate Observations

Marco Mussi [*] [1] Simone Drago [*] [1] Marcello Restelli [1] Alberto Maria Metelli [1]

Abstract

In several real-world sequential decision problems, at every step, the learner is required to select different actions. Every action affects a specific part of the system and generates an observable intermediate effect. In this paper, we introduce the Factored-Reward Bandits (FRBs), a novel setting able to effectively capture and exploit the structure of this class of scenarios, where the reward is computed as the product of the action intermediate observations. We characterize the statistical complexity of the learning problem in the FRBs, by deriving worst-case and asymptotic instance-dependent regret lower bounds. Then, we devise and analyze two regret minimization algorithms. The former, `F-UCB`, is an anytime optimistic approach matching the worst-case lower bound (up to logarithmic factors) but fails to perform optimally from the instance-dependent perspective. The latter, `F-Track`, is a bound-tracking approach, that enjoys optimal asymptotic instance-dependent regret guarantees.

1. Introduction

In several real-world sequential decision-making problems, the learner is required to select, at every interaction, different actions, i.e., an *action vector*, acting on different portions of the system, each producing an *intermediate observation*. In such scenarios, the reward is often a combination of these observations. Consider, for instance, the case in which we want to sell a product on an e-commerce website. Our goal is to maximize the overall revenue derived from the sales of a given item. In this business process, we have to choose (i) the *price* at which to sell the product and (ii) how much *budget* to invest in advertising. On the one hand, the price we set determines the propensity of the users to buy a given item, i.e., the *conversion rate*, representing for each price, the fraction of the customers that will buy the item (Broder & Rusmevichientong, 2012; Den Boer, 2015). On the other hand, the advertising budget we invest influences the number of potential customers that will be exposed to such an item, i.e., the number of *impressions* we are able to generate with the advertisement campaign (Feldman et al., 2007). Thus, every time we select a *price-budget* pair (i.e., *action vector*), we observe a noisy realization of the conversion rate, which depends on the price, and a noisy realization of the expected number of impressions, which depends on the budget we invest in advertising (i.e., *intermediate observations*). Thus, our objective is to maximize the revenue (i.e., *reward*) that is computed as the *product* between the price, the conversion rate, and the impressions (which will give us our income) subtracting the invested advertising budget.[1]

This scenario can be, in principle, addressed as a standard Multi-Armed Bandit (MAB, Lattimore & Szepesvári, 2020) by looking at the reward (i.e., revenue) only and considering price-budget couples as actions. However, with such an approach, intermediate observations (i.e., the *conversion rate* – consequence of the price we set – and the *impressions* we generate – a consequence of the adv budget we invest) that could provide useful information would be ignored with a possible detrimental effect on the learning process. Indeed, if we look just at the reward and disregard this *factored* structure, the learning problem will: (i) present an unnecessarily large action space, including all the possible combinations of action components (e.g., price and budget pairs), and (ii) suffer a possibly amplified effect of the noise in the reward due to the product of the noisy intermediate observations (e.g., *impressions* times *conversion rate*).

A notion of *factored bandits* has been studied in (Zimmert & Seldin, 2018) in which the expected reward is a *general* function of the action components. No intermediate observations are considered and the noise is applied to the final reward only. Thus, this setting ultimately fails to model the real-world scenarios we are interested in, where the intermediate observations play a crucial role and are combined with a *specific* function (i.e., the product). As we shall see later in the paper, this specificity, motivated by the considered real-world scenarios, will allow us to obtain tighter and more detailed performance guarantees.

[*]Equal contribution [1]Politecnico di Milano, Milan, Italy. Correspondence to: Marco Mussi <marco.mussi@polimi.it>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] The formalization of this example and an additional motivating example are reported in Appendix A.

Contributions In this paper, we propose the novel setting of the *Factored-Reward Bandits* (FRBs) to model sequential decision-making problems in which the agent is required to play an action vector $\mathbf{a} = (a_1, \dots, a_d)^\intercal$ consisting of d action components. Each action component a_i provides a noisy intermediate observation x_i whose product forms the reward $r = x_1 x_2 \cdots x_d$. We study this setting from computational and statistical perspectives and propose two regret minimization algorithms endowed with theoretical guarantees. The contributions are summarized as follows:

- In Section 2, we introduce the FRB setting, describe the *feedback and noise models*, and the learning problem.
- In Section 3, we study the *statistical complexity* of the learning problem in the FRB setting by deriving regret *lower bounds*. First, in Theorem 3.1, we present the *worst-case* regret lower bound of order $\Omega(\sigma d \sqrt{kT})$, being σ the subgaussian proxy, d the number of action components, k the number of possible choices for each action component, and T is the learning horizon.[2] This result highlights how the complexity of the problem scales linearly with d and its derivation makes use of technical tools from the multitask bandits literature. In Theorem 3.2, we show that dependence on σ^d (exponential in d) is unavoidable when intermediate observations are not present, motivating their crucial role. Second, we present the *instance-dependent* asymptotic regret lower bound which is first formulated as a linear program of $\mathcal{O}(k^d)$ variables (Theorem 3.3) and, subsequently, elaborated in a more explicit form (Theorem 3.4), whose derivation makes use of the *rearrangement inequalities* (Hardy et al., 1952) and that enjoys a computational complexity of $\mathcal{O}(dk \log k)$. Qualitatively, this result shows how the different action components choices need to *coordinate* to match the lower bound.
- In Section 4, we provide a novel intuitive *optimistic anytime regret minimization algorithm*, `Factored Upper Confidence Bound` (`F-UCB`), in which optimism is applied to every action component *independently*. Then, we characterize its *worst-case* regret which has order $\widetilde{\mathcal{O}}(\sigma d \sqrt{kT})$, matching the lower bound up to logarithmic factors (Theorem 4.1). Then, we empirically study its *instance-dependent* regret, revealing that it does not match the lower bound (Theorem 4.3). This confirms how *coordination* between action components is necessary.
- In Section 5, we design and analyze a novel algorithm, *`Factored Track`* (`F-Track`). `F-Track` is based on *tracking* the bound (Lattimore & Szepesvari, 2017), and succeeds in matching the instance-dependent lower bound in the asymptotic regime (Theorem 5.1). Its analysis reveals, once more, the need for coordinating the action components to achieve the optimal performance.

Appendix B discusses additional related works. Numerical simulations are provided in Appendix E. The proofs of all the statements are reported in Appendix C.

2. Factored Reward Bandits

In this section, we introduce the *Factored-Reward Bandits* (FRBs), the learner-environment interaction, the assumptions, and we present the learning problem.[3]

Problem Formulation Let $T \in \mathbb{N}$ be the learning horizon. In a FRB, at every round $t \in [\![T]\!]$, the learner chooses an *action vector* $\mathbf{a}(t) = (a_1(t), \dots, a_d(t))^\intercal$ in the action space $\mathcal{A} := [\![k_1]\!] \times \cdots \times [\![k_d]\!]$, where for every $i \in [\![d]\!]$ we have that $k_i \in \mathbb{N}_{\geqslant 2}$ is the number of options of the i^{th} *action component* $a_i(t)$ of the vector, and $d \in \mathbb{N}_{\geqslant 1}$ is the action vector dimension (i.e., the number of components that the learner must select at every round t). As an effect of the action, the learner observes a vector of d *intermediate observations* $\mathbf{x}(t) = (x_1(t), \dots, x_d(t))^\intercal$ and receives as reward the product of the intermediate observations $r(t) = \prod_{i \in [\![d]\!]} x_i(t)$. The i^{th} component $x_i(t)$ of the intermediate observation vector $\mathbf{x}(t)$ is the effect of the i^{th} action component $a_i(t)$ in the action vector $\mathbf{a}(t)$. Specifically, every component $i \in [\![d]\!]$ of the intermediate observation vector $\mathbf{x}(t)$ is independent of the others and sampled from a distribution $x_i(t) \sim \nu_{i,a_i(t)}$, so that, $\mathbf{x}(t) \sim \boldsymbol{\nu}_{\mathbf{a}(t)} := \otimes_{i \in [\![d]\!]} \nu_{i,a_i(t)}$. Thus, we will denote an FRB as $\underline{\boldsymbol{\nu}} := \otimes_{i \in [\![d]\!]} \otimes_{a_i \in [\![k_i]\!]} \nu_{i,a_i}$. Furthermore, we can write $x_i(t) = \mu_{i,a_i(t)} + \epsilon_i(t)$, where $\mu_{i,a_i(t)}$ is the *expected intermediate observation* of the i^{th} action component $a_i(t)$, and $\epsilon_i(t)$ is σ^2-subgaussian random noise, independent conditioned to the past and the other noise realizations $\epsilon_j(t)$ for $j \in [\![d]\!] \backslash \{i\}$. As customary, we assume bounded expected values for the intermediate observations, i.e., $\mu_{i,a_i} \in [0,1]$ for every $i \in [\![d]\!]$ and $a_i \in [\![k_i]\!]$, and all intermediate observation components $x_i(t)$ characterized by the same known subgaussian proxy σ.[4]

Learning Problem An optimal action vector is $\mathbf{a}^* = (a_1^*, \dots, a_d^*)^\intercal \in \arg\max_{\mathbf{a}=(a_1,\dots,a_d)^\intercal \in \mathcal{A}} \prod_{i \in [\![d]\!]} \mu_{i,a_i}$ and, since all expected intermediate observations are non-negative, we can factorize the optimization problem observing that $a_i^* \in \arg\max_{a_i \in [\![k_i]\!]} \mu_{i,a_i}$ for every $i \in [\![d]\!]$. We denote with $\mu_i^* = \mu_{i,a_i^*}$ the expected intermediate observation of the optimal i^{th} action component. We define the suboptimality gap related to the i^{th} action component as $\Delta_{i,a_i} := \mu_i^* - \mu_{i,a_i}$ for $a_i \in [\![k_i]\!]$, and the suboptimality gap related to the action vector $\mathbf{a} = (a_1, \dots, a_d)^\intercal \in \mathcal{A}$ as $\Delta_{\mathbf{a}} := \prod_{i \in [\![d]\!]} \mu_i^* - \prod_{i \in [\![d]\!]} \mu_{i,a_i}$.

[2] In the following, we provide more general results in which each action component i can have a different number k_i of choices.

[3] Let $a, b \in \mathbb{N}$ with $a \leqslant b$, we introduce the symbols: $[\![a,b]\!] := \{a, a+1, \dots, b-1, b\}$ and $[\![b]\!] := [\![1,b]\!]$. A zero-mean random variable ξ is σ^2-subgaussian if $\mathbb{E}[e^{\lambda \xi}] \leqslant e^{\frac{\lambda^2 \sigma^2}{2}}$, for every $\lambda \in \mathbb{R}$.

[4] The extension with different known subgaussian proxies σ_i for every component $i \in [\![d]\!]$ is straightforward.

Let $\underline{\boldsymbol{\nu}}$ be an FRB, $\mathfrak{A}$ be a learning algorithm, and $T \in \mathbb{N}$ be the learning horizon, we define its *cumulative regret* as:

$$R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}}) := T \prod_{i \in [\![d]\!]} \mu_i^* - \sum_{t \in [\![T]\!]} \prod_{i \in [\![d]\!]} \mu_{i,a_i(t)} = \sum_{t \in [\![T]\!]} \Delta_{\mathbf{a}(t)}. \quad (1)$$

The goal of the learner consists in minimizing the *expected cumulative regret* $\mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})]$, where the expectation is taken w.r.t. the randomness of the observations and the possible randomness of the algorithm $\mathfrak{A}$.

3. Regret Lower Bounds

In this section, we provide lower bounds to the expected regret that any learning algorithm suffers when addressing the learning problem in a FRB, both in the minimax (Section 3.1) and the instance-dependent (Section 3.2) cases.

3.1. Worst-Case Lower Bound

We present the worst-case lower bound that every algorithm suffers and discuss the role of the structure of the FRB.

Theorem 3.1 (Worst-Case Lower Bound)**.** *For every algorithm $\mathfrak{A}$, there exists an FRB $\underline{\boldsymbol{\nu}}$ such that for:*

$$T \geqslant 2\left(1 - 2^{-\frac{1}{d-1}}\right)^{-2} \sigma^2 \max_{i \in [\![d]\!]} k_i = \mathcal{O}\left(\sigma^2 d^2 k\right), \quad (2)$$

$\mathfrak{A}$ suffers an expected cumulative regret of at least:

$$\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geqslant \frac{\sigma}{4\sqrt{2}} \sum_{i \in [\![d]\!]} \sqrt{k_i T}.$$

In particular, if $k_i =: k$ for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geqslant \Omega(\sigma d \sqrt{kT})$.

Proof Sketch. The challenge is the structure of the regret in a FRB. We lower-bound the regret $R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})$ as a sum of the regrets $R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})$ that an algorithm $\mathfrak{A}$ would have suffered by playing d *parallel MABs*. Choosing $\mu_i^* = 1$:

$$\begin{aligned} R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}}) &= \sum_{t \in [\![T]\!]} \left(1 - \prod_{i \in [\![d]\!]} \left(1 - \Delta_{i,a_i(t)}\right)\right) \\ &\geqslant \frac{1}{2} \sum_{i \in [\![d]\!]} \sum_{t \in [\![T]\!]} \Delta_{i,a_i(t)} =: \frac{1}{2} \sum_{i \in [\![d]\!]} R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}}). \end{aligned}$$

This derivation leverages an ad-hoc technical Lemma C.2, which holds for sufficiently small suboptimality gaps, i.e., $\Delta_{i,a_i} \leqslant 1 - 2^{-\frac{1}{d-1}}$. This condition gives rise to the constraint on the minimum time horizon (Equation 2), since the suboptimality gaps will be chosen $\propto T^{-1/2}$. Indeed, intuitively, if the suboptimality gaps Δ_{i,a_i} are too large (depending on d) we will have $1 - \prod_{i \in [\![d]\!]}(1 - \Delta_{i,a_i(t)}) \ll \sum_{i \in [\![d]\!]} \Delta_{i,a_i}$ making the instances more distinguishable and, consequently, reducing the regret. The result is obtained by showing that regret component satisfies $R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}}) \geqslant \Omega(\sigma\sqrt{k_i T})$ redesigning for the subgaussian case the solution designed for Bernoulli rewards from the *multitask bandit* literature (Wang et al., 2021, Theorem 10). □

To understand the beneficial effect of: (i) the factored structure and (ii) the intermediate observations, it is worth comparing the result of Theorem 3.1 with the regret lower bounds of common settings. If we remove (i), we are in the presence of a MAB with $\mathcal{A} = [\![k_1]\!] \times \cdots \times [\![k_d]\!]$ as action space.[5] It is worth noting that, even in this case, the reward $r(t) = \prod_{i \in [\![d]\!]} x_i(t)$ is the *product* of d subgaussian random variables which is not, in general, subgaussian (see Lemma D.1). Nevertheless, $r(t)$ is guaranteed to preserve a finite variance of order at least $\underline{\sigma}^2 = \sigma^{2d}$ (see Lemma D.3). Thus, we can look at the setting as a *heavy-tailed* MAB with finite variance (Bubeck et al., 2013) with $\prod_{i \in [\![d]\!]} k_i$ actions, leading to a regret of order $\Omega(\underline{\sigma}\sqrt{\prod_{i \in [\![d]\!]} k_i T})$, which becomes $\Omega(\sigma^d \sqrt{k^d T})$ when $k_i = k$ for every $i \in [\![d]\!]$.

It is natural to wonder if (i) is enough to break the exponential dependence in d (on both σ and k). This setting is similar, but not exactly overlapping, to that of Zimmert & Seldin (2018), in which a general "factored" structure is considered without intermediate observations and assuming that the subgaussian noise is applied to the reward directly. Nevertheless, (Zimmert & Seldin, 2018) provide neither worst-case lower bound nor worst-case regret analysis of the proposed algorithm. The following result shows that (i) only is enough to remove the exponential dependence in d on k but not on σ, which remains unavoidable without (ii).

Theorem 3.2 (Worst-Case Lower Bound without Intermediate Observations)**.** *For every algorithm $\mathfrak{A}^\dagger$ that ignores the intermediate observations $\mathbf{x}(t)$ and observes the reward $r(t)$ only, there exists an FRB $\underline{\boldsymbol{\nu}}$ such that for:*

$$T \geqslant 4(\min_{i \in [\![d]\!]} k_i - 1)/d,$$

$\mathfrak{A}^\dagger$ suffers an expected cumulative regret of at least:

$$\mathbb{E}\left[R_T(\mathfrak{A}^\dagger, \underline{\boldsymbol{\nu}})\right] \geqslant \frac{\sigma^d}{8} \sqrt{\frac{(\min_{i \in [\![d]\!]} k_i - 1)T}{d}}.$$

In particular, if $k_i =: k$ for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\mathfrak{A}^\dagger, \underline{\boldsymbol{\nu}})\right] \geqslant \Omega(\sigma^d \sqrt{kT/d})$.

Thus, Theorem 3.2 shows that the exponential dependence of d on σ is maintained even with the factored structure. This is particularly significant when $\sigma > 1$, a regime in which the function $\sigma^d/\sqrt{d}$ is exponentially increasing in d. This motivates the interest in studying this setting combining factored structure (i) and intermediate observations (ii).

Remark 3.1 (About the independence of the intermediate observations)**.** *The formulation of the FRB in Section 2 assumes that the components $x_i(t)$ of the observation vector $\mathbf{x}(t)$ are* independent*. This is necessary to treat the problem with appropriate advantages over standard MABs on the combinatorial action space $\mathcal{A}$. Indeed, if we rule out the independence assumption, we can always define a FRB in which $\mathbf{x}(t) = (y(t), 1, \ldots, 1)^T$, where $y(t) \sim \nu_{1,\mathbf{a}(t)}$. This*

[5] Note that makes no sense to consider (ii) without (i).

corresponds to a standard σ^2-subgaussian MAB with $\mathcal{A}$ as action space and arm distributions $\nu_{1,\mathbf{a}}$. Nevertheless, it is possible to relax the independence *assumption, by requiring* non-correlation *among the intermediate observations.*

3.2. Instance-Dependent Lower Bound

We present the instance-dependent lower bound that every algorithm suffers on a specific instance $\underline{\boldsymbol{\nu}}$ of the FRB setting.

Theorem 3.3 (Instance-Dependent Lower Bound)**.** *For every consistent*[6] *algorithm $\mathfrak{A}$ and FRB $\underline{\boldsymbol{\nu}}$ with unique optimal arm $\mathbf{a}^* \in \mathcal{A}$ it holds that:*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]}{\log T} \geqslant \underline{C}(\underline{\boldsymbol{\nu}}), \tag{3}$$

where $\underline{C}(\underline{\boldsymbol{\nu}})$ is defined as the solution to the following optimization problem:

$$\min_{(L_{\mathbf{a}})_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}}} \sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}} L_{\mathbf{a}}\Delta_{\mathbf{a}} \tag{4}$$

$$\text{s.t.} \quad L_{i,j} = \sum_{\substack{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\} \\ a_i=j}} L_{\mathbf{a}}, \ \forall i\in[\![d]\!],\ j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{5}$$

$$L_{i,j} \geqslant \frac{2\sigma^2}{\Delta_{i,j}^2},\ \forall i\in[\![d]\!], j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{6}$$

$$L_{\mathbf{a}} \geqslant 0, \quad \forall \mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}. \tag{7}$$

Proof Sketch. Here we provide an informal derivation that captures the intuition, although the formal proof requires some additional technical effort (see Appendix C.1). Thanks to the factored structure, we can show, as for stochastic bandits, that for every $j \in [\![k_i]\!]\setminus\{a_i^*\}$ and $i \in [\![d]\!]$ the expected number of pulls $\mathbb{E}[N_{i,j}(T)]$ is lower bounded by (Constraint 6):

$$L_{i,j} := \frac{\mathbb{E}[N_{i,j}(T)]}{\log T} \geqslant \frac{2\sigma^2}{\Delta_{i,j}^2} \qquad \text{for} \quad T \to +\infty$$

We now want to find the arrangements of the number of pulls of action vectors $N_{\mathbf{a}}(T)$, for every $\mathbf{a} \in \mathcal{A}\setminus\{\mathbf{a}^*\}$, to minimize the cumulative regret. Recalling that $N_{i,j}(T) = \sum_{\mathbf{a}\in\mathcal{A}\,:\,a_i=j} N_{\mathbf{a}}(T)$, we define $L_{i,j} = \sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}\,:\,a_i=j} L_{\mathbf{a}}$ (Constraint 5). Finally, by recalling the decomposition of the regret $\frac{\mathbb{E}[R_T(\mathfrak{A},\underline{\boldsymbol{\nu}})]}{\log T} = \sum_{\mathbf{a}\in\mathcal{A}} L_{\mathbf{a}}\Delta_{\mathbf{a}}$ we get the objective function in Equation (4) to be minimized. Notice that to make the proof fully formal we need to properly manage the asymptotic behavior of the sequences $\mathbb{E}[N_{i,j}(T)]$ and $\mathbb{E}[N_{\mathbf{a}}(T)]$ when $T \to +\infty$. □

The optimization problem in Theorem 3.3 is a Linear Program (LP) with $\prod_{i\in[\![d]\!]} k_i + \sum_{i\in[\![d]\!]} k_i - d - 1$ variables and $\prod_{i\in[\![d]\!]} k_i + 2\sum_{i\in[\![d]\!]} k_i - 2d - 1$ constraints. Constraint (5) establishes the relation between the number of pulls of the action vectors $L_{\mathbf{a}}$ and the number of pulls of the action components $L_{i,j}$. This captures the "information sharing" of the setting in which we obtain a sample for the action component (i, j) whenever we pull an action vector $\mathbf{a}$ such that $a_i = j$. Being a minimization problem, Constraint (6) will be satisfied with equality allowing the removal of variables $L_{i,j}$ and the relative constraints. Thus, the LP can be solved in polynomial time w.r.t. $\prod_{i\in[\![d]\!]} k_i$ (Vaidya, 1989).

Explicit Solution of the LP Program We now illustrate how to solve the LP program with a smaller time complexity of order $\mathcal{O}(\sum_{i\in[\![d]\!]} k_i \log k_i)$. We first provide the intuition and, then, provide the formal argument.

The minimum proportion with which the action component (i, j) is to be pulled (Constraint 6) can be accomplished by pulling different sequences of action vectors $\mathbf{a}$ such that $a_i = j$. *How to "arrange" the pulls of the action vectors to satisfy Constraint* (6) *and minimize the regret?* To start capturing the intuition, consider the simplest setting with $d = 2$, $k_1 = k_2 = 2$, $a_1^* = a_2^* = 1$, $\mu_{1,1} = \mu_{2,1} = 1$ and $\mu_{1,2} = \mu_{2,2} = y \in (0, 1)$. To satisfy Constraint (6), we have to guarantee $L_{1,2} = L_{2,2} = 2\sigma^2(1 - y)^{-2}$ (in the solution the constraint is satisfied with equality) and we have at our disposal 4 action vectors $\mathcal{A} = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$. We can satisfy the constraint in two ways:[7]

(i) playing action $(2, 2)$ (i.e., with both suboptimal components) for a proportion of $2\sigma^2(1 - y)^{-2}$ times, suffering $1 - y^2$ instantaneous regret;

(ii) playing actions $(1, 2)$ and $(2, 1)$ (i.e., with one suboptimal component) for a proportion of $2\sigma^2(1 - y)^{-2}$ *each*, suffering $1 - y$ instantaneous regret;

It is simple to convince that (i) is the choice that minimizes the cumulative regret. Indeed, for $y \in (0, 1)$, we have:

$$\underbrace{2\sigma^2(1-y)^{-2}(1-y^2)}_{\text{case } (i)} \leqslant \underbrace{4\sigma^2(1-y)^{-2}(1-y)}_{\text{case } (ii)}. \tag{8}$$

This intuitive reasoning can be extended to the general case. To this end, let us define the *sorting functions* $\pi_i : [\![k_i]\!] \to [\![k_i]\!]$ for every $i \in [\![d]\!]$ as any bijective function such that:

$$\mu_{i,\pi_i(1)} \leqslant \cdots \leqslant \mu_{i,\pi_i(k_i-1)} \leqslant \mu_{i,\pi_i(k_i)} = \mu_i^*.$$

We claim that in the optimal arrangement the action components need to *coordinate* as illustrated in Figure 1. For every dimension $i \in [\![d]\!]$ (row), we sort the action components in non-decreasing order of $\mu_{i,j}$ according to the sorting function π_i. To every $j \in [\![k_i - 1]\!]$, an interval of length $L_{i,j}$ is associated corresponding to the proportion of pull. Now, we combine the different rows to obtain the "active action vector" (represented by different colors) made by the corresponding action components. Each active action vector will be pulled for a proportion (the colored vertical slices) depending on the $L_{i,j}$ values of the corresponding components. Notice that we can have at most $\sum_{i\in[\![d]\!]} k_i - 1$ active action vectors and the total proportion-

[6] An algorithm $\mathfrak{A}$ is *consistent* if for every FRB $\underline{\boldsymbol{\nu}}$ and $p > 0$, it holds that $\limsup_{T\to+\infty} \mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})]/T^p = 0$.

[7] Any mix between (i) and (ii) is clearly suboptimal.

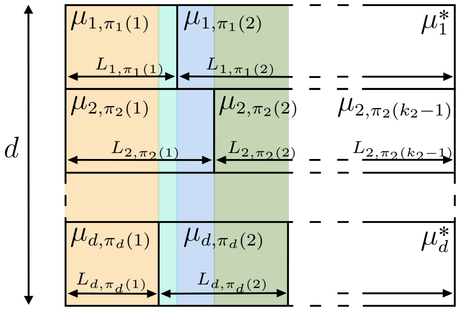


Figure 1. Efficient solution to the LP presented in Theorem 3.3.

tion of the pulls (the width of the full table in Figure 1) is given by $M := \max_{i\in[\![d]\!]} \sum_{j\in[\![k_i-1]\!]} L_{i,j}$. To formally characterize the solution, we introduce, for every $i \in [\![d]\!]$ and $l \in [\![k_i - 1]\!]$, the variables $M_{i,l} := \sum_{l'\in[\![l]\!]} L_{i,\pi_i(l')}$ and $M_{i,k_i} = +\infty$ as the cumulative proportion of pulls of the action components more suboptimal than $(i, \pi_i(l))$, i.e., fixing a row i, the position of the black vertical lines in Figure 1 sorted from left to right. Let us define the sorting function $\boldsymbol{\pi} : [\![K]\!] \to \bigcup_{i\in[\![d]\!]}(\{i\} \times [\![k_i]\!])$, where $K = \sum_{i\in[\![d]\!]} k_i$, as any bijection such that:

$$M_{\boldsymbol{\pi}(1)} \leqslant \cdots \leqslant M_{\boldsymbol{\pi}(K-d)},$$

with the convention $M_{\boldsymbol{\pi}(0)} = 0$, i.e., the position in which we move from one vertical slice to the next one in Figure 1 sorted from left to right. For every $\ell \in [\![K]\!]$, we define the active action vector as $\boldsymbol{\alpha}_\ell = (j_{1,\ell}, \ldots, j_{d,\ell})^\top \in \mathcal{A}$ where:

$$j_{i,\ell} := \pi_i^{-1}\left(\arg\max_{l\in[\![k_i]\!]}\{M_{i,l} \geqslant M_{\boldsymbol{\pi}(\ell)}\}\right).$$

This allows us to prove the following result.

Theorem 3.4 (Instance-Dependent Lower Bound (Explicit))**.** *Let $\underline{C}(\boldsymbol{\nu})$ be the solution of the optimization problem of Theorem 3.3. It holds that:*

$$\underline{C}(\boldsymbol{\nu}) = \sum_{\ell=1}^{K-d}\left(M_{\boldsymbol{\pi}(\ell)} - M_{\boldsymbol{\pi}(\ell-1)}\right)\Delta_{\boldsymbol{\alpha}_\ell},$$

that can be computed in $\mathcal{O}(\sum_{i\in[\![d]\!]} k_i \log k_i)$.

Proof Sketch. We generalize Equation (8) with the *rearrangement inequality* for integrals (Luttinger & Friedberg, 1976), the continuous version of the more known rearrangement inequality for sequences (Hardy et al., 1952). □

4. A Worst-Case Optimal Algorithm

In this section, we present an *optimistic any-time* regret minimization algorithm for the FRB setting. *`Factored Upper Confidence Bound`* (`F-UCB`), whose pseudocode is reported in Algorithm 1, is based on the idea of running a UCB-like exploration (Auer et al., 2002) *independently* for every dimension $i \in [\![d]\!]$ and estimate the expected observation μ_{i,a_i} for every action component $a_i \in [\![k_i]\!]$.

Algorithm 1: `F-UCB`.

```
Input : Exploration Parameter $\alpha$, Subgaussian proxy $\sigma$,
        Action component size $k_i$, $\forall i \in [\![d]\!]$
1 Initialize $N_{i,a_i}(0) \leftarrow 0$, $\widehat{\mu}_{i,a_i}(0) \leftarrow 0$ $\forall a_i \in [\![k_i]\!]$, $i \in [\![d]\!]$
2 for $t \in [\![T]\!]$ do
3     Select $\mathbf{a}(t) \in \arg\max_{\mathbf{a}=(a_1,\ldots a_d)^\top\in\mathcal{A}} \prod_{i\in[\![d]\!]} \mathrm{UCB}_{i,a_i}(t)$
          where $\mathrm{UCB}_{i,a_i}(t) = \widehat{\mu}_{i,a_i}(t-1) + \sigma\sqrt{\frac{\alpha \log t}{N_{i,a_i}(t-1)}}$
4     Play $\mathbf{a}(t)$ and observe $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))^\top$
5     Update $\widehat{\mu}_{i,a_i(t)}(t)$ and $N_{i,a_i(t)}(t)$ for every $i \in [\![d]\!]$
6 end
```

The algorithm requires as input the number of action components k_i for every $i \in [\![d]\!]$, the exploration parameter $\alpha > 2$, and the subgaussian proxy σ. After initializing the variables to keep track of the number of pulls $N_{i,a_i}(t)$ and the sample mean $\widehat{\mu}_{i,a_i}(t)$ for all action components (line 1), the algorithm starts the learner-environment interaction. At every round $t \in [\![T]\!]$, `F-UCB` computes the optimistic action, i.e., the action $\mathbf{a}(t)$ maximizing the optimistic index:

$$\mathbf{a}(t) \in \arg\max_{\mathbf{a}=(a_1,\ldots,a_d)^\top\in\mathcal{A}} \prod_{i\in[\![d]\!]} \mathrm{UCB}_{i,a_i}(t),$$

where $\widehat{\mu}_{i,a_i}(t)$ is the empirical mean of the observations for the i^{th} component of the observation vector determined by the action component a_i, and $N_{i,a_i}(t)$ is the number of times the corresponding component of the action vector has been played (line 3). Then, the algorithm plays it and observes the d-dimensional observation vector $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))^\top$ (line 4). The observation vector is used to incrementally update the sample means of *all* action components involved and the related counters (lines 5). Finally, the algorithm reduces to `UCB1` when $d = 1$.

`F-UCB` enjoys a *time complexity* of $\mathcal{O}(T\sum_{i\in[\![d]\!]} k_i)$ and a *space complexity* of $\mathcal{O}(\sum_{i\in[\![d]\!]} k_i)$. Indeed, at every round $t \in [\![T]\!]$, we need to recompute the index $\mathrm{UCB}_{i,a_i}(t)$ for all $\sum_{i\in[\![d]\!]} k_i$ action components (at least the bonus changes at every round). Note that the computation of the optimistic action is not combinatorial since the optimization can be performed *independently* for every dimension $i \in [\![d]\!]$.

4.1. Worst-Case Regret Analysis

In this section, we provide the worst-case regret analysis of `F-UCB` as summarized in the following result.

Theorem 4.1 (Worst-Case Upper Bound for `F-UCB`)**.** *For any FRB $\underline{\boldsymbol{\nu}}$, `F-UCB` with $\alpha > 2$ suffers an expected regret bounded as:*

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant 4\sigma \sum_{i\in[\![d]\!]} \sqrt{\alpha k_i T \log T} + g(\alpha)\sum_{i\in[\![d]\!]} k_i,$$

where $g(\alpha) = \widetilde{\mathcal{O}}\left((\alpha-2)^{-2}\right)$.[8] *In particular, if $k_i =: k$, for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant \widetilde{\mathcal{O}}(\sigma d\sqrt{kT})$.*

[8] The complete expression is reported in the proof.

Proof Sketch. Under a suitable "good event", we have that $\mu_{i,a_i} \leqslant \mathrm{UCB}_{i,a_i}(t)$ for every $i \in [\![d]\!]$, $a_i \in [\![k_i]\!]$, and $t \in [\![T]\!]$. Thus, the instantaneous regret is bounded as:

$$\prod_{i\in[\![d]\!]} \mu_i^* - \prod_{i\in[\![d]\!]} \mu_{i,a_i(t)}$$
$$= \sum_{l\in[\![d]\!]} \prod_{i\in[\![l-1]\!]} \underbrace{\mu_i^*}_{\in[0,1]} \underbrace{\left(\mu_l^* - \mu_{l,a_l(t)}\right)}_{\leqslant \mathrm{UCB}_{i,a_i(t)}(t) - \mu_{l,a_l(t)}} \prod_{i\in[\![l+1,d]\!]} \underbrace{\mu_{i,a_i(t)}}_{\in[0,1]}$$
$$\leqslant \sum_{l\in[\![d]\!]} \left(\mathrm{UCB}_{l,a_l(t)}(t) - \mu_{l,a_l}\right),$$

where the first line is obtained by summing and subtracting all mixed terms $\prod_{i\in[\![l]\!]} \mu_i^* \prod_{i\in[\![l+1,d]\!]} \mu_{i,a_i(t)}$ and the second by optimism $\mu_l^* \leqslant \mathrm{UCB}_{l,a_l^*}(t) \leqslant \mathrm{UCB}_{l,a_l(t)}(t)$. □

Comparing the upper bound of Theorem 4.1 with the lower bound in Theorem 3.1, we realize that the dependence on the learning horizon T is tight up to logarithmic factors (just like `UCB1`) and the dependence on the number of action components k_i, the number of dimensions d, and the subgaussian proxy σ are tight up to constant factors.

It is worth comparing our results with the ones that could be obtained by applying literature algorithms to our FRB setting. As already mentioned in Section 3, although each intermediate observation $x_i(t)$ is σ^2-subgaussian, their product $r(t)$, i.e., the reward, is not in general. This prevents, for instance, the application of `UCB1` which assumes subgaussian (or bounded) reward. Precisely, for $d = 2$, the reward $r(t) = x_1(t)x_2(t)$ is a *subexponential* random variable, a scenario that can be still approached with the standard sample mean estimator but leveraging the Bernstein's concentration bound (Boucheron et al., 2013). However, for $d \geqslant 3$, as shown in Lemma D.1, the reward $r(t)$ does not admit a moment-generating function and, consequently, displays a *heavy-tailed* behavior (Bubeck et al., 2013). Nevertheless, the reward $r(t)$ random variable maintains a finite variance bounded by $\overline{\sigma}^2 = \left(1+\sigma^2\right)^d - 1$ (see Lemma D.2). This enables the application of algorithms designed for heavy-tailed bandits, such as `Robust-UCB` (Bubeck et al., 2013), able to handle generic distributions with finite variance, by resorting to estimators other than the sample mean. It is easy to verify that by considering the *Median of Means* estimator (Bubeck et al., 2013), we obtain a regret upper bound in the order of $\widetilde{\mathcal{O}}\left(\overline{\sigma}\sqrt{\prod_{i\in[\![d]\!]} k_i T}\right)$. This result is in line with the discussion in Section 3 and, clearly, not optimal. Indeed, the dependence on the product $\prod_{i\in[\![d]\!]} k_i \gg \sum_{i\in[\![d]\!]} k_i$ is because `Robust-UCB` does not exploit the factored property of the FRB setting. Furthermore, the dependence on $\overline{\sigma} = \sqrt{(1+\sigma^2)^d - 1} \geqslant \sigma$ is justified by the fact that the intermediate observations are ignored. Finally, the analysis of `Factored Bandit TEA` (Zimmert & Seldin, 2018) cannot be adapted to our setting since, as already mentioned, the subgaussian noise is applied to the final reward only.

4.2. Instance-Dependent Upper Bound

In this section, we provide the analysis of the instance-dependent regret upper bound for the `F-UCB` algorithm. The following theorem summarizes the result.

Theorem 4.2 (Instance-Dependent Upper Bound for `F-UCB`)**.** *For a given FRB $\underline{\boldsymbol{\nu}}$, `F-UCB` with $\alpha > 2$ suffers an expected regret bounded as:*

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}}),$$

where $\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})$ is defined as the solution to the following optimization problem (where $g(\alpha) = \widetilde{\mathcal{O}}((\alpha-2)^{-2})$):

$$\max_{(N_{\mathbf{a}})_{\mathbf{a}\in\mathcal{A}}} \sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}} N_{\mathbf{a}}\Delta_{\mathbf{a}} \tag{9}$$

$$\text{s.t.} \quad N_{i,j} = \sum_{\substack{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\} \\ a_i = j}} N_{\mathbf{a}}, \ \forall i\in[\![d]\!],\ j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{10}$$

$$N_{i,j} \leqslant \frac{4\alpha\sigma^2 \log T}{\Delta_{i,j}^2} + g(\alpha),\ \forall i\in[\![d]\!],\ j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{11}$$

$$\sum_{\mathbf{a}\in\mathcal{A}} N_{\mathbf{a}} = T \tag{12}$$

$$N_{\mathbf{a}} \geqslant 0,\ \forall \mathbf{a}\in\mathcal{A} \tag{13}$$

The derivation of the LP in Theorem 4.2 follows a similar rationale as that of the instance-dependent lower bound of Theorem 3.3. Since `F-UCB` runs an optimistic UCB strategy *independent* for every action component, we can derive an upper bound on the expected number of pulls for every $i \in [\![d]\!]$ and $j \in [\![k_i]\!]\setminus\{a_i^*\}$ (denoted with $N_{i,j}$ in the LP):

$$\mathbb{E}[N_{i,j}(T)] \leqslant \frac{4\alpha\sigma^2 \log T}{\Delta_{i,j}^2} + g(\alpha),$$

generating Constraint (11), that, since the problem involves a maximization, will be satisfied with equality. To relate the expected number of pulls $\mathbb{E}[N_{\mathbf{a}}(T)]$ of the action vectors $\mathbf{a} \in \mathcal{A}\setminus\{\mathbf{a}^*\}$ (denoted with $N_{\mathbf{a}}$ in the LP) with the ones of the action components $\mathbb{E}[N_{i,j}(T)]$, we use the same argument of Theorem 3.3, producing Constraint (10). Similarly to the LP in Theorem 3.3, the problem is made of $\prod_{i\in[\![d]\!]} k_i + \sum_{i\in[\![d]\!]} k_i - d$ variables and $1+\prod_{i\in[\![d]\!]} k_i + 2\sum_{i\in[\![d]\!]} k_i - 2d$ constraints. We now provide an explicit solution to a *relaxation* of the LP of Theorem 4.2.

Corollary 4.3 (Explicit Instance-Dependent Upper Bound for `F-UCB`)**.** *For a given FRB $\underline{\boldsymbol{\nu}}$, `F-UCB` with $\alpha > 2$ suffers an expected regret bounded by:*

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})$$
$$\leqslant 4\alpha\sigma^2 \log T \sum_{i\in[\![d]\!]} \mu_{-i}^* \sum_{j\in[\![k_i]\!]\setminus\{a_i^*\}} \Delta_{i,j}^{-1} + g(\alpha) \sum_{i\in[\![d]\!]} k_i,$$

where $\mu_{-i}^* = \prod_{l\in[\![d]\!]\setminus\{i\}} \mu_l^* \leqslant 1$ for every $i \in [\![d]\!]$.

Proof Sketch. The result is based on providing a *relaxation* of the objective function of the optimization problem in Theorem 4.2, which is based on the following bound on the suboptimality gaps of the action vector $\mathbf{a} = (a_1, \dots, a_d)^\top$

in terms of the suboptimality gaps of the action components:

$$\Delta_{\mathbf{a}} \leqslant \sum_{i \in [\![d]\!]} \Delta_{i,a_i} \mu^*_{-i}.$$

This allows to upper bound the objective function as:

$$\sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} N_{\mathbf{a}} \Delta_{\mathbf{a}} \leqslant \sum_{i \in [\![d]\!]} \mu^*_{-i} \sum_{j \in [\![k_i]\!] \setminus \{a^*_i\}} N_{i,j} \Delta_{i,j}.$$

By Constraint (11) to upper bound N_{i,a_i}, we get the result. Alternatively, we can drop the constraint $\sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} N_{\mathbf{a}} = T$ and use a *rearrangement inequality* (Hardy et al., 1952) to upper bound the objective function. □

It is worth comparing this instance-dependent regret upper bound of F-UCB with the one achievable with an algorithm for heavy-tailed bandits, such as Robust-UCB (Bubeck et al., 2013). Our result of Corollary 4.3 is of order (neglecting the dependence on α and on constants):

$$\mathcal{O}\left(\sigma^2 \sum_{i \in [\![d]\!]} \mu^*_{-i} \sum_{j \in [\![k_i]\!] \setminus \{a^*_i\}} \frac{\log T}{\Delta_{i,j}} \right). \tag{14}$$

Instead, Robust-UCB, for instance with the *Median of Means* estimator, is characterized by the following instance-dependent regret of order (neglecting constants):

$$\mathcal{O}\left(\overline{\sigma}^2 \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} \frac{\log T}{\Delta_{\mathbf{a}}} \right). \tag{15}$$

where $\overline{\sigma}^2 = (1+\sigma^2)^d - 1 \geqslant \sigma^2$. It is simple to observe that Equation (15) is larger than Equation (14). Indeed, consider the subset of action vectors in which exactly one component is not optimal, i.e., $\mathcal{A}^\circ = \bigcup_{i \in [\![d]\!]} \mathcal{A}^\circ_i$ where $\mathcal{A}^\circ_i := \{\mathbf{a} \in \mathcal{A} : a_i \neq a^*_i,\ a_j = a^*_j, j \in [\![d]\!] \setminus \{i\}\}$. We observe that for every $\mathbf{a} \in \mathcal{A}^\circ_i$, the action vector suboptimality gap is related with equality to that of the suboptimal component:

$$\Delta_{\mathbf{a}} = \prod_{l \in [\![d]\!]} \mu^*_l - \mu_{i,a_i} \prod_{l \in [\![d]\!] \setminus \{i\}} \mu^*_l = \mu^*_{-i} \Delta_{i,a_i}.$$

This allows the conclusion of the following as desired:

$$\sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} \frac{\log T}{\Delta_{\mathbf{a}}} \geqslant \sum_{\mathbf{a} \in \mathcal{A}^\circ} \frac{\log T}{\Delta_{\mathbf{a}}} \geqslant \sum_{i \in [\![d]\!]} \mu^*_{-i} \sum_{j \in [\![k_i]\!] \setminus \{a^*_i\}} \frac{\log T}{\Delta_{i,j}}.$$

Finally, let us compare Corollary 4.3 with the instance-dependent regret upper bound of the Factored Bandit TEA algorithm (Zimmert & Seldin, 2018), although the noise model is different. Theorem 2 of (Zimmert & Seldin, 2018) provides a bound of order (neglecting constants):

$$\mathcal{O}\left(\kappa \sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a^*_i\}} \frac{\log(T \log T) + \log \frac{\log(T \log T)}{\Delta^2_{i,j}}}{\Delta_{i,j}} \right),$$

where κ is such that $\Delta_{\mathbf{a}} \leqslant \kappa \sum_{i \in [\![d]\!]} \Delta_{i,a_i}$. Thus, we can set $\kappa = \max_{i \in [\![d]\!]} \mu^*_{-i}$. This result is slightly worse than ours because of the presence of the larger κ and the additional $\log \log T$ and $\log(1/\Delta^2_{i,j})$ terms.

Remark 4.1 (About Instance-Dependent Optimality of F-UCB)**.** *We argue about the instance-dependent optimality of F-UCB. To this end, we focus on a specific FRB instance with generic $d > 1$ and $k_1 = \cdots = k_d = 2$. We consider Gaussian intermediate observations with expected values $\mu_{i,1} = 1$ and $\mu_{i,2} = 1 - \Delta$ where $\Delta \in (0,1)$ for every $i \in [\![d]\!]$. By applying Theorems 3.3 and 4.2, we deduce that for $T \to +\infty$, we have the lower bound (left) and the F-UCB upper bound (right) on the number of pulls of each suboptimal action component $i \in [\![d]\!]$ bounded as:*

$$\frac{\mathbb{E}[N_{i,2}(T)]}{\log T} \geqslant \frac{2\sigma^2}{\Delta^2} \quad \textit{and} \quad \frac{\mathbb{E}[N_{i,2}(T)]}{\log T} \leqslant \frac{4\alpha\sigma^2}{\Delta^2}.$$

Thanks to Theorem 3.4 and Corollary 4.3, we can compute $\underline{C}(\underline{\boldsymbol{\nu}})$ and upper bound $\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})$:

$$\underline{C}(\underline{\boldsymbol{\nu}}) = \frac{2\sigma^2(1-(1-\Delta)^d)}{\Delta^2} \quad \textit{and} \quad \frac{\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})}{\log T} \leqslant \frac{4d\alpha\sigma^2}{\Delta}.$$

It is immediate to realize the following extreme behaviors:

$$\frac{\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})}{\underline{C}(\underline{\boldsymbol{\nu}}) \log T} \leqslant \frac{2d\alpha\Delta}{1-(1-\Delta)^d} \to \begin{cases} 2\alpha & \Delta \to 0 \\ 2\alpha d & \Delta \to 1 \end{cases}. \tag{16}$$

This suggests that for sufficiently large $\Delta \approx 1$, F-UCB can perform significantly worse than the lower bound, introducing an additional dependence on d. Instead, for sufficiently small $\Delta \approx 0$, F-UCB can match the lower bound up to constant factors.[9] *Clearly, we conducted this analysis employing an* upper bound *to the expected regret of* F-UCB*, which might, in principle, be affected by some analysis artifacts, making it not tight. In Figure 2, we compare the ratio between the* actual *regret obtained by running F-UCB (5 runs) on the proposed FRB example and the instance-dependent lower bound (left) with the ratio between the upper bound and the instance-dependent lower bound computed in Equation (16) (right). We clearly observe that, although the y-scales are different, the behavior confirms a linear dependence of the actual regret of F-UCB on the number of dimensions of the action vector d.*

5. Optimal Asymptotic Instance-Dependent Algorithm

In this section, we provide an algorithm that matches the derived instance-dependent lower bound (Theorem 3.3) in the asymptotic regime. The algorithm, named *Factored Track* (F-Track), whose pseudocode is reported in Algorithm 2, is based on the idea of *tracking the lower bound* (Lattimore & Szepesvari, 2017). The rationale behind the algorithm is that if we want to match the instance-dependent lower bound, we need to properly *coordinate* the choice of the action vectors $\mathbf{a} \in \mathcal{A}$, given that we have a

[9] Indeed, when the suboptimality gaps are close to 0, the instantaneous regret $\prod_{i \in [\![d]\!]} \mu^*_i - \prod_{i \in [\![d]\!]} \mu_{i,a_i(t)}$ approaches the sum of the regrets on each action component $\sum_{i \in [\![d]\!]} (\mu^*_i - \mu_{i,a_i(t)})$.

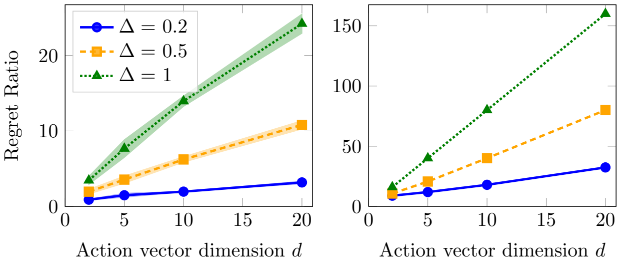


Figure 2. Ratio between the *actual* regret of F-UCB and the instance-dependent lower bound (left) and ratio between the regret upper bound and the instance-dependent lower bound (Equation 16) (right), for different values of d (5 runs, mean ± 2std).

lower bound on the minimum number of pulls for the action components (i, j) (Theorem 3.3). To impose such a structure we must plan in advance our sequence of action vector choices. We devise an algorithm composed of three phases: *warm-up*, *success*, and *recovery*. In the warm-up phase, the algorithm pulls some action vectors in such a way that each action component is pulled at least N_0 times, i.e., $N_{i,j} \geqslant N_0$ (line 3). This can be achieved by round-robing the action components values j of each component i, leading to a number of pulls in the warm-up phase equal to $T_{\text{warm-up}} = N_0 \max_{i \in [\![d]\!]} k_i$. We use these samples to estimate the expected values $\widehat{\mu}_{i,j}(T_{\text{warm-up}})$ and define the confidence interval threshold ϵ_T. Then, we use these values as if they were the true ones $\mu_{i,j}$ to compute the suboptimality gaps $\widehat{\Delta}_{i,j} := \max_{j' \in [\![k_i]\!]} \widehat{\mu}_{i,j'}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(T_{\text{warm-up}})$ (line 6) and, using them, the number of pulls (line 7):

$$\widehat{N}_{i,j} = \frac{2\sigma^2 f_T(1/T)}{\widehat{\Delta}^2_{i,j}}, \quad \forall j \in [\![k_i]\!],\ i \in [\![d]\!]$$

where for every $\delta \in (0, 1)$:

$$f_T(\delta) := \left(1 + \frac{1}{\log T}\right)\left(c \log\log T + \log\left(\frac{1}{\delta}\right)\right),$$

where c is a universal constant and, with them, we compute the number of pulls for every action vector $\widehat{N}_{\mathbf{a}}$ by solving the optimization problem in Theorem 3.3 (line 8). It is worth noting that $f_T(1/T) \approx \log T$ and this form is needed for technical reasons to guarantee that the confidence bounds hold. In the success phase, until we run out of the rounds $t \leqslant T$, we track the lower bound by pulling in a round-robin fashion all arms whose number of pulls $N_{\mathbf{a}}(t) < \widehat{N}_{\mathbf{a}}$ (line 10). If we realize that the estimated expected reward $\widehat{\mu}_{i,j}(t-1)$ are too far from the ones estimated at the end of the warm-up phase $\widehat{\mu}_{i,a_i}(T_{\text{warm-up}})$ based on the threshold ϵ_T, we move to the recovery phase (line 9). In this phase, we play F-UCB until the end of the rounds discarding all the data collected so far (line 12).

The following result shows that F-Track asymptotically matches the lower bound for a proper choice of N_0 and ϵ_T.

Algorithm 2: F-Track.

```
Input :Warm-up sample size $N_0$, Threshold $\epsilon_T$, Action
       component size $k_i,\ \forall i \in [\![d]\!]$,
1  $t \leftarrow 1$
2  while $\min_{i \in [\![d]\!]} \min_{j \in [\![k_i]\!]} N_{i,j}(t) < N_0$ do
3      Pull action vector $\mathbf{a}(t)$ with $a_i(t) = (t-1)$
         mod $k_i + 1$ for all $i \in [\![d]\!]$, $t \leftarrow t + 1$
4  end
5  $T_{\text{warm-up}} \leftarrow t - 1$
6  Estimate the suboptimality gaps $\forall i \in [\![d]\!],\ j \in [\![k_i]\!]$ :
       $\widehat{\Delta}_{i,j} := \max_{j' \in [\![k_i]\!]} \widehat{\mu}_{i,j'}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(T_{\text{warm-up}})$
7  Compute the number of pulls $\widehat{N}_{i,j} = 2\sigma^2 f_T(1/T)\widehat{\Delta}^{-2}_{i,j}$
     for every action component $i \in [\![d]\!]$ and $j \in [\![k_i]\!]$
8  Compute the number of pulls $\widehat{N}_{\mathbf{a}}$ for every action vector
     $\mathbf{a} \in \mathcal{A}$ by solving the LP in Theorem 3.3
9  while $t \leqslant T$ and
     $\max_{i \in [\![d]\!], j \in [\![k_i]\!]} |\widehat{\mu}_{i,j}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(t-1)| \leqslant 2\epsilon_T$
     do
10     Pull action vector $\mathbf{a}(t) \in \arg\min\{N_{\mathbf{a}}(t) : \mathbf{a} \in \mathcal{A}$
         and $N_{\mathbf{a}}(t) \leqslant \widehat{N}_{\mathbf{a}}\}$, $t \leftarrow t + 1$
11 end
12 Discard all data and play F-UCB until $t = T$
```

Theorem 5.1 (Instance-Dependent Upper Bound for F-Track)**.** *For any FRB $\underline{\boldsymbol{\nu}}$,* F-Track *run with:*

$$N_0 = \left\lceil \sqrt{\log T} \right\rceil \quad \text{and} \quad \epsilon_T = \sqrt{\frac{2\sigma^2 f_T(1/\log T)}{N_0}},$$

suffers an expected regret of:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}\left[R_T(\texttt{F-Track}, \underline{\boldsymbol{\nu}})\right]}{\log T} = \underline{C}(\underline{\boldsymbol{\nu}}).$$

6. Discussion and Conclusions

In this paper, we introduced the Factored-Reward Bandits, a novel setting to represent decision-making problems in which the learner is required to perform a set of actions, whose effects can be observed, and the reward is the product of those effects. We characterized the inherent complexity through worst-case and instance-dependent lower bounds, and we discussed the performances of current solutions. To address the regret minimization problem, we proposed two algorithms using the intermediate observations to reduce the complexity of learning in this setting. The first F-UCB is an optimistic solution that we proved minimax optimal (up to logarithmic factors). Such a solution deals with action components independently of the others and we have illustrated how, without coordination, we cannot reach instance-dependent optimality. To overcome this issue, we propose F-Track, an algorithm able to perform planning on the action components, and we proved its asymptotically instance-dependent optimality. As future lines of research, we plan to investigate the possibility of developing an algorithm able to guarantee both non-asymptotic instance-dependent optimality and to consider functions for aggregating intermediate observations different from the product.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 - 15 March 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011.

Agrawal, R. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Broder, J. and Rusmevichientong, P. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Bubeck, S. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille, 2010.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Combes, R., Talebi, M. S., Proutière, A., and Lelarge, M. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2116–2124, 2015.

Combes, R., Magureanu, S., and Proutière, A. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1763–1771, 2017.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Den Boer, A. V. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1): 1–18, 2015.

Feldman, J., Muthukrishnan, S., Pal, M., and Stein, C. Budget optimization in search-based advertising auctions. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, pp. 40–49, 2007.

Hardy, G. H., Littlewood, J. E., and Pólya, G. *Inequalities*. Cambridge University Press, 1952.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. Stochastic rank-1 bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 392–401. PMLR, 2017.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 535–543. PMLR, 2015.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvari, C. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 728–737. PMLR, 2017.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Luttinger, J. and Friedberg, R. A new rearrangement inequality for multiple integrals. *Archive for Rational Mechanics and Analysis*, 61:45–64, 1976.

Magureanu, S., Combes, R., and Proutière, A. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 35, pp. 975–999. JMLR, 2014.

Pinelis, I. Product of three or more independent subgaussian varibles. MathOverflow, 2021.

Vaidya, P. M. Speeding-up linear programming using fast matrix multiplication. In *Annual symposium on foundations of computer science*, pp. 332–337. IEEE Computer Society, 1989.

Wang, Z., Zhang, C., Singh, M. K., Riek, L. D., and Chaudhuri, K. Multitask bandit learning through heterogeneous feedback aggregation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pp. 1531–1539. PMLR, 2021.

Yu, J. Y. and Mannor, S. Unimodal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 41–48. Omnipress, 2011.

Zimmert, J. and Seldin, Y. Factored bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2840–2849, 2018.

A. Examples

In this appendix, we first formalize the example described in Section 1 using the formalism of the FRB setting (Appendix A.1). Then, we present an additional example of a higher dimensional problem that can be generalized by the FRB setting (Appendix A.2).

A.1. Formalization of the Example of Section 1

Consider the case of joint pricing and advertising described in Section 1. In this scenario, at every round $t \in [\![T]\!]$, we must select a vector of dimension $d = 2$. Suppose that the first action component is the advertising *budget*, and the second action component is the selling *price*. We have k_1 advertising budgets over which we want to choose and k_2 prices at which we can sell our item.

At every round t, we select the budget $a_1(t)$ and the price $a_2(t)$. Then, we observe a realization of the impressions we generate due to the budget $a_1(t)$ we invested: $x_1(t) = \mu_{1,a_1(t)} + \epsilon_1(t)$, and a realization of the conversion rate due to the price $a_2(t)$ we set: $x_2(t) = \mu_{2,a_2(t)} + \epsilon_2(t)$.

The reward is equal to $r(t) = a_2(t)x_1(t)x_2(t) - a_1(t)$, corresponding to the return for each sales (the price, considering the turnover as target), multiplied by the fraction of users willing to buy and by the number of customers exposed to the price (i.e., the impressions), minus the budget invested in advertising. Note that the operations of multiplying by the selling price and subtracting the advertising budget do not increase the statistical complexity of the learning problem, as after we select an action, such quantities are deterministic. However, to deal with this more elaborated formulation, we have to take care of it in the choice of the optimal action $\mathbf{a}^*$:

$$\mathbf{a}^* \in \operatorname*{arg\,max}_{\mathbf{a}=(a_1,\,a_2)^\intercal \in \mathcal{A}} a_2 \prod_{i \in [\![2]\!]} \mu_{i,a_i} - a_1. \tag{17}$$

Run this problem on `F-UCB` Moving to the F-UCB, we can easily adapt the formulation of Equation (17) to the one required by the algorithm:

$$\mathbf{a}(t) \in \operatorname*{arg\,max}_{\mathbf{a}=(a_1,\,a_2)^\intercal \in \mathcal{A}} a_2 \prod_{i \in [\![2]\!]} \mathrm{UCB}_{i,a_i}(t) - a_1.$$

In practice, as we have done in Section 4, we can replace the real value with our optimistic estimator. Clearly, the analysis of the regret continues to hold with a multiplicative factor $\max_{a_2 \in [\![k_2]\!]} |a_2|$.

A.2. Additional Example

We present an additional example of problems that can be generalized through the FRB setting related to *manufacturing processes*.

Consider the problem in which we run a manufacturing firm that has to set up the production line for a product. The goal in this scenario is to optimize the following trade-off: maximize the production yield (i.e., the number of items that come out of the production line undamaged) while minimizing the production cost.

Considering the item we want to manufacture, let us define a batch size B and a production line consisting of d stages. Assume that each stage has a $1:1$ production rate (i.e., 1 input corresponds to 1 output). For each stage $i \in [\![d]\!]$, we have to select a method to fulfill the stage among a set of k_i available alternatives. Each alternative will have an aleatoric impact on the percentage of faulty outputs, and a deterministic cost of production.

As such, at every round t, we select an action vector $\mathbf{a}(t) = (a_1(t), a_2(t), \dots, a_d(t))$, with $a_i(t) \in [\![k_i]\!], \forall i \in [\![d]\!]$. At every stage i, we then observe a percentage of undamaged outputs defined as:

$$x_i(t) = \mu_{i,a_i(t)} + \epsilon_i(t),$$

where:

- $\mu_{i,a_i(t)} \in [0, 1]$ is the expected percentage of faultless products due to selecting action $a_i(t)$,

- $\epsilon_i(t)$ is a σ^2-subgaussian random noise, independent conditioned to the past and the other noise realizations $\epsilon_j(t)$ for $j \in [\![d]\!] \backslash \{i\}$.

We can model the reward function as:

$$r(t) = B \prod_{i=1}^{d} x_i(t) - \sum_{i=1}^{d} c_i(a_i(t)),$$

where $c_i(a_i(t))$ is the (deterministic and known) cost associated with the selection of action $a_i(t)$. Observe that B is a known and fixed quantity, and $c_i(a_i(t))$ are deterministic and known to the learner. As such, they do not increase the complexity of the learning problem. For this reason, this scenario can be generalized through the `F-UCB` setting.

B. Additional Related Works

In this section, we discuss the related works from the *action structure* perspective and the works that present a *notion of factored structure*.Then, we compare the most significant related algorithms with our work from the theoretical perspective.

Action Structure Originally, multi-armed bandit frameworks focused on independent arms with no inherent structure (Lai & Robbins, 1985). However, in recent decades, various bandit models with several kinds of structure have emerged, such as linear (Dani et al., 2008; Abbasi-Yadkori et al., 2011), Lipschitz (Agrawal, 1995; Magureanu et al., 2014) and unimodal (Yu & Mannor, 2011) bandits. These contributions aim to incorporate diverse forms of structure into the arms being considered. Combes et al. (2017) introduced a generalization of structured bandits, accommodating a wide range of structural concepts among arms. Their work offers a statistically efficient (at least in the general case) algorithm for handling generic structures, at the expense of solving a semi-infinite linear program at each time step. The necessity of choosing several actions at a time in a structured manner has been widely studied in the field of combinatorial bandits (Cesa-Bianchi & Lugosi, 2012; Kveton et al., 2015; Combes et al., 2015).

Notions of Factored Bandits Among the several kinds of structure, Zimmert & Seldin (2018) is the most similar to the work we propose from the point of view of the action structure, although the two works differ from the feedback perspective. Both works employ an action structure in which an action component a_i is selected for each problem dimension $i \in [\![d]\!]$. The action components are combined with a general function that obeys a *uniform identifiability* assumption under which the performance of each action vector can only improve when any action component is switched with the optimal one. However, in the work of Zimmert & Seldin (2018) the feedback comprises a single observation of the subgaussian reward $r(\mathbf{a}_t)$ applied to the aggregated expected reward, whereas, in our work, the feedback comprises one noisy observation for every action component. This peculiarity of our work implies that the reward obtained as the product over all the dimensions is not subgaussian anymore (Lemma D.1). (Zimmert & Seldin, 2018) generalizes (Katariya et al., 2017) to the case of more than two dimensions.

B.1. Comparison of the Theoretical Results

In Table 1, we summarize our setting with the one of Heavy-Tails Bandits (Bubeck et al., 2013) and the Factored Bandits (Zimmert & Seldin, 2018). We also analyze and compare both our solutions with `Robust-UCB` (Bubeck et al., 2013) and `TEA` (Zimmert & Seldin, 2018) from the instance-dependent point of view. Then, in Table 2 we compare worst-case lower and upper bounds from the worst-case perspective.

Table 1. Comparison with the instance-dependent guarantees of (Bubeck et al., 2013) and (Zimmert & Seldin, 2018). †This result holds for $T \to \infty$. ‡The authors consider $\sigma = 1$.

	Setting Characteristics		Lower Bound	Upper Bound	Match			
	Factored Structure	Intermediate Feedback			σ	d	k	T
Robust-UCB (Bubeck et al., 2013)	✗	✗	$\Omega\left(\underline{\sigma}^2 \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} \frac{\log T}{\Delta_{\mathbf{a}}}\right)$	$\mathcal{O}\left(\overline{\sigma}^2 \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} \frac{\log T}{\Delta_{\mathbf{a}}}\right)$	✗	✓	✓	✓
TEA (Zimmert & Seldin, 2018)	✓	✗	$\Omega\left(\sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} \frac{\log T}{\Delta_{i,j}}\right)^{\dagger}$	$\mathcal{O}\left(\sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} \frac{\log(T \log T) + \log \frac{\log(T \log T)}{\Delta_{i,j}^2}}{\Delta_{i,j}}\right)$	✓‡	✓	✓	✓
This Work F-UCB	✓	✓	Theorem 3.3†	Theorem 4.2	✓	✗	✓	✓
F-Track				Theorem 5.1†	✓	✓	✓	✓

Table 2. Comparison with the worst-case guarantees of (Bubeck et al., 2013) (Zimmert & Seldin 2018 do not provide worst-case bounds).

	Setting Characteristics		Lower Bound	Upper Bound	Match			
	Factored Structure	Intermediate Feedback			σ	d	k	T
Robust-UCB (Bubeck et al., 2013)	✗	✗	$\Omega\left(\underline{\sigma}\sqrt{k^d T}\right)$	$\mathcal{O}\left(\overline{\sigma}\sqrt{k^d T}\right)$	✗	✓	✓	✓
This Work (F-UCB)	✓	✓	Theorem 3.1	Theorem 4.1	✓	✓	✓	✓

C. Proofs and Derivations

In this section, we provide proofs of the statements discussed in the main paper (Section C.1) and some technical lemmas needed in order to prove them (Section C.2).

C.1. Proofs of the Theorems

Theorem 3.1 (Worst-Case Lower Bound)**.** *For every algorithm $\mathfrak{A}$, there exists an FRB $\underline{\boldsymbol{\nu}}$ such that for:*

$$T \geqslant 2\left(1 - 2^{-\frac{1}{d-1}}\right)^{-2} \sigma^2 \max_{i \in [\![d]\!]} k_i = \mathcal{O}\left(\sigma^2 d^2 k\right), \tag{2}$$

$\mathfrak{A}$ suffers an expected cumulative regret of at least:

$$\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geqslant \frac{\sigma}{4\sqrt{2}} \sum_{i \in [\![d]\!]} \sqrt{k_i T}.$$

In particular, if $k_i =: k$ for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geqslant \Omega(\sigma d \sqrt{kT})$.

Proof. Consider an scenario in which $\mu_{\mathbf{a}*} = 1$ and $\Delta_{i,j} \leqslant \overline{\Delta} = 1 - 2^{-1/(d-1)}, \forall i \in [\![d]\!], j \in [\![k_i]\!]$, then Lemma C.3 allow us to rewrite the expected regret as:

$$\begin{aligned}
\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] &= \mathbb{E}\left[\sum_{t \in [\![T]\!]} \left(1 - \prod_{i \in [\![d]\!]} \left(1 - \Delta_{i, a_i(t)}\right)\right)\right] \\
&\geqslant \frac{1}{2} \mathbb{E}\left[\sum_{t \in [\![T]\!]} \sum_{i \in [\![d]\!]} \Delta_{i, a_i(t)}\right] \\
&= \frac{1}{2} \sum_{i \in [\![d]\!]} \mathbb{E}\left[\sum_{t \in [\![T]\!]} \Delta_{i, a_i(t)}\right] \\
&= \frac{1}{2} \sum_{i \in [\![d]\!]} \mathbb{E}\left[R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right],
\end{aligned} \tag{18}$$

where $R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})$ is the expected regret generated by pulling suboptimal arms on the component $i \in [\![d]\!]$. This fact implies that if we take sufficiently small $\Delta_{i,j} < \overline{\Delta}, \forall i \in [\![d]\!], j \in [\![k_i]\!]$, we can analyze the expected regret $R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})$ we pay for each action component $i \in [\![d]\!]$ independently and then summing up the regret we pay as shown above. We will see how the condition of the sufficiently small $\Delta_{i,j}$ implies that we have to add a condition on the minimum time budget T for which this lower bound holds.

We can define a set of $\prod_{i \in [\![d]\!]} k_i$ FRB *base instances* as follows. Given a vector $(h_1, \ldots, h_d)^\intercal \in [\![k_1]\!] \times \cdots \times [\![k_d]\!]$ identifying an instance, we define the expected rewards of such an instance as follows, for $\Delta \in (0, 1/2)$:

$$\mu_{i,j} = \begin{cases} 1 & \text{if } j = h_i \\ 1 - \Delta & \text{if } j \in [\![k_i]\!] \backslash \{h_i\} \end{cases}, \quad \forall i \in [\![d]\!]. \tag{19}$$

We refer as $\underline{\boldsymbol{\nu}}_{(h_1, \ldots, h_d)}$ to the instance in which expected values are characterized by the vector $(h_1, \ldots, h_d)^\intercal \in [\![k_1]\!] \times \cdots \times [\![k_d]\!]$ as in Equation (19).

We now focus on bounding the regret of a single component $i \in [\![d]\!]$. In particular, we focus on component $i = 1$ for the sake of simplicity in the presentation. Then, we can extend the same reasoning to all the others. Let us define a set of *helper instances* which are needed for the analysis. For all the components different from the first, we consider as before a vector $(h_2, \ldots, h_d)^\intercal \in [\![k_2]\!] \times \cdots \times [\![k_d]\!]$ which characterize the instance $\underline{\boldsymbol{\nu}}_{(0, h_2, \ldots, h_d)}$ defined as follows:

$$\mu_{1,j} = 1 - \Delta, \quad \forall j \in [\![k_1]\!] \qquad \mu_{i,j} = \begin{cases} 1 & \text{if } j = h_i \\ 1 - \Delta & \text{if } j \in [\![k_i]\!] \backslash \{h_i\} \end{cases}, \quad \forall i \in [\![2, d]\!]. \tag{20}$$

We now need to introduce some additional objects. Given a vector $(h_1, h_2, \ldots, h_d)^\intercal \in (\{0\} \cup [\![k_1]\!]) \times [\![k_2]\!] \times \cdots \times [\![k_d]\!]$, we call $\mathbb{P}_{(h_1, h_2, \ldots, h_d)}$ the distribution induced by the history of the pulls and the related rewards for the d components over time horizon T in instance $\underline{\boldsymbol{\nu}}_{(h_1,h_2,\dots,h_d)}$. We denote with $\mathbb{P}_h$ for $h \in \{0\} \cup [\![k_1]\!]$ the distribution induced by the history averaged over the other dimensions, formally: $\mathbb{P}_h = \frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \mathbb{P}_{(h,h_2,\dots,h_d)}$, and with $\mathbb{E}_h$ the expectation over $\mathbb{P}_h$.

Coming back to the proof, given the definition of the base instances (Equation 19), the expected regret $\mathbb{E}\left[R_T^{(1)}(\mathfrak{A}, \underline{\boldsymbol{\nu}}_{(h_1,\dots,h_d)})\right]$ related to the first component is given by:

$$\begin{aligned}\mathbb{E}\left[R_T^{(1)}(\mathfrak{A}, \underline{\boldsymbol{\nu}}_{(h_1,\dots,h_d)})\right] &= \Delta \sum_{j\in[\![k_1]\!]\setminus\{h_1\}} \mathbb{E}\left[N_{1,j}(T)\right] \\ &= \Delta \left(T - \mathbb{E}\left[N_{1,h_1}(T)\right]\right).\end{aligned}$$

We now want to use Lemma C.4 in order to obtain the following condition:

$$\frac{1}{k_1} \sum_{h\in[\![k_1]\!]} \mathbb{E}_h[T - N_{1,h}(T)] \geqslant \frac{T}{4}. \tag{21}$$

To apply Lemma C.4, we need an upper bound on the total variation d_{TV} that we can compute $\forall h \in [\![k_1]\!]$ as follows:

$$\begin{aligned}
d_{\text{TV}} &= \frac{1}{2}\left\|\mathbb{P}_0 - \mathbb{P}_h\right\|_1 \\
&= \frac{1}{2}\left\|\frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \left(\mathbb{P}_{(0,h_2,\dots,h_d)} - \mathbb{P}_{(h,h_2,\dots,h_d)}\right)\right\|_1 \\
&\leqslant \frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \frac{1}{2}\left\|\mathbb{P}_{(0,h_2,\dots,h_d)} - \mathbb{P}_{(h,h_2,\dots,h_d)}\right\|_1 && (22)\\
&\leqslant \frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \sqrt{\frac{1}{2} D_{\text{KL}}\left(\mathbb{P}_{(0,h_2,\dots,h_d)} \| \mathbb{P}_{(h,h_2,\dots,h_d)}\right)} && (23)\\
&= \frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \sqrt{\frac{1}{2}\mathbb{E}_{(0,h_2,\dots,h_d)}[N_{1,h}(T)] D_{\text{KL}}\left(p_0 \| p_h\right)} && (24)\\
&= \frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \sqrt{\frac{1}{2}\mathbb{E}_{(0,h_2,\dots,h_d)}[N_{1,h}(T)] \frac{\Delta^2}{2\sigma^2}} && (25)\\
&\leqslant \sqrt{\frac{1}{\prod_{i\in[\![2,d]\!]} k_i} \sum_{(h_2,h_3,\dots,h_d)\in[\![k_2]\!]\times\cdots\times[\![k_d]\!]} \frac{1}{2}\mathbb{E}_{(0,h_2,\dots,h_d)}[N_{1,h}(T)] \frac{\Delta^2}{2\sigma^2}} && (26)\\
&\leqslant \frac{1}{4}\sqrt{\frac{\Delta^2}{2\sigma^2}\mathbb{E}_0[N_{1,h}(T)]}, && (27)
\end{aligned}$$

where line (22) is the triangle inequality for norms, line (23) is due the Pinsker's inequality, line (24) is due to the divergence decomposition lemma (Lattimore & Szepesvári, 2020, Lemma 15.1) considering that all the component different from the first are equal, line (25) is derived by the expression of D_{KL} between Gaussian distributions, line (26) is due to the Jensen's inequality, and line (27) is obtained by marginalizing w.r.t. the first component.

Given this upper bound to the total variation, we can finally apply Lemma C.4 considering $m = k_1$ and $B = \frac{2\sigma^2 k_1}{\Delta^2}$. What we get is:

$$\frac{1}{k_1} \sum_{i\in[\![k_1]\!]} \mathbb{E}_h\left[\frac{2\sigma^2 k_1}{\Delta^2} - N_{1,h}(T)\right] \geqslant \frac{\sigma^2 k_1}{2\Delta^2}. \tag{28}$$

We can now select the value of Δ in order to have in Equation (28) a bound on T:

$$T = \frac{2\sigma^2 k_1}{\Delta^2}.$$

This implies a choice of Δ in the form of:

$$\Delta = \sqrt{\frac{2\sigma^2 k_1}{T}}.$$

Given such a choice of Δ and the bound given by Equation (21), we get that the regret of the first action component can be bounded as:

$$\begin{aligned}\mathbb{E}\left[R_T^{(1)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] &\geqslant \Delta\left(T - \mathbb{E}\left[N_{1,h_1}(T)\right]\right)\\ &\geqslant \sqrt{\frac{2\sigma^2 k_1}{T}}\frac{T}{4}\\ &= \sqrt{\frac{\sigma^2 k_1 T}{8}}\\ &= \frac{1}{2\sqrt{2}}\sigma\sqrt{k_1 T}.\end{aligned}$$

The same reasoning can be done for all the others $d-1$ action components and the bound of Equation (18):

$$\begin{aligned}\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] &\geqslant \frac{1}{2}\sum_{i\in[\![d]\!]}\mathbb{E}\left[R_T^{(i)}(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]\\ &\geqslant \frac{1}{4\sqrt{2}}\sigma\sum_{i\in[\![d]\!]}\sqrt{k_i T}.\end{aligned}$$

The last point needed is to check that the condition of the choices we made on the Δ is compliant for all the dimensions $i \in [\![d]\!]$ with the one of Lemma C.3, i.e., all the Δs are less than $\overline{\Delta}$ defined as:

$$\overline{\Delta} = \sqrt{\frac{2\sigma^2 \max_{i\in[\![d]\!]} k_i}{T}}.$$

This implies a lower bound on the T for which this bound holds:

$$\sqrt{\frac{2\sigma^2 \max_{i\in[\![d]\!]} k_i}{T}} \leqslant 1 - 2^{-1/(d-1)}.$$

Isolating T we get:

$$T \geqslant \frac{2\sigma^2 \max_{i\in[\![d]\!]} k_i}{\left(1 - 2^{-1/(d-1)}\right)^2}.$$

We highlight that the lower bound on the horizon T is quadratic in d. Indeed, for $d \geqslant 2$ we have:

$$\left(1 - 2^{-\frac{1}{d-1}}\right)^{-2} = \left(1 - e^{-\frac{\log 2}{d-1}}\right)^{-2} \leqslant \left(\frac{1}{2(d-1)}\right)^{-2} = 4(d-1)^2 = \mathcal{O}(d^2),$$

having exploited the fact that $1 - e^{-x\log 2} \geqslant x/2$ as $x \in [0,1]$, having set $x = \frac{1}{d-1} \in [0,1]$ for $d \geqslant 2$. Thus, we require a mild (quadratic) condition on $T \geqslant \mathcal{O}(d^2 \max_{i\in[\![d]\!]} k_i)$. We remark that even for standard bandits, the minimax lower bound requires the constraint $T \geqslant \mathcal{O}(k)$, being k the number of arms (Theorem 15.3, Lattimore & Szepesvári, 2020).

This concludes the proof. □

Theorem 3.2 (Worst-Case Lower Bound without Intermediate Observations). *For every algorithm $\mathfrak{A}^\dagger$ that ignores the intermediate observations $\mathbf{x}(t)$ and observes the reward $r(t)$ only, there exists an FRB $\underline{\boldsymbol{\nu}}$ such that for:*

$$T \geqslant 4(\min_{i\in[\![d]\!]} k_i - 1)/d,$$

$\mathfrak{A}^\dagger$ suffers an expected cumulative regret of at least:

$$\mathbb{E}\left[R_T(\mathfrak{A}^\dagger, \underline{\boldsymbol{\nu}})\right] \geqslant \frac{\sigma^d}{8}\sqrt{\frac{(\min_{i\in[\![d]\!]} k_i - 1)T}{d}}.$$

In particular, if $k_i =: k$ for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\mathfrak{A}^\dagger, \underline{\boldsymbol{\nu}})\right] \geqslant \Omega(\sigma^d\sqrt{kT/d})$.

Proof. For simplicity, we consider d even. We consider the following base instance $\underline{\boldsymbol{\nu}}$, parametrized by $\sigma > 1$ and

$\Delta \in [0, 1/4]$ with $\Delta \leqslant \sigma^d$, defined for all $i \in [\![d]\!]$ and $j \in [\![k_i]\!]\backslash\{1\}$:

$$\nu_{i,1} = \begin{cases} \sigma & \text{w.p. } \frac{1}{2} + \frac{\Delta^{1/d}}{2\sigma} \\ -\sigma & \text{w.p. } \frac{1}{2} - \frac{\Delta^{1/d}}{2\sigma} \end{cases}, \qquad \nu_{i,j} = \begin{cases} \sigma & \text{w.p. } \frac{1}{2} \\ -\sigma & \text{w.p. } \frac{1}{2} \end{cases}. \tag{29}$$

It is clear that $\mu_{i,1} = \Delta^{1/d}$ and $\mu_{i,j} = 0$. Consequently, the optimal arm is $(1, \dots, 1)^\top$ with performance $\mu^* = \Delta$ and all the other arms have performance 0. Furthermore, the variance of the suboptimal arm components is given by σ^2 which is also the subgaussian proxy, while for the optimal arm components, the variance is smaller. Consider now for every $i \in [\![d]\!]$:

$$j_i^* \in \operatorname*{arg\,min}_{j \in [\![k_i]\!]\backslash\{1\}} \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[N_{i,j}(T)] \implies \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[N_{i,j_i^*}(T)] \leqslant \frac{T}{k_i - 1}. \tag{30}$$

We construct the alternative instance $\underline{\boldsymbol{\nu}}$ which is equal to $\underline{\boldsymbol{\nu}}'$ except for the the components (i, j_i^*) for $i \in [\![d]\!]$:

$$\nu_{i,j_i^*} = \begin{cases} \sigma & \text{w.p. } \frac{1}{2} + \frac{(2\Delta)^{1/d}}{2\sigma} \\ -\sigma & \text{w.p. } \frac{1}{2} - \frac{(2\Delta)^{1/d}}{2\sigma} \end{cases}, \tag{31}$$

enforcing $\Delta \leqslant \sigma^d/2$. In this alternative instance, the optimal arm is $(j_1^*, \dots j_d^*)^\top$, with performance $(\mu^*)' = 2\Delta$.

We are considering algorithms that do not observe individual components. Therefore, the distribution of the product of the individual components has to be computed. Since they will be used in the computation of the KL-divergence, we just consider the two most dissimilar ones:

$$\nu_\dagger^\otimes = \begin{cases} \sigma^d & \text{w.p. } \frac{1}{2} + \frac{\Delta}{\sigma^d} \\ -\sigma^d & \text{w.p. } \frac{1}{2} - \frac{\Delta}{\sigma^d} \end{cases}, \qquad \nu_\ddagger^\otimes = \begin{cases} \sigma^d & \text{w.p. } \frac{1}{2} \\ -\sigma^d & \text{w.p. } \frac{1}{2} \end{cases}, \tag{32}$$

where the probability of the first case in which we play, for instance, $(1, \dots, 1)^\top$ in the base instance is obtained by the following reasoning: we get σ^d if the number of σ realizations is even (being d even). Thus, we have:

$$\mathbb{P}(\{\sigma^d\}) = \sum_{l=0}^{d} \mathbb{1}\{l \text{ is even}\} \binom{d}{j} \left(\frac{1}{2} + \frac{(2\Delta)^{1/d}}{2\sigma^d}\right)^j \left(\frac{1}{2} - \frac{(2\Delta)^{1/d}}{2\sigma^d}\right)^{d-j} = \frac{1}{2} + \frac{\Delta}{\sigma^d}. \tag{33}$$

The KL divergence becomes, using reverse Pinsker inequality:

$$D_{\text{KL}}(\nu_\dagger^\otimes, \nu_\ddagger^\otimes) \leqslant \frac{1}{\frac{1}{2} - \frac{\Delta}{\sigma^d}} D_{\text{TV}}(\nu_\dagger^\otimes, \nu_\ddagger^\otimes) = 4\left(\frac{\Delta}{\sigma^d}\right)^2 = \frac{4\Delta^2}{\sigma^{2d}}. \tag{34}$$

requiring $\Delta \leqslant \sigma^d/4$.

Let us now lower bound the regret with Bretagnolle-Huber's inequality:

$$\max\{\mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})], \mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}}')]\} \geqslant \frac{\Delta T}{4} \exp\left(-\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}\left[\sum_{t=1}^{T} \mathbb{1}\{\exists i \in [\![d]\!] \,:\, a_i(t) = j_i^*\} D_{\text{KL}}(\nu_{\mathbf{a}(t)}^\otimes \| (\nu')_{\mathbf{a}(t)}^\otimes)\right]\right) \tag{35}$$

$$\geqslant \frac{\Delta T}{4} \exp\left(-\sum_{i \in [\![d]\!]} \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[N_{i,j_i^*}(T)] \frac{4\Delta^2}{\sigma^{2d}}\right) \tag{36}$$

$$\geqslant \frac{\Delta T}{4} \exp\left(-\frac{4dT\Delta^2}{\sigma^{2d}(k^* - 1)}\right), \tag{37}$$

being $k^* = \min_{i \in [\![d]\!]} k_i$. We set $\Delta = \sqrt{\frac{\sigma^{2d}(k^*-1)}{4dT}}$ with $T \geqslant 4(k^* - 1)/d$. □

Theorem 3.3 (Instance-Dependent Lower Bound)**.** *For every consistent*[10] *algorithm $\mathfrak{A}$ and FRB $\underline{\boldsymbol{\nu}}$ with unique optimal arm $\mathbf{a}^* \in \mathcal{A}$ it holds that:*

$$\liminf_{T \to +\infty} \frac{\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]}{\log T} \geqslant \underline{C}(\underline{\boldsymbol{\nu}}), \tag{3}$$

where $\underline{C}(\underline{\boldsymbol{\nu}})$ is defined as the solution to the following optimization problem:

$$\min_{(L_{\mathbf{a}})_{\mathbf{a} \in \mathcal{A}\backslash\{\mathbf{a}^*\}}} \sum_{\mathbf{a} \in \mathcal{A}\backslash\{\mathbf{a}^*\}} L_{\mathbf{a}} \Delta_{\mathbf{a}} \tag{4}$$

[10] An algorithm $\mathfrak{A}$ is *consistent* if for every FRB $\underline{\boldsymbol{\nu}}$ and $p > 0$, it holds that $\limsup_{T \to +\infty} \mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})]/T^p = 0$.

$$\text{s.t.} \quad L_{i,j} = \sum_{\substack{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\} \\ a_i = j}} L_{\mathbf{a}}, \ \forall i \in \llbracket d \rrbracket, \ j \in \llbracket k_i \rrbracket \setminus \{a_i^*\} \tag{5}$$

$$L_{i,j} \geqslant \frac{2\sigma^2}{\Delta_{i,j}^2}, \ \forall i \in \llbracket d \rrbracket, j \in \llbracket k_i \rrbracket \setminus \{a_i^*\} \tag{6}$$

$$L_{\mathbf{a}} \geqslant 0, \quad \forall \mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}. \tag{7}$$

Proof. The proof of this statement is divided into two parts. Part one is dedicated to finding a lower bound on the expected number of pulls of every action component $N_{i,j}(T)$ for each action component $i \in \llbracket d \rrbracket, \ j \in \llbracket k_i \rrbracket \setminus \{a_i^*\}$. Part two is dedicated to understanding how these pulls of the action components can be combined in action vectors in the best way possible.

Part 1: Lower bounding the expected number of pulls for each action component

The proof of the expected number of pulls of a sub-optimal action $j \in \llbracket k_i \rrbracket \setminus \{a_i^*\}$ of action component $i \in \llbracket d \rrbracket$ is inspired by the proof of the asymptotic number of pulls of sub-optimal arms presented in Theorem 16.2 of (Lattimore & Szepesvári, 2020).

We call $\mathcal{M}_{mn}$ the set of distributions referring to the m^{th} component ($m \in \llbracket d \rrbracket$) and the n^{th} arm ($n \in \llbracket k_m \rrbracket$). Then, consider P_{mn} as a specific distribution taken from $\mathcal{M}_{mn}$ to model the reward observations of arm n of component m in a given instance of the setting.

Let $\underline{\boldsymbol{\nu}}$ be an instance of the FRB setting with d components and k_i actions for every $i \in \llbracket d \rrbracket$. We start by selecting a component $\underline{i}$ and a sub-optimal arm $\underline{j}$. Let $\varepsilon > 0 \in \mathbb{R}$ be arbitrary constant. We define a new instance of the FRB setting $\underline{\boldsymbol{\nu}}'$ such that $P'_{ij} = P_{ij}, \forall i \in \llbracket d \rrbracket \setminus \{\underline{i}\}, \forall j \in \llbracket k_i \rrbracket$, and $P'_{\underline{i}j} = P_{\underline{i}j}, \forall j \in \llbracket k_{\underline{i}} \rrbracket \setminus \{\underline{j}\}$, and $P'_{\underline{i},\underline{j}} \in \mathcal{M}_{\underline{i},\underline{j}}$ be such that $D_{KL}(P_{\underline{i},\underline{j}}, P'_{\underline{i},\underline{j}}) \leqslant d_{\underline{i},\underline{j}} + \varepsilon$ and $\mu'_{\underline{i},\underline{j}} > \mu^*_{\underline{i}}$. d_{mn} represents the KL divergence between P_{mn} and P^*_m. The newly defined instance $\underline{\boldsymbol{\nu}}'$ is then identical to $\underline{\boldsymbol{\nu}}$ for every arm of every component different from $\underline{i}$, and in the $\underline{i}^{\text{th}}$ component every arm is identical except for arm $\underline{j}$, which is sub-optimal in $\underline{\boldsymbol{\nu}}$ and is optimal in $\underline{\boldsymbol{\nu}}'$. Following the original proof, we can define, for any event $\mathcal{E}$:

$$\mathbb{P}_{\underline{\boldsymbol{\nu}}}(\mathcal{E}_{\underline{i},\underline{j}}) + \mathbb{P}_{\underline{\boldsymbol{\nu}}'}(\mathcal{E}^{\complement}_{\underline{i},\underline{j}}) \geqslant \frac{1}{2} \exp\left(-\mathbb{E}_{\underline{\boldsymbol{\nu}}} \left[N_{\underline{i},\underline{j}}(T) \right] \left(d_{\underline{i},\underline{j}} + \varepsilon \right) \right).$$

Now, let $\mathcal{E}_{\underline{i},\underline{j}} = \{N_{\underline{i},\underline{j}}(T) > T/2\}$, and let $R_T = R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})$ and $R'_T = R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}}')$. Then:

$$R_T + R'_T \geqslant \frac{T}{2} \left(\mathbb{P}_{\underline{\boldsymbol{\nu}}}(\mathcal{E}_{\underline{i},\underline{j}}) f_{\underline{i}}(\boldsymbol{\mu}) \Delta_{\underline{i},\underline{j}} + \mathbb{P}_{\underline{\boldsymbol{\nu}}'}(\mathcal{E}^{\complement}_{\underline{i},\underline{j}}) f_{\underline{i}}(\boldsymbol{\mu}) (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}}) \right),$$

where $f_{\underline{i}}(\boldsymbol{\mu})$ is obtained by the following observation. Since at every round $t \in \llbracket T \rrbracket$, in which we pull $(\underline{i}, \underline{j})$ we suffer the instantaneous regret in the base instance:

$$\prod_{i \in \llbracket d \rrbracket} \mu_i^* - \mu_{\underline{i},\underline{j}} \prod_{i \in \llbracket d \rrbracket \setminus \{\underline{i}\}} \mu_{i,j(t)} \geqslant (\mu_i^* - \mu_{\underline{i},\underline{j}}) \prod_{i \in \llbracket d \rrbracket \setminus \{\underline{i}\}} \mu_i^* = \Delta_{\underline{i},\underline{j}} \prod_{i \in \llbracket d \rrbracket \setminus \{\underline{i}\}} \mu_i^* \tag{38}$$

and in the alternative instance:

$$\mu'_{\underline{i},\underline{j}} \prod_{i \in \llbracket d \rrbracket \setminus \{\underline{i}\}} \mu_i^* - \prod_{i \in \llbracket d \rrbracket} \mu_{i,j(t)} \geqslant (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}}) \prod_{i \in \llbracket d \rrbracket \setminus \{\underline{i}\}} \mu_i^*, \tag{39}$$

we define:

$$f_{\underline{i}}(\boldsymbol{\mu}) := \prod_{i \in \llbracket d \rrbracket \neq \{\underline{i}\}} \mu_i^*. \tag{40}$$

Since the term $f_{\underline{i}}(\boldsymbol{\mu})$ multiplies both $\Delta_{\underline{i},\underline{j}}$ and $(\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}})$, it is straightforward to continue the original proof and write:

$$R_T + R'_T \geqslant \frac{T}{4} f_{\underline{i}}(\boldsymbol{\mu}) \min\{\Delta_{\underline{i},\underline{j}}, (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}})\} \exp\left(-\mathbb{E}_{\underline{\boldsymbol{\nu}}} \left[N_{\underline{i},\underline{j}}(T) \right] \left(d_{\underline{i},\underline{j}} + \varepsilon \right) \right).$$

Rearranging and dividing by $\log T$, we obtain:

$$\frac{\mathbb{E}_{\underline{\boldsymbol{\nu}}}[N_{\underline{i},\underline{j}}(T)]}{\log(T)} \geqslant \frac{\log(T) + \log\left(\frac{f_{\underline{i}}(\boldsymbol{\mu})}{4} \min\{\Delta_{\underline{i},\underline{j}}, (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}})\} \right) - \log(R_T + R'_T)}{(d_{\underline{i},\underline{j}} + \varepsilon) \log(T)} \tag{41}$$

$$= \frac{1}{d_{\underline{i},\underline{j}} + \varepsilon} + \frac{\log\left(\frac{f_{\underline{i}}(\boldsymbol{\mu})}{4} \min\{\Delta_{\underline{i},\underline{j}}, (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}})\}\right) - \log(R_T + R'_T)}{(d_{\underline{i},\underline{j}} + \varepsilon)\log(T)} \tag{42}$$

$$\geqslant \frac{2\sigma^2}{\Delta^2_{\underline{i},\underline{j}}} - h_{\underline{i},\underline{j}}(T), \tag{43}$$

by letting $\varepsilon \to 0$, having exploited the expression of KL-divergence between Gaussians and having set:

$$h_{\underline{i},\underline{j}}(T) := \max\left\{0, \frac{\log\left(\frac{f_{\underline{i}}(\boldsymbol{\mu})}{4} \min\{\Delta_{\underline{i},\underline{j}}, (\mu'_{\underline{i},\underline{j}} - \mu^*_{\underline{i}})\}\right) - \log(R_T + R'_T)}{d_{\underline{i},\underline{j}} \log T}\right\}. \tag{44}$$

Notice that $\limsup_{T\to+\infty} h_{i,j}(T) = 0$ under consistency.

Now, iterating this reasoning over $\underline{i} \in [\![d]\!]$ and over $\underline{j} \in [\![k_i]\!]$, we get the lower bound on the expected number of pulls for all the arms of all the action components.

Part 2: Understanding how the pulls we have to perform on the action components can be combined

From Part 1 of this proof, we have a result on the expectation of the minimum number of pulls. We can now define the quantity:

$$L_{i,j}(T) := \frac{\mathbb{E}[N_{i,j}(T)]}{\log T}, \qquad \forall i \in [\![d]\!],\ j \in [\![k_i]\!].$$

This quantity can be lower bounded as:

$$L_{i,j}(T) \geqslant \frac{2\sigma^2}{\Delta^2_{ij}} - h_{i,j}(T), \qquad \forall i \in [\![d]\!],\ j \in [\![k_i]\!]\backslash\{a^*_i\}.$$

Now, we want to understand how these pulls of the action's suboptimal components influence the regret. We chose to look at the asymptotic expected regret, defined as follows:

$$\frac{\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]}{\log T} = \sum_{\mathbf{a}\in\mathcal{A}} \frac{\mathbb{E}\left[N_{\mathbf{a}}(T)\right]}{\log T} \Delta_{\mathbf{a}},$$

and we denote:

$$L_{\mathbf{a}}(T) := \frac{\mathbb{E}[N_{\mathbf{a}}(T)]}{\log T},\ \ \forall \mathbf{a} \in \mathcal{A}.$$

The regret becomes defined as:

$$\frac{\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]}{\log T} = \sum_{\mathbf{a}\in\mathcal{A}} L_{\mathbf{a}}(T)\Delta_{\mathbf{a}},$$

Now, we want to look at how the pulls of the action vectors $L_{\mathbf{a}}$ and the ones of the action components are related. We can easily observe that the following relation occurs:

$$L_{i,j}(T) = \sum_{\mathbf{a}\in\mathcal{A}:a_i=j} L_{\mathbf{a}}(T),\ \ \forall i \in [\![d]\!],\ j \in [\![k_i]\!].$$

Given that, we can write an optimization problem in which we search for the best combination of pulls of the action vector satisfying the constraints on the minimum number of pulls of the action components.

$$\min_{L_{\mathbf{a}}(T), L_{i,j}(T)} \sum_{\mathbf{a}\in\mathcal{A}\backslash\{\mathbf{a}^*\}} L_{\mathbf{a}}(T)\Delta_{\mathbf{a}} \tag{45}$$

$$\text{s.t. } L_{i,j}(T) = \sum_{\mathbf{a}\in\mathcal{A}:a_i=j} L_{\mathbf{a}}(T), \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!]\backslash\{a^*_i\} \tag{46}$$

$$L_{i,j}(T) \geqslant \frac{2\sigma^2}{\Delta^2_{i,j}} - h_{i,j}(T), \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!]\backslash\{a^*_i\} \tag{47}$$

$$L_{\mathbf{a}}(T) \geqslant 0, \quad \forall \mathbf{a} \in \mathcal{A}\backslash\{\mathbf{a}^*\}. \tag{48}$$

Now, to simplify notation, we define $x(\mathbf{a}) = L_{\mathbf{a}}(T)$, remove the variables $L_{i,j}$ since constraint (47) will be satisfied with equality, and reformulate in the unconstrained form using the indicator function $I_{\mathcal{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ +\infty & \text{otherwise} \end{cases}$:

$$\inf_{x(\boldsymbol{a})} f_T(x) := \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} x(\mathbf{a}) \Delta_{\mathbf{a}} + \sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} I_{\mathbb{R}_{\geqslant 0}} \left(\sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} x(\boldsymbol{a}) - \frac{2\sigma^2}{\Delta_{i,j}^2} + h_{i,j}(T) \right) + \sum_{\mathbf{a} \in \mathcal{A}} I_{\mathbb{R}_{\geqslant 0}}(x(\mathbf{a})). \tag{49}$$

With this notation, we want to characterize the value of the optimization problem as the horizon T grows to infinity, i.e., $\liminf_{T \to +\infty} \inf_{x(\boldsymbol{a})} f_T(x)$. Notice that this is exactly what we need to obtain a lower bound to $\liminf_{T \to +\infty} \frac{\mathbb{E}[R_T(\mathfrak{A}, \boldsymbol{\nu})]}{\log T}$.

In the following, we show that:

$$\liminf_{T \to +\infty} \inf_{x(\boldsymbol{a})} f_T(x) = \inf_{x(\boldsymbol{a})} f_\infty(x), \tag{50}$$

where f_∞ is defined as follows:

$$f_\infty(x) := \sum_{\mathbf{a} \in \mathcal{A}} x(\mathbf{a}) \Delta_{\mathbf{a}} + \sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} I_{\mathbb{R}_{\geqslant 0}} \left(\sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} x(\boldsymbol{a}) - \frac{2\sigma^2}{\Delta_{i,j}^2} \right) + \sum_{\mathbf{a} \in \mathcal{A}} I_{\mathbb{R}_{\geqslant 0}}(x(\mathbf{a})), \tag{51}$$

corresponding to the optimization problem in which we remove the $h_{i,j}(T)$ function from the right-hand side of the constraint. First of all, we observe that for every x and T, we have that $f_T(x) \leqslant f_\infty(x)$. It follows that $\inf_{x(\mathbf{a})} f_T(x) \leqslant \inf_{x(\mathbf{a})} f_\infty(x)$ and, consequently, $\liminf_{T \to +\infty} \inf_{x(\boldsymbol{a})} f_T(x) \leqslant \inf_{x(\boldsymbol{a})} f_\infty(x)$. Thus, it remains to prove that $\liminf_{T \to +\infty} \inf_{x(\boldsymbol{a})} f_T(x) \geqslant \inf_{x(\boldsymbol{a})} f_\infty(x)$. Since the optimization problem is linear and feasible (for sufficiently large T), there must exist x_T^* such that $\inf_{x(\boldsymbol{a})} f_T(x) = f_T(x_T^*)$ for every finite T, but also for $T = \infty$. Now, consider for a fixed x:

$$\liminf_{T \to +\infty} f_T(x) = \sum_{\mathbf{a} \in \mathcal{A}} x(\mathbf{a}) \Delta_{\mathbf{a}} + \sum_{\mathbf{a} \in \mathcal{A}} I_{\mathbb{R}_{\geqslant 0}}(x(\mathbf{a})) + \liminf_{T \to +\infty} \sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} I_{\mathbb{R}_{\geqslant 0}} \left(\sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} x(\boldsymbol{a}) - \frac{2\sigma^2}{\Delta_{i,j}^2} + h_{i,j}(T) \right) \tag{52}$$

$$\geqslant \sum_{\mathbf{a} \in \mathcal{A}} x(\mathbf{a}) \Delta_{\mathbf{a}} + \sum_{\mathbf{a} \in \mathcal{A}} I_{\mathbb{R}_{\geqslant 0}}(x(\mathbf{a})) + \sum_{i \in [\![d]\!]} \sum_{j \in [\![k_i]\!] \setminus \{a_i^*\}} \liminf_{T \to +\infty} I_{\mathbb{R}_{\geqslant 0}} \left(\sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} x(\boldsymbol{a}) - \frac{2\sigma^2}{\Delta_{i,j}^2} + h_{i,j}(T) \right) \tag{53}$$

$$= f_\infty(x), \tag{54}$$

uniformly since $\limsup_{T \to +\infty} h_{i,j}(T) = 0$ and $I_{\mathbb{R}_{\geqslant 0}}$ is a decreasing function in its argument, having also exploited that $\liminf_n (a_n + b_n) \geqslant \liminf_n a_n + \liminf_n b_n$. Indeed, let $c = \sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} x(\boldsymbol{a}) - \frac{2\sigma^2}{\Delta_{i,j}^2}$ and $y_T = h_{i,j}(T)$, we have to compute $\liminf_{T \to +\infty} I_{\mathbb{R}_{\geqslant 0}}(c + y_T)$. Since $0 \leqslant y_T$ and $\limsup_{T \to +\infty} y_T = 0$, we have $\lim_{T \to +\infty} y_T = 0$. If $c \neq 0$, there exists $T(c)$ such that for $T \geqslant T(c)$, we have that $y_T \leqslant |c|/2$. Consequently, $\liminf_{T \to +\infty} I_{\mathbb{R}_{\geqslant 0}}(c + y_T) = I_{\mathbb{R}_{\geqslant 0}}(c)$. If, instead, $c = 0$, we have to compute $\lim_{T \to +\infty} I_{\mathbb{R}_{\geqslant 0}}(y_T)$; being $I_{\mathbb{R}_{\geqslant 0}}$ right continuous and $y_T \geqslant 0$ we have that $\lim_{T \to +\infty} I_{\mathbb{R}_{\geqslant 0}}(y_T) = 0$.

This, combined with the fact $f_T(x) \leqslant f_\infty(x)$ leads to $\liminf_{T \to +\infty} f_T(x) = f_\infty(x)$, uniformly. Thus, we have that for every $\varepsilon > 0$ there exists $T(\varepsilon) > 0$ such that for every $T \geqslant T_0(\varepsilon)$ we have uniformly:

$$\left| \inf_{T' \geqslant T} f_{T'}(x) - f_\infty(x) \right| \leqslant \varepsilon. \tag{55}$$

Consequently, we have:

$$\inf_{T' \geqslant T} \inf_{x(\mathbf{a})} f_{T'}(x) = \inf_{T' \geqslant T} f_{T'}(x_{T'}^*) \geqslant f_\infty(x_{T'}^*) - \varepsilon \geqslant f_\infty(x_\infty^*) - \varepsilon = \inf_{x(\mathbf{a})} f_\infty(x(\mathbf{a})) - \varepsilon. \tag{56}$$

This concludes the proof. □

Theorem 3.4 (Instance-Dependent Lower Bound (Explicit)). *Let $\underline{C}(\boldsymbol{\nu})$ be the solution of the optimization problem of Theorem 3.3. It holds that:*

$$\underline{C}(\boldsymbol{\nu}) = \sum_{\ell=1}^{K-d} \left(M_{\boldsymbol{\pi}(\ell)} - M_{\boldsymbol{\pi}(\ell-1)} \right) \Delta_{\boldsymbol{\alpha}_\ell},$$

that can be computed in $\mathcal{O}(\sum_{i \in [\![d]\!]} k_i \log k_i)$.

Proof. Let $M = \max_{i \in [\![d]\!]} M_{i,k_i-1}$. For every $i \in [\![d]\!]$, let us define a non-negative function function $f_i : \mathbb{R} \to \{\mu_{i,j}\}_{j \in [\![k_i]\!]} \cup \{0\}$ such that:

$$\int_{\mathbb{R}} \mathbb{1}\{f_i(x) = \mu_{i,j}\} \mathrm{d}x = L_{i,j} \qquad \forall j \in [\![k_i]\!] \backslash \{a_i^*\}, \tag{57}$$

$$\int_{\mathbb{R}} \mathbb{1}\{f_i(x) = \mu_{i,a_i^*}\} \mathrm{d}x = M - M_{i,k_i-1}. \tag{58}$$

Clearly, f_i is not uniquely defined. Any function f_i satisfying these conditions is measurable (by definition, since the pre-image of any $\mathcal{Y} \subseteq \{\mu_{i,j}\}_{j \in [\![k_i]\!]} \cup \{0\}$ is measurable) and correspond to a possible arrangement of a proportion of pulls of the arm components of dimension i. Specifically, all functions satisfying these conditions are called "equimesurable" meaning that for every f_i, g_i fulfilling the conditions, we have that $\{x : f_i(x) \geqslant y\} = \{x : g_i(x) \geqslant y\}$ for every $y \in \mathbb{R}$. We call this set of functions $\mathcal{F}_i$.

A possible arrangement of the proportion of the pulls for component $i \in [\![d]\!]$, corresponds to a function $f_i \in \mathcal{F}_i$ such that $f_i(x) = 0$ for $x < 0$ or $x > M$. Thus, to minimize the regret as in the optimization problem of Theorem 3.3, we maximize the reward as follows:

$$\sup_{f_i \in \mathcal{F}_i,\, f_i(x) = 0 \text{ for } x < 0 \text{ or } x > M,\, i \in [\![d]\!]} \int_{\mathbb{R}^d} \prod_{i \in [\![d]\!]} f_i(x_i) \mathrm{d}x_i \leqslant \sup_{f_i \in \mathcal{F}_i,\, i \in [\![d]\!]} \int_{\mathbb{R}^d} \prod_{i \in [\![d]\!]} f_i(x_i) \mathrm{d}x_i. \tag{59}$$

Let f_i^* be the *symmetric decreasing rearrangement* of f_i for every $i \in [\![d]\!]$, which, in our specific case, is a piecewise constant symmetric function. Define $x_0 = 0$, $x_{i,1} = (M - M_{i,k_i-1})/2$, $x_{i,l+1} = x_{i,l} + L_{i,\pi_i(k_i-l)}/2$ for $l \in [\![k_i]\!]$, we have:

$$f_i^*(x) = \sum_{l \in [\![k_i]\!]} \mu_{i,\pi_i(k_i-l+1)} \mathbb{1}\{|x| \in [x_{i,l-1}, x_{i,l})\}. \tag{60}$$

From the rearrangement inequality for multiple integrals (Luttinger & Friedberg, 1976), we have:

$$\sup_{f_i \in \mathcal{F}_i,\, i \in [\![d]\!]} \int_{\mathbb{R}^d} \prod_{i \in [\![d]\!]} f_i(x_i) \mathrm{d}x_i = \int_{\mathbb{R}^d} \prod_{i \in [\![d]\!]} f_i^*(x_i) \mathrm{d}x_i. \tag{61}$$

Let us observe that the product of $\int_{\mathbb{R}^d} \prod_{i \in [\![d]\!]} f_i^*(x_i) \mathrm{d}x_i$ actually leads to the solution depicted in the statement of the theorem.

Concerning the computational complexity, we observe that it is dominated by the sorting in each dimension $i \in [\![d]\!]$. □

Theorem 4.1 (Worst-Case Upper Bound for F-UCB)**.** *For any FRB $\underline{\boldsymbol{\nu}}$, F-UCB with $\alpha > 2$ suffers an expected regret bounded as:*

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant 4\sigma \sum_{i \in [\![d]\!]} \sqrt{\alpha k_i T \log T} + g(\alpha) \sum_{i \in [\![d]\!]} k_i,$$

where $g(\alpha) = \widetilde{\mathcal{O}}\left((\alpha - 2)^{-2}\right)$.[11] *In particular, if $k_i =: k$, for every $i \in [\![d]\!]$, we have $\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant \widetilde{\mathcal{O}}(\sigma d \sqrt{kT})$.*

Proof. The proof is composed of two parts. In the first part, we define the probability, given the chosen confidence bounds, that the good event holds, i.e., the probability that all the confidence bounds are valid. The goal is to find an upper bound on the probability that the good event does not hold along the whole time horizon T. In the second part, we aim to characterize the regret under the good event for a specific round $t \in [\![T]\!]$. Finally, we join the two parts to find an upper bound on the expected cumulative regret.

Part 1: Upper bounding the bad event over time horizon T

We start by defining our good event $\mathcal{E}_t$ at round $t \in [\![T]\!]$, which implies that all the confidence bounds of interest hold, i.e., we are not making a severe underestimate of the expected value of the optimal action components, and severely overestimating the expected values of the suboptimal ones. Formally:

$$\mathcal{E}_t := \left\{ \forall i \in [\![d]\!], \forall a_i \in [\![k_i]\!] \backslash \{a_i^*\} : \widehat{\mu}_{i,a_i}(t) - \mu_{i,a_i} \leqslant \sigma \sqrt{\frac{\alpha \log t}{N_{i,a_i}(t)}} \right\}$$

[11] The complete expression is reported in the proof.

$$\cap \left\{ \forall i \in [\![d]\!] : \mu_{i,a_i^*} - \widehat{\mu}_{i,a_i^*}(t) \leqslant \sigma \sqrt{\frac{\alpha \log t}{N_{i,a_i^*}(t)}} \right\}.$$

We now want to find an upper bound of the probability of the bad event $\mathcal{E}_t^{\complement}$:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_t^{\complement}\right) &\leqslant \mathbb{P}\left(\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\} : \widehat{\mu}_{i,a_i}(t) - \mu_{i,a_i} > \sigma \sqrt{\frac{\alpha \log t}{N_{i,a_i}(t)}} \right) + \\
&\quad + \mathbb{P}\left(\exists i \in [\![d]\!] : \mu_{i,a_i^*} - \widehat{\mu}_{i,a_i^*}(t) > \sigma \sqrt{\frac{\alpha \log t}{N_{i,a_i^*}(t)}} \right) \\
&\leqslant \mathbb{P}\left(\underbrace{\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\}, \exists s \in [\![t]\!] : \widehat{\mu}_{i,a_i}[s] - \mu_{i,a_i(t)} > \sigma \sqrt{\frac{\alpha \log t}{s}}}_{\text{(A)}} \right) \\
&\quad + \mathbb{P}\left(\underbrace{\exists i \in [\![d]\!], \exists s \in [\![t]\!] : \mu_{i,a_i^*} - \widehat{\mu}_{i,a_i^*}[s] > \sigma \sqrt{\frac{\alpha \log t}{s}}}_{\text{(B)}} \right),
\end{aligned} \tag{62}$$

having highlighted with the symbols $\widehat{\mu}_{i,a_i}[s]$ and $\widehat{\mu}_{i,a_i^*}[s]$ the dependence of the estimators on the number of pulls s. We now bound (A) and (B) separately. Similar to the proof of Theorem 2.2 proposed by Bubeck (2010), we use a peeling argument together with Hoeffding's maximal inequality. We apply the peeling argument with a geometric grid over the time interval $[1, t]$ to bound the probability of term (A). Given $\beta \in (0, 1)$, we note that if $s \in \{1, \dots, t\}$, then $\exists j \in \left\{0, \dots, \frac{\log t}{\log 1/\beta}\right\} : \beta^{j+1} t < s \leqslant \beta^j t$. As such, we obtain:

$$\begin{aligned}
\mathbb{P}((\text{A})) &= \mathbb{P}\left(\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\}, \exists s \in [\![t]\!] : \widehat{\mu}_{i,a_i}[s] - \mu_{i,a_i} > \sigma \sqrt{\frac{\alpha \log t}{s}} \right) \\
&= \mathbb{P}\left(\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\}, \exists s \in [\![t]\!] : \sum_{l=1}^{s} \left(x_{i,a_i}[l] - \mu_{i,a_i(t)} \right) > \sigma \sqrt{\alpha s \log t} \right) \\
&\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\}, \exists s : \beta^{j+1} t < s \leqslant \beta^j t, \sum_{l=1}^{s} \left(x_{i,a_i}[l] - \mu_{i,a_i(t)} \right) > \sigma \sqrt{\alpha s \log t} \right) \\
&\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists i \in [\![d]\!], \exists a_i \in [\![k_i]\!] \backslash \{a_i^*\}, \exists s : \beta^{j+1} t < s \leqslant \beta^j t, \sum_{l=1}^{s} \left(x_{i,a_i}[l] - \mu_{i,a_i(t)} \right) > \sigma \sqrt{\alpha \beta^{j+1} t \log t} \right),
\end{aligned}$$

having denoted with $x_{i,a_i}[l]$ the l-sample used to compute the sample mean $\widehat{\mu}_{i,a_i}[s]$. Applying a union bound on the summations on i and a_i, and Hoeffding's maximal inequality, we obtain:

$$\begin{aligned}
\mathbb{P}((\text{A})) &\leqslant \sum_{i \in [\![d]\!]} \sum_{a_i \in [\![k_i]\!] \backslash \{a_i^*\}} \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(- \frac{\left(\sqrt{\sigma^2 \alpha \beta^{j+1} t \log t} \right)^2}{2 \sigma^2 \beta^j t} \right) \\
&= \sum_{i \in [\![d]\!]} \sum_{a_i \in [\![k_i]\!] \backslash \{a_i^*\}} \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(- \frac{\alpha \beta \log t}{2} \right) \\
&= \sum_{i \in [\![d]\!]} \sum_{a_i \in [\![k_i]\!] \backslash \{a_i^*\}} \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} t^{-\frac{\alpha \beta}{2}} \\
&\leqslant \sum_{i \in [\![d]\!]} \sum_{a_i \in [\![k_i]\!] \backslash \{a_i^*\}} \left(\frac{\log t}{\log \frac{1}{\beta}} + 1 \right) t^{-\frac{\alpha \beta}{2}}.
\end{aligned}$$

Applying the same procedure, we can bound the probability of term (B) in Equation (62) to obtain:

$$\mathbb{P}\left((\mathrm{B})\right) \leqslant \sum_{i \in \llbracket d \rrbracket} \left(\frac{\log t}{\log \frac{1}{\beta}} + 1 \right) t^{-\frac{\alpha\beta}{2}}.$$

As such, we can write the upper bound of the probability of the bad event as:

$$\mathbb{P}\left(\mathcal{E}_t^{\complement}\right) = \mathbb{P}\left((\mathrm{A})\right) + \mathbb{P}\left((\mathrm{B})\right) \leqslant \sum_{i \in \llbracket d \rrbracket} k_i \left(\frac{\log t}{\log \frac{1}{\beta}} + 1 \right) t^{-\frac{\alpha\beta}{2}}.$$

Let us now bound the sum of the probabilities of the bad event over the horizon T:

$$\begin{aligned}
\sum_{t \in \llbracket T \rrbracket} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) &\leqslant \sum_{i \in \llbracket d \rrbracket} k_i \sum_{t \in \llbracket T \rrbracket} \left(\frac{\log t}{\log \frac{1}{\beta}} + 1 \right) t^{-\frac{\alpha\beta}{2}} \\
&\leqslant \sum_{i \in \llbracket d \rrbracket} k_i \int_1^T \left(\frac{\log t}{\log \frac{1}{\beta}} + 1 \right) t^{-\frac{\alpha\beta}{2}} \mathrm{d}t && (63) \\
&= \sum_{i \in \llbracket d \rrbracket} k_i \left(\left[\left(\frac{\log t}{\log 1/\beta} + 1 \right) \left(\frac{2}{2-\alpha\beta} t^{1-\frac{\alpha\beta}{2}} \right) \right]_1^{+\infty} - \frac{4}{(2-\alpha\beta)\log 1/\beta} \int_1^{+\infty} t^{-\frac{\alpha\beta}{2}} \mathrm{d}t \right) && (64) \\
&= \sum_{i \in \llbracket d \rrbracket} k_i \left(-\frac{2}{2-\alpha\beta} - \frac{4}{(2-\alpha\beta)^2 \log(1/\beta)} \left[t^{1-\frac{\alpha\beta}{2}} \right]_1^{+\infty} \right) && (65) \\
&= \sum_{i \in \llbracket d \rrbracket} k_i \left(-\frac{2}{2-\alpha\beta} + \frac{4}{(2-\alpha\beta)^2 \log(1/\beta)} \right), && (66)
\end{aligned}$$

where line (63) is obtained by bounding the summation with the integral, line (64) is obtained via integration by parts, and the first term of line (65) is obtained by imposing $\alpha\beta > 2$. Substituting now $\beta = \frac{4}{\alpha+2}$, which verifies $\beta \in (0, 1)$ if $\alpha > 2$, we obtain:

$$\sum_{t \in \llbracket T \rrbracket} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) \leqslant \left(\frac{\alpha+2}{\alpha-2} + \frac{(\alpha+2)^2}{(\alpha-2)^2} \frac{1}{\log\left(\frac{\alpha+2}{4}\right)} \right) \sum_{i \in \llbracket d \rrbracket} k_i = \widetilde{\mathcal{O}}\left((\alpha-2)^2\right) \sum_{i \in \llbracket d \rrbracket} k_i.$$

Part 2: Upper bounding the instantaneous regret at time t under the good event

We can now bound the instantaneous regret at time t supposing the good event holds. We define the regret R_t at time t as the difference in expectation between the optimal action and the one performed by F-UCB, formally:

$$\begin{aligned}
R_t &= \prod_{i \in \llbracket d \rrbracket} \mu_i^* - \prod_{i \in \llbracket d \rrbracket} \mu_{i,a_i(t)} && (67) \\
&= \sum_{l \in \llbracket d \rrbracket} \underbrace{\prod_{i \in \llbracket l-1 \rrbracket} \mu_l^*}_{\in[0,1]} \left(\mu_l^* - \mu_{l,a_l(t)} \right) \underbrace{\prod_{i \in \llbracket l+1, d \rrbracket} \mu_{i,a_i(t)}}_{\in[0,1]} && (68) \\
&\leqslant \sum_{l \in \llbracket d \rrbracket} \left(\mu_l^* - \mu_{l,a_l(t)} \right) && (69) \\
&= \sum_{l \in \llbracket d \rrbracket} \left(\mu_l^* - \mu_{l,a_l(t)} \pm \mathrm{UCB}_{l,a_l(t)}(t) \right) && (70) \\
&\leqslant \sum_{l \in \llbracket d \rrbracket} \left(\mathrm{UCB}_{l,a_l(t)}(t) - \mu_{l,a_l(t)} \right) && (71) \\
&= \sum_{l \in \llbracket d \rrbracket} \left(\widehat{\mu}_{l,a_l(t)}(t) + \beta_{l,a_l(t)}(t) - \mu_{l,a_l(t)} \right) && (72) \\
&\leqslant 2 \sum_{l \in \llbracket d \rrbracket} \beta_{l,a_l(t)}(t), && (73)
\end{aligned}$$

where line (68) is obtained by summing and subtracting all mixed terms, line (69) follows from bounding the left and right products with 1 being all factors (including the middle one) made of non-negative terms, line (71) comes from the optimism under the good event, having denoted with $\beta_{l,a_l}(t)$ the exploration bonus.

Upper bound of the expected cumulative regret $R(\texttt{F-UCB}, T)$

Recalling that we call R_t the instantaneous regret under the good event, can now compute an upper bound on the expected cumulative regret as:

$$
\begin{aligned}
\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] &\leqslant \sum_{t\in[\![T]\!]} \left(1 \cdot \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + R_t \cdot \mathbb{P}\left(\tilde{\mathcal{E}}_t\right)\right) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + \sum_{t\in[\![T]\!]} R_t \cdot \mathbb{P}\left(\tilde{\mathcal{E}}_T\right) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + \sum_{t\in[\![T]\!]} R_t \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + \sum_{t\in[\![T]\!]} 2 \sum_{i\in[\![d]\!]} \beta_{i,a_i(t)}(t) \\
&= \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2 \sum_{t\in[\![T]\!]} \sum_{i\in[\![d]\!]} \sigma \sqrt{\frac{\alpha \log t}{N_{i,a_i(t)}}} \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{t\in[\![T]\!]} \sum_{i\in[\![d]\!]} \sqrt{\frac{1}{N_{i,a_i(t)}}} \\
&= \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sum_{a_i\in[\![k_i]\!]} \sum_{j\in[\![N_{i,a_i}(T)]\!]} \sqrt{\frac{1}{j}} \qquad (74) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sum_{a_i\in[\![k_i]\!]} \sum_{j\in[\![T/k_i]\!]} \sqrt{\frac{1}{j}} \qquad (75) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sum_{a_i\in[\![k_i]\!]} \int_1^{T/k_i} \sqrt{\frac{1}{j}} \text{dj} \qquad (76) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sum_{a_i\in[\![k_i]\!]} \left(1 + 2\sqrt{\frac{T}{k_i}} - 2\right) \\
&\leqslant \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 2\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sum_{a_i\in[\![k_i]\!]} 2\sqrt{\frac{T}{k_i}} \\
&= \sum_{t\in[\![T]\!]} \mathbb{P}\left(\mathcal{E}_t^{\complement}\right) + 4\sigma\sqrt{\alpha \log T} \sum_{i\in[\![d]\!]} \sqrt{k_i T} \\
&\leqslant \left(\frac{\alpha+2}{\alpha-2} + \frac{(\alpha+2)^2}{(\alpha-2)^2} \frac{1}{\log\left(\frac{\alpha+2}{4}\right)}\right) \sum_{i\in[\![d]\!]} k_i + 4\sigma\sqrt{\alpha T \log T} \sum_{i\in[\![d]\!]} \sqrt{k_i}.
\end{aligned}
$$

where line (74) is obtained by rewriting the series over the arms and the number of pulls for each arm, line (75) is derived by considering the worst case, i.e., when all the arms are pulled equally (this is the worst case because we are looking at a concave function), and line (76) is obtained by bounding the summation with the corresponding integral. This concludes the proof.

□

Theorem 4.2 (Instance-Dependent Upper Bound for $\texttt{F-UCB}$)**.** *For a given FRB $\underline{\boldsymbol{\nu}}$, $\texttt{F-UCB}$ with $\alpha > 2$ suffers an expected regret bounded as:*

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leqslant \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}}),$$

where $\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})$ is defined as the solution to the following optimization problem (where $g(\alpha) = \widetilde{\mathcal{O}}((\alpha-2)^{-2})$):

$$\max_{(N_{\mathbf{a}})_{\mathbf{a}\in\mathcal{A}}} \sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}} N_{\mathbf{a}}\Delta_{\mathbf{a}} \tag{9}$$

$$\text{s.t.} \quad N_{i,j} = \sum_{\substack{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}\\ a_i=j}} N_{\mathbf{a}},\ \forall i\in[\![d]\!],\ j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{10}$$

$$N_{i,j} \leqslant \frac{4\alpha\sigma^2 \log T}{\Delta_{i,j}^2} + g(\alpha),\ \forall i\in[\![d]\!],\ j\in[\![k_i]\!]\setminus\{a_i^*\} \tag{11}$$

$$\sum_{\mathbf{a}\in\mathcal{A}} N_{\mathbf{a}} = T \tag{12}$$

$$N_{\mathbf{a}} \geqslant 0,\ \forall \mathbf{a}\in\mathcal{A} \tag{13}$$

Proof. The proof of this statement is divided into two parts. The first part is dedicated to finding an upper bound on the expected number of pulls for each action component N_{ij}. The second part is dedicated to understanding how these pulls can be combined to find an upper bound on the regret.

Part 1: Upper bounding the expected number of pulls for each action component

The proof of the expected number of pulls for σ^2-subgaussian variables comprises three parts, extending and following the proof of Theorem 2.2 proposed by Bubeck (2010).

Given an instance $\underline{\boldsymbol{\nu}}$ of FRB, consider a component $i \in [\![d]\!]$, and a suboptimal action $a_i \in [\![k_i]\!]\setminus\{a_i^*\}$, which suffers a suboptimality gap of Δ_{i,a_i}. In this part, we show that if $I_{i,t} = a_i$ (i.e., the action selected for component i at time t is a_i), then one of the three following equations is true:

$$\text{UCB}_{i,a_i^*}(t) \leqslant \mu_i^*, \tag{77}$$

or

$$\hat{\mu}_{i,a_i}(t-1) > \mu_{i,a_i} + \sigma\sqrt{\frac{\alpha \log t}{N_{i,a_i}(t-1)}}, \tag{78}$$

or

$$N_{i,a_i}(t-1) < \frac{4\sigma^2\alpha \log T}{\Delta_{i,a_i}^2}, \tag{79}$$

where: $\text{UCB}_{i,a_i^*}(t)$ is the confidence bound of the optimal arm for component i at time t, having pulled such an arm for $N_{i,a_i^*}(t-1)$ times in the previous rounds, and $\hat{\mu}_{i,a_i,N_{i,a_i}(t-1)}$ is the estimated value of the mean of arm a_i of component i after $N_{i,a_i}(t-1)$ pulls. For absurd, if we assume that the three equations are false, then we have:

$$\begin{aligned}
\text{UCB}_{i,a_i^*}(t) &> \mu_i^* \\
&= \mu_{i,a_i} + \Delta_{i,a_i} \\
&\geqslant \mu_{i,a_i} + 2\sqrt{\frac{\sigma^2\alpha \log t}{N_{i,a_i}(t-1)}} \\
&\geqslant \hat{\mu}_{i,a_i,N_{i,a_i}(t-1)} + \sqrt{\frac{\sigma^2\alpha \log t}{N_{i,a_i}(t-1)}} \\
&= \text{UCB}_{i,a_i}(t-1),
\end{aligned}$$

which implies that $a_i(t) \neq a_i$. Now, we bound the probability that Equation (77) or Equation (78) hold true. Similar to the original proof, we use a peeling argument together with Hoeffding's maximal inequality, which is a consequence of Azuma-Hoeffding inequality. Note that:

$$\mathbb{P}(\text{Eq. (77) is true}) \leqslant \mathbb{P}\left(\exists s \in \{1,\ldots,t\} : \hat{\mu}_{i,a_i^*}[s] + \sqrt{\frac{\sigma^2\alpha \log t}{s}} \leqslant \mu_i^*\right)$$

$$= \mathbb{P}\left(\exists s \in \{1, \dots, t\} : \sum_{l=1}^{s}(x_{i,a_i^*}[l] - \mu_i^*) \leqslant -\sqrt{\sigma^2 \alpha s \log t}\right)$$

We now apply the peeling argument with a geometric grid over the time interval $[1, t]$. More precisely, given $\beta \in (0, 1)$, we note that if $s \in \{1, \dots, t\}$, then $\exists j \in \left\{0, \dots, \frac{\log t}{\log 1/\beta}\right\} : \beta^{j+1} t < s \leqslant \beta^j t$.

As such, we get:

$$\begin{aligned}\mathbb{P}(\text{Eq. (77) is true}) &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1} t < s \leqslant \beta^j t, \sum_{l=1}^{s}(x_{i,a_i^*}[l] - \mu_i^*) \leqslant -\sqrt{\sigma^2 \alpha s \log t}\right)\\ &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1} t < s \leqslant \beta^j t, \sum_{l=1}^{s}(x_{i,a_i^*}[l] - \mu_i^*) \leqslant -\sqrt{\sigma^2 \alpha \beta^{j+1} t \log t}\right)\end{aligned}$$

We now bound this last term using Hoeffding's maximal inequality, which gives:

$$\begin{aligned}\mathbb{P}(\text{Eq. (77) is true}) &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(-\frac{\left(\sqrt{\sigma^2 \alpha \beta^{j+1} t \log t}\right)^2}{2\sigma^2 \beta^j t}\right)\\ &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(-\frac{\alpha \beta \log t}{2}\right)\\ &\leqslant \left(\frac{\log t}{\log 1/\beta} + 1\right) \frac{1}{t^{\frac{\beta\alpha}{2}}}.\end{aligned}$$

Using the same arguments, it can be proven that:

$$\mathbb{P}(\text{Eq. (78) is true}) \leqslant \left(\frac{\log t}{\log 1/\beta} + 1\right) \frac{1}{t^{\frac{\beta\alpha}{2}}}.$$

We can now write:

$$\begin{aligned}\mathbb{E}\left[N_{i,a_i}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_{i,t} = a_i\}}\right] &\leqslant u + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{\{I_{i,t} = a_i \text{ and Eq. (79) is false}\}}\right]\\ &= u + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{\{\text{Eq. (77) or Eq. (78) is true}\}}\right]\\ &\leqslant u + \sum_{t=u+1}^{T} \left(\mathbb{P}(\text{Eq. (77) is true}) + \mathbb{P}(\text{Eq. (78) is true})\right),\end{aligned}$$

where $u = \lceil \frac{4\sigma^2 \alpha \log T}{\Delta_{i,a_i}^2} \rceil$.

We can now upper bound the probability of Equations (77) and (78) holds:

$$\begin{aligned}&\sum_{t=u+1}^{T} (\mathbb{P}(\text{Eq. (77) is true}) + \mathbb{P}(\text{Eq. (78) is true}))\\ &\qquad\leqslant 2 \sum_{t=u+1}^{T} \left(\frac{\log t}{\log 1/\beta} + 1\right) \frac{1}{t^{\frac{\beta\alpha}{2}}}\end{aligned}$$

$$\leqslant 2\int_1^{+\infty}\left(\frac{\log t}{\log 1/\beta}+1\right)\frac{1}{t^{\frac{\beta\alpha}{2}}}dt$$

$$= 2\left[\left(\frac{\log t}{\log 1/\beta}+1\right)\left(\frac{2}{2-\alpha\beta}t^{1-\frac{\alpha\beta}{2}}\right)\right]_1^{+\infty} - \frac{4}{(2-\alpha\beta)\log 1/\beta}\int_1^{+\infty} t^{-\frac{\alpha\beta}{2}}\,\mathrm{d}t \tag{80}$$

$$= -\frac{4}{2-\alpha\beta} - \frac{8}{(2-\alpha\beta)^2\log 1/\beta}\left[t^{1-\frac{\alpha\beta}{2}}\right]_1^{+\infty} \tag{81}$$

$$= -\frac{4}{2-\alpha\beta} + \frac{8}{(2-\alpha\beta)^2\log 1/\beta},$$

where line (80) is obtained via integration by parts and the first term of line (81) is obtained imposing $\alpha\beta > 2$. Substituting now $\beta = \frac{4}{\alpha+2}$, which verifies $\beta \in (0,1)$ if $\alpha > 2$, we obtain:

$$\begin{aligned}\sum_{t=u+1}^{T}\left(\mathbb{P}(\text{Eq. (77) is true}) + \mathbb{P}(\text{Eq. (78) is true})\right) &\leqslant -\frac{4}{2-\frac{4\alpha}{\alpha+2}} + \frac{8}{\left(2-\frac{4\alpha}{\alpha+2}\right)^2}\frac{1}{\log\left(\frac{\alpha+2}{4}\right)} \\ &= -\frac{2(\alpha+2)}{2-\alpha} + \frac{2(\alpha+2)^2}{(2-\alpha)^2}\frac{1}{\log\left(\frac{\alpha+2}{4}\right)} \\ &= \frac{2(\alpha+2)}{\alpha-2} + \frac{2}{\log\left(\frac{\alpha+2}{4}\right)}\left(\frac{\alpha+2}{\alpha-2}\right)^2.\end{aligned}$$

Rearranging the upper bound on the expected number of pulls given the three cases presented above, we get:

$$\mathbb{E}[N_{i,j}(T)] \leqslant \frac{4\alpha\sigma^2\log T}{\Delta_{i,j}^2} + \frac{2(\alpha+2)}{\alpha-2} + \frac{2}{\log\left(\frac{\alpha+2}{4}\right)}\left(\frac{\alpha+2}{\alpha-2}\right)^2.$$

We set $g(\alpha) = \frac{2(\alpha+2)}{\alpha-2} + \frac{2}{\log\left(\frac{\alpha+2}{4}\right)}\left(\frac{\alpha+2}{\alpha-2}\right)^2 = \widetilde{\mathcal{O}}\left((\alpha-2)^{-2}\right)$.

Part 2: Upper bounding the expected cumulative regret

We now have to understand how the pulls defined in part 1 can be combined. We want to look at the worst combination in which we can pull the suboptimal action components.

We recall that regret can be defined by highlighting the dependence on the pulls of the action vectors:

$$\mathbb{E}[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})] = \sum_{\mathbf{a}\in\mathcal{A}} N_{\mathbf{a}}\Delta_{\mathbf{a}}.$$

As before, we can bind the pulls of the action components N_{ij} and the action vectors $N_{\mathbf{a}}$ as follows:

$$\mathbb{E}[N_{i,j}(T)] = \sum_{\mathbf{a}\in\mathcal{A}:a_i=j} N_{\mathbf{a}}, \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!].$$

We know that the pulls cannot be negative, and that the total number of pulls of the action vectors sums to T, so we impose these additional constraints. Now, acting on the number of pulls $N_{\mathbf{a}}$, $\forall \mathbf{a} \in \mathcal{A}$ we want to find the worst-case in which we can combine action components in action vectors. So, we solve a maximization problem on the regret defined as a function of the number of pulls, given the constraints defined above, and the upper bound on the expected number of pulls of the action components N_{ij}, $\forall i \in [\![d]\!]$, $j \in [\![k_i]\!]\backslash\{a_i^*\}$ defined in Part 1 of this proof. □

Corollary 4.3 (Explicit Instance-Dependent Upper Bound for F-UCB)**.** *For a given FRB $\underline{\boldsymbol{\nu}}$,* F-UCB *with $\alpha > 2$ suffers an expected regret bounded by:*

$$\begin{aligned}\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] &\leqslant \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}}) \\ &\leqslant 4\alpha\sigma^2\log T \sum_{i\in[\![d]\!]} \mu_{-i}^* \sum_{j\in[\![k_i]\!]\backslash\{a_i^*\}} \Delta_{i,j}^{-1} + g(\alpha)\sum_{i\in[\![d]\!]} k_i,\end{aligned}$$

where $\mu_{-i}^* = \prod_{l\in[\![d]\!]\backslash\{i\}}\mu_l^* \leqslant 1$ for every $i \in [\![d]\!]$.

Proof. In order to obtain a relaxed solution of the optimization problem in Theorem 4.2, we first derive the following upper bound to the suboptimality gaps of the action vector $\mathbf{a} = (a_1, \dots, a_d)^\top$:

$$\Delta_{\mathbf{a}} = \prod_{i \in \llbracket d \rrbracket} \mu_i^* - \prod_{i \in \llbracket d \rrbracket} \mu_{i,a_i} = \prod_{i \in \llbracket d \rrbracket} \mu_i^* \left(1 - \prod_{i \in \llbracket d \rrbracket} \frac{\mu_{i,a_i}}{\mu_i^*} \right) \tag{82}$$

$$\leqslant \prod_{i \in \llbracket d \rrbracket} \mu_i^* \left(1 - \min_{i \in \llbracket d \rrbracket} \frac{\mu_{i,a_i}}{\mu_i^*} \right) \tag{83}$$

$$= \prod_{i \in \llbracket d \rrbracket} \mu_i^* \max_{i \in \llbracket d \rrbracket} \left(1 - \frac{\mu_{i,a_i}}{\mu_i^*} \right) \tag{84}$$

$$\leqslant \prod_{i \in \llbracket d \rrbracket} \mu_i^* \sum_{i \in \llbracket d \rrbracket} \left(1 - \frac{\mu_{i,a_i}}{\mu_i^*} \right) \tag{85}$$

$$= \sum_{i \in \llbracket d \rrbracket} (\mu_i^* - \mu_{i,a_i}) \prod_{j \in \llbracket d \rrbracket \setminus \{j\}} \mu_j^* \tag{86}$$

$$= \sum_{i \in \llbracket d \rrbracket} \Delta_{i,a_i} \mu_{-i}^*, \tag{87}$$

where line (83) follows from observing that $\prod_{i \in \llbracket d \rrbracket} \frac{\mu_{i,a_i}}{\mu_i^*} \leqslant \min_{i \in \llbracket d \rrbracket} \frac{\mu_{i,a_i}}{\mu_i^*}$ since $\frac{\mu_{i,a_i}}{\mu_i^*} \in [0, 1)$, line (86) comes from defining $\mu_{-i}^* := \prod_{j \in \llbracket d \rrbracket \setminus \{j\}} \mu_j^* \leqslant 1$. Thus, by considering the objective function in the optimization problem of Theorem 4.2, we have:

$$\sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} N_{\mathbf{a}} \Delta_{\mathbf{a}} \leqslant \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} N_{\mathbf{a}} \sum_{i \in \llbracket d \rrbracket} \Delta_{i,a_i} \mu_{-i}^* \tag{88}$$

$$= \sum_{i \in \llbracket d \rrbracket} \mu_{-i}^* \sum_{j \in \llbracket k_i \rrbracket} \sum_{\mathbf{a} \in \mathcal{A} \,:\, a_i = j} N_{\mathbf{a}} \Delta_{i,a_i} \tag{89}$$

$$= \sum_{i \in \llbracket d \rrbracket} \mu_{-i}^* \sum_{a_i \in \llbracket k_i \rrbracket \setminus \{a_i^*\}} N_{i,a_i} \Delta_{i,a_i}. \tag{90}$$

By using the Constraint (11) to upper bound N_{i,a_i} and recalling that $\Delta_{i,j} \leqslant 1$, we get the result.

□

Theorem 5.1 (Instance-Dependent Upper Bound for F-Track)**.** *For any FRB $\underline{\boldsymbol{\nu}}$, F-Track run with:*

$$N_0 = \left\lceil \sqrt{\log T} \right\rceil \quad \textit{and} \quad \epsilon_T = \sqrt{\frac{2\sigma^2 f_T(1/\log T)}{N_0}},$$

suffers an expected regret of:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}\left[R_T(\texttt{F-Track}, \underline{\boldsymbol{\nu}})\right]}{\log T} = \underline{C}(\underline{\boldsymbol{\nu}}).$$

Proof. **Preliminary Results** Let us introduce the symbol:

$$\epsilon_{i,j}(t, \delta) := \sqrt{\frac{2\sigma^2 f_T(\delta)}{N_{i,j}(t)}}. \tag{91}$$

Consider the event $\mathcal{E}(\delta) := \{\exists i \in \llbracket d \rrbracket,\, \exists j \in \llbracket k_i \rrbracket,\, \exists t \in \llbracket T_{\text{warm-up}}, T \rrbracket \geqslant 1 \,:\, |\widehat{\mu}_{i,j}(t) - \mu_{i,j}| > \epsilon_{i,j}(t, \delta)\}$ and let us bound its probability:

$$\mathbb{P}(\mathcal{E}(\delta)) \leqslant \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket} \mathbb{P}\left(\exists t \in \llbracket T_{\text{warm-up}}, T \rrbracket \,:\, |\widehat{\mu}_{i,j}(t) - \mu_{i,j}| > \epsilon_{i,j}(t, \delta)\right) \tag{92}$$

$$= \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket} \mathbb{P}\left(\exists s \in \llbracket T \rrbracket \,:\, |\widehat{\mu}_{i,j}[s] - \mu_{i,j}| > \sqrt{\frac{2\sigma^2 f_T(\delta)}{s}}\right) \tag{93}$$

$$\leqslant \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket} \delta = k\delta, \tag{94}$$

where line (92) follows from a union bound over the values of i and j, line (93) follows by rewriting the probability by highlighting the dependence of the estimator on the number of samples s, and line (94) follows from Lemma C.1, recalling that $s(\widehat{\mu}_{i,j}[s] - \mu_{i,j})$ is a martingale difference sequence and it is σ^2-subgaussian.

We will make use of the following two instantiations of event $\mathcal{E}(\delta)$:

$$\mathcal{E}_1 := \mathcal{E}(1/\log T) \qquad \text{and} \qquad \mathcal{E}_2 := \mathcal{E}(1/T). \tag{95}$$

Clearly, from the previous result, we have that $\mathbb{P}(\mathcal{E}_1) \leqslant k/\log T$ and $\mathbb{P}(\mathcal{E}_2) \leqslant k/T$.

We start decomposing the regret over the phases of the algorithm:

$$\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[R(\texttt{F-Track}, T)] = \underbrace{\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}\left[\sum_{t \in \textit{warm-up}} \Delta_{\mathbf{a}(t)}\right]}_{=:\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{warm-up}}(T)]} + \underbrace{\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}\left[\sum_{t \in \textit{success}} \Delta_{\mathbf{a}(t)}\right]}_{=:\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{success}}(T)]} + \underbrace{\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}\left[\sum_{t \in \textit{recovery}} \Delta_{\mathbf{a}(t)}\right]}_{=:\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T)]}, \tag{96}$$

where, with little abuse of notation, we denoted with $t \in \textit{phase}$ denotes the rounds in which phase *phase* is active. We proceed to analyze the three components separately.

Part 1: Regret in Warm-Up Phase $\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{warm-up}}(T)]$ We start by analyzing the regret in the *warm-up* phase, whose duration is given by $T_{\text{warm-up}} = N_0 \max_{i \in \llbracket d \rrbracket} k_i = \lceil \sqrt{\log T} \rceil \max_{i \in \llbracket d \rrbracket} k_i$. Thus, the corresponding expected cumulative regret can be bounded as follows:

$$\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[R_{\text{warm-up}}(T)] \leqslant \Delta_{\mathbf{max}} \left\lceil \sqrt{\log T} \right\rceil \max_{i \in \llbracket d \rrbracket} k_i = \mathcal{O}\left(\sqrt{\log T}\right), \tag{97}$$

where $\Delta_{\mathbf{max}} = \max_{\mathbf{a} \in \mathcal{A}} \Delta_{\mathbf{a}}$ and the Big-O notation retains the dependence on T only. Thus, its contribution to the regret is asymptotically negligible:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{warm-up}}(T)]}{\log T} = 0. \tag{98}$$

Part 2: Regret in the Recovery Phase $\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T)]$ We move to the analysis of the regret in the *recovery* phase. We start by showing that if event $\mathcal{E}_1$ does not hold, then, the recovery phase never activates. Indeed, under $\mathcal{E}_1^{\complement}$ simultaneously for all $i \in \llbracket d \rrbracket$, $j \in \llbracket k_i \rrbracket$, and $t \in \llbracket T_{\text{warm-up}}, T \rrbracket$ we have that:

$$|\widehat{\mu}_{i,j}(t) - \mu_{i,j}| \leqslant \epsilon_{i,j}(t, 1/\log T), \tag{99}$$

which implies simultaneously for all $i \in \llbracket d \rrbracket$, $j \in \llbracket k_i \rrbracket$, and $t \in \llbracket T_{\text{warm-up}}, T \rrbracket$ that:

$$|\widehat{\mu}_{i,j}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(t-1)| \leqslant |\widehat{\mu}_{i,j}(T_{\text{warm-up}}) - \mu_{i,j}| + |\widehat{\mu}_{i,j}(t-1) - \mu_{i,j}| \tag{100}$$

$$\leqslant \epsilon_{i,j}(T_{\text{warm-up}}, 1/\log T) + \epsilon_{i,j}(t-1, 1/\log T) \tag{101}$$

$$\leqslant 2\epsilon_{i,j}(T_{\text{warm-up}}, 1/\log T), \tag{102}$$

being $\epsilon_{i,j}(t, 1/\log T)$ a decreasing in t. Recalling that $N_{i,j}(T_{\text{warm-up}}) \geqslant N_0$, we have:

$$2\epsilon_{i,j}(T_{\text{warm-up}}, 1/\log T) = 2\sqrt{\frac{2\sigma^2 f_T(1/\log T)}{N_{i,j}(T_{\text{warm-up}})}} \leqslant 2\sqrt{\frac{2\sigma^2 f_T(1/\log T)}{N_0}} = 2\epsilon_T. \tag{103}$$

Thus, we conclude that the termination condition of the while loop never activates and, consequently, the recovery phase activates only when $\mathcal{E}_1$ holds, i.e., with probability at most $1/\log T$.

In the recovery phase, our `F-Track` algorithm plays `F-UCB` that, from Corollary 4.3, is proved to suffer logarithmic regret of the form:

$$\rho(T) := 4\alpha\sigma^2 \log T \sum_{i \in \llbracket d \rrbracket} \mu^*_{-i} \sum_{j \in \llbracket k_i \rrbracket \setminus \{a_i^*\}} \Delta_{i,j}^{-1} + g(\alpha) \sum_{i \in \llbracket d \rrbracket} k_i = \mathcal{O}(\log T). \tag{104}$$

Thus, we have that the cumulative regret of the recovery phase is bounded by:

$$\mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T)] = \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T) | \mathcal{E}_1^{\complement}] \, \mathbb{P}(\mathcal{E}_1^{\complement}) + \mathop{\mathbb{E}}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T) | \mathcal{E}_1] \, \mathbb{P}(\mathcal{E}_1) \leqslant 0 + \frac{\rho(T)}{\log T} = \mathcal{O}(1). \tag{105}$$

Consequently, its contribution to the expected cumulative regret is asymptotically negligible. Indeed:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{recovery}}(T)]}{\log T} = 0. \tag{106}$$

Part 3: Regret in the Success Phase $\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{success}}(T)]$ We conclude with the most challenging part consisting of bounding the regret in the success phase. The cumulative regret in the success phase needs to be further decomposed as follows:

$$\mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{\text{success}}(T)] = \mathbb{E}_{\underline{\boldsymbol{\nu}}}\left[\mathbf{1}\{\mathcal{E}_1^{\complement}\} \sum_{t \in success} \Delta_{\mathbf{a}(t)}\right] + \mathbb{E}_{\underline{\boldsymbol{\nu}}}\left[\mathbf{1}\{\mathcal{E}_1 \wedge \mathcal{E}_2^{\complement}\} \sum_{t \in success} \Delta_{\mathbf{a}(t)}\right] + \mathbb{E}_{\underline{\boldsymbol{\nu}}}\left[\mathbf{1}\{\mathcal{E}_2\} \sum_{t \in success} \Delta_{\mathbf{a}(t)}\right] \tag{107}$$

We analyze each term separately.

<u>Part 3.1: Regret under $\mathcal{E}_1^{\complement}$</u> In what follows, all estimated quantities are estimated with the samples available at the end of the warm-up phase and, thus, we will omit the dependence on $T_{\text{warm-up}}$. We show that asymptotically, during the success phase and under event $\mathcal{E}_1^{\complement}$, the algorithm suffers the optimal regret. To this end, we need to introduce some auxiliary tools. For every $i \in [\![d]\!]$, let us define a *sorting function* as any bijective function $\pi_i : [\![k_i]\!] \to [\![k_i]\!]$ such that:

$$\mu_{i,\pi_i(1)} \leqslant \cdots \leqslant \mu_{i,\pi_i(k_i)}. \tag{108}$$

If all $\mu_{i,j}$ are different, the sorting function is unique. Furthermore, for every $i \in [\![d]\!]$ and $j \in [\![k_i]\!] \backslash \{\pi_i(k_i)\}$ (i.e., excluding the action component with maximum expected reward), let us denote:

$$N_{i,j} = \frac{2\sigma^2 f_T(1/T)}{\Delta_{i,j}^2}, \tag{109}$$

where $\Delta_{i,j} = \mu_{i,\pi_i(k_i)} - \mu_{i,j}$. Let us notice that $N_{i,j}$ corresponds approximately to the minimum number of pulls of component (i,j) prescribed by the lower bound in Theorem 3.3 and denoted with $L_{i,j} = \frac{2\sigma^2 \log T}{\Delta_{i,j}^2}$. Given the definition of $f_T(1/T)$, we have that $L_{i,j}/N_{i,j} \to 1$ as $T \to +\infty$. Given the sorting function, it is clear that also:

$$N_{i,\pi_i(1)} \leqslant \cdots \leqslant N_{i,\pi_i(k_i)}. \tag{110}$$

Let us define:

$$\beta_i := f_T(1/T)^{-1} \min_{l,l' \in [\![k_i]\!] \,:\, N_{i,\pi_i(l)} \neq N_{i,\pi_i(l')}} \left| N_{i,\pi_i(l)} - N_{i,\pi_i(l')} \right|. \tag{111}$$

It is clear that if for every $i \in [\![b]\!]$ and $j \in [\![k_i]\!]$ we have we have $|\widehat{N}_{i,j} - N_{i,j}| \leqslant \beta_i f_T(1/T)/4$, then, for any sorting function $\widehat{\pi}_i$ of the estimated quantities $\underline{N}_{i,j}$, there exist a sorting function π_i of the true quantities $N_{i,j}$ such that $\widehat{\pi}_i = \pi_i$.

Let us define for every $i \in [\![d]\!]$ and $j \in [\![k_i]\!]$:

$$M_{i,j} := \sum_{l=1}^{j} N_{i,\pi_i(l)}. \tag{112}$$

We define now a sorting function $\pi : [\![k]\!] \to \bigcup_{i \in [\![d]\!]} (\{i\} \times [\![k_i]\!])$ as any bijection such that:

$$M_{\pi(1)} \leqslant \cdots \leqslant M_{\pi(k)}, \tag{113}$$

and convene (with a little abuse of notation) that $M_{\pi(0)} = 0$. It is clear that $M_{\pi(k)} = M_{\pi(k-1)} = \cdots = M_{\pi(k-d+1)} = T$. Let $l \in [\![k]\!]$, we define the *active action* as:

$$\boldsymbol{\alpha}(l) := (j_1, \ldots, j_d) \quad \text{where} \quad j_i \text{ s.t. } \pi(l') = (i, j_i) \text{ and } l' = \min\{l'' \geqslant l \text{ and } \pi(l'') = (i, \cdot)\} \text{ with } i \in [\![d]\!]. \tag{114}$$

We can now rewrite the regret with this notation:

$$\sum_{\mathbf{a} \neq \mathbf{a}^*} N_{\mathbf{a}} \Delta_{\mathbf{a}} = \sum_{l=1}^{k-d} \left(M_{\pi(l)} - M_{\pi(l-1)} \right) \Delta_{\boldsymbol{\alpha}(l)}, \tag{115}$$

having observed that for the $k-d+1$ terms we play the optimal action and the successive ones are zero. Furthermore, given the relation between $L_{i,j}$ and $N_{i,j}$, we have that:

$$\frac{\sum_{\mathbf{a} \neq \mathbf{a}^*} N_{\mathbf{a}}}{f_T(1/T)} = \underline{C} \qquad \text{and} \qquad \limsup_{T \to +\infty} \frac{\sum_{\mathbf{a} \neq \mathbf{a}^*} N_{\mathbf{a}}}{\log T} = \underline{C}. \tag{116}$$

Let us now define:

$$\beta := f_T(1/T)^{-1} \min_{l,l' \in [\![k]\!] \,:\, M_{\pi(l)} \neq M_{\pi(l')}} \left| M_{\pi(l)} - M_{\pi(l')} \right|. \tag{117}$$

It is clear that if for every $i \in [\![b]\!]$ and $j \in [\![k_i]\!]$ we have $|\widehat{M}_{i,j} - M_{i,j}| \leqslant \beta f_T(1/T)/4$, for every sorting function $\widehat{\pi}$ of the estimated quantities $\widehat{M}_{i,j}$, there exist a sorting function π of the true quantities $M_{i,j}$ such that $\widehat{\pi} = \pi$. If this is the case, then, the *active action* $\widehat{\boldsymbol{\alpha}}(l)$ induced by $\widehat{\pi}$ must be the same as $\boldsymbol{\alpha}(l)$ since the active action depends on the sorting function only.

We now show that we can always guarantee $|\widehat{N}_{i,j} - N_{i,j}| \leqslant (\beta_i f_T(1/T))/4$ and $|\widehat{M}_{i,j} - M_{i,j}| \leqslant (\beta f_T(1/T))/4$ for sufficiently large T. First of all, let us ensure that we identify the optimal component for every $i \in [\![d]\!]$. This is guaranteed whenever for every $j \in [\![k_i]\!]$ we have:

$$|\widehat{\mu}_{i,j} - \mu_{i,j}| \leqslant \epsilon_{i,j}(T_{\text{warm-up}}, 1/\log T) \leqslant \epsilon_T \leqslant \Delta_{\min}/4, \tag{118}$$

where $\Delta_{\min} = \min_{i \in [\![d]\!]} \min_{j \in [\![k_i]\!] \setminus \{\pi_i(k_i)\}} \mu_{i,\pi_i(k_i)} - \mu_{i,j}$. The inequality is satisfied for sufficiently large T since:

$$\epsilon_T = \sqrt{\frac{2\sigma^2 f_T(1/\log T)}{\lceil \sqrt{\log T} \rceil}} = \mathcal{O}\left(\sqrt{\frac{\sigma^2 \log\log T}{\sqrt{\log T}}} \right) \to 0 \qquad \text{as} \qquad T \to +\infty. \tag{119}$$

Under this condition, we have that $\pi_i(k_i) = \widehat{\pi}_i(k_i)$ and, consequently:

$$\widehat{\Delta}_{i,j} = \widehat{\mu}_{i,\pi(k_i)} - \widehat{\mu}_{i,j} \qquad \text{and} \qquad \Delta_{i,j} = \mu_{i,\pi(k_i)} - \mu_{i,j}. \tag{120}$$

Thus, under event $\mathcal{E}_1^\complement$, we have $|\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leqslant 2\epsilon_T$. Let us now consider $i \in [\![k]\!]$ and $j \in [\![k_i]\!] \setminus \{\pi_i(k_i)\}$, we have:

$$\left| \widehat{N}_{i,j} - N_{i,j} \right| = \left| \frac{2\sigma^2 f_T(1/T)}{\widehat{\Delta}_{i,j}^2} - \frac{2\sigma^2 f_T(1/T)}{\Delta_{i,j}^2} \right| \tag{121}$$

$$= 2\sigma^2 f_T(1/T) \frac{(\Delta_{i,j} + \widehat{\Delta}_{i,j})|\Delta_{i,j} - \widehat{\Delta}_{i,j}|}{\Delta_{i,j}^2 \widehat{\Delta}_{i,j}^2} \tag{122}$$

$$\leqslant 8\sigma^2 f_T(1/T) \frac{(2\Delta_{\max} + \Delta_{\min}/2)}{\Delta_{\min}^4} \epsilon_T, \tag{123}$$

where $\Delta_{\max} = \max_{i \in [\![d]\!]} \max_{j,j' \in [\![k_i]\!]} |\mu_{i,j} - \mu_{i,j'}|$ and having observed that $\widehat{\Delta}_{i,j} \geqslant \Delta_{i,j} - 2\epsilon_T \geqslant \Delta_{\min} - \Delta_{\min}/2 = \Delta_{\min}/2$ and $\widehat{\Delta}_{i,j} \leqslant \Delta_{i,j} + 2\epsilon_T \leqslant \Delta_{\max} + \Delta_{\min}/2 = \Delta_{\min}/2$. Thus, the difference can go below $\beta_i f_T(1/T)$ for sufficiently large T. Let us now move to the $M_{i,j}$ variables. For sufficiently large T such that the sorting function π_i coincide with their estimated counterparts $\widehat{\pi}_i$, we have that for $i \in [\![d]\!]$ and $j \in [\![k_i]\!]$:

$$\left| M_{i,j} - \widehat{M}_{i,j} \right| = \left| \sum_{l=1}^{j} N_{i,\pi_i(l)} - \sum_{l=1}^{j} \widehat{N}_{i,\widehat{\pi}_i(l)} \right| \tag{124}$$

$$\leqslant \sum_{l=1}^{j} \left| \underline{N}_{i,\pi_i(l)} - \widehat{N}_{i,\pi_i(l)} \right| \tag{125}$$

$$\leqslant 8\sigma^2 j f_T(1/T) \frac{(2\Delta_{\max} + \Delta_{\min}/2)}{\Delta_{\min}^4} \epsilon_T. \tag{126}$$

Similarly, as before, we can conclude that this difference can be made smaller than β for sufficiently large T, and, consequently, make the estimated sorting function $\widehat{\pi}$ equal the true counterpart π.

Under these conditions, we can bound the cumulative regret under $\mathcal{E}_1^\complement$:

$$\sum_{t \in \mathit{success}} \Delta_{\mathbf{a}(t)} = \sum_{\mathbf{a} \neq \mathbf{a}^*} \widehat{N}_{\mathbf{a}} \Delta_{\mathbf{a}} \tag{127}$$

$$= \sum_{l=1}^{k-d} \left(\widehat{M}_{\widehat{\pi}(l)} - \widehat{M}_{\widehat{\pi}(l-1)} \right) \Delta_{\widehat{\boldsymbol{\alpha}}(l)} \tag{128}$$

$$= \sum_{l=1}^{k-d} \left(\widehat{M}_{\pi(l)} - \widehat{M}_{\pi(l-1)} \right) \Delta_{\boldsymbol{\alpha}(l)} \tag{129}$$

$$= \sum_{l=1}^{k-d} \left(\widehat{M}_{\pi(l)} - M_{\pi(l)} + M_{\pi(l-1)} - \widehat{M}_{\pi(l-1)} \right) \Delta_{\boldsymbol{\alpha}(l)} + \sum_{l=1}^{k-d} \left(M_{\pi(l)} - M_{\pi(l-1)} \right) \Delta_{\boldsymbol{\alpha}(l)} \tag{130}$$

$$\leqslant 2\Delta_{\mathbf{max}} \sum_{l=1}^{k-d} \left| \widehat{M}_{\pi(l)} - M_{\pi(l)} \right| + \underline{C} f_T(1/T) \tag{131}$$

$$\leqslant 8\sigma^2 (k-d) \max_{i \in \llbracket d \rrbracket} k_i f_T(1/T) \frac{(2\Delta_{\max} + \Delta_{\min}/2)}{\Delta_{\min}^4} \epsilon_T + \underline{C} f_T(1/T) \tag{132}$$

$$= \mathcal{O}(\epsilon_T f_T(1/T)) + \underline{C} f_T(1/T), \tag{133}$$

where we used Equation 126. Thus, recalling that $\epsilon_T \to 0$ for $T \to +\infty$, we have:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1^\complement\} \sum_{t \in \mathit{success}} \Delta_{\mathbf{a}(t)} \right]}{\log T} = \underline{C}. \tag{134}$$

Consequently, its contribution to the asymptotic regret is exactly $\underline{C}$.

Part 3.2: Regret under $\mathcal{E}_1 \wedge \mathcal{E}_2^\complement$ In this case, we have to prove that the regret remains logarithmic. We consider two cases:

Case 1 We perform the analysis in the first case under the following conditions:

$$\forall i \in \llbracket d \rrbracket : \quad \pi_i(k_i) = \widehat{\pi}_i(k_i) \quad \text{and} \quad \forall j \in \llbracket k_i \rrbracket \backslash \{\pi_i(k_i)\} : \quad \widehat{\Delta}_{i,j} \geqslant \Delta_{\min}/4. \tag{135}$$

In such a case, it is simple to show that the regret is at most logarithmic. Indeed, being the optimal arm correctly identified ($\pi_i(k_i) = \widehat{\pi}_i(k_i)$) we have:

$$\sum_{\mathbf{a} \neq \mathbf{a}*} \widehat{N}_{\mathbf{a}} \Delta_{\mathbf{a}} \leqslant 2\Delta_{\mathbf{max}} \sum_{l=1}^{k-d} \widehat{M}_{\widehat{\pi}(l)} \tag{136}$$

$$\leqslant 2\Delta_{\mathbf{max}} \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket \backslash \{\pi_i(k_i)\}} \widehat{N}_{i,\pi_i(j)} \tag{137}$$

$$\leqslant 4\sigma^2 f_T(1/T) \Delta_{\mathbf{max}} \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket \backslash \{\pi_i(k_i)\}} \widehat{\Delta}_{i,\pi_i(j)}^{-2} \tag{138}$$

$$\leqslant 64 k \sigma^2 f_T(1/T) \Delta_{\mathbf{max}} \Delta_{\min}^{-2} = \mathcal{O}(\log T), \tag{139}$$

where we observed that since the optimal arm is correctly identified, the following inequality holds: $\sum_{l=1}^{k-d} \widehat{M}_{\widehat{\pi}(l)} \leqslant \sum_{i \in \llbracket d \rrbracket} \sum_{j \in \llbracket k_i \rrbracket \backslash \{\pi_i(k_i)\}} \widehat{N}_{i,\pi_i(j)}$.

Case 2 If the condition in Equation (135) is violated, we show that the success phase stops after a logarithmic number of rounds. Consider the smallest round $t_{i,j}$ in which for a given $i \in \llbracket k \rrbracket$ and $j \in \llbracket k_i \rrbracket \backslash \{\widehat{\pi}_i(k_i)\}$, it holds that:

$$N_{i,j}(t_{i,j}) \geqslant \min \left\{ \frac{2\sigma^2 f_T(1/T)}{\widehat{\Delta}_{i,j}^2}, \frac{128\sigma^2 f_T(1/T)}{\Delta_{\min}^2} \right\}. \tag{140}$$

Since the `F-Track` algorithm in the success phase proceeds with the round robin of at most k arms, we have that:

$$t_{i,j} \leqslant k \min \left\{ \frac{2\sigma^2 f_T(1/T)}{\widehat{\Delta}_{i,j}^2}, \frac{128\sigma^2 f_T(1/T)}{\Delta_{\min}^2} \right\} \leqslant \frac{128 k \sigma^2 f_T(1/T)}{\Delta_{\min}^2} =: t^* = \mathcal{O}(\log T). \tag{141}$$

Now, we consider two sub-cases.

Case 2.1 In the first sub-case, we deal with the case in which some optimal components are not correctly identified:

$$\exists i \in \llbracket d \rrbracket : \quad \pi_i(k_i) \neq \widehat{\pi}_i(k_i) \tag{142}$$

In such a case, at most at round t^*, we have that:

$$\widehat{\mu}_{i,\pi_i(k_i)}(t) \geqslant \mu_{i,\pi_i(k_i)}(t) - \sqrt{\frac{2\sigma^2 f_T(1/T)}{N_{i,\pi_i(k_i)}(t)}} \tag{143}$$

$$\geqslant \mu_{i,\pi_i(k_i)}(t) - \max \left\{ \widehat{\Delta}_{i,\pi_i(k_i)}, \Delta_{\min}/8 \right\} \tag{144}$$

$$\geqslant \mu_{i,\pi_i(k_i)}(t) - \widehat{\Delta}_{i,\pi_i(k_i)} - \Delta_{\min}/8 \tag{145}$$

$$\geqslant \mu_{i,\widehat{\pi}_i(k_i)}(t) + \Delta_{i,\widehat{\pi}_i(k_i)} - \Delta_{\min}/8 - \widehat{\Delta}_{i,\pi_i(k_i)} \tag{146}$$

$$\geqslant \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(t) - \sqrt{\frac{2\sigma^2 f_T(1/T)}{N_{i,\widehat{\pi}_i(k_i)}(t)}} + \Delta_{i,\widehat{\pi}_i(k_i)} - \Delta_{\min}/8 - \widehat{\Delta}_{i,\pi_i(k_i)} \tag{147}$$

$$\geqslant \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(t) - \max\{0, \Delta_{\min}/8\} + \Delta_{i,\widehat{\pi}_i(k_i)} - \Delta_{\min}/8 - \widehat{\Delta}_{i,\pi_i(k_i)} \tag{148}$$

$$\geqslant \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(t) - 3/4\Delta_{\min} + \widehat{\mu}_{i,\pi_i(k_i)}(T_{\text{warm-up}}) - \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(T_{\text{warm-up}}). \tag{149}$$

where line (143) follows from the fact that event $\mathcal{E}_2$ does not hold, line (144) follows from Equation (140) with $j = \pi_i(k_i)$, line (145) is obtained with $\max a, b \leqslant a + b$ for $a, b \geqslant 0$, line (146) is obtained from the definition of $\Delta_{i,\widehat{\pi}_i(k_i)}$, line (147) follows from the fact that event $\mathcal{E}_2$ does not hold, line (148) follows from Equation (140) with $j = \widehat{\pi}_i(k_i)$ (whose estimated $\widehat{\Delta}_{i,\widehat{\pi}_i(k_i)} = 0$, and line (149) is obtained from the definition of $\widehat{\Delta}_{i,\pi_i(k_i)}$ and from $\Delta_{i,\widehat{\pi}_i(k_i)} \geqslant \Delta_{\min}$.

This implies that at this round:

$$\widehat{\mu}_{i,\pi_i(k_i)}(t) - \widehat{\mu}_{i,\pi_i(k_i)}(T_{\text{warm-up}}) + \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(T_{\text{warm-up}}) - \widehat{\mu}_{i,\widehat{\pi}_i(k_i)}(t) \geqslant 3/4\Delta_{\min} \geqslant 4\epsilon_T, \tag{150}$$

where the latter holds for sufficiently large T. Thus, we have that the success phase stops after at most t^* rounds, leading to a regret of:

$$\sum_{t \in \textit{success}} \Delta_{\mathbf{a}(t)} \leqslant \Delta_{\mathbf{max}} \frac{32k\sigma^2 f_T(1/T)}{\Delta^2_{\min}} = \mathcal{O}(\log T). \tag{151}$$

Case 2.2 In the first sub-case, we deal with the case holding under the condition:

$$\forall i \in [\![d]\!]: \quad \pi_i(k_i) = \widehat{\pi}_i(k_i) \quad \text{and} \quad \exists i \in [\![d]\!]: \quad \exists j \in [\![k_i]\!] \backslash \{\pi_i(k_i)\}: \quad \widehat{\Delta}_{i,j} < \Delta_{\min}/4 \tag{152}$$

At round t^*, for the (i, j) fulfilling the second part of the condition:

$$\widehat{\mu}_{i,\pi_i(k_i)}(t) - \widehat{\mu}_{i,\pi_i(k_i)}(T_{\text{warm-up}}) + \widehat{\mu}_{i,j}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(t) \tag{153}$$

$$\geqslant \widehat{\mu}_{i,\pi_i(k_i)}(t) - \widehat{\mu}_{i,j}(t) - \widehat{\Delta}_{i,j} \tag{154}$$

$$\geqslant \mu_{i,\pi_i(k_i)}(t) - \sqrt{\frac{2\sigma^2 f_T(1/T)}{N_{i,\pi_i(k_i)}(t)}} - \mu_{i,j}(t) - \sqrt{\frac{2\sigma^2 f_T(1/T)}{N_{i,j}(t)}} - \widehat{\Delta}_{i,j} \tag{155}$$

$$\geqslant - \max\{0, \Delta_{\min}/8\} - \max\{\widehat{\Delta}_{i,j}, \Delta_{\min}/8\} + \Delta_{i,j} - \widehat{\Delta}_{i,j} \tag{156}$$

$$\geqslant \Delta_{\min}/4, \tag{157}$$

having exploited $\widehat{\Delta}_{i,j} \leqslant \Delta_{\min}/4$ and $\Delta_{i,j} \geqslant \Delta_{\min}$. Thus, for sufficiently large T, we have that $4\epsilon_T \leqslant \Delta_{\min}/4$ and, consequently, the success phase ends.

Part 3.3: Regret under $\mathcal{E}_2$ We conclude by bounding the regret under event $\mathcal{E}_2$, In this case, we proceed with the following trivial bound, recalling that $\Pr(\mathcal{E}_2) \leqslant 1/T$.

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_2\} \sum_{t \in \textit{success}} \Delta_{\mathbf{a}(t)}\right] \leqslant \Delta_{\mathbf{max}} T\, \mathbb{P}(\mathcal{E}_2) \leqslant \Delta_{\mathbf{max}} = \mathcal{O}(1). \tag{158}$$

Consequently, its contribution to the asymptotic regret is negligible. □

C.2. Technical Lemmas

Lemma C.1. *Let $T \in \mathbb{N}$, $\epsilon > 0$. Let $X_1, \ldots, X_T$ be a martingale difference sequence adapted to the filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$, such that for every $t \in [\![T]\!]$, it holds that $\mathbb{E}[e^{\lambda X_t}] \leqslant e^{(\sigma^2\lambda^2)/2}$ a.s. for every $\lambda \in \mathbb{R}$. Then, for every $\delta \in (0, 1)$ it holds that:*

$$\mathbb{P}\left(\exists t \in [\![T]\!] : \sum_{s=1}^{t} X_s \geqslant \sqrt{2\left(1 + (\log T)^{-1}\right) \max\left\{\epsilon, t\sigma^2\right\} \left(\log\left(1 + \left\lceil \frac{\log(T\sigma^2/\epsilon)}{\log(1 + (\log T)^{-1})} \right\rceil\right) + \log\left(\frac{1}{\delta}\right)\right)}\right) \leqslant \delta. \tag{159}$$

Furthermore, for sufficiently large T, it holds that:

$$\mathbb{P}\left(\exists t \in [\![T]\!] : \sum_{s=1}^{t} X_s \geqslant \sqrt{2\sigma^2 t f_T(\delta)}\right) \leqslant \delta, \tag{160}$$

where:

$$f_T(\delta) := \left(1 + \frac{1}{\log T}\right)\left(c \log\log T + \log\left(\frac{1}{\delta}\right)\right), \tag{161}$$

and $c > 0$ is a universal constant.

Proof. The first statement is obtained from Lemma 14 of (Lattimore & Szepesvari, 2017) considering that the inequality employed in Equation (19) of that proof applies for σ^2-subgaussian random variables and not for Gaussian variables only. The second statement is obtained by setting $\epsilon = \sigma^2$ and bounding $\frac{1}{\log(1+(\log T)^{-1})} \leqslant \log T$ and $\log(1+\lceil(\log T)^2\rceil) \leqslant c \log\log T$ for some universal constant c (≈ 2). □

Lemma C.2. *Let $\overline{x} \in [0,1)$, $d \in \mathbb{N}$, then if $x_i \in [0,\overline{x})$, $\forall i \in \llbracket d \rrbracket$, it holds:*

$$1 - \prod_{i \in \llbracket d \rrbracket} (1 - x_i) \geqslant (1 - \overline{x})^{d-1} \sum_{i \in \llbracket d \rrbracket} x_i.$$

Proof. We prove this statement by induction.

First, we can observe how for $d = 1$ this result trivially holds:

$$1 - (1 - x_1) = x_1.$$

We can now make the inductive step on d:

$$\begin{aligned}
1 - \prod_{i \in \llbracket d \rrbracket} (1 - x_i) &= 1 - (1 - x_d) \prod_{i \in \llbracket d-1 \rrbracket} (1 - x_i) \\
&= 1 - (1 - x_d) \prod_{i \in \llbracket d-1 \rrbracket} (1 - x_i) \pm x_d \\
&= (1 - x_d)\left(1 - \prod_{i \in \llbracket d-1 \rrbracket} (1 - x_i)\right) + x_d \qquad (162) \\
&\geqslant (1 - x_d)\left((1 - \overline{x})^{d-2} \sum_{i \in \llbracket d-1 \rrbracket} x_i\right) + x_d \\
&\geqslant (1 - \overline{x})^{d-1} \sum_{i \in \llbracket d \rrbracket} x_i,
\end{aligned}$$

where line (162) is the inductive step on d. □

Lemma C.3. *In a FRB, considering $\mu_{\mathbf{a}^*} = 1$, if $\Delta_{i,j} \leqslant \overline{\Delta} = 1 - \frac{1}{2^{1/(d-1)}}$, $\forall i \in \llbracket d \rrbracket, j \in \llbracket k_i \rrbracket$, the regret can be bounded as:*

$$R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}}) = \sum_{t \in \llbracket T \rrbracket} \left(1 - \prod_{i \in \llbracket d \rrbracket} \left(1 - \Delta_{i,a_i(t)}\right)\right) \geqslant \frac{1}{2} \sum_{t \in \llbracket T \rrbracket} \sum_{i \in \llbracket d \rrbracket} \Delta_{i,a_i(t)}.$$

Proof. We prove this statement by looking at a single time t. We can rewrite Lemma C.2 as:

$$1 - \prod_{i \in \llbracket d \rrbracket} (1 - \Delta_{i,a_i(t)}) \geqslant (1 - \overline{\Delta})^{d-1} \sum_{i \in \llbracket d \rrbracket} \Delta_{i,a_i(t)},$$

if $\Delta_{i,j} \leqslant \overline{\Delta} \in [0,1)$, $\forall i \in \llbracket d \rrbracket, j \in \llbracket k_i \rrbracket$.
We make a choice we want to transform this result in order to have:

$$1 - \prod_{i \in \llbracket d \rrbracket} (1 - \Delta_{i,a_i(t)}) \geqslant \frac{1}{2} \sum_{i \in \llbracket d \rrbracket} \Delta_{i,a_i(t)}.$$

This can be done by imposing:

$$\begin{aligned}
\frac{1}{2} &\leqslant (1 - \overline{\Delta})^{d-1} \\
\frac{1}{2^{1/(d-1)}} &\leqslant (1 - \overline{\Delta})
\end{aligned}$$

$$\overline{\Delta} \leqslant 1 - \frac{1}{2^{1/(d-1)}}.$$

□

Lemma C.4 (Wang et al. 2021)**.** *Suppose m, B are positive integers and $m \geqslant 2$; there are $m+1$ probability distributions $\mathbb{P}_0, \mathbb{P}_1, \dots \mathbb{P}_m$, and m random variables $N_1, \dots, N_m$, such that: (i) Under any of the P_i's, $N_1, \dots, N_m$ are non-negative and $\sum_{i \in \llbracket m \rrbracket} N_i \leqslant B$ with probability 1; (ii) $\forall i \in \llbracket m \rrbracket$, $d_{\text{TV}} \leqslant \frac{1}{4}\sqrt{\frac{m}{B}\mathbb{E}_0[N_i]}$. Then:*

$$\frac{1}{m}\sum_{i \in \llbracket m \rrbracket} \mathbb{E}_i[B - N_i] \geqslant \frac{B}{4}.$$

Proof. For the proof of this Lemma, we refer the reader to Lemma 24 of (Wang et al., 2021). □

D. Additional Theorems and Lemmas

In this section, we provide additional Theorems and Lemmas useful in the discussion of the work.

Lemma D.1. *The product $X_1 X_2 \cdots X_n$ of $n \geqslant 3$ independent σ^2-subgaussian random variables is not subgaussian.*

Proof. The proof follows the one proposed by (Pinelis, 2021).

The proof of this statement can be done by verifying that the moment-generating function of the product of n independent Gaussian distributions with unit variance ($X_i \sim \mathcal{N}(0,1),\ \forall i \in \llbracket n \rrbracket$) is unbounded:

$$\mathbb{E}\left[\exp\left(c \prod_{i \in \llbracket n \rrbracket} X_i\right)\right] = \infty, \qquad \forall c > 0.$$

Let us call X the vector composed of our random variables $X := (X_1, X_2, \dots, X_n)$ and let $(U_1, U_2, \dots U_n)$ be a uniformly distributed unit random vector. For some real $C_n > 0$:

$$\mathbb{E}\left[\exp\left(c \prod_{i \in \llbracket n \rrbracket} X_i\right)\right] \geqslant \mathbb{E}\left[\exp\left(c \prod_{i \in \llbracket n \rrbracket} X_i\right) \mathbb{1}\left\{X_i > \frac{||X||_2}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right\}\right] \tag{163}$$

$$= C_n \int_0^\infty \exp\left(c \underbrace{\frac{1}{(2\sqrt{n})^n} r^n}_{\text{(A)}}\right) \underbrace{r^{n-1} \exp\left(-\frac{r^2}{2}\right)}_{\text{(B)}} \mathrm{d}r \cdot \underbrace{\mathbb{P}\left(U_i > \frac{1}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right)}_{\text{(C)}} \tag{164}$$

$$= C_n \frac{(2\sqrt{n})^n}{cn} \int_0^\infty \underbrace{\exp\left(c \frac{1}{(2\sqrt{n})^n} r^n\right) \frac{cn}{(2\sqrt{n})^n} r^{n-1}}_{g'(r)} \underbrace{\exp\left(-\frac{r^2}{2}\right)}_{f(r)} \mathrm{d}r \cdot \mathbb{P}\left(U_i > \frac{1}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right)$$

$$= C_n \frac{(2\sqrt{n})^n}{cn} \left(\left[\exp\left(c \frac{1}{(2\sqrt{n})^n} r^n\right) \exp\left(-\frac{r^2}{2}\right)\right]_0^\infty + \int_0^\infty \exp\left(\underbrace{c \frac{1}{(2\sqrt{n})^n} r^n}_{(D)} - \frac{r^2}{2}\right) r \,\mathrm{d}r\right)$$

$$\cdot \mathbb{P}\left(U_i > \frac{1}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right) \tag{165}$$

$$\geqslant C_n \frac{(2\sqrt{n})^n}{cn} \left([\infty - 0] + \int_0^\infty \exp\left(-\frac{r^2}{2}\right) r \,\mathrm{d}r\right) \cdot \mathbb{P}\left(U_i > \frac{1}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right) \tag{166}$$

$$= C_n \frac{(2\sqrt{n})^n}{cn} \left([\infty - 0] - \left[\exp\left(-\frac{r^2}{2}\right)\right]_0^\infty\right) \cdot \mathbb{P}\left(U_i > \frac{1}{2\sqrt{n}}, \forall i \in \llbracket n \rrbracket\right)$$

$$\overset{\substack{C_n > 0 \\ n \geqslant 3 \\ c \geqslant 0}}{=} \infty.$$

The inequality in Equation (163) follows from the fact that the event inside the indicator function happens with a probability $\leqslant 1$. Equation (164) is a rewriting of the previous line under the assumption that the indicator function evaluates to 1. We can rewrite the expected value as an integral over the positive real numbers since, according to the indicator function, every random variable X_i must be greater than $\frac{||X||_2}{2\sqrt{n}}$, which is a positive quantity.

Term (A) is a substitution of $\prod_{i\in[\![n]\!]} X_i$ with $\frac{r}{2\sqrt{n}}$ repeated n times, which comes from the indicator function. r is the integration variable and represents the Euclidean norm of vector X.

Term (B) represents the probability density of the Euclidean norm of a Gaussian vector $X \sim \mathcal{N}(0, \mathbf{I}_n)$.

Finally, term (C) represents the probability of the indicator function evaluating to 1. Considering the vector Y whose elements are $Y_i = X_i/||X||_2$, then $||Y||_2 = 1$. The probability that $Y_i > \frac{1}{2\sqrt{n}}, \forall i \in [\![n]\!]$ can be thought of as the probability that the point defined by Y in the n-dimensional space is located on the surface of the n-dimensional hyper-sphere of radius 1 in the region induced by the condition $Y_i > \frac{1}{2\sqrt{n}}$.

Equation (165) is an integration by parts of the two functions $f(r)$ and $g'(r)$ identified in the line above.

Equation (166) holds under the assumption that $n \geqslant 3$ and $c > 0$. First, the term:

$$\left[\exp\left(c\frac{1}{(2\sqrt{n})^n}r^n\right)\exp\left(-\frac{r^2}{2}\right)\right]_0^\infty \overset{\substack{n\geqslant 3\\ c>0}}{=} \infty - 0$$

under such an assumption. Second, we can write:

$$\exp\left(c\frac{1}{(2\sqrt{n})^n}r^n - \frac{r^2}{2}\right) \geqslant \exp\left(-\frac{r^2}{2}\right) \Rightarrow \int_0^\infty \exp\left(c\frac{1}{(2\sqrt{n})^n}r^n - \frac{r^2}{2}\right)\,\mathrm{d}r \geqslant \int_0^\infty \exp\left(-\frac{r^2}{2}\right)\,\mathrm{d}r$$

The final result then holds under the further assumption that $C_n > 0$. □

Lemma D.2 (Variance of the product of independent random variables)**.** *Let X_1, X_2, $\ldots X_n$ independent random variables. The variance of their product is:*

$$\mathbb{V}\mathrm{ar}[X_1X_2\cdots X_n] = \prod_{i\in[\![n]\!]}\left(\mathbb{V}\mathrm{ar}[X_i] + (\mathbb{E}[X_i])^2\right) - \prod_{i\in[\![n]\!]}(\mathbb{E}[X_i])^2$$

Proof.

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X_1X_2\cdots X_n] &= \mathbb{E}[(X_1X_2\cdots X_n)^2] - (\mathbb{E}[X_1X_2\cdots X_n])^2\\ &= \mathbb{E}[X_1^2X_2^2\cdots X_n^2] - (\mathbb{E}[X_1])^2(\mathbb{E}[X_2])^2\cdots(\mathbb{E}[X_n])^2\\ &= \mathbb{E}[X_1^2]\,\mathbb{E}[X_2^2]\cdots\mathbb{E}[X_n^2] - (\mathbb{E}[X_1])^2(\mathbb{E}[X_2])^2\cdots(\mathbb{E}[X_n])^2\\ &= \prod_{i\in[\![n]\!]}\left(\mathbb{V}\mathrm{ar}[X_i] + (\mathbb{E}[X_i])^2\right) - \prod_{i\in[\![n]\!]}(\mathbb{E}[X_i])^2\end{aligned}$$

□

Lemma D.3. *Let X_1, X_2, $\ldots, X_n$ independent subgaussian random variables with expected value $\mu_i \in [0,1]$ and subgaussianity parameter $\sigma_i \in [0,+\infty)$. The variance of the product $X_1X_2\cdots X_n$ is bounded by:*

$$\prod_{i\in[\![d]\!]}\sigma_i^2 \leqslant \mathbb{V}\mathrm{ar}[X_1X_2\cdots X_n] \leqslant \prod_{i\in[\![n]\!]}\left(1+\sigma_i^2\right) - 1$$

Proof. Now, we want to find the worst combination of $\mu_i, i \in [\![n]\!]$, i.e., the combination of expected values which maximizes the variance of the product of such random variables. To do so, we can consider a single $\bar{i} \in [\![n]\!]$, and look at the behavior of the first derivative when we change $\mu_{\bar{i}} \in [0,1]$. We recall from Lemma D.2 that:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X_1X_2\cdots X_n] &= \prod_{i\in[\![n]\!]}\left(\mathbb{V}\mathrm{ar}[X_i] + (\mathbb{E}[X_i])^2\right) - \prod_{i\in[\![n]\!]}(\mathbb{E}[X_i])^2\\ &= \prod_{i\in[\![n]\!]}\left(\sigma_i^2 + \mu_i^2\right) - \prod_{i\in[\![n]\!]}\mu_i^2\end{aligned}$$

$$= \left(\sigma_{\bar{i}}^2 + \mu_{\bar{i}}^2\right) \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \left(\sigma_i^2 + \mu_i^2\right) - \mu_{\bar{i}}^2 \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \mu_i^2, \tag{167}$$

$$= \mu_{\bar{i}}^2 \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \left(\sigma_i^2 + \mu_i^2\right) - \mu_{\bar{i}}^2 \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \mu_i^2 + \sigma_{\bar{i}}^2 \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \left(\sigma_i^2 + \mu_i^2\right) \tag{168}$$

$$= \mu_{\bar{i}}^2 \left(\underbrace{\prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \left(\sigma_i^2 + \mu_i^2\right)}_{\mathtt{A}} - \underbrace{\prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \mu_i^2}_{\mathtt{B}} \right) + \underbrace{\sigma_{\bar{i}}^2 \prod_{i \in [\![n]\!] \setminus \{\bar{i}\}} \left(\sigma_i^2 + \mu_i^2\right)}_{\mathtt{C}} \tag{169}$$

where lines (167), (168) and (169) are no other than an algebraic step to make explicit in the product the dependence on $\mu_{\bar{i}}$. Now we want to look at the worst case scenario for the variance, i.e., the value of $\mu_{\bar{i}}$ that maximize it.

Recalling the constraints on μ_i which is assumed to be bounded in $[0, 1]$ and σ_i^2 that is defined over $[0, +\infty]$, it trivial to see that term A is predominant over term B and so the worst case for element $\bar{i}$ is to consider $\mu_{\bar{i}} = 1$, no matter the other values of $\mu_i, i \in [\![n]\!] \setminus \{\bar{i}\}$. The term C is not relevant as $\mu_{\bar{i}}$ does not appear. This reasoning applies for all the possible values of $\bar{i} \in [\![n]\!]$, and so the worst case variance is when all the μ_i are equal to 1, for all the components $i \in [\![n]\!]$.

Given that, the variance of the product of independent random variables with expected values in $\mu_i \in [0, 1]$ and variance σ_i^2 can be bounded as:

$$\mathbb{V}\mathrm{ar}[X_1 X_2 \cdots X_n] \leqslant \prod_{i \in [\![n]\!]} \left(1 + \sigma_i^2\right) - 1.$$

A symmetric reasoning leads to the lower bound.

This concludes the proof. □

E. Numerical Validation

In this appendix, we provide numerical simulations to validate the proposed solutions. First, in Appendix E.1, we validate `F-UCB` against bandit baselines in several scenarios. Then, in Appendix E.2, we compare the two algorithms we propose (i.e., `F-UCB` and `F-Track`) in different scenarios to highlight their peculiarities. Finally, in Appendix E.3, we evaluate the proposed algorithms' behavior in the case in which the noise affecting intermediate observations is partially correlated. The code of the experiments can be found at `https://github.com/marcomussi/FRB`.

E.1. Comparison of `F-UCB` against Bandit Baselines

In this part, we show the effectiveness of `F-UCB` against bandit baselines.

Baselines The first baseline we consider is `UCB1` (Auer et al., 2002), which is designed for stochastic bandits. We consider the anytime version of the algorithm, proposed by Bubeck (2010). Due to its characteristics, we expect it to perform in a comparable manner to `F-UCB` for $d = 1$, with its performance degrading as the dimensionality grows. As an additional baseline, we consider a *robust* version of `UCB` algorithm designed for heavy-tail (HT) distributions (Bubeck et al., 2013) considering the *Median of Means* estimator (`RUCB-MoM`). Due to the capability of this algorithm to handle non-subgaussian noise, we expect it to converge for any problem dimensionality, although at a slower rate. Finally, we consider the `TEA` algorithm, proposed by Zimmert & Seldin (2018). Since this algorithm provides theoretical guarantees for handling only subgaussian noise applied to the reward, we expect it to have a performance that degrades when $d > 1$. For all the baselines, we consider the values of the hyperparameters as prescribed in the respective original papers.

Setting For the sake of simplicity in the presentation of the results, we consider the scenario in which all the problem dimensions present the same number of actions (i.e., $k_1 = \cdots = k_d =: k$). Moreover, we consider the setting in which the intermediate observations are drawn from Gaussian distributions with mean $\mu_{i,a_i(t)}$ for every action component $a_i(t)$ in position i of the action vector $\mathbf{a}$, formally $x_i(t) \sim \mathcal{N}(\mu_{i,a_i(t)}, \sigma^2)$, $\forall i \in [\![d]\!]$. We consider values of $k \in [\![3, 5]\!]$, and values of $d \in [\![4]\!]$. We draw the expected values $\mu_{i,j}$ for $i \in [\![d]\!]$ and $j \in [\![k]\!]$ from a uniform distribution in the range $[0.7, 1]$. We fix a value of $\sigma = 0.1$. It is worth noting that the results in the following paragraph are not comparable among the different k and d, mostly for what concerns the comparison between different values of d. We evaluate the performances in terms of cumulative regret with $T = 10^4$, averaged over 50 trials.

Results In Figure 3, we present the cumulative regret for the `F-UCB` algorithm and the other bandit baselines. The value of k increases with the columns, and the value of d increases with the rows of the figure. The following comments are valid for all the considered values of k, as no unexpected or relevant behaviors are present when we increase the number of actions for each action component. We observe that for $d = 1$, `F-UCB` achieves a cumulative regret that matches that of `UCB1`. This is expected, as `F-UCB` collapses to `UCB1` for $d = 1$. `RUCB-MoM` achieves a sublinear regret, although higher than the previous algorithms, whereas `TEA` suffers a cumulative regret that is linear in the considered time horizon. The behavior changes for $d = 2$. `F-UCB` achieves a low cumulative regret. The cumulative regret of `UCB1`, instead, constantly increases over the time horizon. `RUCB-MoM` continues to achieve a sublinear regret, however it is higher, due to the increased cardinality of the equivalent action space and the incremented effect of the noise. The behavior of `TEA` remains the same as for $d = 1$. For $d \geqslant 3$, we observe a stabilization of the behavior. `F-UCB` manages to achieve a cumulative regret that scales well as d and k increase. `UCB1` now suffers a linear regret, `RUCB-MoM` a sublinear regret worse with the increase of d, and `TEA` behaves as in the previous cases.

E.2. Comparison of `F-UCB` and `F-Track`

In this part, we provide numerical simulations intended to compare `F-UCB` and `F-Track` in different scenarios. As discussed in Remark 4.1 and shown Figure 2, the performances of `F-UCB` decrease when the number d of dimensions increases and when the suboptimality gaps are large. The goal of this part is to (i) verify once again this fact and (ii) observe if `F-Track` is able to mitigate such a phenomenon.

Setting We consider the scenario in which the number of arms is constant across all dimensions, i.e., $k_i = k, \forall i \in [\![d]\!]$. Given our goal to verify the algorithms' behavior over the action vector dimensionality d and the suboptimality gaps dimension, we fixed the other parameters. We consider a scenario in which we have $k = 2$ and observations affected Gaussian i.i.d. noise with $\sigma = 0.5$. We evaluate the two algorithms for $d \in \{2, 5, 10, 20, 30\}$. For what concerns the expected values, for all the dimensions, we enforce the first arm to be the best one, with expected value $\mu_{i,1} = \mu_i^* = 1, \forall i \in [\![d]\!]$. The suboptimal arms have all the same expected values $\mu_{i,2} = 1 - \Delta_{i,2}, \forall i \in [\![d]\!]$. Such a value $\Delta_{i,2}$ has been tested in the set $\Delta_{i,2} \in \{0.5, 0.7, 0.9\}$. We evaluate the performances in terms of regret, averaged over 10 runs with target time horizons $T \in [10^4, 10^5]$. We remark that `F-UCB` is an anytime algorithm and can be run once to obtain the entire curve of the cumulative regret. Instead, `F-Track` requires the knowledge of the horizon to compute the correct values of N_0 and ϵ_T. As such, we repeated the experiment for `F-Track` several times, each with a different time horizon up to the maximum T.

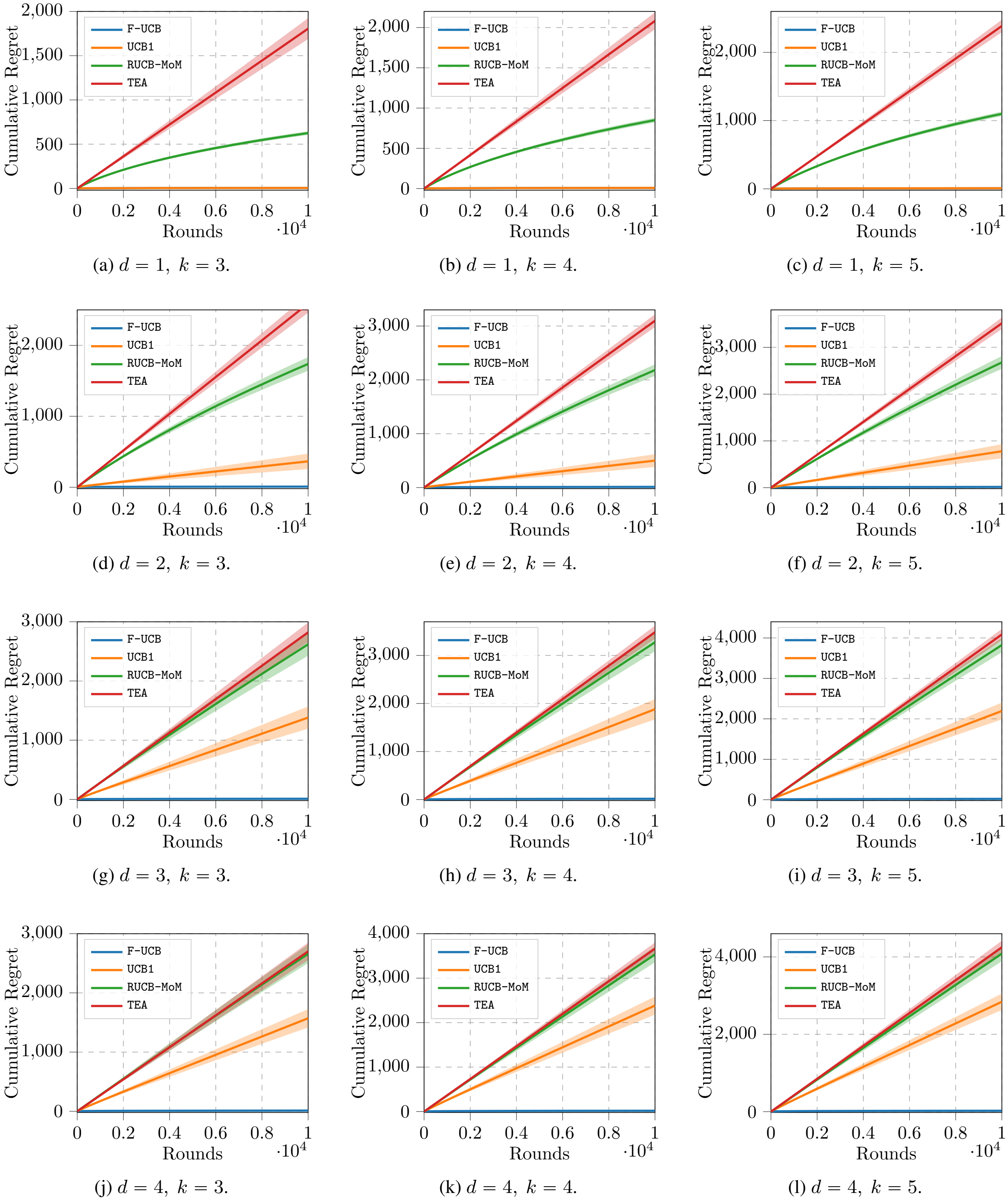
Results In Figure 4, we present the cumulative regret for `F-UCB` and `F-Track` in the above-mentioned setting. First, we observe that for small values of d (i.e., $d \in \{2, 5\}$), `F-UCB` outperforms `F-Track` for all the values of $\Delta_{i,2}$. This behavior is less evident when we move to $d = 10$, where the performances become comparable, with an advantage for `F-UCB` for smaller values of $\Delta_{i,2}$, while for larger value of the suboptimality gap, `F-Track` is better. The results turn in favor of `F-Track` when d becomes larger (i.e., $d \in \{20, 30\}$), and such an advantage further increases when $\Delta_{i,2}$ is large.

E.3. Robustness to Correlated Noise

In this part, we provide numerical simulations intended to compare `F-UCB` and `F-Track` when there is a correlation between the noises affecting the different dimensions. As discussed in Remark 3.1, in our setting, we require that the observations must be non-correlated. Otherwise, the problem cannot be factored properly given that, in general, if there is a correlation between the noises, we have that:

$$\mathbb{E}\left[\prod_{i \in [\![d]\!]} x_i(t)\right] \neq \prod_{i \in [\![d]\!]} \mathbb{E}\left[x_i(t)\right]. \tag{170}$$

Setting We consider the scenario in which the number of arms is constant across all dimensions, i.e., $k_i = k, \forall i \in [\![d]\!]$. We consider $k = 2$ and $d = 10$. For what concerns the expected values, for all the dimensions, we enforce the first arm to be the best one, with expected value $\mu_{i,1} = \mu_i^* = 1, \forall i \in [\![d]\!]$. The suboptimal arms have all the same expected values $\mu_{i,2} = 0.5, \forall i \in [\![d]\!]$. In order to evaluate the behavior of the algorithms in the presence of correlation in the noise of intermediate observations, we introduce a term $\alpha \in [0, 1]$ to control the interdependence of the intermediate observations.



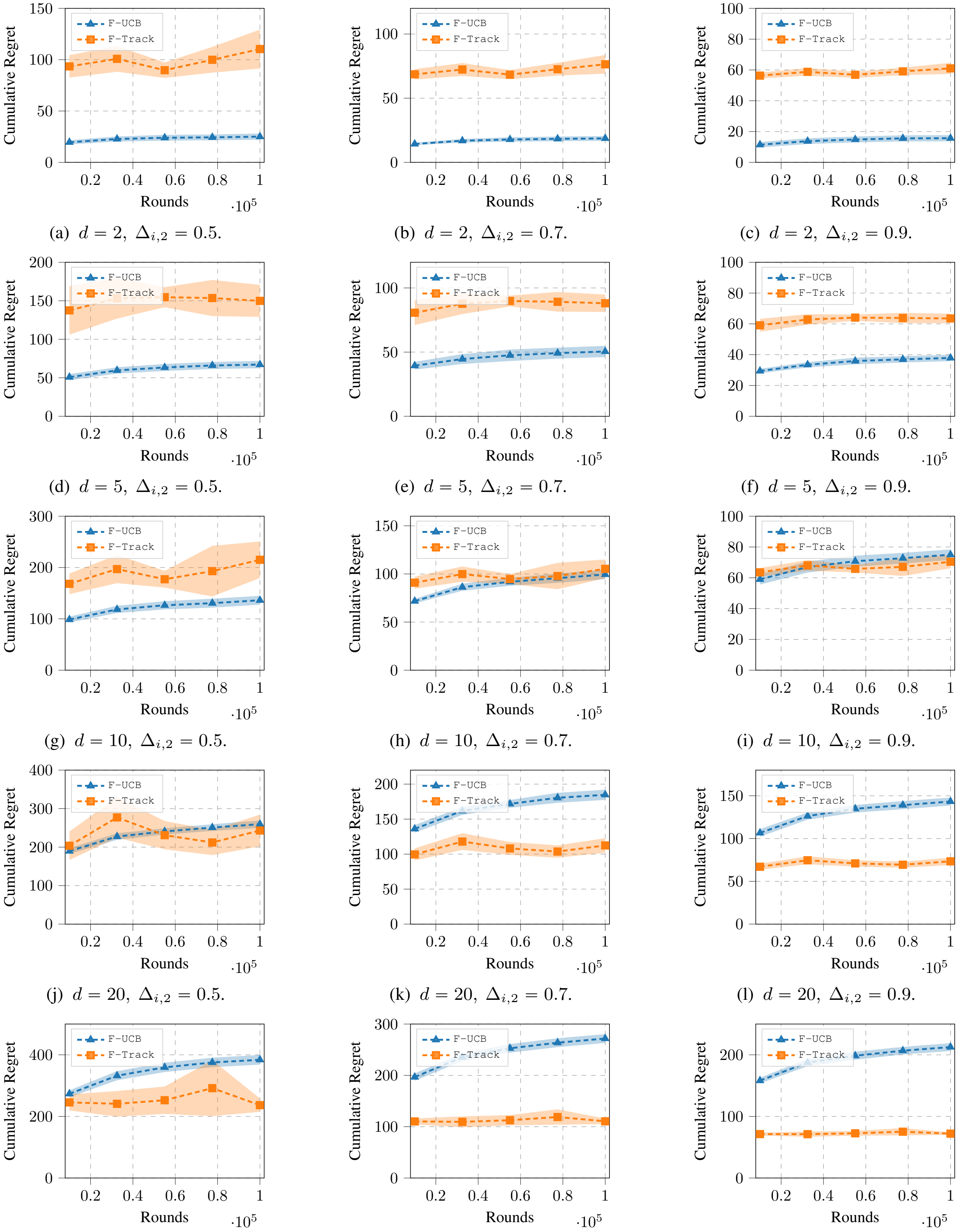
(a) $d = 1,\ k = 3.$ (b) $d = 1,\ k = 4.$ (c) $d = 1,\ k = 5.$

(d) $d = 2,\ k = 3.$ (e) $d = 2,\ k = 4.$ (f) $d = 2,\ k = 5.$

(g) $d = 3,\ k = 3.$ (h) $d = 3,\ k = 4.$ (i) $d = 3,\ k = 5.$

(j) $d = 4,\ k = 3.$ (k) $d = 4,\ k = 4.$ (l) $d = 4,\ k = 5.$

Figure 3. Performance of F-UCB, UCB1, RUCB-MoM and TEA considering $k \in [\![3, 5]\!]$ and $d \in [\![4]\!]$ (50 runs, mean ± std).



(a) $d = 2,\ \Delta_{i,2} = 0.5$.

(b) $d = 2,\ \Delta_{i,2} = 0.7$.

(c) $d = 2,\ \Delta_{i,2} = 0.9$.

(d) $d = 5,\ \Delta_{i,2} = 0.5$.

(e) $d = 5,\ \Delta_{i,2} = 0.7$.

(f) $d = 5,\ \Delta_{i,2} = 0.9$.

(g) $d = 10,\ \Delta_{i,2} = 0.5$.

(h) $d = 10,\ \Delta_{i,2} = 0.7$.

(i) $d = 10,\ \Delta_{i,2} = 0.9$.

(j) $d = 20,\ \Delta_{i,2} = 0.5$.

(k) $d = 20,\ \Delta_{i,2} = 0.7$.

(l) $d = 20,\ \Delta_{i,2} = 0.9$.

(m) $d = 30,\ \Delta_{i,2} = 0.5$.

(n) $d = 30,\ \Delta_{i,2} = 0.7$.

(o) $d = 30,\ \Delta_{i,2} = 0.9$.

Figure 4. Cumulative regret of `F-UCB` and `F-Track` considering $k = 2$, $\sigma = 0.5$, $d \in \{2, 5, 10, 20, 30\}$, and $\Delta_{i,2} \in \{0.5, 0.7, 0.9\}$, $\forall i \in [\![d]\!]$ (10 runs, mean ± 2std).

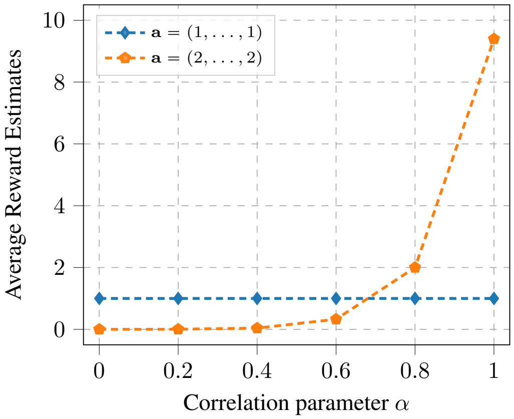
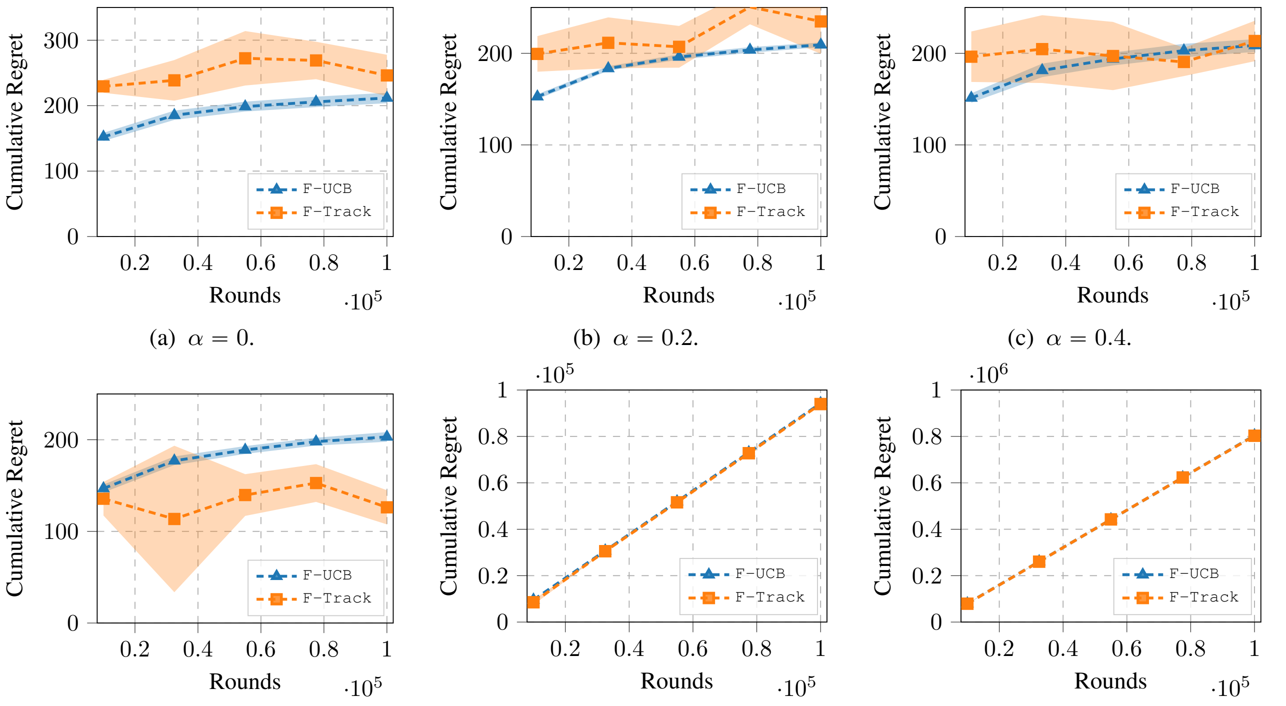


Figure 5. Monte Carlo estimates of the expected values for the tested values of the correlation parameter $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ (10^6 Monte Carlo simulations).

The additive noise applied to the observations $x_i(t)$ is defined as $\alpha\,\eta(t) + (1 - \alpha)\epsilon_i(t)$, where $\eta(t), \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$. The noise term $\eta(t)$ is applied to all the dimensions, whereas the $\epsilon_i(t)$ terms are individual and applied to the single dimensions $i \in [\![d]\!]$. Given this formulation, if $\alpha = 0$ the intermediate observations are independent, while if $\alpha = 1$, the intermediate observations are fully correlated. For values of $\alpha \in (0, 1)$, the noise term in the intermediate observations will comprise a correlated term and an independent term. We consider the case in which the Gaussian noise with $\sigma = 0.5$ (for both the independent and correlated components) affects *only action components* $a_i = 2$ (i.e., those with expected value $\mu_{i,2} = 0.5$) for $i \in [\![d]\!]$. We consider values of $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. We evaluate the performances in terms of cumulative regret averaged over 10 runs with target time horizons $T \in [10^4, 10^5]$.

Results Before commenting on the results, we observe that the presence of correlated noise over action components $a_i = 2$ has the effect of changing the optimal vector action depending on the value of α. In Figure 5, we plot the value of the expected reward of the action vectors $(1, \ldots, 1)$ and $(2, \ldots, 2)$ estimated using 10^6 Monte Carlo simulations for the values of α under analysis. We consider just the two action vectors $(1, \ldots, 1)$ and $(2, \ldots, 2)$, given that all the other combinations of action components will give intermediate results (and are suboptimal). We first observe that, given that all the observations of the action vector $(1, \ldots, 1)$ are not influenced by any noise, its expected reward is stable over α. On the other hand, for action vector $(2, \ldots, 2)$, affected by noise, we see how as the correlation increases, the expected reward increases itself and overtakes the one of action vector $(1, \ldots, 1)$.

Moving to the simulations, Figure 6 shows a comparison of the performances of `F-UCB` and `F-Track` when we vary correlation parameter α. First, we observe how the two algorithms present a consistent behavior over the different values of α. They are able to achieve satisfactory performances (i.e., sublinear regret) up to $\alpha = 0.6$. Then, the regret degenerates to linear. This is consistent with what we observed in Figure 5, as these algorithms look at the expected values of the single action components, but in this case, the noise correlation altered the optimal arm, which is no longer the one with the highest product of the expected observations.



(a) $\alpha = 0$.

(b) $\alpha = 0.2$.

(c) $\alpha = 0.4$.

(d) $\alpha = 0.6$.

(e) $\alpha = 0.8$.

(f) $\alpha = 1$.

Figure 6. Cumulative regret of F-UCB and F-Track considering $k = 2$, $\sigma = 0.5$, $d = 5$, $\Delta_{i,2} = 0.5, \forall i \in [\![d]\!]$, and correlation parameter $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ (10 runs, mean ± 2std).